λ-PSD: Scalable Approximate SNR-Optimised Polynomial Stein Discrepancies

Minh-Long Nguyen [1 2] **Thanh-Long Vu** [1 3] **Christopher Drovandi** [1 4 2] **Leah F. South** [1 2] **Trung-Tin Nguyen** [1 4 2]

Abstract

Polynomial Stein discrepancies (PSD) provide a scalable alternative to kernel Stein methods for measuring sample quality and goodness-of-fit testing, but their statistical properties remain poorly understood. We show that increasing polynomial degree primarily amplifies signal without adequately controlling variance, rather than directly optimising the signal-to-noise ratio (SNR). Under suitable assumptions, this might lead to a failure mode in which the SNR^2 can provably decay exponentially with polynomial degree. Motivated by this observation, we reformulate Stein discrepancy construction as an explicit SNR^2 maximisation problem, yielding a Rayleigh quotient over Stein features. This perspective motivates λ-PSD, an approximate scalable covariance-aware reweighting scheme defined in a low-dimensional subspace. Under Gaussian settings, we show that λ-PSD avoids the exponential SNR^2 collapse and achieves a stable SNR^2. Empirically, λ-PSD substantially improves test power while retaining linear-time complexity in the number of samples, highlighting the importance of SNR-aware design for scalable Stein discrepancies.

Keywords: Computational statistics · Stein's method · sample quality measure · hypothesis testing · scalable Monte Carlo

1. Introduction

Assessing sampling quality remains a fundamental problem in modern Bayesian computation (Bhattacharya et al., 2024). In many settings, the target distribution p is known only up to a normalising constant, while the approximation q is represented through samples. Stein discrepancies provide a principled framework for directly measuring discrepancy between p and q using only score evaluations of the target distribution (Gorham & Mackey, 2015). Among these, kernel Stein discrepancy (KSD; Chwialkowski et al. 2016; Liu et al. 2016; Gorham & Mackey 2017) have strong theoretical guarantees but typically require quadratic computational complexity in the number of samples n, limiting scalability in large-scale settings (Jitkrittum et al., 2017). Recent work on polynomial Stein discrepancy (PSD; Srinivasan et al. 2025) addresses this limitation by replacing kernel witness functions with polynomial Stein features of order Q, yielding computationally efficient linear-time discrepancy measures (see Section 2 for further discussion of linear-time Stein discrepancies). PSD is particularly effective at detecting moment convergence under the Bernstein-von Mises limit (Srinivasan et al., 2025). However, its statistical power is highly sensitive to the choice of polynomial order Q, and the statistical mechanisms underlying this behaviour remain poorly understood.

1.1. Contributions

Our main contributions are summarised as follows:

- We identify and theoretically characterise a previously unstudied limitation of PSD, showing that the asymptotic scaling of SNR^2 may decay exponentially with polynomial degree under Gaussian settings.
- We introduce an SNR^2-optimised Stein discrepancy framework that reformulates discrepancy construction as a Rayleigh quotient optimisation problem, leading to the proposed λ-PSD estimator, which achieves a stable SNR^2 under Gaussian settings; we further discuss extensions and generalisations of the λ-PSD framework.
- We demonstrate through simulation studies that λ-PSD substantially improves robustness to the choice of Q and achieves competitive goodness-of-fit testing performance while retaining linear-time complexity.

2. Related Works

(Kernel) Stein Discrepancies Stein discrepancies provide a principled framework for directly measuring discrepancy

[1]School of Mathematical Sciences, Queensland University of Technology, Australia [2]Centre for Data Science, Queensland University of Technology, Australia [3]Quantium, Brisbane, Australia [4]ARC Centre of Excellence for the Mathematical Analysis of Cellular Systems (MACSYS). Correspondence to: Minh-Long Nguyen <minhlong.nguyen@hdr.qut.edu.au>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

between p and q using only score evaluations of the target distribution (Gorham & Mackey, 2015). By combining Stein operators (Stein, 1972) with rich function classes, they construct integral probability metrics (Müller, 1997) that goes to zero when the sampling distribution matches the target. Stein discrepancies have since been widely applied to goodness-of-fit testing, importance sampling, and model training, to just name a few (Liu et al., 2016; Liu & Wang, 2016; Liu & Lee, 2017; Grathwohl et al., 2020).

A large body of work has focused on kernel Stein discrepancy (KSD), which combines the first-order Langevin-Stein operator with reproducing kernel Hilbert spaces to obtain computationally tractable discrepancy measures (Liu et al., 2016; Chwialkowski et al., 2016; Gorham & Mackey, 2017). Under suitable kernels and regularity conditions on p, KSD is convergence determining, meaning that vanishing discrepancy implies weak convergence of the empirical distribution to the target distribution (Gorham & Mackey, 2017). However, standard KSD estimators require quadratic computational complexity due to pairwise kernel evaluations, limiting scalability in large-sample settings (Jitkrittum et al., 2017).

Goodness-of-Fit Testing In goodness-of-fit testing[1], many Stein-based methods, particularly KSD-based tests, rely on asymptotic null distributions or (wild) bootstrap procedures (Leucht & Neumann, 2013; Chwialkowski et al., 2016; Liu et al., 2016). A smaller class of methods instead uses asymptotically normal test statistics, enabling z-test-style calibration (Grathwohl et al., 2020; Repasky et al., 2023). While some Stein-based tests rely on heuristic constructions, such as the widely used median heuristic for Gaussian kernels (Garreau et al., 2017), many approaches explicitly optimise test power through adaptive feature or kernel design (Jitkrittum et al., 2017; Grathwohl et al., 2020).

Linear-Time Stein Discrepancies. Standard KSD estimators typically incur quadratic computational complexity due to pairwise kernel evaluations. This has motivated scalable alternatives such as the finite-set Stein discrepancy (FSSD; Jitkrittum et al. 2017), random-feature Stein discrepancy (RFSD; Huggins & Mackey 2018) and polynomial Stein discrepancy (PSD; (Srinivasan et al., 2025), all of which achieve (near) linear-time complexity through finite-dimensional feature constructions. FSSD relies on carefully tuned test locations and kernel bandwidths, typically requiring sample splitting (Jitkrittum et al., 2017), while RFSD depends on several example-specific design choices and may perform poorly under indirect sampling (Srinivasan et al., 2025). Unlike most methods discussed here, PSD specifically targets moment discrepancies. Its power need not deteriorate with increasing d when the mismatch is distributed across many coordinates (Srinivasan et al., 2025). However, PSD is sensitive to the choice of Q, and the mechanisms underlying this behaviour remain poorly understood.

[1]We consider the hypothesis testing problem $H_0 : q = p$ against $H_1 : q \neq p$.

3. Polynomial Stein Discrepancy in $\mathcal{L}_1$

We first introduce some common notation used throughout this section. Let p denote the target probability measure on $\mathbb{R}^d$ with continuously differentiable density. We assume the score function $s_p := \nabla \log p$ can be evaluated, although direct integration under p is generally intractable (Gorham & Mackey, 2017). Let q denote another probability measure on $\mathbb{R}^d$ with continuously differentiable density, known through samples $\mathbf{x} = x_1, \ldots, x_n$. We are interested in quantifying the discrepancy between q and p through $\mathbf{x}$. Most results in this section are stated under the assumption that $x_1, \ldots, x_n$ are i.i.d. samples; although this assumption can often be relaxed, it does not restrict the methods to i.i.d. sampling settings.

We denote by $\langle \cdot, \cdot \rangle_F$ the Frobenius inner product between two matrices of the same dimension. The corresponding Frobenius norm is denoted by $\| \cdot \|_F$. For a vector v, $\mathrm{diag}(v)$ denotes the diagonal matrix with entries of v on its diagonal. Let $\mathrm{vec}(\cdot)$ denote the vectorisation operator that stacks the columns of a matrix into a single vector. Finally, let $z_{[i]}$ denote the ith component of a vector z.

3.1. Definitions

We consider the first-order Langevin-Stein operator

$$\mathcal{L}_1 h(x) = s_p(x)^\top h(x) + \nabla \cdot h(x), \tag{1}$$

for a vector-valued test function $h : \mathbb{R}^d \to \mathbb{R}^d$. To obtain a computationally efficient class of test functions, we restrict h to a non-interacting polynomial basis

$$h(x) = \sum_{i=1}^{Q} A_i x^{\odot i} + b, \tag{2}$$

where $A_i \in \mathbb{R}^{d \times d}$, $b \in \mathbb{R}^d$, and $\odot$ denotes the Hadamard product. This class admits a degenerate case when $Q = 0$, in which the test function reduces to the constant vector field $h(x) = b$.

It is possible to show that

$$\mathcal{L}_1 h(x) = \sum_{i=1}^{Q} \langle A_i, \Phi_i(x) \rangle_F + b^\top s_p(x), \tag{3}$$

where the feature maps $\Phi_i(x) \in \mathbb{R}^{d \times d}$ are given by

$$\Phi_i(x) = s_p(x)(x^{\odot i})^\top + i\,\mathrm{diag}(x^{\odot (i-1)}). \tag{4}$$

We define the following polynomial Stein discrepancy, denoted by $\mathrm{PSD}_{\mathcal{L}_1}$, to distinguish it from the $\mathrm{PSD}_{\mathcal{L}_2}$ introduced in Srinivasan et al. (2025), which is based on the second-order Langevin-Stein operator. Although the constructions differ, analogous generalisations are expected to hold. To the best of our knowledge, neither this $\mathcal{L}_1$ formulation nor its statistical properties have been previously derived or analysed.

Definition 3.1. The polynomial Stein discrepancy (PSD) associated with $\mathcal{L}_1$ is defined as

$$\mathrm{PSD}_{\mathcal{L}_1}(q \,\|\, p) := \sup_{\{A_i\}_{i=1}^{Q},\, b} \mathbb{E}_q\big[\mathcal{L}_1 h(x)\big], \tag{5}$$

such that $\sum_{i=1}^{Q} \|A_i\|_F^2 + \|b\|_2^2 \leq 1$.

Under regularity condition on tails of p (see South et al. (2022a,b)) [2], if $q = p$, then $\mathbb{E}_p[\mathrm{PSD}_{\mathcal{L}_1}(q \,\|\, p)] = 0$; hence, deviations from zero may indicate a discrepancy between p and q.

The maximiser of $\mathrm{PSD}_{\mathcal{L}_1}(q \,\|\, p)$ admits a closed-form expression:

Proposition 3.2. *Let $\mu_i = \mathbb{E}_q[\Phi_i(x)] \in \mathbb{R}^{d\times d}$ for $i \in \{1, \ldots, Q\}$, and $\mu_b = \mathbb{E}_q[s_p(x)] \in \mathbb{R}^d$. The maximiser of $\mathrm{PSD}_{\mathcal{L}_1}(q \,\|\, p)$ is given by*

$$A_i^\star = \frac{\mu_i}{Z}, \qquad b^\star = \frac{\mu_b}{Z}, \tag{6}$$

where the normalisation constant Z is also the optimal discrepancy value:

$$\mathrm{PSD}_{\mathcal{L}_1}(q \,\|\, p) = Z = \sqrt{\sum_{k=1}^{Q} \|\mu_k\|_F^2 + \|\mu_b\|_2^2}. \tag{7}$$

In practice, we use the empirical $\widehat{\mathrm{PSD}}_{\mathcal{L}_1}(q \,\|\, p)$, obtained by replacing the population moments $\{\mu_i, \mu_b\}$ with their respective sample estimates $\{\hat{\mu}_i, \hat{\mu}_b\}$.

In the degenerate case $Q = 0$, the optimization reduces to $\sup_{\|b\|_2 \leq 1} b^\top \mu_b$, and the empirical discrepancy simplifies to the ℓ_2-norm of the empirical average score:

$$\widehat{\mathrm{PSD}}_{\mathcal{L}_1}(q \,\|\, p) = \|\hat{\mu}_b\|_2 = \left\| \frac{1}{n} \sum_{j=1}^{n} s_p(x_j) \right\|_2. \tag{8}$$

3.2. Time Complexity

The empirical Stein moments $\hat{\mu}_i = \frac{1}{n}\sum_{j=1}^{n} \Phi_i(x_j)$ and $\hat{\mu}_b = \frac{1}{n}\sum_{j=1}^{n} s_p(x_j)$ allow for a closed-form evaluation of the discrepancy. Specifically, the total computational complexity to evaluate $\widehat{\mathrm{PSD}}_{\mathcal{L}_1}$ is $\mathcal{O}(Qnd^2)$. While we focus on this formulation, a scalable $\mathcal{O}(Qnd)$ variant can be obtained by restricting A_i to be diagonal.

[2] that we assume to hold through-out this work

3.3. Characterisation under the Bernstein-von-Mises limit

Similar to $\mathrm{PSD}_{\mathcal{L}_2}$, the proposed $\mathrm{PSD}_{\mathcal{L}_1}$ captures moment convergence under the Bernstein-von Mises limit. However, due to its lower-parameter construction, it characterises at most $Qd^2 + d$ moment conditions, compared to the $O\left(\binom{Q+d}{d}\right)$ conditions captured by $\mathrm{PSD}_{\mathcal{L}_2}$ (Srinivasan et al., 2025).

Theorem 3.3. *Let $p = \mathcal{N}(\mu_p, \Sigma_p)$ with positive-definite Σ_p, so $s_p(x) = -\Sigma^{-1}(x - \mu_p)$. For $Q \geq 1$, define*

$$h(x) = \sum_{i=1}^{Q} A_i\big(x^{\circ i}\big) + b, \qquad A_i \in \mathbb{R}^{d\times d},\ b \in \mathbb{R}^d. \tag{9}$$

Then for any q with finite moments up to order $Q+1$,

$$\mathrm{PSD}_{\mathcal{L}_1}(q \,\|\, p) = 0 \iff \begin{cases} \mathbb{E}_q[X] = \mu_p, \\ \mathbb{E}_q\left[\left(X_{[l]} - \mu_{p[l]}\right) X_{[j]}^{i}\right] = \\ i\,\Sigma_{lj}\,\mathbb{E}_q\left[X_{[j]}^{i-1}\right], \end{cases} \tag{10}$$

for all $l, j \in \{1, \ldots, d\},\ i = 1, \ldots, Q$.

Proof. Proof is given in Appendix A. □

Corollary 3.4. *Let $p = \mathcal{N}(\mu_p, \Sigma_p)$ with positive-definite Σ_p and $h(x) = Ax + b$ (i.e., $Q = 1$). Then for any q with finite second moments,*

$$\mathrm{PSD}_{\mathcal{L}_1}(q \,\|\, p) = 0 \iff \begin{cases} \mathbb{E}_q[X] = \mu_p, \\ \mathbb{E}_q[(X - \mu_p)(X - \mu_p)^\top] = \Sigma_p. \end{cases}$$

Proof. The result follows directly from the previous theorem by setting $Q = 1$. □

3.4. Goodness-of-Fit Testing via $\mathrm{PSD}_{\mathcal{L}_1}(q \,\|\, p)$

The empirical squared discrepancy $\widehat{\mathrm{PSD}}^2_{\mathcal{L}_1}$ can be expressed as a V-statistic. Let $\xi(x) \in \mathbb{R}^{Qd^2+d}$ be a concatenated feature vector consisting of the flattened feature maps $\mathrm{vec}(\Phi_i(x))$ for $i = 1, \ldots, Q$ and the score $s_p(x)$. We can then write the squared discrepancy as the norm of an empirical mean:

$$\widehat{\mathrm{PSD}}^2_{\mathcal{L}_1}(q \,\|\, p) = \left\| \frac{1}{n} \sum_{j=1}^{n} \xi(x_j) \right\|_2^2 = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} k(x_j, x_k), \tag{11}$$

where $k(x, x') = \xi(x)^\top \xi(x')$ is a linear kernel in the augmented feature space. This formulation identifies the empirical PSD as a V-statistic of rank $Qd^2 + d$. By excluding the diagonal terms ($j = k$), one obtains the corresponding U-statistic:

$$\widehat{\mathrm{PSD}}^2_{\mathcal{L}_1,U} = \frac{1}{n(n-1)} \sum_{j \neq k} \xi(x_j)^\top \xi(x_k), \tag{12}$$

which provides an unbiased estimator of the squared population discrepancy (Liu et al., 2016). Although this estimator has quadratic computational complexity in n, it can be reformulated as a linear-time expression; see Equation 7. This formulation facilitates the use of U-statistic asymptotic theory for goodness-of-fit testing (Chwialkowski et al., 2016; Srinivasan et al., 2025). In practice, implementation typically relies on the wild bootstrap procedure (Leucht & Neumann, 2013), which offers robustness in finite samples and handles potential dependencies (Chwialkowski et al., 2016; Srinivasan et al., 2025).

However, the statistical power of such tests is sensitive to the choice of Q. Empirical evidence suggests a non-monotonic relationship in certain regimes: while a low order Q_1 or a high order Q_2 may yield acceptable results, an intermediate choice $Q_1 < Q_3 < Q_2$ can result in significantly degraded performance (Srinivasan et al., 2025). Beyond these local fluctuations, we observe a more fundamental challenge: as the polynomial order Q increases globally, the statistical power consistently degrades (e.g., when p and q differ through moment mismatches of order at most $Q^* < Q$).

To characterise this failure mode, we analyse the asymptotic behaviour of SNR^2 for $\widehat{\mathrm{PSD}}_{\mathcal{L}_1,U}$, focusing on high-order moment contributions under a standard Gaussian target $p = \mathcal{N}(0, I_d)$ and a slightly perturbed sampler $q = \mathcal{N}(\delta\mathbf{1}, \sigma^2 I_d)$.

Theorem 3.5 (SNR^2 of $\widehat{\mathrm{PSD}}_{\mathcal{L}_1,U}$)**.** *Assume $\mathbb{E}_q[\|\xi(X)\|_2^2] < \infty$. For the U-statistic estimator $\widehat{D}^2_U$ of degree Q, the SNR^2 satisfies*

$$\mathrm{SNR}(Q)^2 = \frac{\mathbb{E}_q\big[\widehat{D}^2_U\big]^2}{\mathbb{V}_q\big[\widehat{D}^2_U\big]} = \frac{\big(\sum_{i=1}^Q \|\mu_i\|_F^2 + \|\mu_b\|_2^2\big)^2}{\frac{4(n-2)}{n(n-1)}\zeta_{1,Q} + \frac{2}{n(n-1)}\zeta_{2,Q}}, \tag{13}$$

where $\zeta_{1,Q} = \mathbb{V}_x\big[\sum_{i=1}^Q \langle \Phi_i(x), \mu_i\rangle_F + s_p(x)^\top \mu_b\big]$ and $\zeta_{2,Q} = \mathbb{E}_{x,x'}\big[(\xi(x)^\top \xi(x'))^2\big] - \mathbb{E}_q[\xi(x)^\top \xi(x')]^2$.

Proof. This is based on Hoeffding decomposition (Hoeffding, 1992) of variance of U-statistic. See Appendix B. □

Remark 3.6 (Moment-Dominant Scaling for SNR^2 of $\widehat{\mathrm{PSD}}_{\mathcal{L}_1,U}$ under Gaussianity)**.** *Let $p = \mathcal{N}(0, I_d)$ and $q = \mathcal{N}(\delta\mathbf{1}, \sigma^2 I_d)$ with fixed (n, d). In regimes where $n\zeta_{1,Q} \gg \zeta_{2,Q}$, and under a moment-dominance assumption*[3]*,*

$$\mathrm{SNR}(Q)^2 \asymp \frac{\big((Q+1)!!\big)^2}{\big((2Q+1)!!\big)} \tag{14}$$

Proof. Sketch of proof is given in Appendix B. □

Remark 3.6 provides critical practical implications for both $\widehat{\mathrm{PSD}}_{\mathcal{L}_1,U}$ and $\widehat{\mathrm{PSD}}_{\mathcal{L}_2,U}$. Empirically, although their performance depends on Q, simply increasing Q is not effective, as the SNR of the corresponding U-statistic deteriorates. In this sense, existing PSD frameworks optimise the *wrong* objective: by focusing solely on maximising the signal, they suffer from uncontrolled variance inflation in higher polynomial orders. A robust GoF test should instead optimise the SNR (or equivalently SNR^2) instead (Sutherland et al., 2017; Jitkrittum et al., 2017; Grathwohl et al., 2020).

3.4.1. z-TEST FOR $\mathrm{PSD}_{\mathcal{L}_1}$

In this work, we employ a split-sample testing framework to bypass the $\mathcal{O}(BnQd^2)$ computational overhead associated with bootstrapping high-dimensional U-statistics. Given samples $\{x_j\}_{j=1}^n \sim q$, we split the data into training and test sets. The training set is used to estimate the optimal $\{A_1^\star, \cdots, A_Q^\star, b^\star\}$, while the test set is used to construct a one-sample z-statistic from $u(x) = \sum_{i=1}^Q \langle A_i^*, \Phi_i(x)\rangle_F + b^{*\top} s_p(x)$ for all x in test set. This construction enables a one-sample two-tailed z-test, which circumvents the need for a B-iteration bootstrap (see Algorithm 3). The validity of this test follows from the standard central limit theorem.

Proposition 3.7. *Let $\{x_j\}_{j=1}^n$ be i.i.d. samples and let $\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{te}}$ be the training and test splits with $|\mathcal{D}_{\mathrm{te}}| \to \infty$. Let $h^\star = h^\star(\mathcal{D}_{\mathrm{tr}})$ denote the test function estimated on the training set. Assume $\mathbb{E}_p[(\mathcal{L}_1 h^\star(X))^2] < \infty$ almost surely. Then under the null hypothesis $H_0 : \{x_j\}_{j=1}^n \sim p$, the test statistic z from Algorithm 1 satisfies*

$$z \xrightarrow{d} \mathcal{N}(0, 1). \tag{15}$$

Consequently, the test that rejects H_0 when $|z| > z_{1-\alpha/2}$ has asymptotic size α.

Proof. The result follows a standard CLT argument, conditionally on $\mathcal{D}_{\mathrm{tr}}$, together with Stein's identity under H_0. □

Crucially, this z-statistic inherits the same exponential SNR decay as the U-statistic.

Remark 3.8. *Let z_Q denote the split-sample z-statistic constructed using the theoretically optimal coefficients $\{A_i^\star, b^\star\}$*

[3]This assumes that under expansion of the numerator and denominator, the highest-order moment terms dominate all lower-order moment contributions in magnitude.

and polynomial degree Q. Under the settings of Remark 3.6 and the moment-dominance argument, the SNR^2 of z_Q deteriorates as $Q \to \infty$. In particular,

$$\mathrm{SNR}(Q)^2 \asymp \frac{\big((Q+1)!!\big)^2}{(2Q+1)!!}, \tag{16}$$

which decays exponentially in Q.

Proof. This follows the same arguments as the sketch of proof of Remark 3.6. □

This confirms that the degradation of SNR^2 is a fundamental property of the polynomial Stein basis that persists even under computationally efficient split-sample testing.

Practical Considerations for Dependent Sampling Under dependent sampling, the test statistic is not necessarily asymptotically standard Gaussian under the CLT, as the variance is typically misspecified and requires correction (e.g., via MCMC covariance estimators; see Brooks et al. 2011).

Nevertheless, in our simulation studies we include examples with dependent sampling, and observe that methods derived under the i.i.d. assumption can still perform reasonably well in this setting, despite the aforementioned limitations.

4. SNR^2-Optimised $\mathrm{PSD}_{\mathcal{L}_1}$

Definition 4.1 (SNR^2-optimised $\mathrm{PSD}_{\mathcal{L}_1}$)**.** The SNR^2-optimised polynomial Stein discrepancy is defined as

$$\text{o-PSD}_{\mathcal{L}_1}(q \,\|\, p) := \sup_h \frac{\mathbb{E}_q^2[\mathcal{L}_1 h(x)]}{\mathbb{V}_q[\mathcal{L}_1 h(x)]}. \tag{17}$$

Since the objective is scale-invariant, no norm constraint is required.

Let $\theta \in \mathbb{R}^{Qd^2+d}$ denote the stacked parameters $\{A_1, \ldots, A_Q, b\}$ and let $\xi(x)$ be defined in Section 3.4, so that $\mathcal{L}_1 h(x) = \theta^\top \xi(x)$. Define

$$\mu = \mathbb{E}_q[\xi(x)], \qquad \Sigma = \mathbb{COV}_q[\xi(x)].$$

Then the objective reduces to the generalised Rayleigh quotient

$$J(\theta) = \frac{(\theta^\top \mu)^2}{\theta^\top \Sigma \theta} = \frac{\theta^\top (\mu\mu^\top)\theta}{\theta^\top \Sigma \theta}. \tag{18}$$

Proposition 4.2. *Given positive-definite Σ, the objective $J(\boldsymbol{\theta})$ is maximised by $\boldsymbol{\theta}^\star \propto \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. The resulting squared discrepancy* $\text{o-PSD}^2_{\mathcal{L}_1}(q \,\|\, p)$ *is $\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.*

Proof. Proof is given in Appendix B. □

While theoretically optimal, this formulation presents significant challenges. The feature vector dimension is $\mathcal{O}(Qd^2 + d)$. In high-dimensional regimes, the empirical covariance $\widehat{\boldsymbol{\Sigma}}$ is rank-deficient, rendering the unregularised estimator non-identifiable. Furthermore, the $\mathcal{O}(Q^3 d^6)$ inversion cost is often prohibitive. Thus, without structural regularisation, the naive SNR^2-optimised test is both statistically ill-posed and computationally infeasible. In this work, we do not attempt to compute the estimator for $\boldsymbol{\theta}$ that directly optimises the SNR^2.

4.1. Scalable Approximate SNR-Optimised $\mathrm{PSD}_{\mathcal{L}_1}$

To obtain a tractable approximation, we restrict attention to the span of the unweighted $\mathrm{PSD}_{\mathcal{L}_1}$ solution and introduce scalar weights to reweight each Stein feature.

Proposition 4.3 (λ-$\mathrm{PSD}_{\mathcal{L}_1}$)**.** *Let $\{A_1^\star, \ldots, A_Q^\star, b^\star\}$ denote the theoretical maximiser of $\mathrm{PSD}_{\mathcal{L}_1}$, and define*

$$\Psi(X) := \begin{pmatrix} \langle A_1^\star, \Phi_1(X)\rangle_F \\ \vdots \\ \langle A_Q^\star, \Phi_Q(X)\rangle_F \\ (b^\star)^\top s_p(X) \end{pmatrix} \in \mathbb{R}^{Q+1}.$$

Further define $\mu_\Psi = \mathbb{E}_q[\Psi(X)]$ and $\Sigma_\Psi = \mathbb{COV}_q(\Psi(X))$.

The approximate SNR^2-optimised polynomial Stein discrepancy (λ-PSD$_{\mathcal{L}_1}$) is

$$\lambda\text{-PSD}_{\mathcal{L}_1}(q \,\|\, p) := \sup_{\lambda \in \mathbb{R}^{Q+1}} \frac{\lambda^\top \mu_\Psi}{\sqrt{\lambda^\top \Sigma_\Psi \lambda}},$$

whose maximiser satisfies $\lambda^\star \propto \Sigma_\Psi^{-1}\mu_\Psi$.

Given optimal coefficients $A_1^\star, \ldots, A_Q^\star, b^\star$, potentially estimated from a sample of size n_1, the optimal weight vector $\lambda^\star$ can be estimated in $\mathcal{O}(Qn_2d^2 + Q^3)$ time using an additional sample of size n_2.

Theorem 4.4 (SNR Stability of λ-$\mathrm{PSD}_{\mathcal{L}_1}$)**.** *Let $p = \mathcal{N}(0, I_d)$ and $q = \mathcal{N}(\delta\mathbf{1}, \sigma^2 I_d)$ with $\|\delta\|_2 > 0$ and $\sigma^2 > 0.5$. Let $\Psi(x) = (\psi_1(x), \ldots, \psi_{Q+1}(x))^\top$ be the vector of Stein features spanning the subspace $\mathcal{V}_Q = \mathrm{span}\{\Psi(x)\}$ ($Q \geq 1$). The $\mathrm{SNR}(Q)^2$ of the optimally weighted statistic $u(x) = \lambda^{\star\top}\Psi(x)$ satisfies:*

$$0 < \mathrm{SNR}(Q)^2 \leq \chi^2(p\|q). \tag{19}$$

Proof. Proof is provided in Appendix B. □

4.2. Optimisation of λ

In practice, we replace the theoretically optimal coefficients $\{A_i^\star, b^\star\}$ of $\mathrm{PSD}_{\mathcal{L}_1}$ with empirical estimates $\{\hat{A}_i^\star, \hat{b}^\star\}$. Although the optimal $\lambda^\star$ admits the closed-form estimator $\widehat{\Sigma}_\Psi^{-1}\widehat{\mu}_\Psi$, this approach is often numerically unstable. As Q

increases, the empirical covariance matrix $\widehat{\Sigma}_\Psi$ associated with higher-order Stein features may become ill-conditioned, particularly in small-sample regimes. Regularisation is therefore highly desirable.

In this work, we adopt a simple and computationally efficient stochastic optimisation strategy based on the cross-entropy (CE) method (Rubinstein, 1997; Rubinstein & Kroese, 2004). This framework adaptively learns a sampling distribution over λ that concentrates on high-SNR directions. Using a sparse parameterisation, such as a Bernoulli-Gaussian prior (see Algorithm 2), the CE method induces implicit regularisation by pruning components that contribute little to the signal. This approach scales favourably with Q and improves interpretability by highlighting which moment orders drive the discrepancy.

4.3. Goodness-of-Fit Testing for λ-$\mathrm{PSD}_{\mathcal{L}_1}$

Given samples $\{x_j\}_{j=1}^n \sim q$, we split the data into training and test sets, where the training set is further divided into two subsets. The first training subset is used to estimate the Stein coefficients $\{A_i^\star\}_{i=1}^Q$ and $b^\star$, while the second is used to estimate the optimal reweighting parameters $\lambda^\star$ given $\{A_i^\star\}_{i=1}^Q$ and $b^\star$. The resulting statistic $u(x) = \sum_{i=1}^Q \lambda_i^\star \langle A_i^\star, \Phi_i(x)\rangle_F + \lambda_{Q+1}^\star b^{\star\top} s_p(x)$ is then evaluated on the test set, followed by a one-sample z-test; full implementation details are provided in Algorithm 3 in Appendix C. Under similar mild conditions, the test statistic is asymptotically Gaussian.

4.4. Extensions

The scalar reweighting framework also admits a natural multi-resolution extension. In the most granular setting, one may introduce separate weights for every coefficient in $\{A_i^\star\}_{i=1}^Q$ and $b^\star$, yielding a total of $Qd^2 + d$ parameters (equivalent to o-$\mathrm{PSD}_{\mathcal{L}_1}$). More structured alternatives are also possible, such as assigning weights per moment order and dimension, resulting in $Qd+d$ parameters, or employing a hierarchical factorisation with separate weights for moment order and coordinate, requiring only $(Q+1)+d$ parameters. All such formulations can still be written as Rayleigh quotient problems over expanded Stein feature spaces, and Theorem 4.4 can be extended accordingly. These structured parameterisations may be particularly effective when the discrepancy is concentrated in a small subset of dimensions. The resulting optimisation problems are closely related to sparse generalised eigenvalue literature (see Tan et al. 2018; Cai et al. 2021). We leave these extensions for future work.

5. Experiments

This section benchmarks our proposed methods against several well-established Stein discrepancy methods from the literature. The compared methods include:

- Gauss KSD: Standard quadratic-time KSD with a Gaussian kernel via the median tune heuristic (Garreau et al., 2017).
- IMQ KSD: Standard quadratic-time KSD with the recommended inverse multiquadric (IMQ) kernel using hyperparameters $c = 1$ and $\beta = -0.5$ (Gorham & Mackey, 2017).
- FSSD-opt: FSSD with a Gaussian kernel and optimised test locations (Jitkrittum et al., 2017); here, 20% of the samples are used to tune the kernel bandwidth.
- RFSD: RFSD with base IMQ kernel and 10 random features (Huggins & Mackey, 2018; Srinivasan et al., 2025).

Apart from PSD-based methods, the kernel-based methods use bootstrap calibration with $B = 500$ simulated null test statistics. Significance level $\alpha = 0.05$ is used throughout this section. Code to reproduce these results is available at `https://github.com/olivervu25/weighted-psd`. This builds on existing implementations from Jitkrittum et al. 2017; Huggins & Mackey 2018; Srinivasan et al. 2025. All experiments were run on a single core of a high-performance computing workstation.

5.1. Experiment 1: Standard Multivariate Gaussian Target

We consider the target distribution p as

$$p(x) = \mathcal{N}(0_d, I_d),$$

and consider the following sampling distributions:

- Standard Gaussian: $q_1(x) = \mathcal{N}(0_d, I_d)$.
- Perturbed Gaussian: $q_2(x) = \mathcal{N}(0_d, \Sigma_d)$, where
$$\Sigma_d = \mathrm{Diag}(1.7, 1, \ldots, 1).$$
- Student-t: $q_3(x)$: independent Student's t distributions with 5 degrees of freedom.
- Laplace: $q_4(x)$ independent $\mathrm{Laplace}(0, 1/\sqrt{2})$ distributions; each marginal has the same first and second moments as a standard Gaussian.

In Experiment 1, we set $n = 1000$, except for the Student-t example where $n = 2000$, following Huggins & Mackey (2018); Srinivasan et al. (2025). We use 200 replicates with significance level $\alpha = 0.05$, and consider dimensions

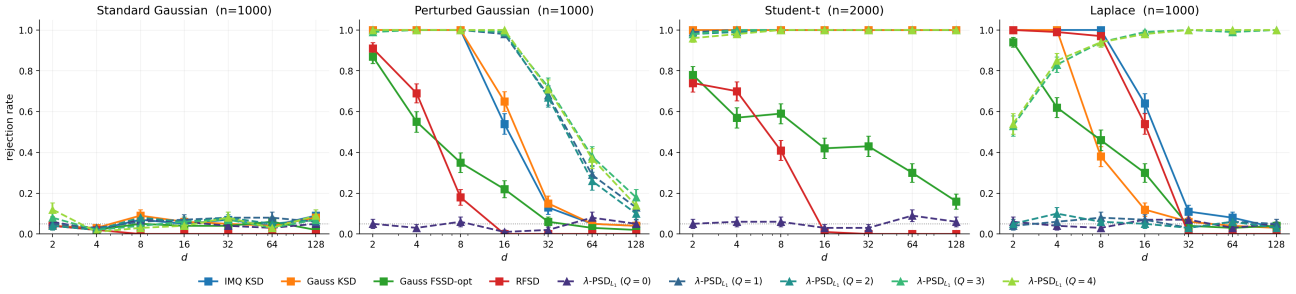

Figure 1. Experiment 1: Rejection rate vs. d. Dashed/triangles: λ-PSD$_{\mathcal{L}_1}$; solid/squares: kernel baselines. Dotted line: $\alpha = 0.05$.

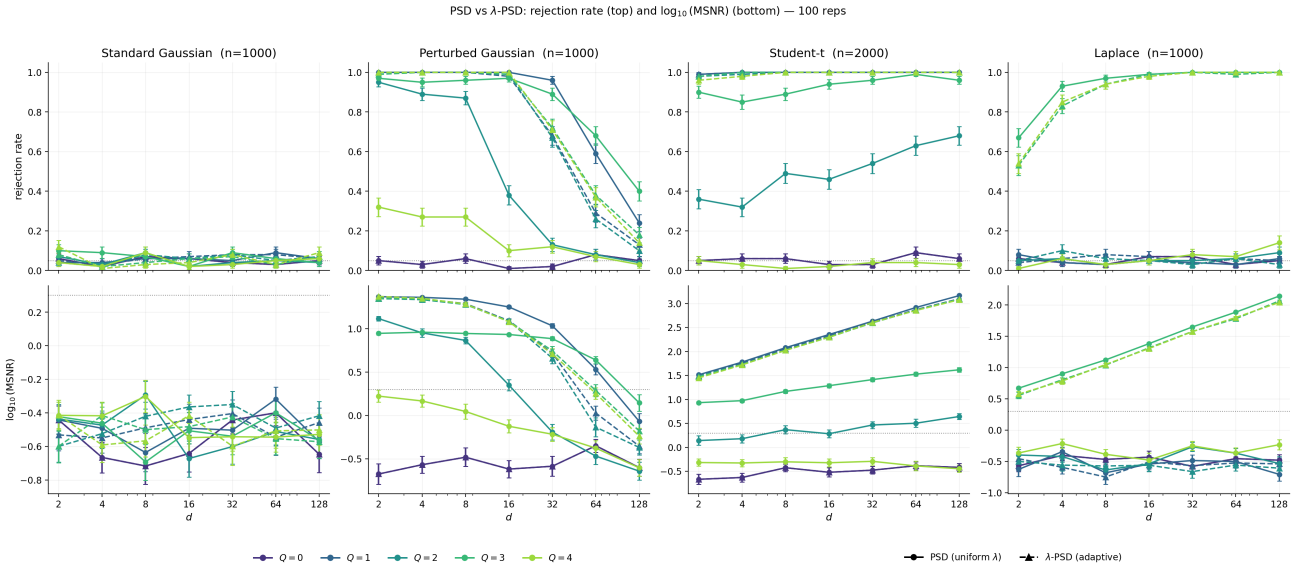

Figure 2. Experiment 1: Rejection rate (top) and $\log_{10}(\mathrm{MSNR})$ (bottom) vs. d. Solid/circles: PSD$_{L_1}$; dashed/triangles: λ-PSD$_{L_1}$. Colour encodes Q. Dotted line: $\log_{10}(1.96)$.

$d \in \{2, 4, 8, 16, 32, 64, 128\}$. Results are shown in Figures 1 and 2. Figure 2 additionally compares PSD$_{\mathcal{L}_1}$ and λ-PSD$_{\mathcal{L}_1}$ through the estimated mean SNR (MSNR), computed as the average absolute z-statistic.

Figures 1 and 2 show that the proposed λ-PSD$_{\mathcal{L}_1}$ achieves strong power across all alternatives while maintaining correct type-I error under the Gaussian null, outperforming IMQ KSD, Gaussian KSD, RFSD, and FSSD-opt in moderate and high dimensions, especially for heavy-tailed targets. Compared with the unweighted PSD$_{\mathcal{L}_1}$, the adaptive version is more stable across Q, avoiding the sensitivity to over-parameterisation seen in the unweighted statistic. For the Laplace example, as expected, low-order choices such as $Q \in \{0, 1, 2\}$ are insufficient to detect the mismatch, but once Q exceeds the effective complexity of the alternative, both rejection power and $\log_{10}(\mathrm{MSNR})$ remain stably high across Q and dimension d. See Appendix D for additional results demonstrating the superior robustness of λ-PSD$_{\mathcal{L}_1}$ with respect to increasing Q.

5.2. Experiment 2: Gaussian-Bernoulli Restricted Boltzmann Machine

In Experiment 2, following Liu et al. 2016; Jitkrittum et al. 2017; Srinivasan et al. 2025, we consider a Gaussian-Bernoulli restricted Boltzmann machine (RBM) with non-normalised target density p. The empirical samples are generated from the same RBM after perturbing the weight matrix $B \in \mathbb{R}^{50\times 40}$ with independent Gaussian noise of standard deviation σ. When $\sigma = 0$, the null hypothesis $H_0 : p = q$ holds exactly, whereas for $\sigma > 0$ the task is to detect deviations induced by the perturbation. We use $n = 1000$ samples generated via Gibbs sampling with 2000 warm-up iterations, and report null rejection rates across a range of perturbation levels based on 200 independent repetitions. The results are presented in Table 1.

Both PSD$_{\mathcal{L}_1}$ and λ-PSD$_{\mathcal{L}_1}$ generally outperform the linear-time methods RFSD and FSSD-opt, while achieving performance competitive with the quadratic-time KSD methods. Moreover, λ-PSD$_{\mathcal{L}_1}$ is noticeably less sensitive to the

choice of truncation level Q than $\mathrm{PSD}_{\mathcal{L}_1}$, particularly at $\sigma^2 = 0.02$, where rejection rates remain consistently high across different values of Q. These results suggest that our proposed methods may also be effective in dependent sampling and non-Gaussian settings.

Table 1. Null rejection rates at perturbation levels in the RBM example

Perturbation - σ^2	0.00	0.02	0.04	0.06
RFSD	0.01	0.28	0.91	0.99
FSSD-opt	0.04	0.65	0.94	0.99
IMQ KSD	0.04	0.98	1.00	1.00
Gauss KSD	0.03	0.95	1.00	1.00
$\mathrm{PSD}_{\mathcal{L}_1}$ $(Q=0)$	0.07	0.23	0.84	0.96
$\mathrm{PSD}_{\mathcal{L}_1}$ $(Q=1)$	0.05	0.98	1.00	1.00
$\mathrm{PSD}_{\mathcal{L}_1}$ $(Q=2)$	0.03	0.54	0.99	1.00
$\mathrm{PSD}_{\mathcal{L}_1}$ $(Q=3)$	0.04	0.97	1.00	1.00
$\mathrm{PSD}_{\mathcal{L}_1}$ $(Q=4)$	0.05	0.60	0.99	1.00
λ-$\mathrm{PSD}_{\mathcal{L}_1}$ $(Q=1)$	0.05	0.89	1.00	1.00
λ-$\mathrm{PSD}_{\mathcal{L}_1}$ $(Q=2)$	0.04	0.84	1.00	1.00
λ-$\mathrm{PSD}_{\mathcal{L}_1}$ $(Q=3)$	0.03	0.86	1.00	1.00
λ-$\mathrm{PSD}_{\mathcal{L}_1}$ $(Q=4)$	0.04	0.83	1.00	1.00

5.3. Runtime Analysis

We further investigate the computational scalability of sparse λ-$\mathrm{PSD}_{\mathcal{L}_1}$ under the standard Gaussian setting. In the first scenario, we fix the dimension at $d = 16$ and vary the sample size over $n \in \{100, 300, 1000, 3000, 10000\}$. In the second scenario, we fix the sample size at $n = 1000$ and vary the dimension over $d \in \{2, 4, 8, \ldots, 1024\}$. We consider polynomial orders $Q \in \{0, 1, 2, 3, 4\}$ and report average runtime (in seconds) over 100 independent replicates. Results are shown in Figures 3 and 4.

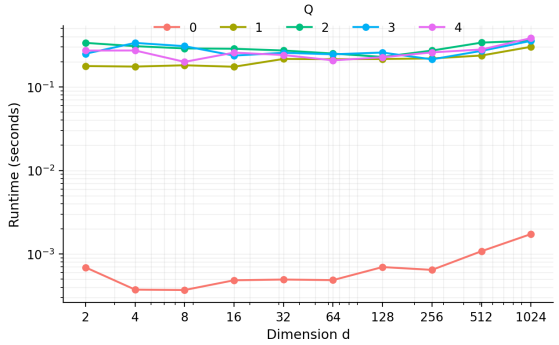

Figure 3. Average runtime of λ-$\mathrm{PSD}_{\mathcal{L}_1}$ as a function of dimension d $(n = 1000)$. Both axes are shown on logarithmic scales.

As observed in Figures 3 and 4, λ-$\mathrm{PSD}_{\mathcal{L}_1}$ scales favourably with both n and d. Despite the $\mathcal{O}(Qd^2n)$ complexity, the method remains practically fast even for $d = 1024$ and $n = 10^4$.

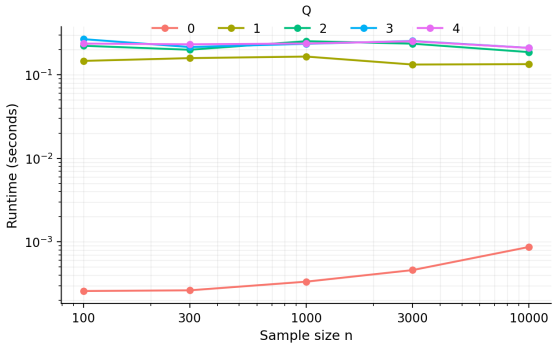

Figure 4. Average runtime of λ-$\mathrm{PSD}_{\mathcal{L}_1}$ as a function of sample size n $(d = 16)$. Both axes are shown on logarithmic scales.

6. Conclusion

We analyse the statistical behaviour of $\mathrm{PSD}_{\mathcal{L}_1}$ from an SNR perspective. Our analysis revealed a previously unstudied failure mode of PSD: although increasing the polynomial order Q enlarges the feature space and increases expressivity, the corresponding goodness-of-fit statistics can suffer severe variance inflation, leading to exponential SNR^2 decay.

Motivated by this observation, we reformulated Stein discrepancy construction as an explicit SNR^2 optimisation problem through a Rayleigh quotient formulation. This led to the proposed λ-$\mathrm{PSD}_{\mathcal{L}_1}$ framework, which introduces covariance-aware reweighting of Stein features within a low-dimensional subspace. Under Gaussian settings, we showed that λ-$\mathrm{PSD}_{\mathcal{L}_1}$ stabilises the SNR^2, avoiding the exponential collapse observed in standard PSD. The resulting procedure remains computationally tractable while improving robustness to the choice of polynomial degree. Empirical studies demonstrated that λ-$\mathrm{PSD}_{\mathcal{L}_1}$ achieves strong goodness-of-fit testing performance across a range of alternatives and dimensions, often outperforming standard kernel Stein methods while retaining linear-time complexity. Like competing approaches, the validity of the associated hypothesis test is asymptotic, and additional adjustments may be required under dependent sampling or small-sample settings.

Beyond the specific construction studied here, our results suggest a broader principle for scalable Stein discrepancy design: optimising the discrepancy alone is insufficient, and effective constructions should directly target the SNR. Future work includes extending the framework to richer reweighting schemes and dependent sampling settings.

References

Bhattacharya, A., Linero, A. R., and Oates, C. J. The ISBA Bulletin. *International Society for Bayesian Analysis*, 31

(3):11–14, 2024.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. *Handbook of Markov chain Monte Carlo*. CRC Press, 2011.

Cai, Y., Fang, G., and Li, P. A note on sparse generalized eigenvalue problem. *Advances in Neural Information Processing Systems*, 34:23036–23048, 2021.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *International Conference on Machine Learning*, pp. 2606–2615. PMLR, 2016.

Garreau, D., Jitkrittum, W., and Kanagawa, M. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

Gelman, A. and Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4): 457–472, 1992.

Gorham, J. and Mackey, L. Measuring sample quality with Stein's method. *Advances in Neural Information Processing Systems*, 28, 2015.

Gorham, J. and Mackey, L. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pp. 1292–1301. PMLR, 2017.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., and Zemel, R. Learning the Stein discrepancy for training and evaluating energy-based models without sampling. In *International Conference on Machine Learning*, pp. 3732–3747. PMLR, 2020.

Hoeffding, W. A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics: Foundations and basic theory*, pp. 308–334. Springer, 1992.

Huggins, J. and Mackey, L. Random feature Stein discrepancies. *Advances in Neural Information Processing Systems*, 31, 2018.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. A linear-time kernel goodness-of-fit test. *Advances in Neural Information Processing Systems*, 30, 2017.

Leucht, A. and Neumann, M. H. Dependent wild bootstrap for degenerate u-and v-statistics. *Journal of Multivariate Analysis*, 117:257–280, 2013.

Liu, Q. and Lee, J. Black-box importance sampling. In *Artificial Intelligence and Statistics*, pp. 952–961. PMLR, 2017.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 29, 2016.

Liu, Q., Lee, J., and Jordan, M. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pp. 276–284. PMLR, 2016.

Müller, A. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29 (2):429–443, 1997.

Repasky, M., Cheng, X., and Xie, Y. Neural Stein critics with staged l2-regularization. *IEEE Transactions on Information Theory*, 69(11):7246–7275, 2023.

Rubinstein, R. Y. Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99(1):89–112, 1997.

Rubinstein, R. Y. and Kroese, D. P. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*, volume 133. Springer, 2004.

South, L. F., Karvonen, T., Nemeth, C., Girolami, M., and Oates, C. J. Semi-exact control functionals from Sard's method. *Biometrika*, 109(2):351–367, 2022a.

South, L. F., Riabiz, M., Teymur, O., and Oates, C. J. Post-processing of MCMC. *Annual Review of Statistics and Its Application*, 9:529–555, 2022b.

Srinivasan, N., Sutton, M., Drovandi, C., and South, L. F. The polynomial Stein discrepancy for assessing moment convergence. In *Forty-second International Conference on Machine Learning*, 2025.

Stein, C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability theory*, volume 6, pp. 583–603. University of California Press, 1972.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*, 2017.

Tan, K. M., Wang, Z., Liu, H., and Zhang, T. Sparse generalized eigenvalue problem: Optimal statistical rates via truncated Rayleigh flow. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(5):1057–1086, 2018.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of mcmc (with discussion). *Bayesian Analysis*, 16(2):667–718, 2021.

A. Proofs of Moment-Convergence Properties under the BvM Limit

Theorem 3.3. Let $p = \mathcal{N}(\mu_p, \Sigma_p)$ with positive-definite Σ_p, so $s_p(x) = -\Sigma^{-1}(x - \mu_p)$. For $Q \geq 1$, define

$$h(x) = \sum_{i=1}^{Q} A_i\big(x^{\circ i}\big) + b, \qquad A_i \in \mathbb{R}^{d\times d},\ b \in \mathbb{R}^d. \tag{20}$$

Then for any q with finite moments up to order $Q+1$,

$$\mathrm{PSD}_{\mathcal{L}_1}(q \,\|\, p) = 0 \iff \begin{cases} \mathbb{E}_q[X] = \mu_p, \\ \mathbb{E}_q\Big[\big(X_{[l]} - \mu_{p[l]}\big)\, X^i_{[j]}\Big] = \\ i\,\Sigma_{lj}\,\mathbb{E}_q\Big[X^{i-1}_{[j]}\Big], \end{cases} \tag{21}$$

for all $l, j \in \{1, \dots, d\},\ i = 1, \dots, Q$.

Proof. By Proposition 3.2, the discrepancy vanishes when every Stein moment vanishes:

$$\mathrm{PSD}_{\mathcal{L}_1}(q \,\|\, p) = 0 \iff \mu_i := \mathbb{E}_q[\Phi_i(X)] = 0 \ \ \forall i = 1, \dots, Q, \quad \mu_b := \mathbb{E}_q[s_p(X)] = 0.$$

Since $p = \mathcal{N}(\mu_p, \Sigma_p)$,

$$s_p(x) = -\Sigma_p^{-1}(x - \mu_p).$$

We can write these two conditions in turn, exploiting the fact that the Gaussian score is affine map.

1. $\mu_b = 0 \iff \mathbb{E}_q[X] = \mu_p$.
   Indeed,
   $$\mu_b = -\Sigma_p^{-1}\mathbb{E}_q[X - \mu_p].$$
   As Σ_p^{-1} is invertible,
   $$\mu_b = 0 \iff \mathbb{E}_q[X - \mu_p] = 0 \iff \mathbb{E}_q[X] = \mu_p.$$
2. For each $i = 1, \dots, Q$, we show that $\mu_i = 0$ is equivalent to a single matrix identity:
   $$\mu_i = 0 \iff \mathbb{E}_q\Big[(X_{[l]} - \mu_{p[l]})X^i_{[j]}\Big] = i\,(\Sigma_p)_{lj}\,\mathbb{E}_q[X^{i-1}_{[j]}]$$
   for all $l, j \in \{1, \dots, d\}$.
   Recall the definition of the feature map $\Phi_i(x)$:
   $$\Phi_i(x) = s_p(x)(x^{\circ i})^\top + i\,\mathrm{Diag}(x^{\circ(i-1)}),$$
   so its (l, j)-th entry splits as
   $$[\Phi_i(x)]_{lj} = [s_p(x)]_l\, x^i_{[j]} + i\,\mathbb{1}(l = j)\, x^{i-1}_{[j]}.$$
   Substituting $[s_p(x)]_l = -\sum_m (\Sigma_p^{-1})_{lm}(x_{[m]} - \mu_{p[m]})$, we have
   $$[\Phi_i(x)]_{lj} = -\sum_{m=1}^{d}(\Sigma_p^{-1})_{lm}(x_{[m]} - \mu_{p[m]})x^i_{[j]} + i\,\mathbb{1}(l = j)x^{i-1}_{[j]}.$$
   Taking expectation of $\Phi_i(x)]_{lj}$ under q (i.e., $[\mu_i]_{lj}$), it becomes
   $$[\mu_i]_{lj} = -\sum_{m=1}^{d}(\Sigma_p^{-1})_{lm}\mathbb{E}_q[(X_{[m]} - \mu_{p[m]})X^i_{[j]}] + i\,\mathbb{1}(l = j)\mathbb{E}_q[X^{i-1}_{[j]}].$$

Define the matrix $M^{(i)} \in \mathbb{R}^{d\times d}$ where $M^{(i)}_{mj} := \mathbb{E}_q[(X_{[m]} - \mu_{p[m]})X^i_{[j]}]$

Then

$$[\mu_i]_{lj} = -(\Sigma_p^{-1}M^{(i)})_{lj} + i\,\mathbb{1}(l=j)\mathbb{E}_q[X^{i-1}_{[j]}].$$

Since $\mu_i = 0$, we have

$$\Sigma_p^{-1}M^{(i)} = i\,\mathrm{diag}\Big(\mathbb{E}_q[X^{i-1}_{[1]}], \cdots, \mathbb{E}_q[X^{i-1}_{[d]}]\Big) \iff M^{(i)} = i\Sigma_p \mathrm{diag}\Big(\mathbb{E}_q[X^{i-1}_{[1]}], \cdots, \mathbb{E}_q[X^{i-1}_{[d]}]\Big),$$

as Σ_p is invertible. Taking the (l,j)-th entry yields

$$\mathbb{E}_q[(X_{[l]} - \mu_{p[l]})X^i_{[j]}] = i\,(\Sigma_p)_{lj}\,\mathbb{E}_q[X^{i-1}_{[j]}].$$

□

B. Proofs of Scaling of SNR^2

Proposition 3.6. Let $p = \mathcal{N}(0, I_d)$ and $q = \mathcal{N}(\delta\mathbf{1}, \sigma^2 I_d)$ with fixed (n, d). In regimes where $n\zeta_{1,Q} \gg \zeta_{2,Q}$, and under the moment-dominance assumption,

$$\mathrm{SNR}^2(Q) \propto \frac{\big((Q+1)!!\big)^2}{\big((2Q+1)!!\big)} \tag{22}$$

Sketch of the proof. For $p = \mathcal{N}(0, I_d)$, $s_p(x) = -x$, so each $\Phi_i(x)$ is a polynomial of degree $i+1$. Hence $\sum_{i=1}^{Q} \|\mu_i\|_F^2 + \|\mu_b\|_2^2$ depends on Gaussian moments up to order $Q+1$, giving growth of order up to $((Q+1)!!)^2$.

For sufficiently large Q, the dominant term of $\zeta_{1,Q}$ is $\mathbb{V}_q\Big[\sum_{i=1}^{Q}\langle\Phi_i(X), \mu_i\rangle_F\Big]$. Since $\Phi_i(X)$ contains degree-$(i+1)$ polynomial features, the variance involves second moments of degree-$(Q+1)$ polynomials, contributing growth of order $(2Q+1)!!$. The coefficients μ_i themselves depend on Gaussian moments up to order $Q+1$, contributing an additional factor of order $((Q+1)!!)^2$. These give $\zeta_{1,Q}$ the growth of order up to $(2Q+1)!!\,((Q+1)!!)^2$

For sufficiently large n, the second term in the denominator vanishes asymptotically. As a result,

$$\mathrm{SNR}^2(Q) \propto \frac{\Big(\big((Q+1)!!\big)^2\Big)^2}{(2Q+1)!!\,((Q+1)!!)^2} = \frac{\big((Q+1)!!\big)^2}{\big((2Q+1)!!\big)}. \tag{23}$$

□

Proposition 4.2. The objective $J(\boldsymbol{\theta})$ is maximised by $\boldsymbol{\theta}^\star \propto \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. The resulting squared discrepancy o-PSD$^2_{\mathcal{L}_1}(q\,\|\,p)$ is $\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$.

Proof. Applying the Cauchy-Schwarz inequality in the $\boldsymbol{\Sigma}$-norm, we have $(\boldsymbol{\theta}^\top\boldsymbol{\mu})^2 = (\boldsymbol{\theta}^\top\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2 \le (\boldsymbol{\theta}^\top\boldsymbol{\Sigma}\boldsymbol{\theta})(\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$. Rearranging terms shows $J(\boldsymbol{\theta}) \le \boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, where equality holds if and only if $\boldsymbol{\theta} \propto \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. □

Theorem 4.4. Let $p = \mathcal{N}(0, I_d)$ and $q = \mathcal{N}(\delta\mathbf{1}, \sigma^2 I_d)$ with $\|\delta\|_2 > 0$ and $\sigma^2 > 0.5$. Let $\Psi(x) = (\psi_1(x), \ldots, \psi_{Q+1}(x))^\top$ be the vector of Stein features spanning the subspace $\mathcal{V}_Q = \mathrm{span}\{\Psi(x)\}$ $(Q \ge 1)$. The $\mathrm{SNR}(Q)^2$ of the optimally weighted statistic $u(x) = \lambda^{\star\top}\Psi(x)$ satisfies:

$$0 < \mathrm{SNR}(Q)^2 \le \chi^2(p\|q). \tag{24}$$

Proof. Define the density residual $R(x) = 1 - \frac{p(x)}{q(x)}$; under the specified q and p, $\mathbb{E}_q[R(X)] = 0$ and $\mathbb{V}_q[R(X)] = \chi^2(p\|q) < \infty$, i.e., Pearson χ^2-divergence.

For any $\mathbf{f} \in \mathcal{V}_Q$, let $g(x) = \mathcal{L}_{1,p}\mathbf{f}(x)$. Under the Gaussian setting, and $R(x)$ are all square-integrable under q. Assuming tail conditions, by Stein's identity, $\mathbb{E}_p[g(X)] = 0$. It is possible to show that:

$$
\begin{aligned}
\mathbb{E}_q[g(X)] &= \int g(x)q(x)\,dx \\
&= \int g(x)\left(1 - \frac{p(x)}{q(x)} + \frac{p(x)}{q(x)}\right) q(x)\,dx \\
&= \int g(x)\left(1 - \frac{p(x)}{q(x)}\right) q(x)\,dx + \int g(x)p(x)\,dx \\
&= \mathbb{E}_q[g(X)R(X)] + \mathbb{E}_p[g(X)] \\
&= \mathbb{E}_q[g(X)R(X)] + 0 && \text{(by Stein's identity)} \\
&= \mathbb{E}_q[g(X)R(X)] - \mathbb{E}_q[g(X)]\mathbb{E}_q[R(X)] && \text{(since } \mathbb{E}_q[R(X)] = 0\text{)} \\
&= \mathbb{COV}_q[g(X), R(X)].
\end{aligned}
$$

Using this identity, the SNR^2 is expressed as:

$$
\begin{aligned}
\text{SNR}(\mathbf{f})^2 &= \frac{\mathbb{E}_q[g(X)]^2}{\mathbb{V}_q[g(X)]} \\
&= \frac{\mathbb{COV}_q[g(X), R(X)]^2}{\mathbb{V}_q[g(X)]} \\
&\leq \frac{\mathbb{V}_q[g(X)]\mathbb{V}_q[R(X)]}{\mathbb{V}_q[g(X)]} && \text{(by Cauchy-Schwarz)} \\
&= \mathbb{V}_q[R(X)] = \chi^2(p\|q).
\end{aligned}
$$

To prove that $\text{SNR}(Q)^2 > 0$, it is sufficient to show that there exists an $\mathbf{f} \in \mathcal{V}_Q$ with non-zero SNR. We focus on the optimal solution of $\text{PSD}_{\mathcal{L}_1}$ with $Q = 1$.

Given $p = \mathcal{N}(0, I_d)$, $q = \mathcal{N}(\delta\mathbf{1}, \sigma^2 I_d)$, and the score function $s_p(x) = -x$, the population Stein moments are:

$$
\begin{aligned}
\mu_b &= \mathbb{E}_q[-x] = -\delta\mathbf{1} \\
\mu_1 &= \mathbb{E}_q[-XX^\top + I_d] \\
&= I_d - \mathbb{E}_q[XX^\top] \\
&= I_d - \left(\mathbb{COV}_q(X) + \mathbb{E}_q[X]\mathbb{E}_q[X]^\top\right) \\
&= I_d - \left(\sigma^2 I_d + (\delta\mathbf{1})(\delta\mathbf{1})^\top\right) \\
&= I_d - \sigma^2 I_d - \delta^2\mathbf{1}\mathbf{1}^\top \\
&= (1-\sigma^2)I_d - \delta^2\mathbf{1}\mathbf{1}^\top
\end{aligned}
$$

The optimal parameters are scaled by $A^\star = \frac{\mu_1}{Z}$ and $b^\star = \frac{\mu_b}{Z}$. Evaluating the operator yields:

$$
\begin{aligned}
\mathcal{L}_1 h^\star(x) &= -x^\top A^\star x - x^\top b^\star + \text{Tr}(A^\star) \\
&\propto -x^\top\left[(1-\sigma^2)I_d - \delta^2\mathbf{1}\mathbf{1}^\top\right]x - x^\top(-\delta\mathbf{1}) + \text{Tr}\left((1-\sigma^2)I_d - \delta^2\mathbf{1}\mathbf{1}^\top\right) \\
&= -(1-\sigma^2)x^\top I_d x + \delta^2 x^\top\mathbf{1}\mathbf{1}^\top x + \delta(\mathbf{1}^\top x) + \text{Tr}\left((1-\sigma^2)I_d\right) - \text{Tr}\left(\delta^2\mathbf{1}\mathbf{1}^\top\right) \\
&= (\sigma^2-1)\|x\|_2^2 + \delta^2(\mathbf{1}^\top x)(\mathbf{1}^\top x) + \delta(\mathbf{1}^\top x) + d(1-\sigma^2) - \delta^2\text{Tr}\left(\mathbf{1}\mathbf{1}^\top\right) \\
&= (\sigma^2-1)\|x\|_2^2 + \delta^2(\mathbf{1}^\top x)^2 + \delta(\mathbf{1}^\top x) + d(1-\sigma^2) - d\delta^2 \\
&= \left[(\sigma^2-1)\|x\|_2^2 + \delta^2(\mathbf{1}^\top x)^2 + \delta(\mathbf{1}^\top x) + d(1-\sigma^2-\delta^2)\right]
\end{aligned}
$$

Given that

$$
\begin{aligned}
&\mathbb{E}_q\left[\|X\|_2^2\right] = d\sigma^2 + d\delta^2 \\
&\mathbb{E}_q\left[(\mathbf{1}^\top X)^2\right] = \mathbb{V}_q(\mathbf{1}^\top X) + \mathbb{E}_q[\mathbf{1}^\top X]^2 = d\sigma^2 + d^2\delta^2 \\
&\mathbb{E}_q\left[\mathbf{1}^\top X\right] = d\delta
\end{aligned}
$$

Substituting these into the bracketed expression yields:

$$\begin{aligned}
\mathbb{E}_{q,h^*}[g(X)] &= \mathbb{E}_q[\mathcal{L}_1 h^\star(X)] \\
&= \mathbb{E}_q\left[(\sigma^2-1)\|X\|_2^2 + \delta^2(\mathbf{1}^\top X)^2 + \delta(\mathbf{1}^\top X) + d(1-\sigma^2-\delta^2)\right] \\
&= (\sigma^2-1)(d\sigma^2+d\delta^2) + \delta^2(d\sigma^2+d^2\delta^2) + \delta(d\delta) + d(1-\sigma^2-\delta^2) \\
&= d\sigma^4 + d\sigma^2\delta^2 - d\sigma^2 - d\delta^2 + d\sigma^2\delta^2 + d^2\delta^4 + d\delta^2 + d - d\sigma^2 - d\delta^2 \\
&= (d\sigma^4 - 2d\sigma^2 + d) + 2d\sigma^2\delta^2 + d^2\delta^4 - d\delta^2 \\
&= d(\sigma^2-1)^2 + d\delta^2(2\sigma^2-1) + d^2\delta^4, \\
&\geq d(\sigma^2-1)^2 + d\delta^2(2\sigma^2-1) + d\delta^4, \\
&\propto (\sigma^2-1)^2 + \delta^2(2\sigma^2-1) + \delta^4 \\
&= (\sigma^2-1)^2 + 2\delta^2(\sigma^2-1) + \delta^4 + \delta^2 \\
&= (\sigma^2-1+\delta^2)^2 + \delta^2,
\end{aligned}$$

which is non-zero unless $\{\delta, \sigma^2\} = \{0, 1\}$. This is also guaranteed by Theorem 3.3; future versions will explicitly include a lower bound in terms of SNR. Consequently,

$$\begin{aligned}
\mathrm{SNR}(Q)^2 &= \max_{\lambda} \frac{\mathbb{E}_q[g(X)]^2}{\mathbb{V}_q[g(X)]} \\
&\geq \frac{\mathbb{E}_{q,h^*}[g(X)]^2}{\mathbb{V}_{q,h^*}[g(X)]} \\
&> 0 \qquad\qquad\qquad\qquad\qquad\qquad \text{(since } \delta \neq 0\text{)}.
\end{aligned}$$

□

C. Pseudocode of Algorithms

Algorithm 1 $\mathrm{PSD}_{\mathcal{L}_1}$ GoF z-test

Input: i.i.d., samples $\{x_j\}_{j=1}^n \sim q$, score function $s_p(\cdot)$
Output: Test statistic and p-value
Randomly split $\{x_j\}_{j=1}^n$ into training set $\mathcal{D}_{\mathrm{tr}}$ and test set $\mathcal{D}_{\mathrm{te}}$
`// Training`
Compute empirical Stein moments $\{\hat{\mu}_i\}, \hat{\mu}_b$ on $\mathcal{D}_{\mathrm{tr}}$ Compute optimal maximiser $\{A_1^\star, \cdots, A_q^\star, b^\star\}$ in closed form (as in Proposition (3.2))
`// Testing`
Evaluate $u_j = \sum_{i=1}^Q \langle A_i^*, \Phi_i(x)\rangle_F + b^{*\top} s_p(x)$ for all $x_j \in \mathcal{D}_{\mathrm{te}}$ Compute the one-sample z-statistic

$$z = \frac{\sqrt{|\mathcal{D}_{\mathrm{te}}|}\,\bar{u}}{s_u} \tag{25}$$

where $\bar{u} = |\mathcal{D}_{\mathrm{te}}|^{-1}\sum_j u_j$ and $s_u^2 = (|\mathcal{D}_{\mathrm{te}}|-1)^{-1}\sum_j (u_j - \bar{u})^2$ Return the z-statistic and the corresponding two-sided p-value

Algorithm 2 Bernoulli-Gaussian Cross Entropy Method (BG-CEM) for λ optimisation

Input: Feature matrix $G \in \mathbb{R}^{n \times (Q+1)}$ (see Algorithm 3), maximum number of iterations $T = 100$, candidates $M = 1000$, elite size $K = 10$, tolerance $\epsilon = 10^{-3}$, patience $p = 3$

Output: Optimised weights $\hat{\lambda}$, sign (i.e., ± 1)

$n \leftarrow \mathrm{nrow}(G)$

Precompute sample mean and sample covariance matrix of G

$$\widehat{\mu_G} \leftarrow \frac{1}{n} G^\top \mathbf{1}_n, \qquad \widehat{\Sigma_G} \leftarrow \frac{1}{n} G^\top \left(I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top \right) G$$

Initialise activation probabilities $p_k = 0.5$, means $\mu_k = 0$, and scales $\sigma_k = 1$ for $k = 1, \ldots, Q+1$

Set best score $s^\star \leftarrow -\infty$, best $\lambda^* \leftarrow$ null

for $t = 1, \ldots, T$ **do**

  for $m = 1, \ldots, M$ **do**

    Sample $z_k^{(m)} \sim \mathrm{Bernoulli}(p_k)$ and $\epsilon_k^{(m)} \sim \mathcal{N}(\mu_k, \sigma_k^2)$ independently

    Form candidate

$$\lambda_k^{(m)} = z_k^{(m)} \cdot \epsilon_k^{(m)}, \quad k = 1, \ldots, Q+1$$

    and normalise $\lambda^{(m)} \leftarrow \lambda^{(m)} / \|\lambda^{(m)}\|$

    Compute objective using precomputed moments

$$s^{(m)} = \frac{(\widehat{\mu_G}^\top \lambda^{(m)})^2}{\lambda^{(m)\top} \widehat{\Sigma_G} \lambda^{(m)}}$$

    Update best solution λ^* if $s^{(m)} > s^\star$

  end

  Select elite set $\mathcal{A}$ of top K candidates according to $s^{(m)}$

  Early stop if score improvement is less than ϵ for p iterations

  Update parameters:

$$p_k \leftarrow \frac{1}{|\mathcal{A}|} \sum_{\lambda \in \mathcal{A}} \mathbf{1}\{|\lambda_k| > 0\}, \quad \mu_k \leftarrow \mathrm{mean}\big(\{\lambda_k : \lambda \in \mathcal{A}\}\big), \quad \sigma_k \leftarrow \mathrm{sd}\big(\{\lambda_k : \lambda \in \mathcal{A}\}\big)$$

end

Return $\{\lambda^*, \mathrm{sign}(\widehat{\mu_G} \lambda^*)\}$

Algorithm 3 λ-PSD$_{\mathcal{L}_1}$ GoF test

Input: Samples $\{x_j\}_{j=1}^n \sim q$, score function $s_p(\cdot)$
Output: Test statistic and p-value
Randomly split $\{x_j\}_{j=1}^n$ into training set $\mathcal{D}_{\text{tr}}$ and test set $\mathcal{D}_{\text{te}}$
Further split $\mathcal{D}_{\text{tr}}$ into $\mathcal{D}_1$ and $\mathcal{D}_2$
`// Training (Stage 1)`
Estimate coefficients $\{\hat{A}_i^*\}_{i=1}^Q$ and $\hat{b}^*$ using $\mathcal{D}_1$
`// Training (Stage 2)`
Initialise $G \in \mathbb{R}^{n_2 \times (Q+1)}$ **for** $k \leftarrow 1$ **to** n_2 **do**
  for $i \leftarrow 1$ **to** Q **do**
    $G_{[k,i]} \leftarrow \langle \hat{A}_i^\star, \Phi_i(x_k) \rangle_F$
  end
  $G_{[k,Q+1]} \leftarrow (\hat{b}^\star)^\top s_p(x_k)$
end
Get $\{\lambda_1^*, \cdots, \lambda_{Q+1}^*\}$ from BG-CEM(G)

`// Testing`
Evaluate $u_j = \sum_{i=1}^Q \lambda_i^* \langle \hat{A}_i^\star, \Phi_i(x) \rangle_F + \lambda_{Q+1}^* (\hat{b}_i^\star)^\top s_p(x)$ for all $x_j \in \mathcal{D}_{\text{te}}$
Compute the one-sample z-statistic

$$z = \frac{\sqrt{|\mathcal{D}_{\text{te}}|}\,\bar{u}}{s_u} \tag{26}$$

where $\bar{u} = |\mathcal{D}_{\text{te}}|^{-1} \sum_j u_j$ and $s_u^2 = (|\mathcal{D}_{\text{te}}| - 1)^{-1} \sum_j (u_j - \bar{u})^2$
Return the z-statistic and the corresponding two-sided p-value

D. Additional Results

We additionally report supplementary results for Experiment 1 in which the dimension is fixed at $d = 16$, while the sample size remains as specified in Section 5. Here, we vary the polynomial order over $Q \in \{0, 1, \ldots, 8\}$. As shown in Figure 5, λ-PSD$_{\mathcal{L}_1}$ maintains relatively stable rejection behaviour and $\log_{10}(\text{MSNR})$ as Q increases, whereas PSD$_{\mathcal{L}_1}$ exhibits substantially larger fluctuations across polynomial orders. These results further support the superior robustness of λ-PSD$_{\mathcal{L}_1}$ with respect to Q.

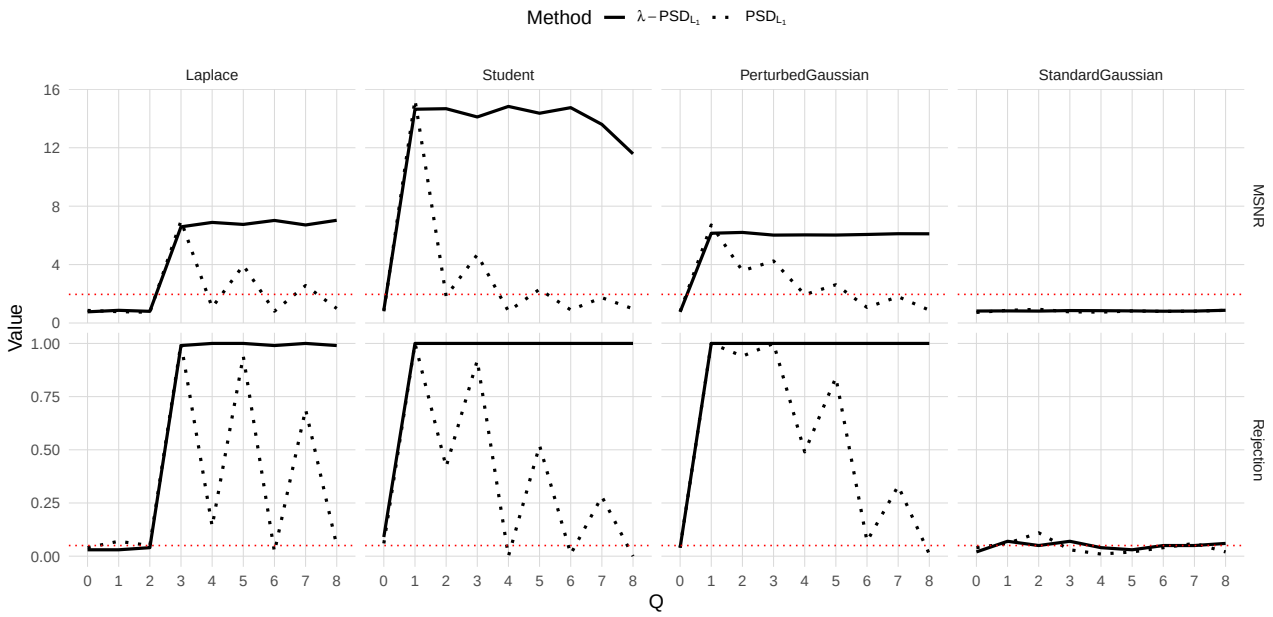

Figure 5. Experiment 1: Comparison of rejection rates and $\log_{10}(\text{MSNR})$ between PSD$_{\mathcal{L}_1}$ and λ-PSD$_{\mathcal{L}_1}$ for varying polynomial orders $Q \in \{0, 1, \ldots, 8\}$ at fixed dimension $d = 16$.