WHEN MARGINALS MATCH BUT STRUCTURE FAILS: COVARIANCE FIDELITY IN GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models are increasingly used to produce synthetic data for scientific analysis, but common evaluation criteria mainly assess likelihood, feature-wise marginal fit, or perceptual quality. These measures may miss changes in covariance structure, which directly affect procedures such as regression and principal component analysis. We study this issue using

$$D_\Sigma(P, Q) = \|\Sigma_P - \Sigma_Q\|_F,$$

the covariance discrepancy between a target distribution P and a generative distribution Q. Three results describe what this quantity can and cannot establish. First, identical univariate marginals can coexist with different copulas and covariance matrices, and the raw covariance discrepancy can grow arbitrarily under rescaling. Second, when marginal variances match, D_Σ exactly determines the difference between two bivariate population regression slopes. Third, covariance perturbation bounds PCA eigenvalue and principal-subspace errors. The resulting Frobenius-based subspace certificate is informative when $D_\Sigma(P, Q)/\gamma_r(P) < 1/2$, where $\gamma_r(P)$ is the relevant population eigengap. Synthetic examples show that matched marginals can conceal a reversal in regression slope and differences in tail dependence that are not summarized by covariance alone. Applications to Fashion-MNIST and gene-expression data show that marginal, covariance, and downstream diagnostics can give different assessments. Observed sample errors are reported separately from theorem-based certificates. Since D_Σ is scale-sensitive and captures only second-order structure, it is intended as a task-linked complement to FID, MMD, copula diagnostics, and direct downstream evaluation, rather than as a general measure of generative quality.

1 INTRODUCTION

Generative models are now used for more than producing visually realistic samples. Synthetic data are also used in scientific studies, predictive modelling, and settings where access to observed data is limited (Chen et al., 2021; Jordon et al., 2022). In these applications, the synthetic sample may be passed directly to regression, dimension reduction, clustering, network analysis, or another downstream procedure. A sample can look plausible and reproduce each variable separately while still changing the relationships among variables. When those relationships determine the analysis, marginal agreement alone may give an incomplete account of model quality.

Several complementary criteria are already used to evaluate generative models. Likelihood-based criteria assess probabilistic fit, although their values do not necessarily track sample quality (Theis et al., 2016). The Fréchet inception distance compares the first two moments of learned image features (Heusel et al., 2017). Maximum mean discrepancy can detect differences between joint distributions when a characteristic kernel is used (Gretton et al., 2012), while precision and recall metrics separate aspects of sample fidelity and coverage (Sajjadi et al., 2018; Kynkäänniemi et al., 2019; Naeem et al., 2020). These measures answer different questions, but none by itself guarantees that a particular downstream analysis will remain stable. The relevant form of fidelity depends on the task. For linear regression and principal component analysis (PCA), covariance is especially important because it enters the target quantities directly. Recent work on synthetic-data selection similarly shows that covariance shift can affect downstream prediction and that covariance matching can improve synthetic-data selection (Rezaei et al., 2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Here, synthetic-data replacement is studied in analyses that depend on covariance structure. Let P denote the target distribution and Q the distribution induced by a generative model. Covariance disagreement is measured by

$$D_\Sigma(P, Q) := \|\Sigma_P - \Sigma_Q\|_F, \tag{1}$$

with a correlation-matrix version used when scale invariance is needed. For a leading PCA subspace, the discrepancy is also considered relative to its population eigengap:

$$R_r(P, Q) := \frac{D_\Sigma(P, Q)}{\gamma_r(P)}, \tag{2}$$

where $\gamma_r(P)$ separates the leading r-dimensional eigenspace from its orthogonal complement.

The two quantities serve different purposes. The value D_Σ measures the size of the covariance discrepancy, whereas R_r relates that discrepancy to the sensitivity of a specified PCA subspace. The form of the Davis–Kahan result used here gives (Davis & Kahan, 1970; Yu et al., 2015)

$$\|\sin\Theta(U_P, U_Q)\|_2 \le 2R_r(P, Q). \tag{3}$$

This certificate is non-vacuous only when $R_r(P, Q) < 1/2$. A larger ratio does not establish that the subspace is unstable. It only shows that this upper bound is inconclusive. For this reason, theoretical identities, non-vacuous perturbation certificates, and observed empirical errors are reported separately throughout the paper.

Contributions. The work makes four contributions.

1. **A task-linked formulation of covariance fidelity.** Marginal fidelity, covariance discrepancy, and downstream stability are placed within one framework. The raw Frobenius discrepancy is reported alongside a correlation-based alternative and, for PCA, an eigengap-relative ratio. These quantities are intended to complement existing generative-model metrics rather than replace them.
2. **A three-part theoretical hierarchy.** First, exact agreement of all univariate marginals is shown not to determine the copula or covariance matrix (Sklar, 1959; Nelsen, 2006). The raw covariance discrepancy can also be made arbitrarily large by rescaling matched marginals. Second, when the marginal variances match, the difference between two bivariate population regression slopes is shown to equal $D_\Sigma/(\sqrt{2}\sigma_X^2)$. Third, Weyl's inequality and a population-eigengap form of the Davis–Kahan theorem are used to bound PCA eigenvalue and principal-subspace errors in terms of covariance perturbation (Davis & Kahan, 1970; Yu et al., 2015).
3. **Constructions that clarify the scope of the framework.** A Gaussian sign-flip construction shows that identical marginals can coexist with regression slopes of opposite signs. A comparison between a Gaussian copula and a t-copula shows that matched marginals can also hide substantial differences in joint-tail probabilities. The second example lies outside the scope of D_Σ and shows why covariance fidelity should not be treated as a complete measure of dependence (Nelsen, 2006).
4. **Empirical illustrations with certificate-aware reporting.** The diagnostics are applied to Fashion-MNIST and two gene-expression settings (Xiao et al., 2017; Srinivasan et al., 2020; Cancer Genome Atlas Network, 2012). Marginal, covariance, RV, regression, and PCA summaries are compared, with the RV-coefficient used as a complementary measure of covariance similarity (Robert & Escoufier, 1976). Sample comparisons are treated as descriptive unless the stated perturbation certificate is non-vacuous. No general architectural conclusion is drawn from a single trained model.

The underlying mathematical tools, including covariance comparison, population regression identities, Weyl's inequality, and Davis–Kahan perturbation theory, are classical (Anderson, 2003; Jolliffe, 2002; Davis & Kahan, 1970). Their use here is narrower: a measurable structural discrepancy is connected to the downstream quantities that it can control. Covariance fidelity is not presented as a universal measure of realism. A small covariance discrepancy does not guarantee correct tail, conditional, nonlinear, temporal, or higher-order dependence. Conversely, a large raw discrepancy can partly reflect differences in variable scale. Distinct structural failures may therefore remain even when several evaluation criteria appear favorable (D'Amour et al., 2022).

Roadmap. Section 2 defines marginal and covariance-level fidelity. Section 3 places the proposed quantities in relation to generative-model evaluation, dependence measures, covariance shift, and matrix perturbation theory. Section 4 gives the three theoretical results. Section 5 presents the synthetic constructions and empirical analyses, and Section 6 discusses the practical interpretation and limitations. Complete proofs, additional experiments, and reproducibility details are provided in the appendices.

2 Problem Setup and Notation

Let P be the target distribution on $\mathbb{R}^d$ and let Q be the distribution produced by a generative model. Both distributions are assumed to have finite second moments. The goal is to determine whether replacing observations from P with samples from Q preserves the structure required by a given statistical analysis.

Marginal fidelity. For coordinate j, let P_j and Q_j denote the corresponding marginal distributions under P and Q. Exact marginal fidelity holds when

$$P_j = Q_j, \qquad j = 1, \dots, d. \tag{4}$$

In practice, marginal agreement may be examined through feature-wise summary statistics or univariate goodness-of-fit measures. Even exact agreement of all such marginals does not identify the joint distribution, since different copulas can share the same collection of univariate marginals (Sklar, 1959; Nelsen, 2006).

Covariance-level dependence fidelity. The covariance matrices of the two distributions are

$$\Sigma_P := \mathrm{Cov}_{X\sim P}(X), \qquad \Sigma_Q := \mathrm{Cov}_{X\sim Q}(X). \tag{5}$$

Their discrepancy is measured by

$$D_\Sigma(P, Q) := \|\Sigma_P - \Sigma_Q\|_F, \tag{6}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Covariance matrices describe second-order variation and linear dependence, but their entries depend on the measurement scales of the coordinates (Anderson, 2003; Pourahmadi, 2013). Accordingly, D_Σ is most meaningful when the variables have common units or have been placed on a comparable scale.

A scale-invariant comparison can be formed from the correlation matrices C_P and C_Q:

$$\widetilde{D}_\Sigma(P, Q) := \|C_P - C_Q\|_F. \tag{7}$$

The value $\widetilde{D}_\Sigma$ is used as a complementary empirical diagnostic. It cannot be substituted directly for D_Σ in a bound written for the original covariance matrices. After coordinate-wise standardization, however, covariance and correlation coincide, so the same matrix perturbation arguments can be applied in the standardized coordinate system.

Downstream functionals. Let $T(P)$ represent a population quantity computed under P. The examples below include regression coefficients, covariance eigenvalues, and leading principal subspaces. Stability is defined relative to the chosen functional: agreement for one quantity does not imply agreement for every procedure that depends on the joint distribution. The results in Sections 4 and 5 therefore connect D_Σ to specific linear and spectral analyses rather than treating it as a general measure of distributional similarity. Standard treatments of these covariance-based procedures can be found in Anderson (2003); Jolliffe (2002).

Eigengap-relative covariance discrepancy. Write the ordered eigenvalues of Σ_P as

$$\lambda_1(\Sigma_P) \geq \cdots \geq \lambda_d(\Sigma_P). \tag{8}$$

For a target dimension $1 \leq r < d$, define the population eigengap by

$$\gamma_r(P) := \lambda_r(\Sigma_P) - \lambda_{r+1}(\Sigma_P). \tag{9}$$

When $\gamma_r(P) > 0$, the covariance perturbation can be expressed relative to this gap:

$$R_r(P, Q) := \frac{D_\Sigma(P, Q)}{\gamma_r(P)}. \tag{10}$$

The Davis–Kahan result used in Theorem 3 gives (Davis & Kahan, 1970; Yu et al., 2015)

$$\|\sin \Theta(U_P, U_Q)\|_2 \leq 2R_r(P, Q), \tag{11}$$

where the columns of U_P and U_Q span the leading r-dimensional eigenspaces of Σ_P and Σ_Q. Since the left-hand side cannot exceed one, the displayed certificate is informative only when $R_r(P, Q) < 1/2$. A ratio at or above $1/2$ does not show that the estimated subspace is unstable. It only makes this particular upper bound inconclusive.

Norm relation and scope. For any matrix A, $\|A\|_2 \leq \|A\|_F$ (Vershynin, 2018). Thus, control of D_Σ also controls the operator-norm perturbation appearing in the eigenvalue and eigenspace results. The resulting guarantees are limited to second-order structure. Tail dependence, nonlinear dependence, and higher-order interactions are not determined by covariance alone (Nelsen, 2006). The copula construction in Appendix D.1 is included to make this boundary explicit.

3 Related Work

Evaluation of generative models. Generative models are evaluated from several perspectives. Likelihood-based criteria measure probabilistic fit, while Wasserstein distance and maximum mean discrepancy (MMD) compare distributions more directly (Theis et al., 2016; Villani, 2009; Gretton et al., 2012). Image-generation studies also rely on the Fréchet inception distance (FID) and precision–recall measures to assess sample quality and coverage (Heusel et al., 2017; Sajjadi et al., 2018; Kynkäänniemi et al., 2019; Naeem et al., 2020). Each metric captures a different part of generative quality, and their limitations have been discussed extensively (Borji, 2022). FID, for example, uses the means and covariances of Inception features and can therefore respond to some second-order differences in that representation. It does not certify covariance preservation in the original variables or in the representation used by a particular downstream analysis. The diagnostic considered here is therefore not proposed as a replacement for existing metrics. It is intended for tasks whose stability can be related directly to covariance perturbation.

Dependence and covariance diagnostics. Copula theory provides a formal separation between univariate marginals and joint dependence (Sklar, 1959; Nelsen, 2006). Other measures describe different aspects of dependence. Distance covariance and the Hilbert–Schmidt independence criterion can detect nonlinear dependence (Székely et al., 2007; Gretton et al., 2005), while the RV-coefficient summarizes similarity between covariance structures (Robert & Escoufier, 1976). These measures are useful for describing structural agreement, but they are not tied to the eigengap of a selected PCA subspace. The quantities D_Σ and R_r are used here because they can be connected to regression and eigenspace perturbation results.

Covariance shift and downstream performance. Models with similar observed performance can behave differently when relevant features of the data-generating process remain underspecified (D'Amour et al., 2022). The recent analysis of Rezaei et al. (2026) is particularly close to the present work. It studies synthetic-data selection through high-dimensional regression and shows that covariance shift can affect predictive generalization. It also identifies settings in which covariance matching is optimal and reports similar empirical patterns for neural networks and generative models. The problem considered here is related but different. Rather than selecting synthetic observations for predictive training, the generated distribution is evaluated as a possible substitute for the observed distribution in a downstream statistical analysis. Covariance is therefore connected to the sensitivity of regression coefficients and principal subspaces. Together, the two perspectives show that covariance agreement matters both for training with synthetic data and for analysing synthetic data after generation.

Matrix perturbation theory. The stability results rely on Weyl's eigenvalue inequality and the Davis–Kahan sin-Θ theorem (Davis & Kahan, 1970). The population-eigengap form used below

follows Yu et al. (2015); related matrix-norm results are standard in high-dimensional probability (Vershynin, 2018). No new matrix perturbation theorem is claimed. These results are used to translate a covariance discrepancy into bounds for specific statistical quantities.

4 Main Results: Dependence Fidelity and Downstream Stability

The argument is organized in three steps. The first result shows that matching all univariate marginals leaves the dependence structure undetermined. The second gives an exact expression for the change in a population regression slope when the marginal variances agree. The third gives eigenvalue and principal-subspace bounds when covariance perturbation is small relative to a population eigengap. Complete proofs are given in Appendix C.

4.1 Marginal Fidelity Does Not Imply Dependence Fidelity

Theorem 1 (Marginal fidelity does not imply dependence fidelity). For every $d \geq 2$, there exist distributions P and Q on $\mathbb{R}^d$ such that

1. $P_j = Q_j$ for every coordinate j;
2. their copulas, $\mathcal{C}_P$ and $\mathcal{C}_Q$, are distinct; and
3. $D_\Sigma(P, Q) > 0$.

Moreover, across a family of matched-marginal pairs obtained by increasing their common marginal scale, $D_\Sigma(P, Q)$ can be made arbitrarily large. If k is a characteristic kernel on the copula domain, then

$$\mathrm{MMD}_k(\mathcal{C}_P, \mathcal{C}_Q) > 0. \tag{12}$$

The construction can be seen directly in two dimensions. Let $\rho \in (0, 1)$ and $\sigma^2 > 0$, and define

$$\Sigma_+ = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad \Sigma_- = \sigma^2 \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}. \tag{13}$$

For $P = \mathcal{N}(0, \Sigma_+)$ and $Q = \mathcal{N}(0, \Sigma_-)$, each coordinate follows $\mathcal{N}(0, \sigma^2)$ under both distributions, but the Gaussian copulas are different (Sklar, 1959; Nelsen, 2006). The covariance discrepancy is

$$D_\Sigma(P, Q) = \|\Sigma_+ - \Sigma_-\|_F = 2\sqrt{2}\,\sigma^2|\rho|. \tag{14}$$

It therefore increases without bound as $\sigma^2 \to \infty$, even though the marginals within each pair remain identical. This statement concerns the raw, scale-sensitive covariance distance. For fixed dimension, the correlation-based discrepancy in equation 7 remains bounded. The same construction extends to $d > 2$ by adding independent coordinates with matching distributions.

Distinct copulas define distinct probability measures on the copula domain. For a characteristic kernel, MMD is zero only when the two measures are equal (Gretton et al., 2012; Sriperumbudur et al., 2010). The final claim of the theorem therefore follows from the copula difference.

4.2 Covariance Discrepancy and Regression Sensitivity

For a distribution P on (X, Y) with $\mathrm{Var}_P(X) > 0$, the population least-squares slope with an intercept is (Anderson, 2003)

$$\beta(P) := \frac{\mathrm{Cov}_P(X, Y)}{\mathrm{Var}_P(X)}. \tag{15}$$

Theorem 2 (Exact regression sensitivity under matched marginal variances). Let P and Q be distributions on $\mathbb{R}^2$ with finite second moments and

$$\mathrm{Var}_P(X) = \mathrm{Var}_Q(X) = \sigma_X^2 > 0, \qquad \mathrm{Var}_P(Y) = \mathrm{Var}_Q(Y) = \sigma_Y^2. \tag{16}$$

Then

$$|\beta(P) - \beta(Q)| = \frac{D_\Sigma(P, Q)}{\sqrt{2}\,\sigma_X^2}. \tag{17}$$

The diagonal entries of $\Sigma_P - \Sigma_Q$ are zero under the matched-variance assumption. Its two off-diagonal entries both equal $\mathrm{Cov}_P(X,Y) - \mathrm{Cov}_Q(X,Y)$, which gives the stated identity. Thus, identical marginals can accompany materially different regression slopes, including slopes with opposite signs. The equality in equation 17 depends on matched marginal variances and should not be used when those variances differ.

4.3 Stability of Eigenvalues and Principal Subspaces

Theorem 3 (PCA stability under covariance perturbation). Let P and Q be distributions on $\mathbb{R}^d$ with finite second moments and covariance matrices Σ_P and Σ_Q. Let their eigenvalues be ordered from largest to smallest. Then, for every $k = 1, \ldots, d$,

$$|\lambda_k(\Sigma_P) - \lambda_k(\Sigma_Q)| \le \|\Sigma_P - \Sigma_Q\|_2 \le D_\Sigma(P,Q). \tag{18}$$

For $1 \le r < d$, let $U_P, U_Q \in \mathbb{R}^{d\times r}$ contain orthonormal bases for the leading r-dimensional eigenspaces of Σ_P and Σ_Q, respectively. Suppose that the population eigengap

$$\gamma_r(P) := \lambda_r(\Sigma_P) - \lambda_{r+1}(\Sigma_P) \tag{19}$$

is positive. Then

$$\|\sin\Theta(U_P,U_Q)\|_2 \le \frac{2\|\Sigma_P - \Sigma_Q\|_2}{\gamma_r(P)} \le \frac{2D_\Sigma(P,Q)}{\gamma_r(P)} = 2R_r(P,Q). \tag{20}$$

The eigenvalue result in equation 18 follows from Weyl's inequality, and equation 20 follows from the population-eigengap form of the Davis–Kahan theorem (Davis & Kahan, 1970; Yu et al., 2015). Since

$$\|\sin\Theta(U_P,U_Q)\|_2 \le 1, \tag{21}$$

the Frobenius-based certificate is informative only when

$$R_r(P,Q) = \frac{D_\Sigma(P,Q)}{\gamma_r(P)} < \frac{1}{2}. \tag{22}$$

A value $R_r(P,Q) \ge 1/2$ does not establish instability. It means only that this bound is inconclusive; the observed principal-subspace error may still be small. The same interpretation is used in the empirical tables and figures.

Together, the three results give a limited hierarchy. Marginal agreement does not control dependence. Under matched marginal variances, covariance discrepancy determines the change in a bivariate population regression slope. When the covariance perturbation is sufficiently small relative to the relevant eigengap, it also provides a sufficient certificate for the stability of the selected principal subspace. Higher-order and tail dependence remain outside the scope of these results.

The operator norm can provide a tighter PCA-specific bound. We retain D_Σ because it gives a common diagnostic for regression and PCA, although the resulting Frobenius-based certificate can be conservative in high dimensions.

5 Synthetic Examples

These examples isolate changes in dependence while keeping the univariate marginals fixed. The first attains the regression identity in Theorem 2; the second contrasts an informative PCA certificate with an inconclusive bound. A tail-dependence example, which lies outside the scope of covariance diagnostics, is given in Appendix D.1.

5.1 Correlation Reversal

Let P and Q be centered bivariate normal distributions with

$$\Sigma_P = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad \Sigma_Q = \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}, \qquad \rho \in (0,1). \tag{23}$$

Both distributions have standard normal marginals, but their population regression slopes are $\beta(P) = \rho$ and $\beta(Q) = -\rho$. Moreover,

$$|\beta(P) - \beta(Q)| = 2|\rho| = \frac{D_\Sigma(P, Q)}{\sqrt{2}}, \qquad D_\Sigma(P, Q) = 2\sqrt{2}|\rho|. \tag{24}$$

Thus, changing only the off-diagonal covariance reverses the regression relationship without changing either marginal distribution. This is the unit-variance case of equation 17. The corresponding covariance matrices and regression slopes are shown in Appendix Figure 2.

5.2 PCA Under a Valid Covariance Perturbation

Consider $P = \mathcal{N}(0, \Sigma_P)$ and $Q_\varepsilon = \mathcal{N}(0, \Sigma_{Q_\varepsilon})$, where

$$\Sigma_P = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \Sigma_{Q_\varepsilon} = \begin{pmatrix} 3 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}, \qquad 0 \le \varepsilon < \sqrt{3}. \tag{25}$$

The stated range makes Σ_{Q_ε} positive definite. The marginal variances remain fixed, the reference eigengap is $\gamma_1(P) = 2$, and

$$D_\Sigma(P, Q_\varepsilon) = \sqrt{2}\varepsilon, \qquad R_1(P, Q_\varepsilon) = \frac{\varepsilon}{\sqrt{2}}, \qquad \sin \angle(u_P, u_{Q_\varepsilon}) \le \sqrt{2}\varepsilon. \tag{26}$$

The certificate is therefore non-vacuous for $\varepsilon < 1/\sqrt{2}$. The exact error is

$$\sin \angle(u_P, u_{Q_\varepsilon}) = \sin\left(\frac{1}{2} \arctan \varepsilon\right). \tag{27}$$

At $\varepsilon = 0.5$, the exact error is 0.230 and the bound is 0.707; at $\varepsilon = 1.6$, the exact error is 0.485, but the bound exceeds one and is uninformative. Figure 1 illustrates this difference. The second case is labelled *inconclusive*, not unstable. Figure 1 compares the exact subspace error with the theoretical bound and shows representative leading eigenvectors.

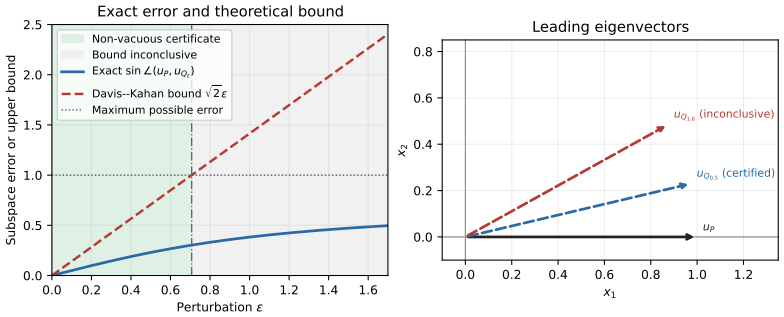


Figure 1: Exact leading-eigenvector error and its Davis–Kahan bound under the valid covariance family in equation 25 (left), with representative eigenvectors (right). The shaded certified region is where the bound is below one.

5.3 Empirical Summary

Table 1: Leading-subspace diagnostics across the empirical examples. A certificate is reported only when $2\widehat{R}_1 < 1$.

Dataset and model	$\widehat{D}_\Sigma$	$\widehat{R}_1$	$2\widehat{R}_1$	Observed error	Certificate
Fashion-MNIST, VAE	4.92	0.639	1.278	0.1940	No
Fashion-MNIST, Gaussian	0.25	0.033	0.065	0.0049	Yes
Alzheimer's, fitted full-covariance Gaussian.	49.35	0.176	0.353	0.0701	Yes
TCGA-BRCA, Splatter	249.79	2.053	4.106	0.9985	No
TCGA-BRCA, Poisson thinning	88.35	0.726	1.452	0.2099	No

6 DISCUSSION

Marginal agreement alone does not preserve dependence-sensitive analyses. Theorem 2 gives an exact regression result under matched variances, while Theorem 3 gives a sufficient PCA certificate when covariance error is small relative to the relevant eigengap. These are task-specific guarantees, not a universal measure of sample quality.

Relation to existing metrics. Feature-wise tests miss dependence changes that leave all marginals fixed. Joint metrics can detect broader differences, but their values do not usually translate into regression or eigenspace perturbation bounds. FID compares moments in an Inception representation, not necessarily in the representation used downstream (Heusel et al., 2017). Characteristic-kernel MMD can distinguish joint distributions without providing an eigengap-relative PCA certificate (Gretton et al., 2012). We therefore use D_Σ alongside marginal, distributional, and perceptual measures rather than as their replacement.

Interpretation and estimation. The diagnostic is model-agnostic. In particular, a diagonal conditional VAE posterior does not force the aggregated latent distribution or generated data to have diagonal covariance; correlations can arise through encoder means and the decoder (Kingma & Welling, 2014). Our Fashion-MNIST result describes one fitted model and is not an architectural impossibility claim.

Empirically, D_Σ is estimated by $\widehat{D}_\Sigma = \|\widehat{\Sigma}_P - \widehat{\Sigma}_Q\|_F$. Its reliability depends on dimension, sample size, and covariance structure. Regularized estimators may be preferable when sample covariance is unstable (Ledoit & Wolf, 2004). Likewise, bootstrap intervals require a resampling scheme appropriate to the estimator and sampling regime. We make no universal claim based on a fixed sample-size ratio.

For PCA, $R_r < 1/2$ yields a non-vacuous certificate under the bound used here; $R_r \geq 1/2$ means only that the bound is inconclusive. Observed subspace angles must then be examined directly. Descriptive alternatives such as the RV coefficient can provide complementary evidence, but rank correlation and two-sample distribution tests should not be interpreted as tests of agreement between diagnostics.

Limitations. Covariance does not characterize nonlinear, conditional, higher-order, or tail dependence. Population perturbation bounds also do not include finite-sample error in estimated covariances and eigengaps. The proposed quantities should therefore be read as transparent, assumption-dependent diagnostics for linear and spectral tasks. Broader claims about synthetic-data reliability require additional dependence and task-specific checks.

GENERATIVE AI USE DISCLOSURE

OpenAI ChatGPT, Claude and Codex were used during preparation of this work. Their use included editing and reorganizing manuscript text; improving grammar and readability; assisting with literature searches and reference verification; and debugging, and documenting code used for statistical analyses, simulations, diagnostic calculations, and scientific figures. The original research question, theoretical results, and scientific conclusions were determined by the authors. Numerical data and results were produced by the documented statistical procedures and executable code, rather than by direct language-model generation. All suggested references were checked against their original publications. AI-assisted code, mathematical content, and scientific interpretations were reviewed, tested, and revised by the authors. The authors take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Complete proofs are given in Appendix C. Data sources, preprocessing steps, experimental settings, diagnostic definitions, and software requirements are documented in Appendix E. Code and saved intermediate results needed to reproduce the reported tables and figures are included in the anonymous supplementary material.

Ethics statement. This study uses publicly available image data and de-identified gene-expression data under their stated access conditions. No new human-participant data were collected, and no individual-level clinical conclusions are drawn. The reported covariance diagnostics should not be treated as general guarantees of synthetic-data quality, privacy, or clinical validity. Privacy, appropriate data use, and the limitations of covariance-based evaluation are discussed in Appendix F.

REFERENCES

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 3rd edition, 2003. ISBN 9780471360919.

Ali Borji. Pros and cons of GAN evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, 2022. doi: 10.1016/j.cviu.2021.103329.

Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. doi: 10.1038/nature11412.

Richard J. Chen, Ming Y. Lu, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021. doi: 10.1038/s41551-021-00751-8.

Antonio Colaprico, Tiago C. Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S. Sabedot, Tathiane M. Malta, Stefano M. Pagnotta, Isabella Castiglioni, Michele Ceccarelli, Gianluca Bontempi, and Houtan Noushmehr. TCGAbiolinks: an R/Bioconductor package for integrative analysis with GDC data. *Nucleic Acids Research*, 44(8):e71, 2016. doi: 10.1093/nar/gkv1507.

Bin Dai and David Wipf. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1903.05789.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://jmlr.org/papers/v23/20-1255.html.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

Stefano Demarta and Alexander J. McNeil. The t copula and related copulas. *International Statistical Review*, 73(1):111–129, 2005. doi: 10.1111/j.1751-5823.2005.tb00254.x.

David Gerard. Data-based RNA-seq simulations by binomial thinning. *BMC Bioinformatics*, 21(1): 206, 2020. doi: 10.1186/s12859-020-3450-9.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory (ALT)*, pp. 63–77, 2005. doi: 10.1007/11564089_7.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002. ISBN 9780387954424. doi: 10.1007/b98835.

James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. Synthetic data – what, why and how? *arXiv preprint arXiv:2205.03257*, 2022. URL https://arxiv.org/abs/2205.03257.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014. URL https://openreview.net/forum?id=33X9fd2-9FyZd.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metrics for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3927–3936, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0234c510bc6d908b28c70ff313743079-Abstract.html.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Christos Louizos and Max Welling. Structured and efficient variational deep learning with matrix Gaussian posteriors. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1708–1716. PMLR, 2016. URL https://proceedings.mlr.press/v48/louizos16.html.

Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. doi: 10.1186/s13059-014-0550-8.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, 2020. URL https://proceedings.mlr.press/v119/naeem20a.html.

Roger B. Nelsen. *An Introduction to Copulas*. Springer, 2nd edition, 2006. ISBN 9780387286785. doi: 10.1007/0-387-28678-0.

Mohsen Pourahmadi. *High-Dimensional Covariance Estimation*. Wiley Series in Probability and Statistics. Wiley, 2013. ISBN 9781118034293. doi: 10.1002/9781118573617.

Parham Rezaei, Filip Kovačević, Francesco Locatello, and Marco Mondelli. High-dimensional analysis of synthetic data selection. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=Y54P2BBPPh.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286. PMLR, 2014. URL https://proceedings.mlr.press/v32/rezende14.html.

P. Robert and Yves Escoufier. A unifying tool for linear multivariate statistical methods: The RV-coefficient. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 25(3):257–265, 1976. doi: 10.2307/2347233.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, volume 31, pp. 5228–5237, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/f7696a9b362ac5a51c3dc8f098b73923-Abstract.html.

Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

Karpagam Srinivasan, Brad A. Friedman, Ainhoa Etxeberria, Melanie A. Huntley, Marcel P. van der Brug, Oded Foreman, Jonathan S. Paw, Zora Modrusan, Thomas G. Beach, Geidy E. Serrano, and David V. Hansen. Alzheimer's patient microglia exhibit enhanced aging and unique transcriptional activation. *Cell Reports*, 31(13):107843, 2020. doi: 10.1016/j.celrep.2020.107843.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010. URL https://jmlr.org/papers/v11/sriperumbudur10a.html.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007. doi: 10.1214/009053607000000505.

Lucas Theis, Aaron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations (ICLR)*, 2016. URL https://arxiv.org/abs/1511.01844.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. ISBN 9781108415194. doi: 10.1017/9781108231596.

Cédric Villani. The Wasserstein distances. In *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*, pp. 93–111. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9_6.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. URL https://arxiv.org/abs/1708.07747.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. doi: 10.1093/biomet/asv008.

Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1):174, 2017. doi: 10.1186/s13059-017-1305-0.

A Appendix

B Appendix Overview

The supplementary material is organized as follows.

- **Appendix A** provides an overview of the supplementary material.
- **Appendix B** contains the proofs of Theorems 1–3.
- **Appendix C** contains the synthetic and empirical analyses.
- **Appendix D** reports the data, implementation, and reproducibility details.
- **Appendix E** discusses ethical considerations, broader impacts, and limitations.

Appendix C is organized according to the question addressed in each analysis:

1. a Gaussian and t-copula example illustrating a tail-dependence difference outside the scope of D_Σ;
2. a Fashion-MNIST comparison of covariance and PCA-subspace fidelity;
3. a small-sample analysis using Alzheimer's gene-expression data;
4. an additional analysis of the Fashion-MNIST VAE and the limits of architectural interpretation;
5. a TCGA-BRCA generator comparison;
6. a cross-dataset empirical summary; and
7. a comparison with MMD and FID.

C Proofs of Main Results

The proofs follow the order of the results in Section 4. The first result gives a counterexample to marginal sufficiency. The second gives an exact regression identity under matched marginal variances. The third gives eigenvalue and principal-subspace perturbation bounds.

C.1 Auxiliary Facts

Copula uniqueness. If a multivariate distribution has continuous marginal distributions, its copula is unique by Sklar's theorem (Sklar, 1959; Nelsen, 2006).

Characteristic kernels. If k is characteristic (Sriperumbudur et al., 2010; Gretton et al., 2012), then

$$\mathrm{MMD}_k(\mu, \nu) = 0 \quad \Longleftrightarrow \quad \mu = \nu. \tag{28}$$

Therefore, two distinct copula distributions have strictly positive MMD when such a kernel is used on the copula domain.

C.2 Proof of Theorem 1

Proof. Fix $\rho \in (0, 1)$ and $\sigma^2 > 0$. In dimension two, define

$$P = \mathcal{N}(0, \Sigma_+), \qquad Q = \mathcal{N}(0, \Sigma_-), \tag{29}$$

where

$$\Sigma_+ = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \qquad \Sigma_- = \sigma^2 \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}. \tag{30}$$

Both matrices are positive definite because their eigenvalues are $\sigma^2(1 + \rho)$ and $\sigma^2(1 - \rho)$.

The two coordinate marginals are $\mathcal{N}(0, \sigma^2)$ under both P and Q. Thus, $P_j = Q_j$ for $j = 1, 2$. Their covariance difference is

$$\Sigma_+ - \Sigma_- = 2\sigma^2\rho \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{31}$$

and hence

$$D_\Sigma(P, Q) = \|\Sigma_+ - \Sigma_-\|_F = 2\sqrt{2}\,\sigma^2\rho > 0. \tag{32}$$

The marginal distributions are continuous and identical. Gaussian copulas are determined by their correlation matrices (Nelsen, 2006), which here have correlations ρ and $-\rho$. Since $\rho \neq 0$, the copulas are distinct. It follows that

$$\mathrm{MMD}_k(\mathcal{C}_P, \mathcal{C}_Q) > 0 \tag{33}$$

for any characteristic kernel k on the copula domain.

To obtain an arbitrarily large raw covariance discrepancy, fix any $\rho \in (0, 1)$ and let $\sigma^2 \to \infty$, equation 32 then tends to infinity, while the marginals of P and Q remain exactly matched for every value of σ^2. This step concerns the scale-sensitive covariance distance; it does not imply that the correlation-based discrepancy is unbounded.

For $d > 2$, append $d - 2$ independent coordinates having the same distributions under P and Q. Equivalently, use block-diagonal covariance matrices with Σ_+ or Σ_- in the first block and the same positive diagonal block in the remaining coordinates. All univariate marginals still match, the first two-dimensional copulas remain different, and the Frobenius discrepancy remains the value in equation 32. This proves the result for every $d \geq 2$.

□

C.3 Proof of Theorem 2

Proof. For a distribution P on (X, Y), consider population least squares with an intercept:

$$R_P(a, b) = \mathbb{E}_P\left[(Y - a - bX)^2\right]. \tag{34}$$

For a fixed b, minimization over a gives

$$a = \mathbb{E}_P[Y] - b\mathbb{E}_P[X]. \tag{35}$$

Substitution centers both variables, so differentiation with respect to b yields

$$\beta(P) = \frac{\mathrm{Cov}_P(X, Y)}{\mathrm{Var}_P(X)}. \tag{36}$$

This is the standard population least-squares identity (Anderson, 2003). The same argument gives the corresponding expression for $\beta(Q)$. No zero-mean assumption is needed when an intercept is included.

Let

$$\Delta_{12} := \mathrm{Cov}_P(X, Y) - \mathrm{Cov}_Q(X, Y). \tag{37}$$

Under the matched-variance assumptions of Theorem 2,

$$|\beta(P) - \beta(Q)| = \frac{|\Delta_{12}|}{\sigma_X^2}. \tag{38}$$

The covariance matrices differ only in their off-diagonal entries:

$$\Sigma_P - \Sigma_Q = \begin{pmatrix} 0 & \Delta_{12} \\ \Delta_{12} & 0 \end{pmatrix}. \tag{39}$$

Therefore,

$$D_\Sigma(P, Q) = \|\Sigma_P - \Sigma_Q\|_F = \sqrt{2}\,|\Delta_{12}|. \tag{40}$$

Combining equation 38 and equation 40 gives

$$|\beta(P) - \beta(Q)| = \frac{D_\Sigma(P, Q)}{\sqrt{2}\,\sigma_X^2}, \tag{41}$$

which is the claimed identity. □

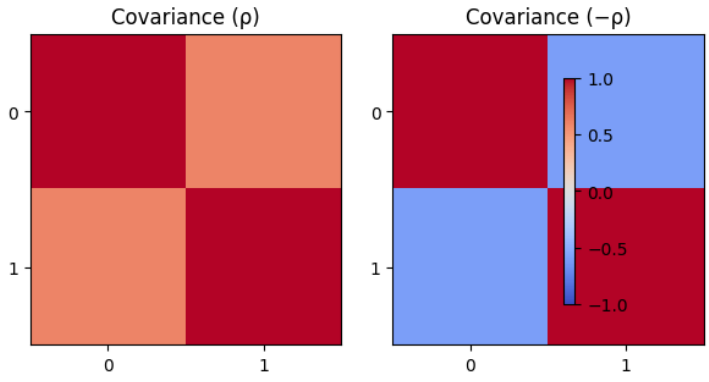


Figure 2: Matched-marginal Gaussian distributions with opposite covariance signs. Although both coordinates have the same standard-normal marginal distributions under P and Q, changing the off-diagonal covariance from ρ to $-\rho$ reverses the population regression slope.

C.4 Proof of Theorem 3

Proof. Let

$$\Delta := \Sigma_Q - \Sigma_P, \qquad \varepsilon := \|\Delta\|_2. \tag{42}$$

Because the operator norm is bounded by the Frobenius norm,

$$\varepsilon \le \|\Delta\|_F = D_\Sigma(P, Q). \tag{43}$$

For every $k = 1, \ldots, d$, Weyl's eigenvalue perturbation inequality gives (Yu et al., 2015)

$$|\lambda_k(\Sigma_P) - \lambda_k(\Sigma_Q)| \le \varepsilon \le D_\Sigma(P, Q), \tag{44}$$

which proves the eigenvalue statement.

Now fix $1 \le r < d$ and write

$$\gamma_r(P) = \lambda_r(\Sigma_P) - \lambda_{r+1}(\Sigma_P) > 0. \tag{45}$$

Two cases are considered. If $2\varepsilon/\gamma_r(P) \ge 1$, then

$$\|\sin\Theta(U_P, U_Q)\|_2 \le 1 \le \frac{2\varepsilon}{\gamma_r(P)}, \tag{46}$$

so the stated operator-norm bound holds, although it is not informative.

Suppose instead that $2\varepsilon/\gamma_r(P) < 1$. Weyl's inequality implies

$$\lambda_{r+1}(\Sigma_Q) \le \lambda_{r+1}(\Sigma_P) + \varepsilon. \tag{47}$$

Consequently,

$$\lambda_r(\Sigma_P) - \lambda_{r+1}(\Sigma_Q) \ge \gamma_r(P) - \varepsilon > \frac{\gamma_r(P)}{2}. \tag{48}$$

The operator-norm Davis–Kahan sin-Θ theorem can therefore be applied to the leading r-dimensional eigenspaces (Davis & Kahan, 1970; Yu et al., 2015), giving

$$\|\sin\Theta(U_P, U_Q)\|_2 \le \frac{\varepsilon}{\gamma_r(P) - \varepsilon} < \frac{2\varepsilon}{\gamma_r(P)}. \tag{49}$$

Together with the first case and equation 43, this yields

$$\|\sin\Theta(U_P, U_Q)\|_2 \le \frac{2\|\Sigma_P - \Sigma_Q\|_2}{\gamma_r(P)} \le \frac{2D_\Sigma(P, Q)}{\gamma_r(P)} = 2R_r(P, Q). \tag{50}$$

This is the population-eigengap form of the Davis–Kahan bound.

Since every principal-angle sine is at most one, the Frobenius-based certificate is non-vacuous when

$$R_r(P, Q) = \frac{D_\Sigma(P, Q)}{\gamma_r(P)} < \frac{1}{2}. \tag{51}$$

If $R_r(P, Q) \ge 1/2$, this particular bound is inconclusive. It does not imply that the two principal subspaces are unstable. □

D Synthetic and Empirical Examples

This appendix separates two questions. Appendix D.1 gives a synthetic example with exactly matched univariate marginals and different tail dependence. Appendix D.2 examines covariance and principal-subspace fidelity for two generators on Fashion-MNIST. The Fashion-MNIST marginals are assessed empirically and are not assumed to match exactly.

D.1 Gaussian and t Copulas: Matched Marginals with Different Tail Dependence

This example illustrates the part of joint dependence that cannot be inferred from univariate marginals. It also marks a boundary of the covariance diagnostic: D_Σ measures second-order structure, whereas tail dependence is a copula-level property.

Construction. Let (X_1, X_2) have standard normal marginals under two joint distributions. Under P, dependence is generated by a Gaussian copula with parameter $\rho \in (0, 1)$. Under Q, dependence is generated by a t copula with parameter ρ and $\nu > 2$ degrees of freedom, followed by the probability integral transform and the standard-normal quantile transform (Nelsen, 2006). Thus,

$$P_j = Q_j = \mathcal{N}(0, 1), \qquad j = 1, 2. \tag{52}$$

The parameter ρ describes the latent copula correlation. After transformation to normal marginals, it need not equal the Pearson correlation of (X_1, X_2) under Q.

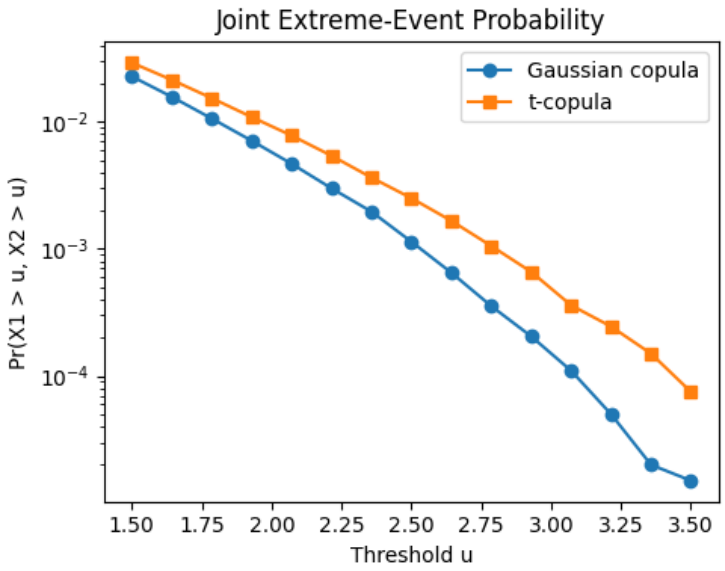


Figure 3: Estimated joint exceedance probabilities for Gaussian and t copulas with identical standard-normal marginals and the same latent copula correlation parameter. The t copula produces larger joint upper-tail probabilities over the displayed thresholds.

Tail dependence. For $|\rho| < 1$, the Gaussian copula has upper-tail dependence coefficient zero, whereas the t copula has the positive coefficient (Demarta & McNeil, 2005) The t copula has

$$\lambda_U = 2T_{\nu+1}\left(-\sqrt{\frac{(\nu+1)(1-\rho)}{1+\rho}}\right) > 0, \tag{53}$$

where $T_{\nu+1}$ is the cumulative distribution function of a univariate t distribution with $\nu+1$ degrees of freedom. The two copulas are therefore distinct. For a characteristic kernel k on the copula domain (Sriperumbudur et al., 2010; Gretton et al., 2012),

$$\mathrm{MMD}_k(\mathcal{C}_P, \mathcal{C}_Q) > 0. \tag{54}$$

For threshold u, define the joint upper-tail functional

$$T_u(R) := \Pr_R(X_1 > u, X_2 > u), \qquad R \in \{P, Q\}. \tag{55}$$

Although the marginal exceedance probabilities are identical, the simulated values of $T_u(P)$ and $T_u(Q)$ separate as u increases.

Simulation. We generated 10^5 observations from each distribution and evaluated T_u over the reported threshold grid. This sample size reduces Monte Carlo error but does not eliminate it. Figure 3 reports the estimated joint exceedance probabilities, and Figure 4 reports the empirical marginal cumulative distribution functions.

This construction establishes a failure of marginal-only evaluation. It should not be described as a covariance-matched counterexample because using the same latent copula parameter does not ensure equal Pearson covariance after the marginal transformation. A strict demonstration of information beyond D_Σ would require calibration of the two models to the same data-space covariance. The present result supports the narrower conclusion that D_Σ does not itself quantify tail dependence.

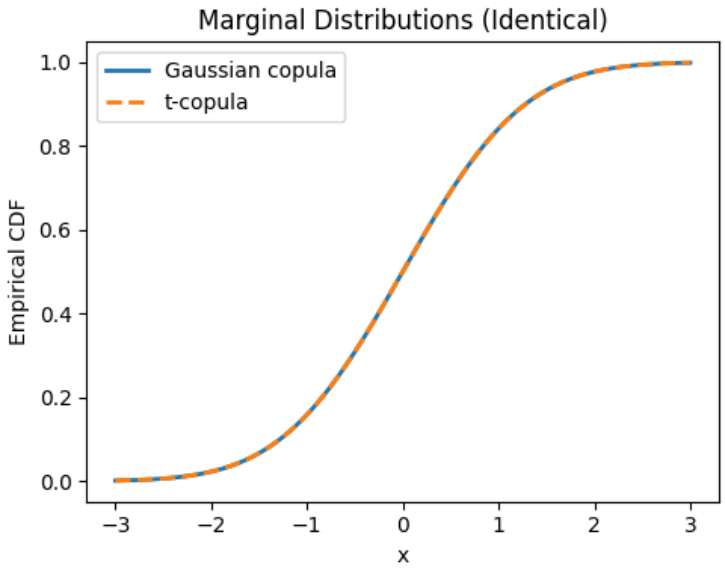


Figure 4: Empirical marginal CDFs for the Gaussian-copula and t-copula samples. The curves agree because both constructions use standard-normal marginals.

D.2 Fashion-MNIST: Covariance and PCA Fidelity

This analysis evaluates whether covariance discrepancies observed in generated samples are accompanied by changes in PCA subspaces and regression slopes. Fashion-MNIST contains 60,000 training images of size 28×28 pixels (Xiao et al., 2017). The analysis uses the first $p = 50$ principal-component coordinates obtained from flattened real training images(Jolliffe, 2002). These components explain 86.3% of the variance in the reference sample.

Generators. Two generators are compared. The first is a standard variational autoencoder with a diagonal-Gaussian conditional approximate posterior (Kingma & Welling, 2014), trained for 30 epochs. Its reported final evidence lower bound loss is 237.15. The second is a multivariate Gaussian fitted to the empirical mean and covariance of the real PCA coordinates:

$$X_{\mathrm{G}} \sim \mathcal{N}(\widehat{\mu}_{\mathrm{real}}, \widehat{\Sigma}_{\mathrm{real}}). \tag{56}$$

The Gaussian model matches the fitted reference covariance at the population level, although its finite synthetic sample has nonzero estimation error. Each generator produces 60,000 observations, which are represented in the same fixed PCA basis.

A diagonal conditional posterior in a VAE does not imply that the aggregated latent distribution or the decoded observations have diagonal covariance. The VAE results below are therefore treated as findings for this fitted model. They do not establish an unavoidable architectural effect.

Because the Gaussian is fitted directly in the evaluation PCA space, it is used as an oracle-like structural control rather than as a generator trained under conditions directly comparable to the VAE.

Diagnostics. For each generator, we compute the sample covariance discrepancy D_{Σ}, the correlation-matrix discrepancy $\widetilde{D}_{\Sigma}$, the RV coefficient, feature-wise Kolmogorov–Smirnov distances, empirical regression slopes, and principal-subspace errors

$$\left\| \sin\Theta(\widehat{U}_{\mathrm{real},r}, \widehat{U}_{\mathrm{syn},r}) \right\|_2, \qquad r \in \{1, 2, 3, 5, 10\}. \tag{57}$$

For PCA, the local ratio and Frobenius-based bound are as follows (Davis & Kahan, 1970; Yu et al., 2015):

$$R_r := \frac{D_{\Sigma}}{\widehat{\gamma}_r}, \qquad B_r := 2R_r, \qquad \widehat{\gamma}_r := \widehat{\lambda}_r - \widehat{\lambda}_{r+1}. \tag{58}$$

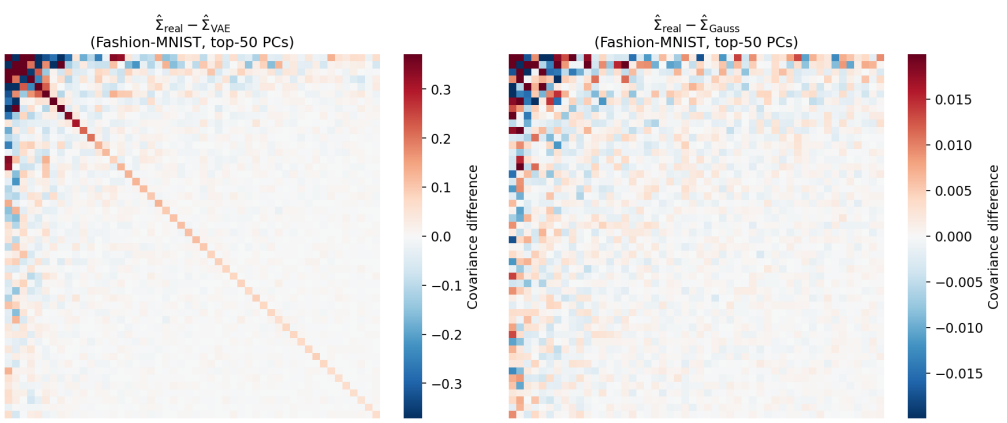


Figure 5: Sample covariance residuals $\widehat{\Sigma}_{\text{real}} - \widehat{\Sigma}_{\text{syn}}$ for the VAE and fitted-Gaussian generators in the top-50 Fashion-MNIST PCA coordinates. The color scales differ between panels and should be read from their respective color bars. The observed discrepancies are 4.92 for the VAE and 0.25 for the Gaussian baseline.

The bound is non-vacuous only when $B_r < 1$, equivalently $R_r < 1/2$.

Covariance discrepancy. The observed Frobenius discrepancies are

$$D_\Sigma(\text{VAE}) = 4.92, \qquad D_\Sigma(\text{Gaussian}) = 0.25. \tag{59}$$

The leading reference eigengap is $\widehat{\gamma}_1 = 7.70$, based on leading sample eigenvalues 19.81, 12.11, 4.11, 3.38, and 2.62. Therefore,

$$R_1(\text{VAE}) = 0.64, \qquad R_1(\text{Gaussian}) = 0.03. \tag{60}$$

For the VAE, the resulting bound is $B_1 = 1.278$ and is vacuous. It provides no stability or instability classification. For the Gaussian baseline, $B_1 = 0.0652$, which is a non-vacuous certificate for the leading empirical eigenspace.

Figure 5 shows the covariance residual matrices. The larger VAE residuals are concentrated among the leading PCA coordinates. This pattern is descriptive; it cannot be attributed to the diagonal conditional posterior without further model comparisons or controlled ablations.

The 500 bootstrap resamples produced central percentile ranges

$$[4.83, 5.04] \quad \text{for the VAE,} \qquad [0.33, 0.56] \quad \text{for the Gaussian baseline,} \tag{61}$$

with reported bootstrap standard errors of 0.056 and 0.059. The Gaussian plug-in estimate, 0.25, falls outside its percentile range. Because the norm is nonnegative and the bootstrap distribution is upward biased near zero, these ranges are reported as resampling summaries rather than calibrated 95% confidence intervals.

The RV coefficients are 0.9835 for the VAE and 0.9999 for the Gaussian baseline. This ordering agrees with the full-data D_Σ ordering, but both RV values are close to one and should be interpreted on their own scale.

Marginal comparisons. The median feature-wise KS distances are 0.090 for the VAE and 0.020 for the Gaussian baseline. Thus, the Gaussian baseline also performs better under this marginal diagnostic. The KS and covariance measures summarize different properties and their numerical magnitudes are not directly comparable. Figure 6 reports the distribution across the 50 PCA coordinates.

Principal-subspace comparison. Table 2 reports observed subspace errors and the Frobenius-based Davis–Kahan bounds. The VAE bound is vacuous at every reported dimension, including $r = 1$ and $r = 2$. Its small observed errors at those two dimensions are empirical findings, not consequences of the bound. At $r = 3$, the observed VAE error increases to 0.6900 as the reference eigengap

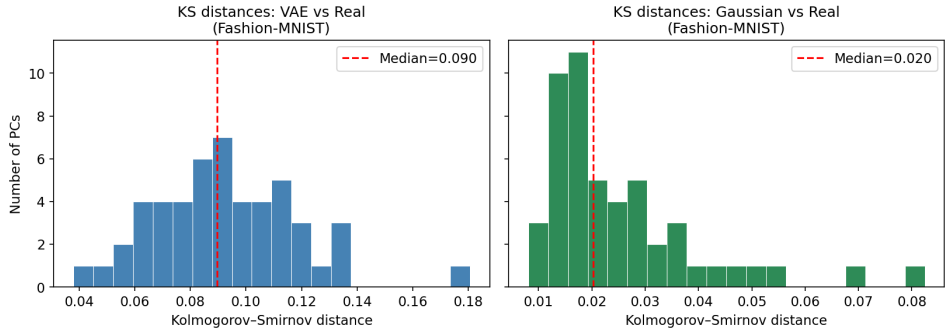


Figure 6: Feature-wise KS distances between real Fashion-MNIST PCA coordinates and samples from the VAE and fitted-Gaussian generators. The median distances are 0.090 and 0.020, respectively. These values measure marginal fit and do not determine covariance or subspace fidelity.

Table 2: Observed PCA subspace errors and Frobenius-based Davis–Kahan bounds on Fashion-MNIST. "Yes" indicates that the reported upper bound is below one. A value above one is vacuous rather than evidence of instability.

	VAE			Gaussian baseline		
r	error	bound	certificate	error	bound	certificate
1	0.1940	1.2780	No	0.0049	0.0652	Yes
2	0.1266	1.2294	No	0.0086	0.0628	Yes
3	0.6900	> 1	No	0.0236	0.6932	Yes
5	0.4817	> 1	No	0.0402	1.9028	No
10	0.4385	> 1	No	0.0284	> 1	No

becomes smaller, but this association does not establish that crossing a theoretical threshold caused the increase.

The Gaussian baseline has non-vacuous certificates at $r = 1$, $r = 2$, and $r = 3$. Its observed errors remain small through $r = 10$, although the Frobenius bound becomes vacuous at $r = 5$ and $r = 10$. A vacuous bound is inconclusive and does not imply an unstable subspace.

Regression comparison. For regressions of PC1 on PC2 through PC10, the reference slopes range from -0.001 to 0.001. The VAE slopes range from -0.35 to 0.66, while the Gaussian-baseline slopes range from -0.030 to 0.013. These comparisons show that the fitted generators produce different off-diagonal relationships in the fixed PCA representation.

Theorem 2 gives an exact identity only when the marginal variances match. Because exact variance matching has not been established for the VAE and reference samples, this experiment should be presented as a descriptive illustration rather than a direct verification of that theorem.

Relationship with the RV coefficient. Across 200 random subsets of 20 PCA coordinates, the reported Spearman correlations between D_Σ and $1 - \mathrm{RV}$ are -0.383 for the VAE and -0.427 for the Gaussian baseline. These negative correlations show that the two measures do not rank the subsets concordantly. The two-sample KS values compare the marginal distributions of standardized metric values; they do not test agreement between the metrics. Figure 9 should therefore be treated as an exploratory sensitivity analysis rather than validation of either diagnostic.

The correlation-matrix discrepancies are $\widetilde{D}_\Sigma = 3.125$ for the VAE and 0.210 for the Gaussian baseline. These results give the same full-data ordering as the raw covariance distance while removing marginal scale differences.

Summary. In this fitted Fashion-MNIST comparison, the VAE has larger covariance and correlation-matrix discrepancies, larger feature-wise KS distances, and larger observed subspace and regression differences than the fitted-Gaussian baseline. The Gaussian baseline receives non-vacuous subspace

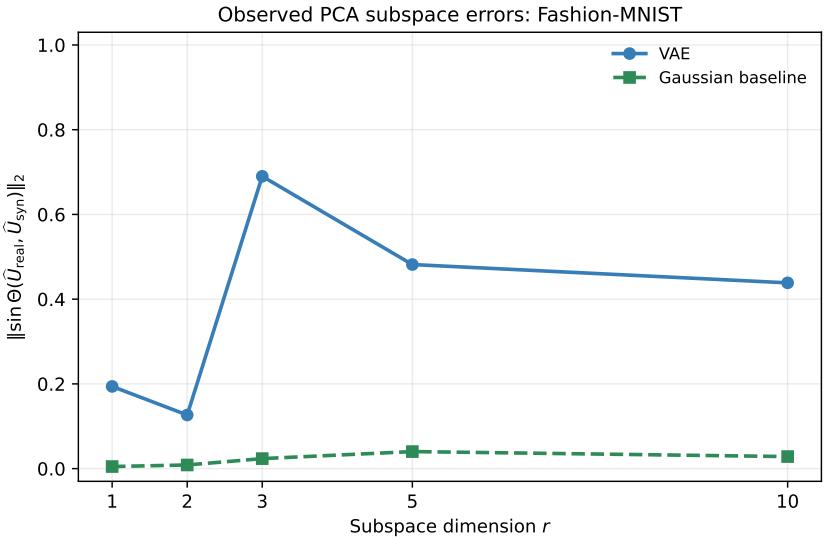


Figure 7: Observed PCA subspace errors across subspace dimensions. The VAE errors are larger than those of the Gaussian baseline, particularly at $r = 3$. The plot reports observed errors only; certificate status is given in Table 2.

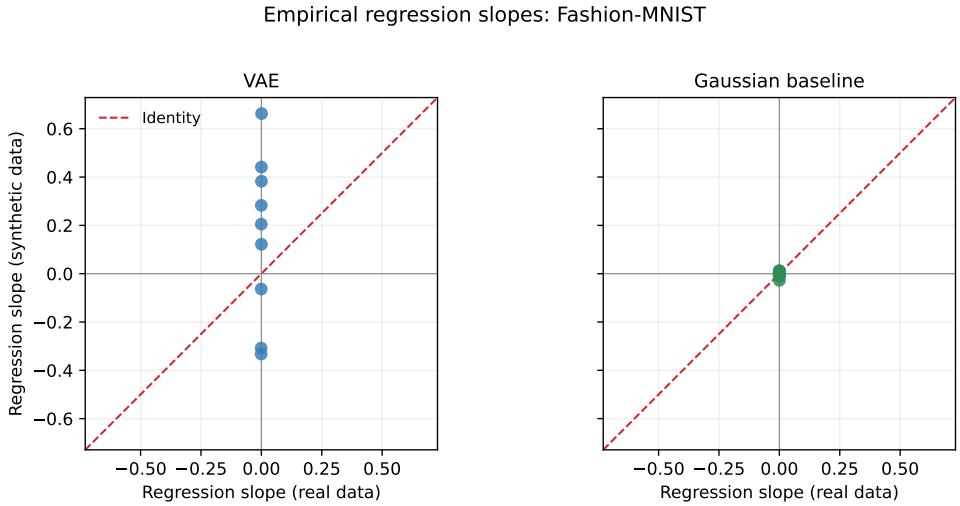


Figure 8: Empirical regression slopes in the real and synthetic Fashion-MNIST PCA representations. The VAE slopes depart more from the identity line than the Gaussian-baseline slopes. The comparison is descriptive because the matched-variance premise of Theorem 2 is not imposed.

certificates for $r = 1, 2, 3$; the remaining Frobenius bounds are inconclusive. The VAE receives no non-vacuous certificate at the reported dimensions.

These results show that the proposed diagnostics can be applied in an image representation. They do not establish that a diagonal conditional posterior necessarily causes covariance loss, nor does the final training loss alone rule out optimization or model-selection effects. An architectural conclusion would require repeated training runs and comparisons with alternative posterior or decoder specifications.

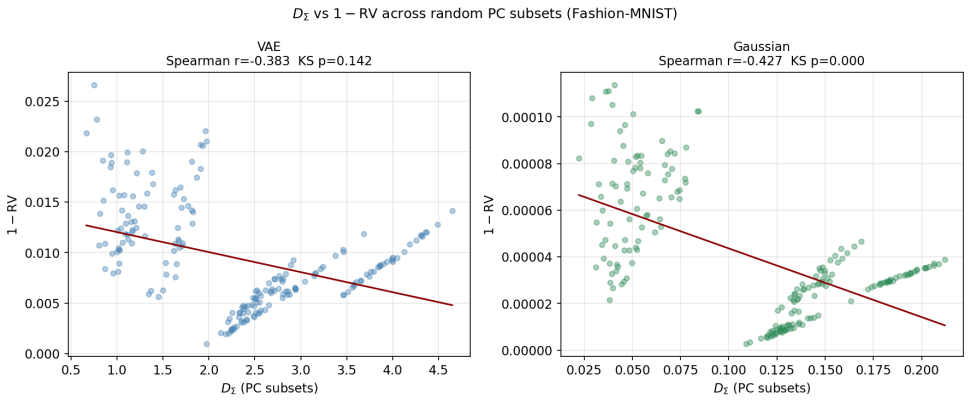


Figure 9: Exploratory comparison of D_Σ and $1 - \mathrm{RV}$ across random PCA-coordinate subsets. The negative Spearman correlations indicate different subset rankings. KS tests of the standardized marginal distributions do not measure agreement between the two diagnostics.

D.3 Alzheimer's Gene Expression: Small-Sample Illustration

This analysis examines the behavior of the sample diagnostics when covariance estimation is less precise. The Alzheimer's gene-expression dataset contains $n = 113$ observations (Srinivasan et al., 2020). We retain the $p = 50$ genes with the largest sample variances, giving $n/p = 2.26$. This ratio indicates that covariance estimates may be noisy, but it is not a universal threshold for reliable or unreliable estimation (Ledoit & Wolf, 2004). All quantities below are therefore treated as descriptive sample estimates.

Generator and diagnostics. The reference data are compared with samples from a multivariate Gaussian fitted to the empirical mean and full sample covariance:

$$X_{\mathrm{syn}} \sim \mathcal{N}\left(\widehat{\mu}_{\mathrm{real}}, \widehat{\Sigma}_{\mathrm{real}}\right). \tag{62}$$

The fitted distribution matches the empirical covariance at the model level, although the covariance of a finite synthetic sample differs because of sampling variation.

The reported estimates, including the RV coefficient (Robert & Escoufier, 1976), are

$$\widehat{D}_\Sigma = 49.35, \qquad \widehat{\widetilde{D}}_\Sigma = 3.69, \qquad \mathrm{RV} = 0.9929, \qquad \mathrm{median(KS)} = 0.26. \tag{63}$$

The raw and correlation-based discrepancies use different scales and should not be compared numerically with the KS or RV measures.

Figure 10 shows the distribution of gene-wise KS distances. Figure 11 shows the sample covariance residual matrix. The heatmap describes differences between the two observed covariance estimates; with this sample size, it should not be interpreted as a precise estimate of the population residual matrix.

Leading-subspace certificate. Using the Davis–Kahan perturbation result (Davis & Kahan, 1970; Yu et al., 2015), the estimated leading eigengap is

$$\widehat{\gamma}_1 = 279.87, \tag{64}$$

so the estimated ratio is

$$\widehat{R}_1 = \frac{\widehat{D}_\Sigma}{\widehat{\gamma}_1} = 0.1763. \tag{65}$$

The corresponding Frobenius-based Davis–Kahan bound is

$$2\widehat{R}_1 = 0.3526, \tag{66}$$

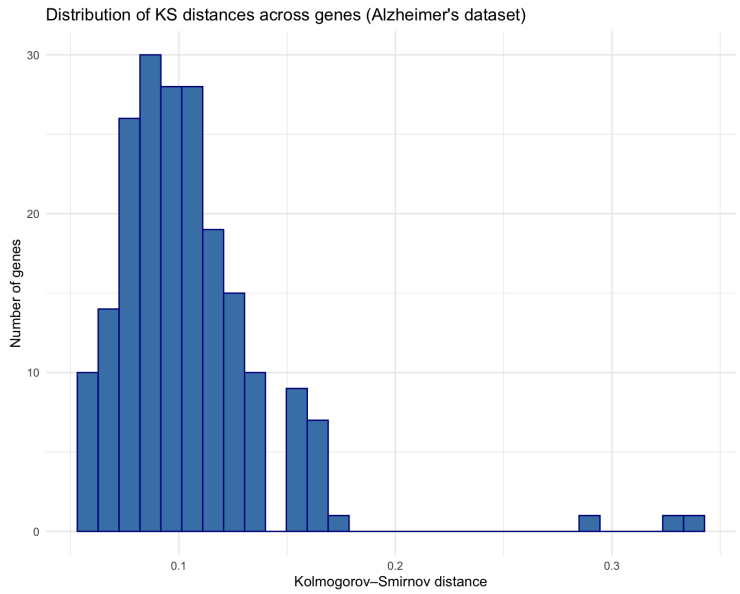


Figure 10: Distribution of gene-wise KS distances for the Alzheimer's dataset. The median value is 0.26. These distances assess marginal fit and do not determine covariance fidelity.

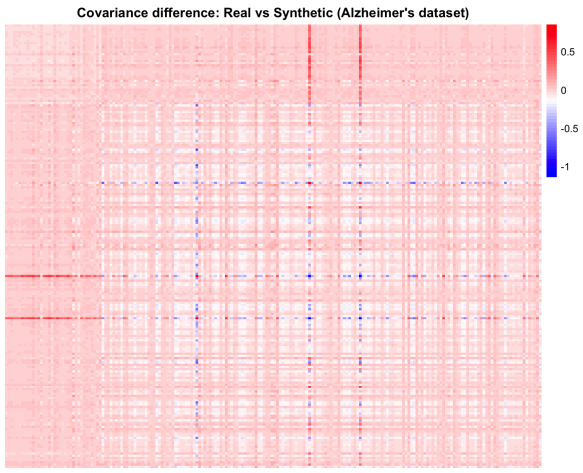


Figure 11: Sample covariance residual matrix $\widehat{\Sigma}_{\text{real}} - \widehat{\Sigma}_{\text{syn}}$ for the Alzheimer's data and the fitted full-covariance Gaussian generator. The visible off-diagonal residuals are descriptive and may contain substantial sampling variation.

not 0.1763. The observed leading-subspace error is 0.0701, which satisfies this within-sample bound. This calculation certifies the relationship between the two estimated covariance matrices. It does not by itself quantify the sampling error between either estimate and its population covariance.

At dimensions $r = 2, 3, 5$, the observed subspace errors are 0.4139, 0.4771, and 0.9997. Local eigengaps and corresponding bounds were not reported for these dimensions, so the larger errors should be described empirically rather than as predictions of Theorem 3.

Bootstrap behavior. Across 500 nonparametric bootstrap resamples (**?**), the reported central 95% bootstrap percentile range for $\widehat{D}_\Sigma$ is $[63.94, 158.51]$, with bootstrap standard deviation 24.52. The plug-in estimate 49.35 lies below this range. Therefore, the percentile range is not presented as a calibrated confidence interval. The result indicates that this resampling procedure is not centered at the observed norm, which is plausible because D_Σ is nonlinear, nonnegative, and sensitive to

Table 3: Descriptive diagnostics for the Alzheimer's dataset ($n = 113$, $p = 50$). The ratio and Davis–Kahan bound refer only to the leading one-dimensional subspace.

Metric	Value	Metric	Value
$\widehat{D}_\Sigma$	49.35	$\widehat{\gamma}_1$	279.87
$\widehat{\widehat{D}}_\Sigma$	3.69	$\widehat{R}_1$	0.1763
RV coefficient	0.9929	$2\widehat{R}_1$	0.3526
Median KS	0.26	Spearman$(D_\Sigma, 1 - \mathrm{RV})$	-0.010
Subspace error, $r = 1$	0.0701	Subspace error, $r = 2$	0.4139
Subspace error, $r = 3$	0.4771	Subspace error, $r = 5$	0.9997

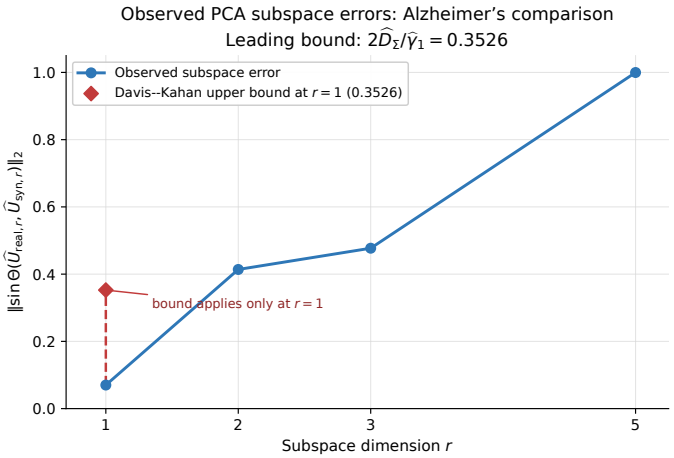


Figure 12: Observed PCA subspace errors for the Alzheimer's comparison. At $r = 1$, the observed error 0.0701 is below the Frobenius-based Davis–Kahan upper bound 0.3526. Results for $r > 1$ are descriptive because their corresponding local-eigengap bounds were not calculated.

covariance-estimation noise. It does not establish a general failure of bootstrap inference or a universal sample-size boundary.

Comparison with the RV coefficient. Across 100 random subsets of 20 genes, the Spearman correlation between D_Σ and $1 - \mathrm{RV}$ is -0.010 with $p = 0.92$. Thus, the two measures produce essentially unrelated subset rankings in this analysis. The reported KS value, $p = 0.581$, compares marginal distributions of standardized metric values; it is not a test of agreement. The near-zero correlation may reflect sampling variability, different scale sensitivity, or both. The present experiment cannot identify the cause.

Regression comparison. The estimated regression slopes from the real and synthetic samples differ in magnitude and sign for several gene pairs. Figure 14 records this empirical comparison. The exact identity in Theorem 2 requires matched marginal variances. Although the fitted Gaussian model uses the empirical variances, finite generated samples do not match them exactly. The figure is therefore an illustration of changed linear relationships rather than a direct verification of the theorem.

Summary. The leading estimated subspace satisfies a non-vacuous within-sample perturbation bound after the missing factor of two is restored. The remaining results show that the covariance, marginal, RV, and subspace diagnostics can give different views of the same small dataset. They do not establish that one diagnostic remains population-valid at $n/p = 2.26$. A stronger small-sample analysis would require repeated sampling (Ledoit & Wolf, 2004), shrinkage estimators, or a simulation with known population covariance.

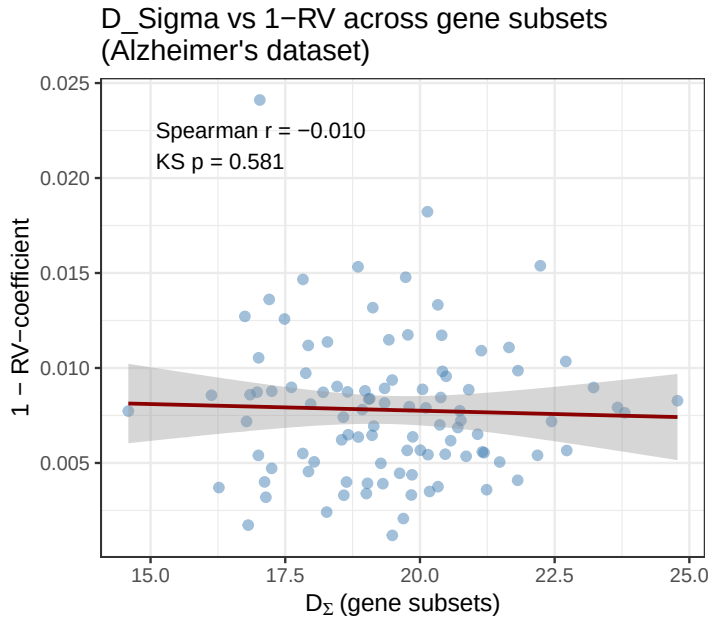


Figure 13: Exploratory comparison of D_Σ and $1 - \mathrm{RV}$ across random gene subsets. The near-zero Spearman correlation shows that the two diagnostics do not rank these subsets similarly. No causal explanation is assigned to this pattern.

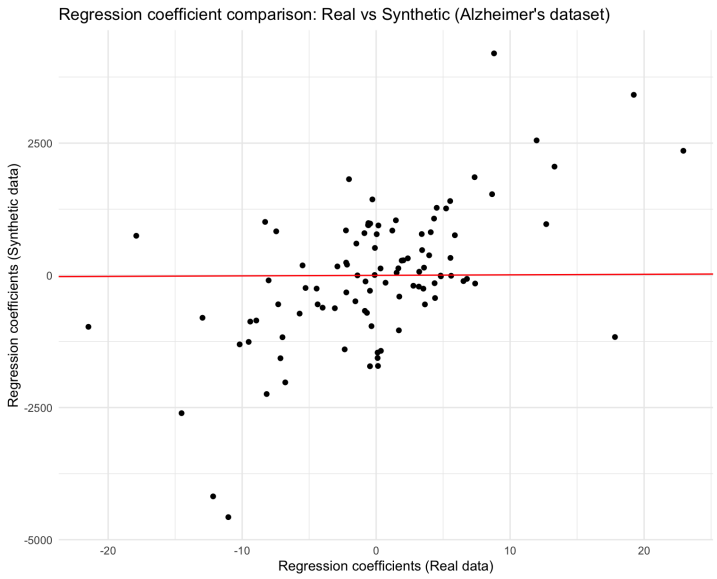


Figure 14: Regression slopes estimated from real and synthetic Alzheimer's data. Departures from the identity line show differences in the fitted linear relationships. Their magnitudes should be interpreted cautiously because the covariance estimates are based on $n = 113$ observations.

D.4 Diagonal VAE Posteriors: Scope of the Architectural Claim

A standard VAE often uses the following conditional approximate posterior (Kingma & Welling, 2014; Rezende et al., 2014):

$$q_\phi(z \mid x) = \mathcal{N}\big(\mu_\phi(x), \mathrm{diag}(\sigma_\phi^2(x))\big)\,. \tag{67}$$

The conditional covariance is diagonal for each fixed x. This fact does not imply that the aggregated posterior or generated observations have diagonal covariance.

By the law of total covariance, the aggregated posterior

$$q_\phi(z) = \mathbb{E}_{p_{\text{data}}(x)}[q_\phi(z \mid x)] \tag{68}$$

has covariance

$$\text{Cov}_{q_\phi}(z) = \mathbb{E}_{p_{\text{data}}(x)}\left[\text{diag}(\sigma_\phi^2(x))\right] + \text{Cov}_{p_{\text{data}}(x)}(\mu_\phi(x)) \,. \tag{69}$$

The second term can be dense. A nonlinear decoder can also produce dense covariance in data space. Consequently, no general lower bound on data-space D_Σ follows from the diagonal form of $q_\phi(z \mid x)$ alone.

Idealized diagonal-covariance comparison. The following construction is useful as a toy model, but it is not a theorem about VAEs. Suppose one distribution has covariance

$$\Sigma_P = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \tag{70}$$

while a diagonal approximation uses

$$\Sigma_Q = \text{diag}(\Sigma_P) = I_2. \tag{71}$$

Both Gaussian distributions have standard-normal marginals, and

$$D_\Sigma(P, Q) = \|\Sigma_P - \Sigma_Q\|_F = \sqrt{2}\,|\rho|. \tag{72}$$

For $\rho = 0.8$, the discrepancy is approximately 1.131.

Under the matched unit variances, Theorem 2 gives

$$|\beta(P) - \beta(Q)| = \frac{D_\Sigma(P, Q)}{\sqrt{2}} = |\rho| = 0.8. \tag{73}$$

This statement concerns the idealized Gaussian pair. It should not be described as a necessary regression effect of a diagonal-posterior VAE.

For PCA, the eigengap of Σ_P is

$$\gamma_1(P) = 2\rho = 1.6. \tag{74}$$

Therefore,

$$R_1(P, Q) = \frac{\sqrt{2}\rho}{2\rho} = \frac{1}{\sqrt{2}} \approx 0.707, \tag{75}$$

and the Frobenius-based Davis–Kahan bound (Davis & Kahan, 1970; Yu et al., 2015) is

$$2R_1(P, Q) = \sqrt{2} \approx 1.414. \tag{76}$$

The bound is vacuous because $R_1 > 1/2$. The condition $D_\Sigma < \gamma_1$ is not enough to make the bound informative.

Moreover, $\Sigma_Q = I_2$ has a repeated leading eigenvalue, so its one-dimensional leading eigenspace is not identifiable. Any unit vector is a valid leading eigenvector. Depending on that arbitrary choice, the sine of the angle relative to the leading eigenvector of Σ_P can range from zero to one. Choosing a coordinate axis gives $1/\sqrt{2}$, but this is neither unique nor the worst case.

Connection to the Fashion-MNIST result. The Fashion-MNIST experiment reports larger covariance discrepancy for one trained VAE than for the fitted-Gaussian baseline. This observation is compatible with several explanations involving the encoder, decoder, latent dimension, regularization, optimization, or model selection (Dai & Wipf, 2019). The experiment does not isolate the diagonal conditional posterior as the cause. Its leading ratio $R_1 = 0.64$ also gives a vacuous Frobenius bound, so it cannot be used to classify the VAE as stable or unstable.

An architectural claim would require repeated training runs and a controlled comparison with alternative posterior structures, such as full-covariance or low-rank families (Louizos & Welling, 2016), while holding the decoder and training procedure fixed. Until such evidence is available, the defensible conclusion is empirical: this fitted VAE did not preserve the covariance structure of the chosen Fashion-MNIST representation as closely as the Gaussian baseline.

Summary. Diagonal conditional covariance is a modeling restriction, but it does not force diagonal aggregate or data-space covariance. The idealized diagonal approximation gives a clear example of regression sensitivity and an inconclusive PCA bound. It should remain separate from the empirical Fashion-MNIST result. The practical role of D_Σ is diagnostic: it measures the covariance discrepancy of generated samples after training without assigning that discrepancy to a particular architectural component.

D.5 Dependence Fidelity in TCGA-BRCA Bulk RNA-seq Data

This analysis applies the sample diagnostics to TCGA breast invasive carcinoma bulk RNA-seq data. The reference matrix contains $n = 1{,}111$ primary-tumor samples and the $p = 100$ genes with the largest sample variances. Raw counts were obtained through TCGAbiolinks (Colaprico et al., 2016). A variance-stabilizing transformation was then applied in DESeq2 (Colaprico et al., 2016; Love et al., 2014). The resulting ratio is $n/p = 11.1$.

Two synthetic-data procedures are compared. Splatter fits gene-wise count models and does not explicitly target the full empirical gene-gene covariance (Zappia et al., 2017). Poisson thinning resamples the observed counts through binomial draws and therefore retains part of the dependence inherited from the reference observations (Gerard, 2020). Poisson thinning is best viewed here as a structure-preserving control rather than as an independently trained generative model. This distinction limits causal comparisons between the two procedures.

Splatter was developed for single-cell RNA-seq and is used here only as a deliberately contrasting gene-wise count baseline, not as a validated simulator for bulk RNA-seq.

Marginal diagnostics. The median gene-wise KS distances are 0.127 for Splatter and 0.110 for Poisson thinning. The difference is 0.017 in absolute terms. Figure 15 reports the distributions across genes. These values describe marginal fit only and do not determine covariance fidelity.

Covariance diagnostics. The reported sample covariance discrepancies are

$$\widehat{D}_\Sigma(\text{Splatter}) = 249.79, \qquad \widehat{D}_\Sigma(\text{Poisson}) = 88.35. \tag{77}$$

The corresponding correlation-matrix discrepancies and RV coefficients are

$$\widehat{\widetilde{D}}_\Sigma(\text{Splatter}) = 26.58, \qquad \text{RV}(\text{Splatter}) = 0.299, \tag{78}$$

$$\widehat{\widetilde{D}}_\Sigma(\text{Poisson}) = 6.54, \qquad \text{RV}(\text{Poisson}) = 0.961. \tag{79}$$

Both scale-sensitive and scale-normalized summaries show a larger sample discrepancy for Splatter. Figure 16 shows the corresponding residual matrices. Different panel color scales must be retained in the caption and color bars.

Leading PCA subspace. The reported leading reference eigengap is $\widehat{\gamma}_1 = 121.67$. Therefore,

$$\widehat{R}_1(\text{Splatter}) = \frac{249.79}{121.67} = 2.053, \qquad 2\widehat{R}_1(\text{Splatter}) = 4.106, \tag{80}$$

$$\widehat{R}_1(\text{Poisson}) = \frac{88.35}{121.67} = 0.726, \qquad 2\widehat{R}_1(\text{Poisson}) = 1.452. \tag{81}$$

Thus, the value 0.726 previously reported for Poisson thinning is the ratio $\widehat{D}_\Sigma/\widehat{\gamma}_1$. Splatter lies outside the small-perturbation condition $\widehat{R}_1 < 1$. Poisson thinning satisfies that condition, but its Frobenius bound exceeds one and is therefore vacuous. Neither comparison receives a non-vacuous leading-subspace certificate from this bound.

The observed leading-subspace errors are 0.9985 for Splatter and 0.2099 for Poisson thinning. The smaller Poisson value is an empirical observation, not a consequence certified by the reported bound. At $r = 2, 3, 5, 10$, the observed errors are shown in Table 4. Because the corresponding local eigengaps are not reported consistently, no theorem-based classification is assigned at those dimensions.

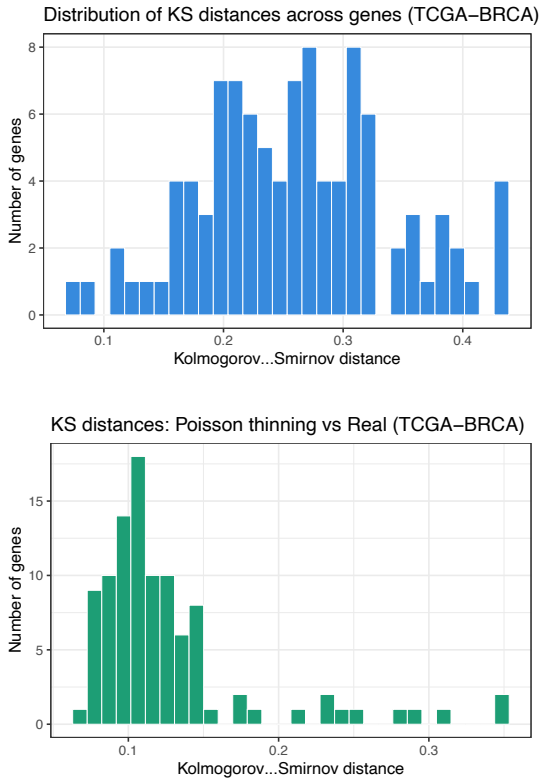


Figure 15: Gene-wise KS distances for Splatter and Poisson-thinned samples relative to TCGA-BRCA. The distributions overlap, although their medians differ. KS distances assess marginal distributions and do not certify joint covariance fidelity.

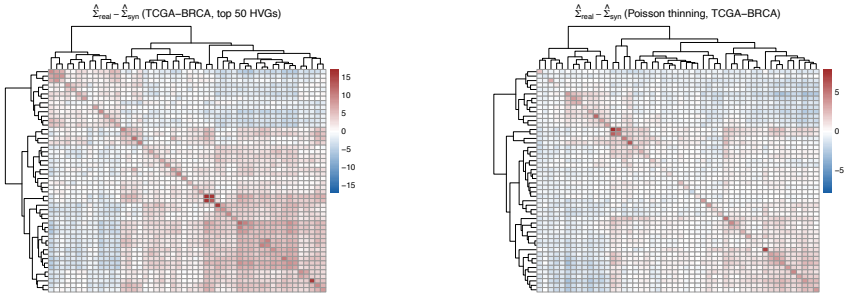


Figure 16: Sample covariance residuals $\widehat{\Sigma}_{\text{real}} - \widehat{\Sigma}_{\text{syn}}$ for Splatter and Poisson thinning on TCGA-BRCA. Splatter has the larger Frobenius discrepancy. The panels use different color ranges and should not be compared by color intensity alone.

Regression comparison. Figure 18 compares slopes estimated from the real and synthetic samples. Splatter departs more strongly from the identity line than Poisson thinning. Theorem 2 gives an exact identity for a two-variable population comparison with matched marginal variances. Those premises are not imposed exactly by the present finite-sample analysis. The plot should therefore be presented as a descriptive downstream comparison rather than an empirical proof of the theorem.

Table 4: Observed PCA subspace errors for TCGA-BRCA. The leading Frobenius bounds are vacuous for both procedures. Values for $r > 1$ are descriptive because matched local eigengaps and bounds are not reported.

r	Splatter	Poisson thinning
1	0.9985	0.2099
2	0.9978	0.4371
3	0.9943	0.4085
5	1.0000	0.8749
10	0.9983	0.9579

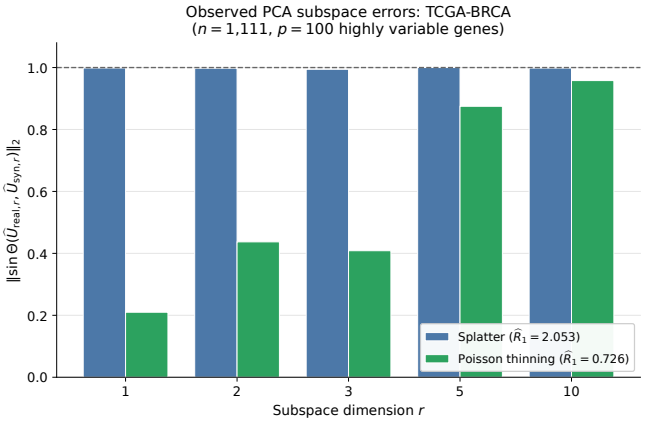


Figure 17: Observed PCA subspace errors for Splatter and Poisson thinning. Splatter is nearly orthogonal to the reference subspace at every displayed dimension. Poisson thinning has smaller errors at the leading dimensions, although its reported leading Frobenius bound is vacuous.

Bootstrap and RV sensitivity analyses. For 500 nonparametric bootstrap resamples, the reported central 95% percentile ranges for $\widehat{D}_\Sigma$ are $[237.10, 265.62]$ for Splatter and $[82.56, 107.78]$ for Poisson thinning, with bootstrap standard deviations 7.28 and 6.60. The plug-in estimates lie within their respective ranges, and the ranges do not overlap. This shows separation under the stated resampling procedure. It should not be described as proof of a population-level difference without a validated inferential argument for the nonlinear norm statistic.

Across 100 random subsets of 30 genes, the reported Spearman correlations between D_Σ and $1 - \mathrm{RV}$ are 0.748 for Splatter and -0.195 for Poisson thinning. The correlations show that the two measures rank subsets similarly for Splatter but not for Poisson thinning. The reported two-sample KS values compare standardized marginal distributions of the two metric values; they do not test agreement or interchangeability. No causal explanation is assigned to the difference between the two subset analyses.

Summary. Splatter has larger covariance and correlation-matrix discrepancies, lower RV similarity, and larger observed subspace and regression differences than Poisson thinning. The comparison is informative because both procedures are evaluated against the same reference data. It does not isolate a single causal mechanism, and Poisson thinning remains a resampling-based control rather than a directly comparable fitted generator. The leading Davis–Kahan bounds are vacuous for both procedures, so the observed difference in subspace error remains empirical.

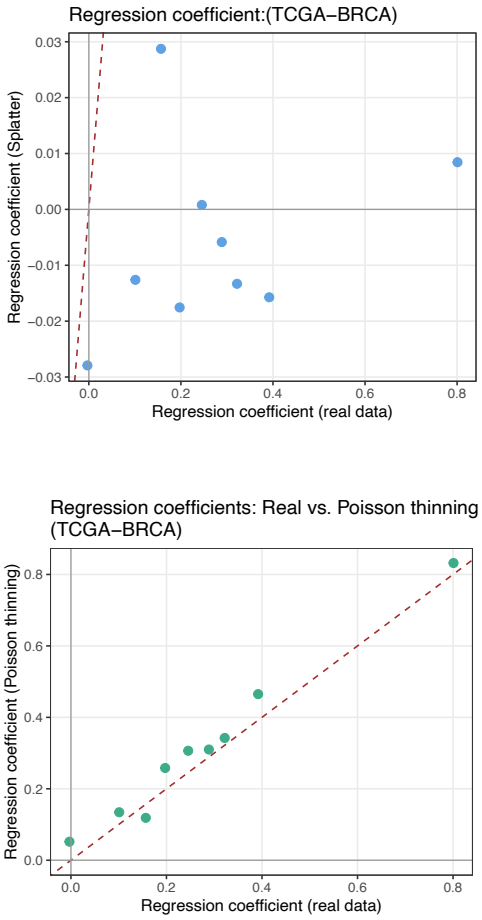


Figure 18: Regression slopes estimated from real and synthetic TCGA-BRCA data. Splatter shows larger departures from the identity line, including sign changes, whereas Poisson thinning follows the reference slopes more closely. The comparison is descriptive because exact finite-sample variance matching is not imposed.

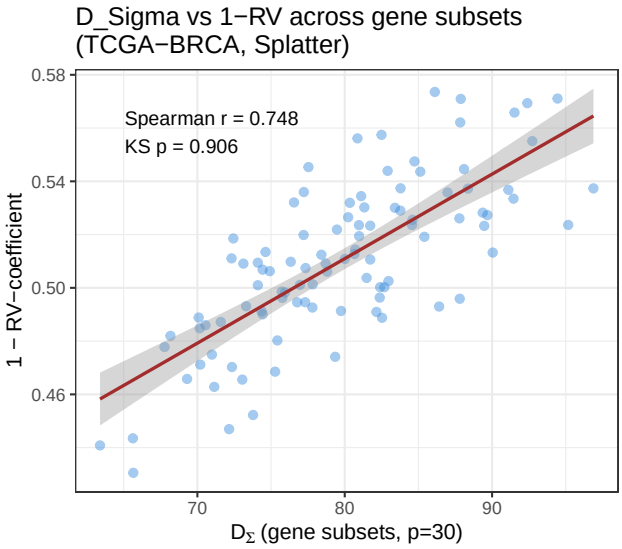


Figure 19: D_Σ versus $1 - \mathrm{RV}$ across 100 random gene subsets of size 30 (TCGA-BRCA, Splatter, $n = 1{,}111$, $p = 100$). Spearman $r = 0.748$ ($p < 0.001$), KS $p = 0.906$. The strong positive concordance and indistinguishable distributions after standardisation confirms that both metrics detect the same widespread covariance failure in the large-divergence unstable regime ($D_\Sigma/\delta = 2.05$). This is the strongest concordance result across all datasets in this study.

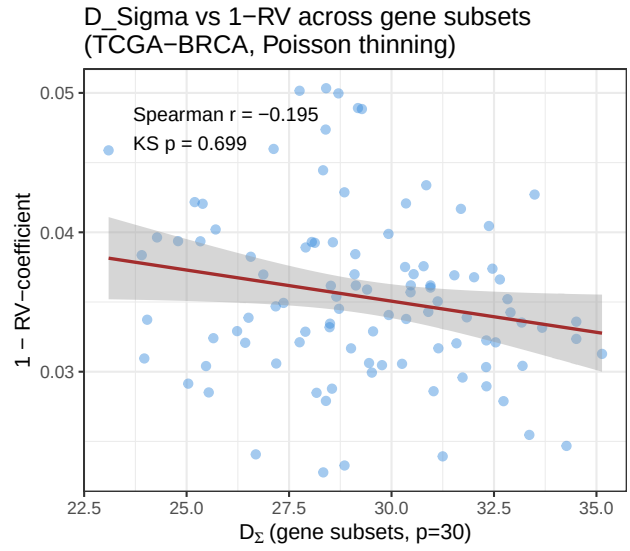


Figure 20: Exploratory comparison of D_Σ and $1 - \mathrm{RV}$ across random gene subsets. The positive rank association for Splatter and weak negative association for Poisson thinning show that the two diagnostics need not rank subsets in the same way.

D.6 Summary Across Datasets

Table 5 collects the reported leading-subspace quantities. It distinguishes the perturbation ratio

$$\widehat{R}_1 = \frac{\widehat{D}_\Sigma}{\widehat{\gamma}_1} \tag{82}$$

from the Frobenius Davis–Kahan bound $2\widehat{R}_1$. A non-vacuous certificate requires $2\widehat{R}_1 < 1$. A value below one for $\widehat{R}_1$ alone does not imply an informative bound.

Marginal and covariance diagnostics. The empirical comparisons show that marginal and covariance diagnostics can rank synthetic procedures differently. The strongest example is TCGA-BRCA, where the KS distributions overlap while the covariance, RV, subspace, and regression summaries differ. Fashion-MNIST is not a marginal-blindness example because the fitted Gaussian baseline also has the smaller median KS distance. The exact impossibility claim remains supplied by Theorem 1 and the synthetic constructions, not by every empirical dataset.

Regression comparisons. In the comparisons examined, procedures with larger sample covariance discrepancies also show larger slope differences. These plots are descriptive because the matched-variance premises of Theorem 2 are not imposed for every finite-sample pair. The theorem supplies the analytic result; the plots illustrate related behavior.

PCA certificates. Non-vacuous leading-subspace certificates are obtained for the Fashion-MNIST Gaussian baseline and the Alzheimer's comparison. The VAE and Poisson-thinning bounds are vacuous, while Splatter falls outside the stated small-perturbation condition. Observed subspace errors remain reportable in all cases, but a vacuous upper bound cannot be used to classify a procedure as stable or unstable.

Overall, the results support using covariance diagnostics alongside marginal measures. They do not establish a universal decision threshold for all generators or downstream tasks. The relevant eigengap and the assumptions of the downstream perturbation result must be checked for each analysis.

D.7 Relation to MMD and FID

MMD on TCGA-BRCA. A characteristic-kernel maximum mean discrepancy is a population metric for equality of joint distributions. In finite samples, its numerical scale depends on the kernel, bandwidth, estimator, sample size, and dimension. The reported Gaussian-kernel MMD values are 0.338 for Splatter and 0.090 for Poisson thinning. These values separate the two procedures, but they do not provide a PCA perturbation certificate.

FID on Fashion-MNIST. The reported FID values, computed from 10,000 images, are 72.02 for the VAE and 192.39 for the fitted-Gaussian baseline. Lower FID indicates that the VAE is closer under the chosen feature representation. This comparison concerns image-feature similarity, whereas D_Σ is computed in the selected PCA representation and concerns second-order structure. The two measures should therefore be reported as complementary rather than contradictory.

FID based on an ImageNet-pretrained Inception network has a domain mismatch for grayscale Fashion-MNIST images. The feature extractor, resizing and channel-conversion steps, implementation, subsampling seed, and estimator must be stated before the numerical values can be reproduced or interpreted. The comparison does not show that either model is generally superior, and it does not identify the VAE posterior family as the cause of the covariance result.

Table 5: Leading-subspace summary. "Non-vacuous" means that the reported Frobenius Davis–Kahan upper bound is below one. The Alzheimer's result is a within-sample comparison and should be interpreted cautiously because $n = 113$ and $p = 50$.

Dataset	Procedure	Median KS	$\widehat{R}_1$	$2\widehat{R}_1$	Error	Certificate
FMNIST	VAE	0.090	0.639	1.278	0.1940	None
FMNIST	Gaussian	0.020	0.0326	0.0652	0.0049	Yes
TCGA	Splatter	0.127	2.053	4.106	0.9985	None
TCGA	Poisson	0.110	0.726	1.452	0.2099	None
Alzheimer's	Full-cov. Gaussian	0.260	0.1763	0.3526	0.0701	Yes

Table 6: Relationship between the theoretical statements and the empirical analyses. Exact theoretical results are separated from descriptive empirical evidence.

Result	Exact statement	Role of empirical analysis
Theorem 1	Matched univariate marginals do not determine the copula or covariance.	Synthetic constructions give exact examples; the datasets show that marginal and covariance diagnostics can disagree in practice.
Theorem 2	Under matched marginal variances, the bivariate population slope difference is determined by the covariance perturbation.	Regression plots are descriptive because exact finite-sample variance matching is not imposed throughout.
Theorem 3	Covariance perturbation bounds eigenvalue error and, under spectral separation, principal-subspace error.	Two reported leading comparisons have non-vacuous Frobenius certificates; the remaining leading bounds are vacuous or outside the assumption.
Remark 1	Covariance does not capture tail or general nonlinear dependence.	The Gaussian versus t-copula construction illustrates a tail-dependence difference under matched marginals.

Table 7: Reported marginal KS, joint MMD, and leading covariance-to-eigengap ratios for TCGA-BRCA. The Poisson-thinning ratio is 0.726.

Procedure	Median KS	MMD	$\widehat{D}_\Sigma/\widehat{\gamma}_1$
Poisson thinning	0.110	0.090	0.726
Splatter	0.127	0.338	2.053

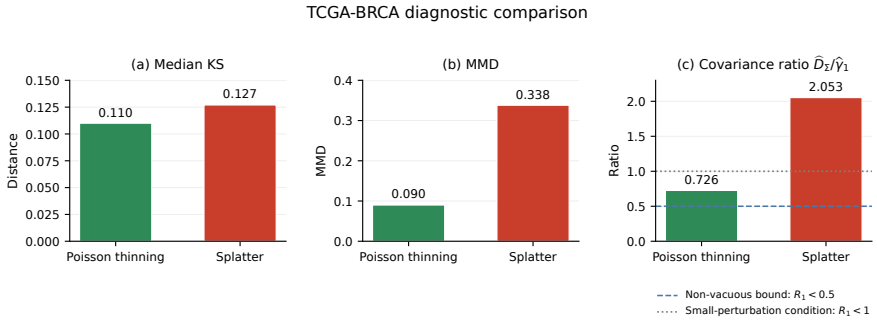


Figure 21: Comparison of median KS distance, MMD, and the leading covariance-to-eigengap ratio for TCGA-BRCA. The metrics have different scales and purposes. Only the ratio enters the reported PCA perturbation calculation, and its corresponding bound is $2\widehat{R}_1$.

E Reproducibility Details

For double-blind review, the code and saved intermediate results are supplied through anonymous supplementary material. After the review period, a permanent versioned repository will be released.

The reproducibility materials include:

- data accession identifiers, download dates, sample filters, gene-selection rules, and transformation code;
- the Fashion-MNIST split, PCA fitting procedure, VAE architecture, optimizer, learning rate, batch size, latent dimension, epoch count, and sample-generation settings;
- Python and R package requirements and recorded software versions;
- random seeds for model initialization, minibatch ordering, synthetic sampling, bootstrap resampling, and random-subset analyses;
- implementations of D_Σ, $\widetilde{D}_\Sigma$, the RV coefficient, subspace error, regression slopes, KS distances, and MMD;
- the MMD estimator and bandwidth-selection rule; and
- scripts for regenerating the reported tables and figures from saved intermediate results.

The analyses use Python with NumPy, SciPy, scikit-learn, and PyTorch, together with R, TCGAbiolinks, DESeq2, Splatter, and the binomial-thinning implementation. Exact package versions are recorded in the supplied environments.

The nonparametric bootstrap uses $B = 500$ resamples. The resulting intervals are reported as central bootstrap percentile ranges rather than calibrated confidence intervals. In the Alzheimer's analysis, the plug-in estimate lies outside the percentile range, and the range is therefore treated only as a resampling summary. Fixed seeds improve repeatability but do not guarantee bitwise agreement across hardware, package versions, or parallel backends.

F Ethical Considerations and Broader Impacts

Fashion-MNIST, TCGA-BRCA, and GSE125050 are existing research resources. No new human-participant data were collected for this study. Their access conditions, citation requirements, and restrictions on redistribution remain applicable. Although the gene-expression data are de-identified, responsible handling and reporting are still required.

Covariance diagnostics may help identify synthetic data that preserve individual-variable distributions while changing linear relationships or principal subspaces. This is relevant when generated data are used for regression, dimension reduction, clustering, network analysis, or other procedures that depend on multivariate structure.

The diagnostic has clear limits. A small covariance discrepancy does not establish correct nonlinear, conditional, temporal, spatial, or tail dependence. A large raw Frobenius discrepancy may also reflect variable scale. The correlation-based version, eigengap, and downstream task should therefore be reported alongside the raw discrepancy when appropriate.

No single statistic should be treated as a general certificate of synthetic-data quality. Covariance diagnostics should be considered together with marginal, higher-order, privacy, and task-specific evaluations. This is particularly important in health and genomic applications, where synthetic data may retain sensitive structure or be used outside the conditions allowed by the source data.