DIFFUSION-SCALE MARTINGALE MMD TESTS ON COMPACT RIEMANNIAN MANIFOLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Two-sample testing on a compact Riemannian manifold requires a kernel that respects geometry and remains sensitive to differences at unknown spatial scales. Diffusion kernels provide an intrinsic family, but smoothing attenuates both signal and noise, and neighboring scales can give redundant tests. We study this tradeoff within martingale maximum mean discrepancy testing. An exact spectral criterion shows that additional diffusion improves local asymptotic efficiency precisely when the average null-noise frequency exceeds the average signal frequency. For a fixed grid, thresholding the covariance pseudoinverse accommodates singularity and yields an asymptotic chi-square null limit with consistently estimated positive rank. Simulations on the circle, torus, and sphere support the predicted frequency dependence and conservative-to-near-nominal size. A matched torus comparison shows how the resampling-free calculation can improve detection by processing more available observations in similar time. These results connect intrinsic scale choice to both statistical power and computational cost.

1 INTRODUCTION

In homing experiments, Gagliardo et al. (2008) reported dispersed initial bearings after olfactory-nerve section, whereas control birds remained oriented toward home. Each bearing is an angle, so the two groups are samples on the circle, and they can differ in mean direction, in concentration around home, or in the shape of the distribution. A test built for one summary, such as the mean direction, can miss a difference in another; a two-sample test of the full directional distribution detects any of them, provided it respects the geometry of the circle rather than treating angles as points on a line. The circle is the simplest compact Riemannian manifold; spheres, tori, rotations and subspaces are further examples. Given independent samples $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_n \sim Q$ on a compact Riemannian manifold M, our target is $H_0 : P = Q$ against $H_1 : P \neq Q$.

Spectral tests have long used Laplacian eigenfunctions to detect departures from uniformity on compact manifolds (Giné, 1975; Jupp, 2008). For two-sample testing, the maximum mean discrepancy (MMD) compares mean features in a reproducing kernel Hilbert space (RKHS) (Gretton et al., 2012a; Sriperumbudur et al., 2011); suitable kernels recover energy distances based on negative-type semimetrics (Sejdinovic et al., 2013). A characteristic kernel detects every distributional difference in principle, while its feature weights determine finite-sample sensitivity. On a curved space, substituting geodesic distance in a Gaussian kernel can lose positive definiteness (Jayasumana et al., 2015). The diffusion kernel supplies a positive definite intrinsic family with a common Laplace–Beltrami eigenbasis (Grigor'yan, 2009; Rosenberg, 1997; De Vito et al., 2021). These eigenfunctions act as Fourier modes with eigenvalues measuring spatial frequency, as in the multiscale representations of diffusion maps (Coifman and Lafon, 2006); here increasing diffusion time suppresses fine variation in both the distributional difference and its sampling noise.

A coarse-scale difference can become easier to detect when diffusion removes high-frequency noise faster than signal; fine-scale differences can lose power under the same smoothing. For manifold data, Cheng and Xie (2024) establish intrinsic-dimension-dependent detection guarantees using bandwidths chosen for a density smoothness class. Within our fixed intrinsic family, the logarithmic derivative of local asymptotic efficiency equals the average null-noise frequency minus the average signal frequency. Covariance between geometric modes accounts for nonuniform sampling.

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.

The quadratic MMD statistic has a degenerate null law, whose critical values are commonly obtained by resampling or spectral approximation. Block MMD averages block statistics (Zaremba et al., 2013); cross-MMD and witness tests use sample splitting for tractable reference laws (Shekhar et al., 2022; Kübler et al., 2022). Martingale MMD (mMMD) instead processes the two samples in a fixed order as pairs (X_i, Y_i): at step i the average feature difference of the first $i-1$ pairs serves as a witness evaluated on the fresh pair. Under the null the fresh pair is independent of the witness and has a mean-zero feature difference, so the increments form a martingale difference sequence; under an alternative each increment has mean equal to the squared MMD. Dividing the sum of increments by the square root of their sum of squares gives a standard normal null limit without resampling or spectral approximation. Chatterjee and Ramdas (2026) establish this self-normalized Gaussian limit, consistency against fixed alternatives, and a chi-square test that combines several kernels when their increment covariance has full rank.

Choosing a kernel to maximize power requires information about the unknown alternative (Gretton et al., 2012b). MMDAgg aggregates kernels with permutation or wild-bootstrap critical values (Schrab et al., 2023); MMD-FUSE combines normalized MMD values through a permutation-calibrated soft maximum (Biggs et al., 2023); DUAL learns kernels that balance power and diversity before separate selection and inference (Zhou et al., 2025). A diffusion grid covers several frequency ranges using known geometry. Shared spectral components also create redundancy: nearby times can give small covariance eigenvalues, while contrasts that coincide on the null support give exact singularity.

We extend full-rank multiple-kernel mMMD to singular diffusion families using a thresholded covariance pseudoinverse. On a predetermined grid, the estimated rank consistently determines the degrees of freedom of the central or noncentral chi-square limit. A spectral iteration computes the finite-kernel statistic in one pass. Experiments connect scale ordering and computation to power: normalized MMD-FUSE can gain at equal sample size, while the cheaper martingale calculation can process more available pairs in similar time. Sections 2 to 5 present the method, theory, experiments and the pigeon application; the appendix contains proofs and additional experiments.

2 METHODOLOGY

2.1 GEOMETRIC SETUP AND DIFFUSION PULLBACK

Let (M, g) be a closed (compact without boundary), connected Riemannian manifold, and let $\mathfrak{m}$ be its Riemannian volume measure normalized to total mass one. Let Δ be the Laplace–Beltrami operator of g, the manifold Laplacian. On a closed manifold $-\Delta$ has a discrete spectrum $0 = \lambda_0 < \lambda_1 \le \lambda_2 \le \cdots$, repeated according to multiplicity, with eigenfunctions φ_j, $-\Delta\varphi_j = \lambda_j\varphi_j$, forming an orthonormal basis of $L^2(\mathfrak{m})$ and $\varphi_0 \equiv 1$ (Rosenberg, 1997). The diffusion semigroup $P_t = e^{t\Delta}$, $t \ge 0$, solves the heat equation $\partial_t u = \Delta u$; for $t > 0$ it has a smooth symmetric heat kernel p_t with $P_t f(x) = \int p_t(x,u) f(u)\, \mathfrak{m}(\mathrm{d}u)$ and $\int p_t(x,u)\, \mathfrak{m}(\mathrm{d}u) = 1$ (Grigor'yan, 2009), and it acts on signed measures by duality, $\int f \, \mathrm{d}(P_t^*\eta) = \int P_t f \, \mathrm{d}\eta$ for $f \in C(M)$.

Let $k : M \times M \to \mathbb{R}$ be continuous, symmetric, and positive definite, with RKHS $\mathcal{H}_k$. Compactness of M implies that k is bounded. For $t \ge 0$, define

$$k_t(x,y) = P_t^x P_t^y k(x,y) = \iint p_t(x,u) p_t(y,v) k(u,v)\, \mathfrak{m}(\mathrm{d}u)\mathfrak{m}(\mathrm{d}v), \tag{1}$$

where the integral form applies for $t > 0$ and $k_0 = k$; each kernel argument is thus diffused before similarity is evaluated. Let $\mathcal{H}_t$ be the RKHS of k_t, $\Phi_t(x) = k_t(x,\cdot)$, $w_t(P,Q) = \int_M \Phi_t(x)(P - Q)(\mathrm{d}x)$ and $\Delta_t^2(P,Q) = \|w_t(P,Q)\|_{\mathcal{H}_t}^2$. The squared MMD $\Delta_t^2(P,Q)$ is the signal at scale t; it is positive for every $P \ne Q$ when k_t is characteristic, which Theorem F.1 establishes whenever the base kernel is.

For a diffusion-kernel base $k = p_\tau$, $\tau > 0$, the semigroup identity gives $k_t = p_a$ with $a = \tau + 2t$ (Proposition 3.1). The implementation takes effective times a_ℓ directly; $e^{-a_\ell\lambda_j}$ weights mode j, and $a_\ell\lambda_1$ expresses time relative to the first nonconstant mode.

2.2 PREDICTABLE-WITNESS STATISTIC

The statistic evaluates the empirical difference between the two laws on a fresh observation pair. Assume $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$ and $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} Q$, independent between samples, and pair them as $Z_i = (X_i, Y_i)$. The pairing is algorithmic: at finite n the statistic depends on the predetermined pairing and order, and unequal groups are first reduced to equal-sized subsamples selected independently of their values. Define $\zeta_i(t) = \Phi_t(X_i) - \Phi_t(Y_i)$ and $\bar{\zeta}_{i-1}(t) = (i-1)^{-1} \sum_{j=1}^{i-1} \zeta_j(t)$. The intrinsic kernel contrast is

$$\begin{aligned} h_t(z, z') &= k_t(x, x') - k_t(x, y') - k_t(y, x') + k_t(y, y') \\ &= \langle \Phi_t(x) - \Phi_t(y), \Phi_t(x') - \Phi_t(y') \rangle_{\mathcal{H}_t}, \end{aligned}$$

for $z = (x, y)$ and $z' = (x', y')$. The row-average predictable-witness increment is

$$D_i(t) = \left\langle \bar{\zeta}_{i-1}(t), \zeta_i(t) \right\rangle_{\mathcal{H}_t} = \frac{1}{i-1} \sum_{j=1}^{i-1} h_t(Z_i, Z_j), \qquad i \geq 2. \tag{2}$$

For $\mathcal{F}_i = \sigma(Z_1, \ldots, Z_i)$, the past witness is measurable and the fresh pair independent of it, giving $\mathbb{E}\{D_i(t) \mid \mathcal{F}_{i-1}\} = \left\langle \bar{\zeta}_{i-1}(t), w_t(P, Q) \right\rangle$.

Under H_0, $w_t(P, Q) = 0$, so the increments are martingale differences. Under an alternative, their unconditional mean is $\Delta_t^2(P, Q)$, which accumulates against null fluctuations. Set

$$S_n(t) = \sum_{i=2}^{n} D_i(t), \qquad V_n(t) = \sum_{i=2}^{n} D_i(t)^2, \qquad T_n(t) = \frac{S_n(t)}{\sqrt{V_n(t)}}.$$

The observed quadratic variation $V_n(t)$ normalizes the same increments, with the convention $T_n(t) = 0$ when $V_n(t) = 0$. The one-sided test rejects when $T_n(t) > z_{1-\alpha}$ and reports the asymptotic p-value $1 - \Phi\{T_n(t)\}$, where Φ is the standard normal distribution function and $z_{1-\alpha}$ its $(1-\alpha)$ quantile. The reference law applies to a sample size and order fixed independently of the values. The row average is asymptotically equivalent to the i^{-1} normalization of mMMD (Chatterjee and Ramdas, 2026); see Proposition G.2.

2.3 FINITE MULTISCALE STATISTIC

To retain sensitivity at several frequency ranges, fix diffusion times $t_1, \ldots, t_L$ before observing the testing samples. Their increments share observations, so aggregation must account for their covariance. Write

$$\boldsymbol{D}_i = (D_i(t_1), \ldots, D_i(t_L))^\top, \qquad \boldsymbol{S}_n = \sum_{i=2}^{n} \boldsymbol{D}_i, \qquad \widehat{\boldsymbol{\Sigma}}_n = \sum_{i=2}^{n} \boldsymbol{D}_i \boldsymbol{D}_i^\top,$$

and set $\widehat{\boldsymbol{G}}_n = \widehat{\boldsymbol{\Sigma}}_n / \log n$. Nearly redundant scales produce small covariance eigenvalues whose inversion can amplify estimation error, so we retain only directions whose estimated variance exceeds a threshold. For a symmetric positive semidefinite matrix $A = \sum_{j=1}^{L} \lambda_j u_j u_j^\top$ with orthonormal eigenvectors u_j and a threshold $\kappa > 0$, let

$$A_\kappa^+ = \sum_{j : \lambda_j > \kappa} \lambda_j^{-1} u_j u_j^\top, \qquad r_\kappa(A) = \sum_{j=1}^{L} \mathbf{1}\{\lambda_j > \kappa\}.$$

At cutoff zero, the same formula defines the Moore–Penrose inverse A^+. Write $\|\cdot\|_{\text{op}}$ for the operator norm. Choose a deterministic sequence $\kappa_n \downarrow 0$. The rank-adaptive (RA) statistic is

$$Q_n^{\text{RA}} = \left(\frac{\boldsymbol{S}_n}{\sqrt{\log n}} \right)^\top (\widehat{\boldsymbol{G}}_n)_{\kappa_n}^+ \left(\frac{\boldsymbol{S}_n}{\sqrt{\log n}} \right), \qquad \widehat{r}_n = r_{\kappa_n}(\widehat{\boldsymbol{G}}_n). \tag{3}$$

The statistic sums squared signal projections, standardized by variance in each retained direction. Its asymptotic p-value is $1 - F_{\chi^2_{\widehat{r}_n}}(Q_n^{\text{RA}})$, with $F_{\chi^2_r}$ the chi-square distribution function; reject

below α. If $\widehat{r}_n = 0$, set $Q_n^{\mathrm{RA}} = 0$ and $p = 1$. Kernel normalization and κ_n jointly specify the retained directions: multiplying every kernel by $c > 0$ is equivalent to replacing the cutoff by κ_n/c^2 (Section B.3). We use the normalized-volume kernels as defined, with no additional rescaling. Retaining every direction gives $\boldsymbol{S}_n^\top \widehat{\boldsymbol{\Sigma}}_n^{-1} \boldsymbol{S}_n$ (Chatterjee and Ramdas, 2026).

A finite-sample alternative obtains critical values for the same Q_n^{RA} from permutations of the pooled observations, recomputing the pairing, covariance and retained rank each time. Section B.5 specifies this option and compares its power and cost with the chi-square reference and MMD-FUSE.

2.4 Spectral computation and tuning

For $L = 1$, the nonnegative MMD signal calls for the one-sided Gaussian test; the aggregate is two-sided, equaling T_n^2 when its direction is retained. For $L > 1$, use the aggregate. Algorithms 1 and 2 give spectral and direct-kernel implementations.

Direct kernel evaluation costs $O(Ln^2 + nL^2)$ arithmetic and $O(n + L^2)$ memory while retaining the observations. For spectral computation, fix a cutoff of J nonconstant modes in complete eigenspaces, independently of the testing observations, and define

$$p_{a_\ell,J}(x,y) = \sum_{j=0}^{J} e^{-a_\ell \lambda_j} \varphi_j(x)\varphi_j(y), \qquad \psi_{a_\ell,J}(x) = \left(e^{-a_\ell \lambda_j/2}\varphi_j(x)\right)_{j=0}^{J}. \tag{4}$$

Let $r_{i\ell} = \psi_{a_\ell,J}(X_i) - \psi_{a_\ell,J}(Y_i)$. Its constant coordinate cancels. Write $D_{i\ell} = D_i(t_\ell)$ for the increment computed with $p_{a_\ell,J}$. Maintaining the past feature sums gives the exact truncated-kernel statistic in $O(nLJ + nL^2)$ time and $O(LJ + L^2)$ working memory. For fixed truncation, consistency concerns alternatives represented in the retained eigenspaces.

Let $R_{i\ell} = \sum_{j=1}^{i} r_{j\ell}$ be the feature sum through pair i, with $R_{0\ell} = 0$. At each scale the iteration first evaluates the fresh pair against the past sum, then updates that sum:

$$D_{i\ell} = \frac{R_{i-1,\ell}^\top r_{i\ell}}{i-1}, \qquad R_{i\ell} = R_{i-1,\ell} + r_{i\ell}, \qquad i \geq 2,$$

starting from $R_{1\ell} = r_{1\ell}$. The past sums and the accumulators $\boldsymbol{S}_n, \widehat{\boldsymbol{\Sigma}}_n$ suffice for all updates; both complete algorithms appear in Section A.

Our predetermined grid spans $a\lambda_1 = 0.025$ to 1, giving first-mode multipliers from 0.98 to 0.37; higher modes decay faster. The common fixed reference 0.10 retains about 90% of first-mode weight. With the stated normalization, $\kappa_n = (\log n)^{-1/4}$ separates null covariance error from positive directions; local rank recovery uses $(\log n)^{-1/8}$. Report the retained rank with the p-value; Section B examines cutoff sensitivity.

3 Theoretical Results

We first establish that diffusion preserves distributional identification, then derive the signal-to-noise ratio that controls local power, and finally account for dependence across scales. Proofs appear in Sections F to J.

3.1 Diffusion-kernel signal and fixed-scale power

Diffusion pullback preserves positive definiteness and compares the smoothed laws through the identity $\mathrm{MMD}_{k_t}(P,Q) = \mathrm{MMD}_k(P_t^*P, P_t^*Q)$. Every multiplier $e^{-t\lambda_j}$ is positive, so P_t^* is injective on finite signed measures and preserves characteristicness of the base kernel (Theorem F.1). Thus smoothing changes discrepancy size while retaining distributional identification. The diffusion-kernel specialization makes its frequency dependence explicit.

Proposition 3.1 (Diffusion-kernel signal)**.** *Take $k = p_\tau$ with $\tau > 0$ and $a = \tau + 2t$. Then $k_t = p_a$ is characteristic. Writing Δ_a for MMD with kernel p_a, for any probability laws P, Q,*

$$\Delta_a^2(P,Q) = \sum_{j \geq 1} e^{-a\lambda_j} \left\{P\varphi_j - Q\varphi_j\right\}^2, \tag{5}$$

where $P\varphi_j = \int \varphi_j \, \mathrm{d}P$.

Each term in (5) is a squared difference in a geometric Fourier coefficient weighted by $e^{-a\lambda_j}$, so increasing a suppresses high frequencies more strongly. A finite implementation retains complete eigenspaces through J, preserving basis invariance; the omitted signal is bounded by four times the uniform kernel truncation error (Section F).

Under H_0, let $C_t = \mathbb{E}\{\zeta_1(t) \otimes \zeta_1(t)\}$ and $\alpha_t = \mathrm{tr}(C_t^2) = \mathbb{E}h_t(Z, Z')^2$, with independent Z, Z' and $(u \otimes v)h = u\langle v, h\rangle$. For a characteristic kernel, $\alpha_t > 0$ exactly when the common law is not a point mass (Proposition G.1). Averaging $i-1$ centered feature differences gives $\mathbb{E}D_i(t)^2 = \alpha_t/(i-1)$; martingale orthogonality yields

$$\mathrm{Var}\{S_n(t)\} = \mathbb{E}V_n(t) = \alpha_t \sum_{i=2}^{n} \frac{1}{i-1} = \alpha_t \log n + O(1).$$

The null fluctuation scale is $\sqrt{\log n}$. Compactness supplies the moment bounds for mMMD (Chatterjee and Ramdas, 2026): $T_n(t) \Rightarrow N(0,1)$ when $\alpha_t > 0$, and $T_n(t) \to \infty$ whenever $\Delta_t(P, Q) > 0$ (Corollary G.3). Local alternatives show how the signal competes with this fluctuation scale.

Local alternatives. For sample-size-dependent laws (P_n, Q_n), let $\zeta_{n,i} = \Phi_t(X_{n,i}) - \Phi_t(Y_{n,i})$ and put $w_n = \mathbb{E}\zeta_{n,1}$, $d_n = \|w_n\|^2$, $\varepsilon_{n,i} = \zeta_{n,i} - w_n$, $C_n = \mathbb{E}(\varepsilon_{n,1} \otimes \varepsilon_{n,1})$ and $\alpha_n = \mathrm{tr}(C_n^2)$. The fixed continuous kernel on compact M bounds all feature differences uniformly.

Theorem 3.2 (Gaussian shift under local alternatives)**.** *Assume*

$$\alpha_n \to \alpha_0 \in (0, \infty), \qquad \frac{nd_n}{\sqrt{\log n}} \to \rho \in [0, \infty). \tag{6}$$

Then

$$T_n(t) \Rightarrow N\left(\frac{\rho}{\sqrt{\alpha_0}}, 1\right). \tag{7}$$

The limiting one-sided power is $1 - \Phi(z_{1-\alpha} - \rho/\sqrt{\alpha_0})$*. In the alternative regime in which* $\alpha_n \to \alpha_0 \in (0, \infty)$*,* $nd_n/\sqrt{\log n} \to \infty$*, and* $nd_n/\log n \to 0$*, one has* $T_n(t) \to \infty$ *in probability.*

Each increment has mean d_n, so the accumulated drift is $(n-1)d_n$. Balancing it against $\sqrt{\log n}$ yields the local MMD scale $(\log n)^{1/4}/\sqrt{n}$: a positive finite ρ produces power strictly between the test level and one. This fixed-kernel comparison complements the varying-bandwidth, density-norm guarantees for manifold testing in Cheng and Xie (2024); the broader mMMD theory also treats varying kernels (Chatterjee and Ramdas, 2026).

3.2 Diffusion time and local efficiency

Let P_0 be the limiting null law and $c_{jk} = \mathrm{Cov}_{P_0}(\varphi_j(X), \varphi_k(X))$. Consider a local spectral direction $\theta = (\theta_j)_{j \geq 1}$ such that $P_n\varphi_j - Q_n\varphi_j = c_n\theta_j$ with $nc_n^2/\sqrt{\log n} \to c \in (0, \infty)$. Let $C_{n,a}$ and $C_{0,a}$ be the centered paired-feature covariances under $P_n \otimes Q_n$ and $P_0 \otimes P_0$. Assume $C_{n,a} \to C_{0,a}$ in Hilbert–Schmidt norm, uniformly on compact positive-time intervals; total-variation convergence of both laws suffices (Lemma E.4). Suppose $\theta_j = \int \varphi_j \, \mathrm{d}\nu$ for a finite signed measure with $\nu(M) = 0$, or directly assume local uniform convergence of the following series and their first derivatives. Set

$$A(a) = \sum_{j \geq 1} e^{-a\lambda_j}\theta_j^2, \qquad N(a) = \sum_{j,k \geq 1} e^{-a(\lambda_j + \lambda_k)}c_{jk}^2. \tag{8}$$

Here $A(a)$ is signal energy and $4N(a)$ is the null variance coefficient. The Gaussian shift is $c\mathcal{E}(a)$ for $\mathcal{E}(a) = A(a)/(2\sqrt{N(a)})$, so maximizing this ratio maximizes local limiting power. This signal-to-noise objective parallels power-based kernel selection (Gretton et al., 2012b); the shared eigenbasis makes its derivative explicit. The covariances c_{jk} account for dependence between modes under nonuniform sampling.

Define the signal and noise average frequencies

$$\bar{\lambda}_{\mathrm{sig}}(a) := \frac{\sum_{j\geq 1} \lambda_j e^{-a\lambda_j}\theta_j^2}{A(a)},$$
$$\bar{\lambda}_{\mathrm{noise}}(a) := \frac{\sum_{j,k\geq 1} \frac{\lambda_j+\lambda_k}{2} e^{-a(\lambda_j+\lambda_k)} c_{jk}^2}{N(a)}.$$

Theorem 3.3 (Exact diffusion-time efficiency criterion)**.** *For every $a > 0$ such that $A(a)N(a) > 0$,*

$$\frac{\mathrm{d}}{\mathrm{d}a} \log \mathcal{E}(a) = \bar{\lambda}_{\mathrm{noise}}(a) - \bar{\lambda}_{\mathrm{sig}}(a). \tag{9}$$

The derivative with respect to pullback time t is twice this difference at $a = \tau + 2t$. Thus the local-efficiency curve has positive derivative at a if and only if the null-noise average frequency exceeds the alternative-signal average frequency at that scale. On an interval where $A(a)N(a) > 0$, finite-interval improvement is determined by the integral of this frequency difference; positivity throughout the interval is sufficient.

The criterion compares the rates at which diffusion removes signal and null variation: low-frequency alternatives benefit when higher-frequency noise decays faster than their signal. For a uniform limiting law, $c_{jk} = \mathbf{1}\{j = k\}$ on the nonconstant modes, so $N(a) = \sum_{j\geq 1} e^{-2a\lambda_j}$. An alternative supported in one eigenspace has $\bar{\lambda}_{\mathrm{sig}}$ equal to that eigenvalue, so its efficiency increases precisely while the diffusion-weighted average null eigenvalue exceeds the signal eigenvalue. Unknown signal frequency motivates retaining a grid of diffusion times. Shared spectral components can make the contrasts redundant; rank adaptation combines their estimable directions. Section 4 compares this combination with MMD-FUSE applied to the same diffusion kernels.

3.3 Rank-adaptive multiscale inference

Under H_0, let Z, Z' be independent paired observations and define the covariance matrix $\mathbf{\Gamma} = (\Gamma_{ab})_{a,b=1}^{L}$ by

$$\Gamma_{ab} = \mathbb{E}\{h_{t_a}(Z, Z')h_{t_b}(Z, Z')\}, \qquad 1 \leq a, b \leq L. \tag{10}$$

Equivalently, $\Gamma_{ab} = \mathrm{tr}(C_{ab}C_{ba})$ for $C_{ab} = \mathbb{E}\{\zeta_1(t_a) \otimes \zeta_1(t_b)\}$.

The matrix $\mathbf{\Gamma}$ is the Gram matrix of the scale-specific null contrasts in $L^2(\mathcal{L}(Z) \otimes \mathcal{L}(Z))$, so its rank counts linearly independent contrasts (Proposition J.3). Diagonalizing $\mathbf{\Gamma}$ and standardizing its positive-eigenvalue coordinates gives independent normal limits whose squared sum explains the chi-square reference and extends full-rank multiple-kernel mMMD (Chatterjee and Ramdas, 2026) to redundant scales.

Theorem 3.4 (Rank-adaptive multiscale null limit)**.** *Suppose H_0 holds and $r = \mathrm{rank}(\mathbf{\Gamma}) \geq 1$. Then*

$$\frac{\boldsymbol{S}_n}{\sqrt{\log n}} \Rightarrow N_L(\mathbf{0}, \mathbf{\Gamma}), \qquad \left\| \widehat{\boldsymbol{G}}_n - \mathbf{\Gamma} \right\|_{\mathrm{op}} = O_{\mathbb{P}}\left((\log n)^{-1/2} \right). \tag{11}$$

Let $\kappa_n \downarrow 0$ satisfy

$$\kappa_n \sqrt{\log n} \to \infty. \tag{12}$$

Then $\widehat{r}_n \to r$ in probability and

$$Q_n^{\mathrm{RA}} \Rightarrow \chi_r^2.$$

The test using the random critical value $\chi^2_{\widehat{r}_n, 1-\alpha}$ has asymptotic level α.

Condition (12) separates covariance error from fixed positive eigenvalues. The degrees of freedom therefore count independent contrasts: repeated times are exactly redundant, and different contrasts can coincide on the null support. Nearby times can instead have small positive eigenvalues that are excluded at finite n. For example, two identical diffusion times produce identical increment sequences, whose covariance has rank one whenever the single-scale null variance is positive; the aggregate then equals T_n^2 once that direction is retained and uses a one-degree-of-freedom reference, although the grid contains two entries.

For local multiscale alternatives, let $d_{n\ell} = \|\mathbb{E}\zeta_{n,1}(t_\ell)\|^2$ and $\boldsymbol{d}_n = (d_{n1}, \ldots, d_{nL})^\top$; center each feature difference at its own mean and let $\mathbf{\Gamma}_n$ be the analogue of (10) formed from the centered differences.

Theorem 3.5 (Multiscale noncentral limit)**.** *Assume L is fixed, $\mathbf{\Gamma}_n \to \mathbf{\Gamma}$, and*

$$\frac{n\boldsymbol{d}_n}{\sqrt{\log n}} \to \boldsymbol{b} \in [0,\infty)^L. \tag{13}$$

Then

$$\frac{\boldsymbol{S}_n}{\sqrt{\log n}} \Rightarrow N_L(\boldsymbol{b},\mathbf{\Gamma}), \qquad \widehat{\boldsymbol{G}}_n \to \mathbf{\Gamma} \quad \textit{in probability.}$$

If $r = \operatorname{rank}(\mathbf{\Gamma}) \geq 1$, the deterministic covariance convergence satisfies

$$\|\mathbf{\Gamma}_n - \mathbf{\Gamma}\|_{\mathrm{op}} = o(\kappa_n), \tag{14}$$

and the rank threshold additionally satisfies

$$\kappa_n (\log n)^{1/4} \to \infty, \tag{15}$$

then

$$Q_n^{\mathrm{RA}} \Rightarrow \chi^2_r\left(\boldsymbol{b}^\top \mathbf{\Gamma}^+ \boldsymbol{b}\right). \tag{16}$$

The noncentrality in (16) sums squared local means divided by their variances along the positive-eigenvalue directions, the signal left after covariance standardization; a null-space component is discarded. Local means perturb the covariance estimate by $O_{\mathbb{P}}\{(\log n)^{-1/4}\}$, motivating $\kappa_n = (\log n)^{-1/8}$ in the local experiments (Section J).

For fixed alternatives and bounded selected kernels, $Q_n^{\mathrm{RA}} \to \infty$ in probability whenever at least one scale has positive squared MMD, provided $\kappa_n \downarrow 0$ and $\kappa_n \log n/n \to 0$; see Proposition J.4.

For positive definite $\mathbf{\Gamma}$ the pseudoinverse is the ordinary inverse; the reductions to single-scale mMMD and Fourier testing on the torus are collected in Section D.

4 Numerical Experiments

The experiments examine whether the asymptotic reference laws control size at practical sample sizes and whether diffusion time changes power according to the spectral mechanism in Theorem 3.3. We use the circle, flat torus, and sphere, with 500 independent replications per design and a nominal level of 0.05. The six effective diffusion times satisfy $a\lambda_1 \in \{0.025, 0.05, 0.10, 0.25, 0.50, 1\}$; the fixed-scale comparison uses $a\lambda_1 = 0.10$. The finite kernels retain complete eigenspaces containing 100, 400, and 440 nonconstant modes, respectively. The rank-adaptive test uses $\kappa_n = (\log n)^{-1/4}$ in these null and fixed-alternative experiments. Section B gives the full study of 95 simulation cells, including local alternatives, rank and truncation sensitivity, and matched competitors.

The null study varies the common law from uniform to concentrated and multimodal, exercising size control under different feature covariances; Section B.2 specifies the distributions. Across these laws at paired sample sizes $n \in \{100, 200, 500, 1000\}$, the default fixed-scale rejection rates ranged from 0.016 to 0.058 and the rank-adaptive rates from 0.026 to 0.050. Six of the 36 fixed-scale Wilson intervals excluded 0.05, all from below. Exact-duplicate diffusion grids and finite-support nulls also gave near-nominal rejection rates, connecting singular-covariance theory to finite samples; Section B reports their retained ranks.

Spectral alternatives reveal why scale choice matters. On the torus we compare the uniform law with density $1 + 0.3\cos(2\pi m x_1)$ for $m \in \{1, 3, 7\}$; on the sphere the alternative density is $1 + 0.3P_\ell(x_3)$ where P_ℓ is the Legendre polynomial of degree $\ell \in \{1, 4, 8\}$. Both densities are relative to normalized Riemannian volume. For the uniform null and a unit signal in the specified eigenspace with eigenvalue λ_{sig}, the finite-basis efficiency is $e^{-a\lambda_{\mathrm{sig}}}/(2\sqrt{\sum_{j=1}^{J} e^{-2a\lambda_j}})$. Figure 1 compares it directly with the $n = 1000$ powers. Its grid maximizers are 1, 0.05 and 0.025 for the low, medium and high frequencies on both manifolds, matching all six observed maxima. The full curves and numerical values appear in Section B.7; with 500 replications, rejection-rate standard errors are at most 0.022, and the high-frequency sphere maximum is close to its neighboring scales. The comparison concerns scale ordering, with the unit signal normalization fixed within each curve.

A matched torus comparison at $n = 40, 100$ and signal strength 0.6 separates aggregation from the reference distribution. At $n = 100$, permutation critical values raise low-frequency rank-adaptive

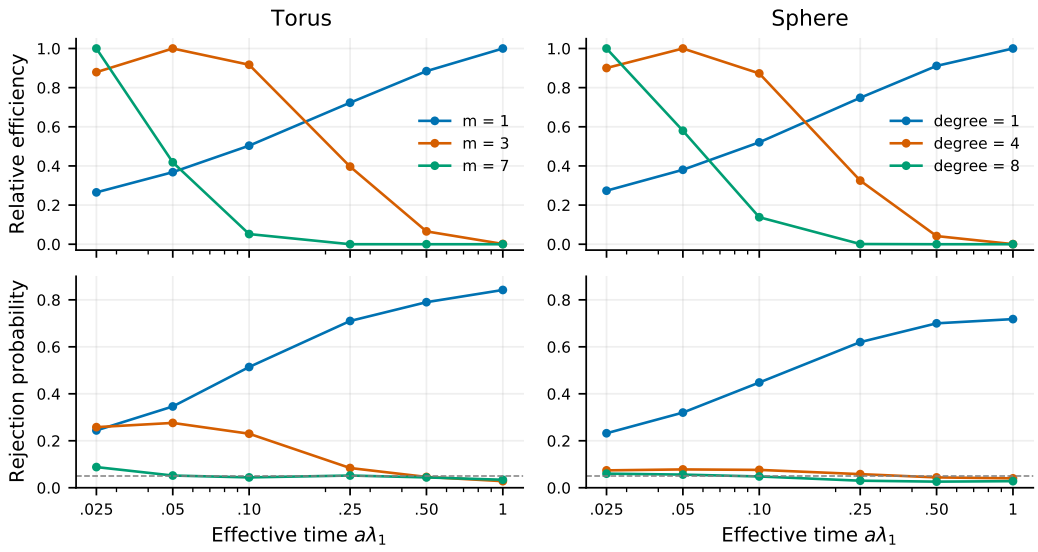


Figure 1: Diffusion-time efficiency and observed power for the same truncated kernels. Top: exact efficiency divided by its grid maximum for each torus frequency m or spherical degree ℓ. Bottom: fixed-scale rejection rates at $n = 1000$, with nominal level 0.05 dashed. The unit spectral direction determines the scale comparison; empirical alternatives use signal strength 0.3.

power from 0.200 to 0.262, while normalized MMD-FUSE (MMD-FUSE-N) reaches 0.488. FUSE also leads at the middle frequency; all three powers remain below 0.07 at the highest frequency. Permutation RA holds the statistic and threshold fixed, isolating the effect of calibration. Section B.5 gives all eight cells, calibration and complete timings; Section B.6 reports application-sized circle null designs.

The larger torus study uses signal strength 0.3, 300 replications per cell and 199 permutations for FUSE. Table 1 compares two resource conditions using the same diffusion family. At equal sample size, FUSE gains 0.153 in high-frequency power over RA (paired 95% interval $[0.088, 0.218]$). With 5000 pairs available, RA can instead process twice as many observations in similar computation time, gaining 0.333 over FUSE on the first 2500 pairs ($[0.268, 0.398]$). An independent null-sample timing study selected the smaller size before power evaluation.

Table 1: Torus frequency-7 comparisons at signal strength 0.3. RA uses chi-square calibration; FUSE uses 199 permutations. Rates use 300 replications. Times include all steps through the p-value: medians over three fresh processes for equal n, or 30 null samples with alternating method order for similar time.

Comparison	Method	n	Null rate	Power	Time (s)
Equal n	RA	5000	0.040	0.613	0.287
	FUSE	5000	0.063	0.767	0.607
Similar time	RA	5000	0.040	0.613	0.251
	FUSE	2500	0.060	0.280	0.237

Both implementations use finite features and bounded blocks; in the similar-time comparison RA's median time is about 6% longer and its gain comes from processing more available pairs, whereas at fixed n FUSE exchanges additional computation for higher power. The timing comparison fixes the kernel family and permutation count and excludes data acquisition; Section C gives all sample sizes, including $n = 20000$, Wilson intervals and timing protocols.

Table 2: Pairwise pigeon-bearing results. Holm adjustment is applied separately to the three rank-adaptive and three permutation p-values.

			Rank-adaptive p		Permutation maxT p	
Contrast	n	$\widehat{r}_n$	Raw	Holm	Raw	Holm
control–trigeminal	40	2	0.223	0.223	0.922	0.922
control–olfactory	27	2	0.006	0.018	0.001	0.003
trigeminal–olfactory	27	3	0.049	0.099	< 0.001	0.003

5 Homing-Pigeon Bearings

The `pigeons` data set of the CRAN `circular` package, version 0.5-2 (Agostinelli and Lund, 2025), records one vanishing bearing relative to home for each of 108 birds: 41 controls, 40 trigeminal-section and 27 olfactory-section birds; we use all complete records, converted from degrees to radians. Bearing-blind seeded subsamples of the larger groups give $n = 40, 27, 27$. Both procedures use the six scales and 100 circle modes, with the RA cutoff from Section 4. We compare RA with symmetric permutation maxT of unbiased quadratic MMD with 9999 balanced relabelings and randomized ranks. Holm adjustment (Holm, 1979) is applied separately to each family of three p-values; Section B.8 specifies subsampling and the reference distributions.

The archive omits release-site, batch and randomization identifiers, so both procedures treat the pooled observations as independent samples and their p-values cannot account for unobserved site structure; the comparisons describe these archived groups without identifying a causal effect.

The control, trigeminal-section and olfactory-section samples had mean directions $6.3°$, $10.4°$ and $54.6°$ and resultant lengths (norms of the mean unit vectors) 0.746, 0.738 and 0.093, so the olfactory group is far less concentrated around home. Table 2 shows no control–trigeminal separation under either procedure and strong control–olfactory evidence after Holm adjustment. For trigeminal–olfactory, the rank-adaptive result was borderline before adjustment and nonsignificant after it ($p = 0.049$, Holm $p = 0.099$), whereas permutation maxT remained small ($p < 0.001$, Holm $p = 0.003$).

The procedures differ in both aggregation and reference distribution. Across 500 order/subsample reruns on the same birds, the fractions rejecting after the same three-comparison Holm adjustment were 0.012, 0.610 and 0.548 in table order (Section B.8), quantifying the sensitivity of the adjusted decisions to pairing and subsampling.

At these sample sizes the six circle null designs give rank-adaptive rejection rates between 0.025 and 0.036 (Section B.6), so the chi-square reference is conservative in the corresponding independent-sample simulations; together with the order sensitivity, this helps explain why the borderline trigeminal–olfactory result depends on the testing procedure.

6 Concluding Remarks

Diffusion-scale martingale MMD makes intrinsic kernel choice interpretable through the signal and null frequency balance. Diffusion pullback preserves identification, and covariance rank determines how a fixed collection of scales can be combined with asymptotic chi-square calibration. The spectral criterion predicts all six observed grid maximizers in the torus and sphere study. With observations available, the computational saving can yield higher power by processing more pairs; at equal sample size, MMD-FUSE-N better detects the harder signals. The pigeon analysis shows how aggregation, reference distribution and observation order jointly affect multiscale decisions.

The chi-square limits use a grid of diffusion times fixed before sampling, with computation based on the known spectrum of the manifold Laplacian. Increasing the number of retained spectral modes would permit detection of increasingly fine alternatives. This requires balancing the uniform kernel tail bound against quadratic-variation error while controlling small covariance eigenvalues. Estimating the Laplacian from an independent geometric sample is another extension. A conditional analysis would track how errors in the estimated eigenspaces affect both the MMD signal and the covariance rank threshold.

AI USE STATEMENT

Generative AI tools were used to revise wording and improve the clarity and readability of the manuscript; to identify potentially relevant literature, references, and search directions, and to assist with research execution, including drafting, debugging, and refining code. For mathematical proofs, AI was used primarily to review and check the authors' arguments for potential errors, inconsistencies, or gaps. AI-assisted manuscript revisions, literature-search suggestions, feedback on research ideas and proofs, and generated or modified code were reviewed and verified by the authors before incorporation into the final work. The authors independently assessed the accuracy and appropriateness of these outputs and retain full responsibility for the methodology, proofs, implementation, analysis, and conclusions presented in the paper.

REPRODUCIBILITY STATEMENT

The code and an accompanying `README` file are submitted with the manuscript. We provide a dedicated reproduction script for each figure and table, allowing the reported results to be reproduced individually. The `README` provides instructions for running the code.

REFERENCES

C. Agostinelli and U. Lund. *circular: Circular Statistics*, version 0.5-2. CRAN, 2025. `https://CRAN.R-project.org/package=circular`.

R. Bhatia. *Matrix Analysis*. Springer, New York, 1997.

F. Biggs, A. Schrab, and A. Gretton. MMD-FUSE: Learning and combining kernels for two-sample testing without data splitting. In *Advances in Neural Information Processing Systems*, volume 36, pages 75151–75188, 2023.

A. Chatterjee and A. Ramdas. A martingale kernel two-sample test. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pages 1–44, 2026.

X. Cheng and Y. Xie. Kernel two-sample tests for manifold data. *Bernoulli*, 30(4):2572–2597, 2024.

R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

E. De Vito, N. Mücke, and L. Rosasco. Reproducing kernel Hilbert spaces on manifolds: Sobolev and diffusion spaces. *Analysis and Applications*, 19(3):363–396, 2021.

A. Gagliardo, P. Ioalè, M. Savini, and M. Wild. Navigational abilities of homing pigeons deprived of olfactory or trigeminally mediated magnetic information when young. *Journal of Experimental Biology*, 211:2046–2051, 2008.

E. Giné. Invariant tests for uniformity on compact Riemannian manifolds based on Sobolev norms. *The Annals of Statistics*, 3(6):1243–1266, 1975.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

A. Grigor'yan. *Heat Kernel and Analysis on Manifolds*. AMS/IP Studies in Advanced Mathematics, volume 47. American Mathematical Society and International Press, 2009.

P. Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, New York, 1980.

J. Hemerik and J. Goeman. Exact testing with random permutations. *TEST*, 27(4):811–825, 2018.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi. Kernel methods on Riemannian manifolds with Gaussian RBF kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(12):2464–2477, 2015.

P. E. Jupp. Data-driven Sobolev tests of uniformity on compact Riemannian manifolds. *The Annals of Statistics*, 36(3):1246–1260, 2008.

J. M. Kübler, W. Jitkrittum, B. Schölkopf, and K. Muandet. A witness two-sample test. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 1403–1419, 2022.

S. Rosenberg. *The Laplacian on a Riemannian Manifold: An Introduction to Analysis on Manifolds*. Cambridge University Press, Cambridge, 1997.

A. Schrab, I. Kim, M. Albert, B. Laurent, B. Guedj, and A. Gretton. MMD aggregated two-sample test. *Journal of Machine Learning Research*, 24(194):1–81, 2023.

D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, New York, 1980.

S. Shekhar, I. Kim, and A. Ramdas. A permutation-free kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 35, pages 18168–18180, 2022.

B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

W. Zaremba, A. Gretton, and M. Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Z. Zhou, X. Tian, L. Peng, C. Lei, A. Schrab, D. J. Sutherland, and F. Liu. DUAL: Learning diverse kernels for aggregated two-sample and independence testing. In *Advances in Neural Information Processing Systems*, volume 38, pages 130661–130695, 2025.

A COMPUTATIONAL AND METHODOLOGICAL DETAILS

This supplement starts with implementation details, followed by the numerical designs and application in Section B and the larger-sample comparisons in Section C. Section D gives special cases. The remaining sections collect preliminary bounds and proofs in the order of the main results.

Algorithm 1 Spectral martingale MMD

Require: Independent samples with declared pairing/order, effective times $a_1, \dots, a_L$, complete spectral cutoff J, normalization, κ_n and level α.

Initialize feature sums $R_\ell = 0$, $\boldsymbol{S} = 0$, $\boldsymbol{\Sigma} = 0$.
for $i = 1, \dots, n$ **do**
  Compute $r_{i\ell}$ at all scales; if $i > 1$, set $D_{i\ell} = R_\ell^\top r_{i\ell}/(i-1)$.
  If $i > 1$, update $\boldsymbol{S} \leftarrow \boldsymbol{S} + \boldsymbol{D}_i$, $\boldsymbol{\Sigma} \leftarrow \boldsymbol{\Sigma} + \boldsymbol{D}_i \boldsymbol{D}_i^\top$.
  Update $R_\ell \leftarrow R_\ell + r_{i\ell}$ at all scales.
end for
For $L = 1$, report T_n and its Gaussian p-value; otherwise form $\widehat{\boldsymbol{G}}_n = \boldsymbol{\Sigma}/\log n$ and report Q_n^{RA}, $\widehat{r}_n$ and the chi-square p-value.

Algorithm 2 Diffusion-scale martingale MMD with rank-adaptive aggregation

Require: Two independent samples with a predetermined pairing/order $Z_i = (X_i, Y_i)$, predetermined kernels $k_{t_1}, \dots, k_{t_L}$, level α, threshold κ_n.

Initialize $\boldsymbol{S} \leftarrow \boldsymbol{0} \in \mathbb{R}^L$ and $\boldsymbol{\Sigma} \leftarrow 0_{L\times L}$.
for $i = 2, \dots, n$ **do**
  for $\ell = 1, \dots, L$ **do**
    $D_{i\ell} \leftarrow (i-1)^{-1} \sum_{j=1}^{i-1} h_{t_\ell}(Z_i, Z_j)$.
  end for
  $\boldsymbol{D}_i \leftarrow (D_{i1}, \dots, D_{iL})^\top$.
  $\boldsymbol{S} \leftarrow \boldsymbol{S} + \boldsymbol{D}_i$ and $\boldsymbol{\Sigma} \leftarrow \boldsymbol{\Sigma} + \boldsymbol{D}_i \boldsymbol{D}_i^\top$.
end for
if $L = 1$ **then**
  $T \leftarrow S_1/\sqrt{\Sigma_{11}}$ with the convention $T = 0$ if $\Sigma_{11} = 0$.
  Reject iff $T > z_{1-\alpha}$.
else
  $\boldsymbol{G} \leftarrow \boldsymbol{\Sigma}/\log n$; compute eigenpairs $(\widehat{\lambda}_j, \widehat{u}_j)$.
  $\widehat{r} \leftarrow \#\{j : \widehat{\lambda}_j > \kappa_n\}$.
  $Q \leftarrow \sum_{\widehat{\lambda}_j > \kappa_n} (\widehat{u}_j^\top \boldsymbol{S})^2/(\widehat{\lambda}_j \log n)$.
  Reject iff $\widehat{r} > 0$ and $Q > \chi^2_{\widehat{r},1-\alpha}$.
end if

The empirical baselines make different uses of the same diffusion features. The fixed-scale martingale test follows one frequency weighting; cross-witness estimates the mean feature difference on one sample split and evaluates it on the other, while block MMD averages within-block U-statistics. These give Gaussian references through sequential prediction, sample splitting, and block averaging, respectively. At the grid level, permutation maxT selects the strongest standardized scale, normalized FUSE uses a soft maximum, and rank aggregation uses a covariance quadratic form. Comparing the latter under chi-square and permutation reference laws holds the aggregation rule fixed.

Quadratic MMD, cross-MMD and direct-kernel mMMD require $O(n^2)$ kernel work per statistic; block and linear variants reduce this cost. Witness learning and learned-kernel aggregation add fitting costs. For L diffusion scales, evaluating the predictable increments costs $O(Ln^2)$ directly or $O(nLJ)$ with J spectral features. Covariance accumulation adds $O(nL^2)$ work. Section B.5 reports complete timings, including permutation calibration.

The R implementation used for the 95-cell study in Section B retains all feature and increment arrays, using $O(nJ + nL + JL + L^2)$ memory; its memory diagnostic is the per-replication R-heap peak. Section C reports process resident memory for the larger-sample Python implementations.

B Supplementary Numerical Evidence

B.1 Design and inferential quantities

We evaluated the circle $\mathbb{S}^1$, the flat torus $\mathbb{T}^2$, and the sphere $\mathbb{S}^2$ on the six predetermined effective times

$$a\lambda_1 \in \{0.025, 0.05, 0.10, 0.25, 0.50, 1\}.$$

The primary finite kernels retain 100 nonconstant circle modes, 400 torus modes through a complete lattice shell, and all 440 spherical harmonics of degrees 1 through 20. The threshold is $\kappa_n = (\log n)^{-1/4}$ in the nonlocal cells except for the declared threshold-sensitivity cells; the eight local spectral cells use $\kappa_n = (\log n)^{-1/8}$, which satisfies the stronger local rank-recovery condition in (15). Each top-level cell contains exactly $R = 500$ Monte Carlo replications. The 95 cells comprise 36 null-calibration cells, 24 fixed spectral alternatives, 8 local spectral alternatives, 6 localized sphere alternatives, 6 rank-stress cells, 4 threshold cells, 8 truncation cells, and 3 matched-comparator cells.

Within each replication, all methods receive the same generated observations and spectral features. The fixed-scale martingale test, split-sample cross-witness test, and block MMD all use $a\lambda_1 = 0.10$; rank adaptation uses the full grid. The cross-witness test constructs the empirical feature-mean difference from the first $\lfloor n/2 \rfloor$ observations in each sample and evaluates it on the remaining observations. It studentizes the difference of the two mean witness evaluations using their held-out sample variances. Block MMD averages unbiased within-block statistics over complete blocks of size $\lfloor \sqrt{n} \rfloor$, discards the final incomplete block, and estimates its standard error from the block statistics. Both comparators use an upper-tail Gaussian critical value. Three dedicated $n = 100$ torus cells additionally compare the unbiased quadratic diffusion-kernel MMD with 1000 balanced relabelings and a symmetric maxT diffusion-grid aggregate: the observed and relabeled grid statistics share the same median/median absolute deviation (MAD) standardization. These three cells use the stated fixed-kernel and maxT constructions; Section B.5 compares normalized MMD-FUSE on the same kernel family. The best fixed scale in the tables is an oracle diagnostic and is not a feasible level-α selector.

Rejection frequency estimates size under a common law and power under an alternative. Retained rank records how many contrast directions enter the test. Individual rejection rates have Wilson intervals; at $R = 500$, their Monte Carlo standard error is at most 0.022. The reproducibility archive contains all 47500 replications and the seed-generation code. Runtime covers the complete replication, while memory is the per-replication R-heap diagnostic described in Section A.

For fixed-kernel and maxT permutations, let B be the number of balanced relabelings, $N_>$ the number of relabeled statistics exceeding the observed statistic, and $N_=$ the number tied with it. The implemented randomized tail probability is $\{N_> + U(N_= + 1)\}/(B + 1)$, with an independent $U \sim \mathrm{Unif}(0, 1)$ (Hemerik and Goeman, 2018, Section 3.4). The maxT statistic standardizes each scale by its median and median absolute deviation over the observed and relabeled statistics together, then takes the maximum across scales. This symmetric construction preserves the exchangeability needed for permutation calibration.

B.2 Sampling distributions

Under each null design, the two samples are independent draws from the same law, with $n \in \{100, 200, 500, 1000\}$ observations per sample. The uniform law on each manifold provides a reference with orthogonal nonconstant Laplace modes. The unimodal and mixture laws change the feature covariance while preserving the equality of the two distributions.

On the circle, angles lie in $[0, 2\pi)$. The unimodal law is von Mises with mean direction 0.4 radians and concentration 2.5. The mixture assigns equal weight to von Mises components with antipodal mean directions $\pi/3$ and $4\pi/3$, both with concentration 5. The finite-support rank experiment instead uses the uniform law on $\{0, \pi/2, \pi, 3\pi/2\}$.

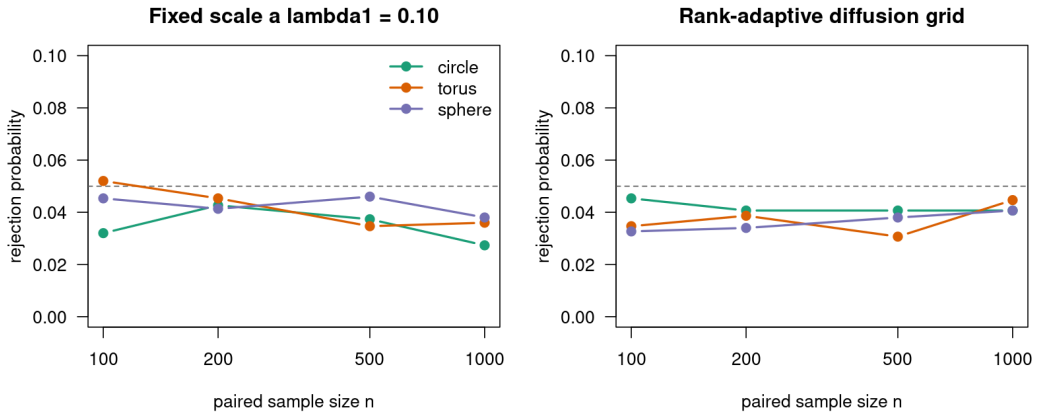


Figure 2: Mean null rejection probability across the three null laws. The dashed line is 0.05; each plotted mean averages three $R = 500$ cells.

On the torus, observations are angular pairs divided by 2π, giving coordinates in $[0, 1)^2$. The unimodal law has independent von Mises angles with mean directions $(0.3, 1.0)$ radians and concentrations $(2.2, 1.8)$. The balanced mixture has component mean-direction pairs $(\pi/3, 2\pi/3)$ and $(4\pi/3, 5\pi/3)$, with concentration 5 in each coordinate. The two angles are independent conditional on a shared component label, so this construction gives two joint modes.

On the sphere, we use von Mises–Fisher densities proportional to $\exp(\kappa\mu^\top x)$, where μ is a unit mean direction and κ is the concentration. The unimodal null has concentration 3 and mean direction proportional to $(0.2, -0.1, 0.97)$. The balanced mixture has concentration 7 in both components, centered at $(0, 0, 1)$ and $(1, 0, 0)$. The localized alternatives change one distributional feature at a time: the mean-shift comparison keeps concentration 5 and rotates the mean direction by 20 degrees from $(0, 0, 1)$ to $(\sin(\pi/9), 0, \cos(\pi/9))$; the concentration comparison keeps mean direction $(0, 0, 1)$ and changes concentration from 2 to 5; the mixture comparison keeps the same two concentration-7 components and changes the weight of the $(0, 0, 1)$ component from 0.5 to 0.7. These comparisons use $n = 200$ and 500.

The fixed spectral alternatives are those in Section 4, with the first sample uniform and the second perturbed in one Fourier or Legendre direction. Local torus alternatives use frequencies 1 and 7 and replace signal strength 0.3 by $\min\{0.8, 2.5(\log n)^{1/4}/\sqrt{n}\}$. This signal strength follows the local scale in Theorem 3.2. Both the fixed and local spectral designs use the four sample sizes above. The exact-duplicate rank grid replaces 0.50 in the default grid by a second 0.25; the nearby-time grid replaces it by 0.251. Both use the uniform circle law; these and the finite-support rank comparisons use $n = 100$ and 500. Truncation comparisons use torus frequencies 1 and 7 at $n = 500$ with $J \in \{10, 25, 50, 100\}$, separating the effect of reduced computation from the loss of represented spectral directions.

B.3 Null rejection rates and rank behavior

The default fixed-scale rejection probability ranged from 0.016 to 0.058 across the 36 null cells, while the rank-adaptive rate ranged from 0.026 to 0.050. More broadly, across all six fixed scales the observed null rates ranged from 0.016 to 0.062. Thirty of the 36 cellwise Wilson intervals for the default scale contained 0.05; all six exclusions were conservative. Thus the study reveals finite-sample conservatism in some designs but no material size inflation. Diagnostics at levels 0.01 and 0.10 gave the same qualitative conclusion.

Rank adaptation maintained near-nominal rejection in the redundant-grid and finite-support designs (Table 4). The exact and nearby duplicates both gave median retained rank 2; the finite-support law gave rank 1. These are the directions retained at the observed sample sizes, while Theorem 3.4 describes recovery of the population rank as the cutoff vanishes. Increasing the exponent in $\kappa_n =$

Table 3: Null rejection probabilities at level 0.05. Each entry is the mean across the uniform, nonuniform, and mixture nulls; brackets give the minimum and maximum across those three cells, not a confidence interval.

Manifold	n	Fixed $a\lambda_1 = 0.10$	Rank-adaptive grid
circle	100	0.032 [0.028,0.034]	0.045 [0.038,0.050]
circle	200	0.043 [0.036,0.050]	0.041 [0.038,0.044]
circle	500	0.037 [0.030,0.042]	0.041 [0.030,0.050]
circle	1000	0.027 [0.016,0.034]	0.041 [0.032,0.046]
torus	100	0.052 [0.044,0.058]	0.035 [0.032,0.038]
torus	200	0.045 [0.042,0.048]	0.039 [0.038,0.040]
torus	500	0.035 [0.022,0.044]	0.031 [0.026,0.038]
torus	1000	0.036 [0.026,0.044]	0.045 [0.044,0.046]
sphere	100	0.045 [0.042,0.050]	0.033 [0.028,0.036]
sphere	200	0.041 [0.032,0.046]	0.034 [0.028,0.038]
sphere	500	0.046 [0.038,0.050]	0.038 [0.030,0.042]
sphere	1000	0.038 [0.028,0.046]	0.041 [0.036,0.044]

Table 4: Rank and cutoff sensitivity at level 0.05, with 500 replications per design. The cutoff is $\kappa_n = (\log n)^{-\gamma}$; all kernel normalizations are held fixed.

Manifold	Design	n	γ	Median $\widehat{r}_n$	Rejection
Circle	Exact duplicate	100	$1/4$	2	0.036
Circle	Exact duplicate	500	$1/4$	2	0.034
Circle	Nearby times	100	$1/4$	2	0.038
Circle	Nearby times	500	$1/4$	2	0.040
Circle	Finite support	100	$1/4$	1	0.044
Circle	Finite support	500	$1/4$	1	0.052
Torus	Uniform	500	$1/8$	3	0.050
Torus	Uniform	500	$1/6$	3	0.054
Torus	Uniform	500	$1/4$	3	0.034
Torus	Uniform	500	$1/3$	4	0.046

$(\log n)^{-\gamma}$ lowers the cutoff: the torus study retained a fourth direction at $\gamma = 1/3$, with rejection rates remaining between 0.034 and 0.054 across the four choices.

Kernel normalization. The normalized-volume convention fixes the overall multiplier of each kernel separately from its diffusion time. To see how this multiplier affects rank selection, multiply kernel ℓ by $c_\ell > 0$ and write $C = \mathrm{diag}(c_1, \ldots, c_L)$. Then $\boldsymbol{D}_i$, $\boldsymbol{S}_n$ and $\widehat{\boldsymbol{G}}_n$ become $C\boldsymbol{D}_i$, $C\boldsymbol{S}_n$ and $C\widehat{\boldsymbol{G}}_n C$. Each single-scale $T_n(t_\ell)$ is unchanged, as is the full-rank quadratic form when all directions are retained. The absolute cutoff can change the retained subspace. For a common multiplier c, its eigenvectors stay fixed and its eigenvalues multiply by c^2, giving exactly the original rank statistic at cutoff κ_n/c^2. Coordinate-specific multipliers can also rotate the retained directions.

We examine this effect on the fixed pigeon samples and orders in Table 2, keeping the six diffusion times, basis and cutoff unchanged. Besides the original normalization, we use common multipliers $1/2$ and 2, and coordinate multipliers $(1/2, 1/2, 1, 1, 2, 2)$ in increasing-time order. A common multiplier of 2 changes the retained ranks from $(2, 2, 3)$ to $(3, 3, 3)$ and the Holm p-values from $(0.223, 0.018, 0.099)$ to $(0.206, 0.043, 0.099)$. A common multiplier of $1/2$ gives ranks $(2, 2, 2)$ and Holm values $(0.223, 0.018, 0.052)$; coordinate-specific scaling gives $(2, 2, 3)$ and $(0.169, 0.024, 0.098)$. The adjusted decisions at 0.05 agree in these cases, although ranks and p-values change. These fixed-data calculations explain why kernel normalization is part of the finite-sample procedure.

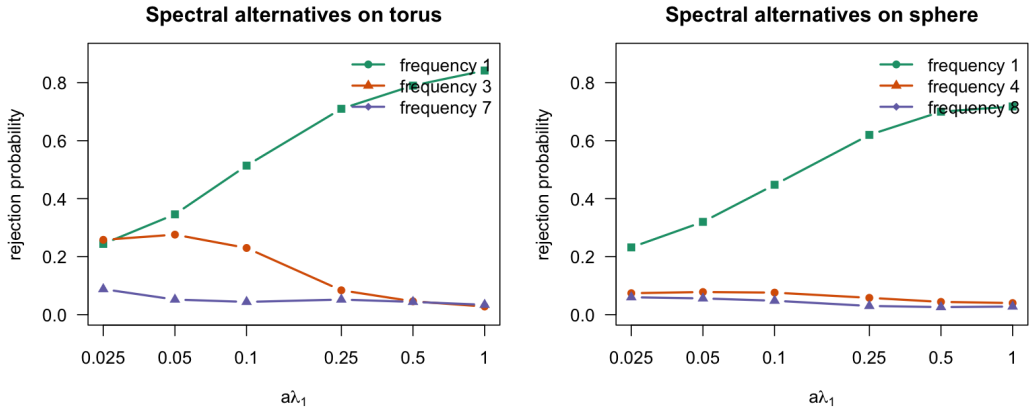


Figure 3: Fixed-scale rejection probabilities at level 0.05 for spectral alternatives with $n = 1000$. Larger diffusion times favor the lowest frequencies, whereas medium and high frequencies peak earlier.

Table 5: Rejection probability at level 0.05 for spectral alternatives of fixed signal strength with $n = 1000$. "Best fixed" is an oracle envelope over the six predetermined scales. Rank denotes chi-square rank aggregation, Cross the split-sample cross-witness test, and Block block MMD, defined in Section B.

Manifold	Freq.	Best $a\lambda_1$	Best fixed	Fixed 0.10	Rank	Cross	Block
torus	1	1	0.842	0.514	0.578	0.638	0.114
torus	3	0.05	0.276	0.230	0.162	0.318	0.078
torus	7	0.025	0.088	0.044	0.072	0.046	0.070
sphere	1	1	0.718	0.448	0.460	0.556	0.140
sphere	4	0.05	0.078	0.076	0.046	0.102	0.054
sphere	8	0.025	0.060	0.048	0.058	0.052	0.036

B.4 Frequency, aggregation, and truncation

Aggregation improves on the default scale for some alternatives, but pays for combining several directions and can discard weak ones through thresholding. The best fixed scale is an oracle envelope over the six scales.

Figure 3 and table 5 give the frequency ordering compared with the exact efficiency in Section B.7. For frequency 1, the observed oracle-best scale was $a\lambda_1 = 1$ on both manifolds; for the medium and high alternatives it moved to 0.05 and 0.025, although the corresponding sphere maxima included near ties at $R = 500$. Power was modest for the highest frequencies even though those modes were present in the primary bases. This is consistent with attenuation by the diffusion weights and the finite signal strength, not a missing-feature artifact. Aggregation was not uniformly best: it avoided committing to one scale, but its rank threshold discarded weak directions. Paired contrasts quantify differences relative to the default fixed scale; comparisons with the oracle envelope or cross-witness marginal rejection rates are descriptive.

The local frequency-1 torus cells retained nontrivial power across sample sizes under the prescribed $(\log n)^{1/4}/\sqrt{n}$ signal strength. For frequency 7 the induced diffusion-kernel MMD noncentrality was too small to distinguish power from size on this grid. These results illustrate the scale in Theorem 3.2; they are not a numerical verification of convergence to a particular noncentral limit.

Table 6 shows high power for changes in mean direction and concentration by $n = 200$. The mixture-weight alternative separates the methods more clearly: the default fixed-scale, rank-adaptive, and cross-witness procedures have similar rejection rates, all substantially above block MMD. By $n = 500$, their power is close to one, while block MMD still rejects in about 72% of replications.

Table 6: Rejection probability at level 0.05 for localized sphere alternatives. Best fixed is the oracle envelope; the remaining method columns follow Table 5.

Alternative	n	Best $a\lambda_1$	Best fixed	Fixed 0.10	Rank	Cross	Block
mean shift	200	0.25	1.000	0.994	0.996	0.994	0.756
mean shift	500	0.05	1.000	1.000	1.000	1.000	1.000
concentration	200	0.25	1.000	0.992	0.996	0.998	0.858
concentration	500	0.025	1.000	1.000	1.000	1.000	0.998
mixture weight	200	1	0.880	0.736	0.670	0.742	0.304
mixture weight	500	1	0.998	0.976	0.988	0.984	0.718

Table 7: Matched $n = 100$ torus comparison at level 0.05. Permutation columns use 1000 balanced relabelings within each replication. Low and high denote torus frequencies 1 and 7 at signal strength 0.3; the Gaussian baselines follow Table 5.

Scenario	Fixed 0.10	Rank	Cross	Block	Perm. fixed	Perm. maxT
null	0.036	0.022	0.048	0.032	0.034	0.052
low	0.068	0.042	0.096	0.044	0.102	0.160
high	0.040	0.052	0.046	0.054	0.052	0.060

The gain from more observations thus extends beyond individual spectral alternatives to localized changes in the spherical distribution.

In Table 7, the null rates ranged from 0.022 for rank adaptation to 0.052 for permutation maxT. The permutation aggregate had the largest low-frequency rejection rate, 0.160, but none of the procedures separated the high-frequency alternative materially from its null performance at $n = 100$. Because the reference distributions differ across methods and Monte Carlo uncertainty is substantial at $R = 500$, these three cells do not support a general ranking.

Truncation provided a negative control. At $n = 500$ and torus frequency 1, decreasing J from 100 to 10 raised the rank-adaptive rejection rate from 0.248 to 0.352 by removing high-frequency noise. In contrast, every $J \leq 100$ truncation omits frequency 7 and gave rejection between 0.042 and 0.064. A smaller kernel can therefore improve power for a represented low-frequency direction while losing identification of alternatives outside its span. Primary truncations were chosen to contain all declared alternatives and complete repeated-eigenvalue shells.

B.5 MMD-FUSE COMPARISON AND PERMUTATION CALIBRATION

This comparison separates the aggregation rule from its reference distribution. At $n = 40$ and 100, the first torus sample is uniform and the second is uniform or has density $1 + 0.6\cos(2\pi m x_1)$ for $m = 1, 3, 7$. Each of the eight cells has 500 independent replications. Every method receives the same observations, six diffusion times and 400 nonconstant modes; the signal strength is common to all three frequencies. The signal-strength-0.3 comparisons are in Tables 5 and 7.

For permutation rank aggregation, draw $B = 199$ independent uniform permutations of the $2n$ pooled observations. The first n permuted values form X, the remaining n form Y, and their positions specify the pairs and order. Recompute every increment, the covariance, retained rank and Q_n^{RA} for each ordering. With $Q^{(0)}$ the observed statistic, report $p = \{1 + \sum_{b=1}^{B} \mathbf{1}(Q^{(b)} \geq Q^{(0)})\}/(B+1)$ and reject for $p \leq 0.05$. Under H_0, the pooled observations are exchangeable, so this rank test has finite-sample level at most 0.05 (Hemerik and Goeman, 2018). The threshold is fixed before testing; we compare exponents $1/4$ and $1/8$ in $\kappa_n = (\log n)^{-\gamma}$.

MMD-FUSE-N (Biggs et al., 2023) forms a soft maximum of normalized unbiased quadratic MMDs with uniform weight on the same six kernels. For pooled observations $z_1, \ldots, z_{2n}$, its kernel normalizer is $\widehat{N}_\ell = \{n(n-1)\}^{-1} \sum_{i \neq j} p_{a_\ell, J}(z_i, z_j)^2$, including the constant mode in $p_{a_\ell, J}$. The statistic is $\lambda^{-1} \log[L^{-1} \sum_{\ell=1}^{L} \exp\{\lambda \widehat{\mathrm{MMD}}^2_{\ell,u} / \sqrt{\widehat{N}_\ell}\}]$. It uses the same 199 permutations and conservative rank p-value. We report the temperature corresponding to the public reference implementation, $\lambda = n(n-1)$, and also the paper's stated experimental choice $\lambda = \sqrt{n(n-1)}$. Both are specified before the runs.

Table 8: Matched torus rejection rates at level 0.05, with 95% Wilson intervals from 500 replications per cell. Rank aggregation uses $\gamma = 1/4$; FUSE uses $\lambda = n(n-1)$.

n	Frequency	RA chi-square	RA permutation	MMD-FUSE-N
40	Null	0.022 [0.012, 0.039]	0.034 [0.021, 0.054]	0.044 [0.029, 0.066]
40	1	0.068 [0.049, 0.094]	0.102 [0.078, 0.132]	0.176 [0.145, 0.212]
40	3	0.024 [0.014, 0.041]	0.054 [0.037, 0.077]	0.082 [0.061, 0.109]
40	7	0.024 [0.014, 0.041]	0.044 [0.029, 0.066]	0.060 [0.042, 0.084]
100	Null	0.028 [0.017, 0.046]	0.042 [0.028, 0.063]	0.030 [0.018, 0.049]
100	1	0.200 [0.167, 0.237]	0.262 [0.225, 0.302]	0.488 [0.444, 0.532]
100	3	0.062 [0.044, 0.087]	0.072 [0.052, 0.098]	0.126 [0.100, 0.158]
100	7	0.040 [0.026, 0.061]	0.066 [0.047, 0.091]	0.060 [0.042, 0.084]

Table 8 shows the strongest separation at low frequency. At $n = 100$, the paired power gain from permuting the rank statistic is 0.062 (95% normal interval $[0.039, 0.085]$); the additional gain from MMD-FUSE-N is 0.226 ($[0.182, 0.270]$). These intervals use differences of rejection indicators on the same samples. At frequency 3, FUSE exceeds permutation rank aggregation by 0.054 ($[0.021, 0.087]$). At frequency 7, FUSE and permutation rank aggregation both have low power, with paired difference -0.006 ($[-0.031, 0.019]$). Null rejection rates under permutation range from 0.030 to 0.044. Changing the rank exponent to $1/8$ or the FUSE temperature to the paper's choice leaves the principal power ordering unchanged (Table 9).

Table 10 measures each default method separately on the first 50 uniform-null samples at each n. Timers run from feature construction through the final p-value, including permutation generation and kernel construction where used; data simulation and compilation warmup are excluded. The NumPy/Numba implementation uses one BLAS thread on an AMD Ryzen Threadripper 3990X. Rank aggregation updates witnesses directly in the 400-dimensional feature space, repeating the covariance and rank calculations for every permutation. FUSE constructs six $2n \times 2n$ Gram matrices once and reuses them for all quadratic statistics. The chi-square reference has the lowest cost here; permuting the order-sensitive rank statistic costs more than combining the cached quadratic statistics.

Table 9: Threshold and temperature sensitivity for the same observations and permutations. Rank aggregation uses $\gamma = 1/8$; FUSE uses $\lambda = \sqrt{n(n-1)}$. Brackets are 95% Wilson intervals.

n	Frequency	RA chi-square	RA permutation	MMD-FUSE-N
40	Null	0.022 [0.012, 0.039]	0.034 [0.021, 0.054]	0.042 [0.028, 0.063]
40	1	0.068 [0.049, 0.094]	0.100 [0.077, 0.129]	0.188 [0.156, 0.225]
40	3	0.026 [0.015, 0.044]	0.060 [0.042, 0.084]	0.082 [0.061, 0.109]
40	7	0.022 [0.012, 0.039]	0.042 [0.028, 0.063]	0.050 [0.034, 0.073]
100	Null	0.030 [0.018, 0.049]	0.040 [0.026, 0.061]	0.036 [0.023, 0.056]
100	1	0.200 [0.167, 0.237]	0.258 [0.222, 0.298]	0.508 [0.464, 0.552]
100	3	0.062 [0.044, 0.087]	0.082 [0.061, 0.109]	0.128 [0.102, 0.160]
100	7	0.040 [0.026, 0.061]	0.056 [0.039, 0.080]	0.050 [0.034, 0.073]

Table 10: Complete method time in milliseconds: median [interquartile range] across 50 samples per size. Rank aggregation uses $\gamma = 1/4$, FUSE uses $\lambda = n(n-1)$, and both permutation methods use 199 draws.

n	RA chi-square	RA permutation	MMD-FUSE-N
40	6.4 [6.3, 6.7]	36.0 [35.8, 36.4]	10.5 [10.4, 10.8]
100	8.8 [8.6, 9.1]	71.9 [71.3, 73.0]	22.1 [21.6, 23.6]

B.6 Null rejection rates at the application sample sizes

The six application-sized null designs use the circular laws in Section B.2, 100 nonconstant modes, the six diffusion times and $\kappa_n = (\log n)^{-1/4}$. Each common law is tested at $n = 27$ and 40, with 1000 independent replications per cell. The corresponding cross-witness rejection rates range from 0.049 to 0.069, and block-MMD rates from 0.026 to 0.044. Rank adaptation is consistently conservative at these sample sizes. These independent-sample null designs assess the reference approximation; release-site structure in the pigeon data is an additional sampling issue.

B.7 Direct spectral-efficiency comparison

The deterministic calculation uses all 400 retained torus modes or 440 spherical harmonics, including multiplicity, and the same six effective times as the Monte Carlo study. Under the uniform null, the nonconstant eigenfunctions are orthonormal, so $N(a) = \sum_{j=1}^{J} e^{-2a\lambda_j}$. Unit energy in the chosen eigenspace gives $A(a) = e^{-a\lambda_{\mathrm{sig}}}$. Multiplying the direction by a fixed coefficient changes efficiency magnitude but preserves its ordering over a. We divide by the largest grid value separately for each direction. The predicted maximizing times are $a\lambda_1 = 1, 0.05, 0.025$ for torus frequencies $1, 3, 7$ and spherical degrees $1, 4, 8$. These are exactly the six observed maxima in Table 5. Small empirical differences near a maximum remain subject to Monte Carlo uncertainty; the comparison does not assert convergence of power at fixed signal strength to a local limiting curve.

B.8 Application order and subsample diagnostics

The trigeminal intervention was bilateral section of the ophthalmic branch; the olfactory intervention was bilateral section of the olfactory nerve. The sampling unit is a bird. We convert bearings to radians and use the six diffusion times in Section 4, retaining 100 nonconstant circular modes and setting $\kappa_n = (\log n)^{-1/4}$. Each pairwise contrast uses every observation from the smaller group and a bearing-blind seeded subsample of the larger group, giving $n = 40$ for control–trigeminal and $n = 27$ for the two olfactory contrasts. We compute the asymptotic rank-adaptive p-value and a symmetric maxT comparison based on the unbiased quadratic MMD with 9999 balanced relabelings. The latter uses the same median/MAD standardization for observed and relabeled statistics and randomized permutation ranks, as detailed in Section B. Holm adjustment is applied separately within each family of three p-values.

Table 11: Application-sized circle null rejection rates at level 0.05. Brackets are 95% Wilson intervals from 1000 replications per cell. Fixed denotes the one-sided martingale test at $a\lambda_1 = 0.10$.

n	Common law	Rank-adaptive [interval]	Fixed [interval]	Median $\widehat{r}_n$
27	Uniform	0.036 [0.026, 0.049]	0.035 [0.025, 0.048]	2
27	Unimodal	0.025 [0.017, 0.037]	0.035 [0.025, 0.048]	2
27	Mixture	0.026 [0.018, 0.038]	0.032 [0.023, 0.045]	2
40	Uniform	0.032 [0.023, 0.045]	0.037 [0.027, 0.051]	2
40	Unimodal	0.032 [0.023, 0.045]	0.044 [0.033, 0.059]	2
40	Mixture	0.025 [0.017, 0.037]	0.039 [0.029, 0.053]	2

Table 12: Exact relative efficiency on the declared grid. Torus frequency m and spherical degree ℓ identify the signal direction. Each row is normalized to maximum 1; tiny entries are reported as < 0.01. The maximizing time agrees with the observed $n = 1000$ maximum in every row.

		Effective time $a\lambda_1$					
Manifold	Frequency	0.025	0.05	0.10	0.25	0.50	1
Torus	1	0.26	0.37	0.50	0.72	0.88	1.00
Torus	3	0.88	1.00	0.92	0.40	0.07	< 0.01
Torus	7	1.00	0.42	0.05	< 0.01	< 0.01	< 0.01
Sphere	1	0.27	0.38	0.52	0.75	0.91	1.00
Sphere	4	0.90	1.00	0.87	0.33	0.04	< 0.01
Sphere	8	1.00	0.58	0.14	< 0.01	< 0.01	< 0.01

The package attributes the data to Gagliardo et al. (2008). Its group counts and circular summaries agree with the pooled first experimental series in that paper, which comprised three unfamiliar release sites; the package documentation's description of two sites is inconsistent with the primary report. Every complete record was eligible for analysis; each unequal contrast used an equal-sized subsample as specified below. Figure 4 displays the archived bearings and their differences across diffusion times; the package and record-level preprocessing code accompany the analysis.

Each main contrast samples and orders record indices without replacement using the fixed seeds in `code/run_application.R`; the two sampled index sequences define the pairing. The three seeds are consecutive integers beginning at 2026082401, in table order. Rerun r adds $1000 + r$ to the corresponding seed. Across 500 predetermined deterministic order/subsample reruns, the rank-adaptive unadjusted $p < 0.05$ rejection fractions were 0.024, 0.842, and 0.802 for control–trigeminal, control–olfactory, and trigeminal–olfactory; the corresponding fixed-scale fractions were 0.000, 0.928, and 0.956. Applying Holm adjustment to the three rank-adaptive p-values with the same rerun index gives rejection fractions 0.012, 0.610 and 0.548. Both olfactory contrasts reject in 199 of the 500 reruns; none of the three rejects in 120. Each rerun reuses the same archived birds, changing only the chosen subsample and pair order. These fractions describe sensitivity to those choices, not repeated-sample power or an aggregated p-value. The primary results in Table 2 use the single predetermined order; the reruns do not select a replacement order.

C Power and computation at larger samples

C.1 Matched size and power

The larger torus comparison retains the six diffusion times, 400 nonconstant modes and rank cutoff $(\log n)^{-1/4}$. The first sample is uniform; the second is uniform or has density $1 + 0.3\cos(2\pi m x_1)$ for $m = 1, 3, 7$. Each of the eight cells at $n = 1000, 5000$ uses 300 independent samples. Two further cells at $n = 20000$, also with 300 independent replications each, examine the null and the remaining high-frequency gap at $m = 7$. Chi-square rank aggregation and MMD-FUSE-N receive the same observations. FUSE uses 199 ordered permutations, the temperature $n(n-1)$ and the conservative rank p-value defined in Section B.5. No tuning is selected from these samples. A preceding null-data cost probe at $n = 500, 2000, 5000$ determined computational feasibility; its data are excluded from the power study.

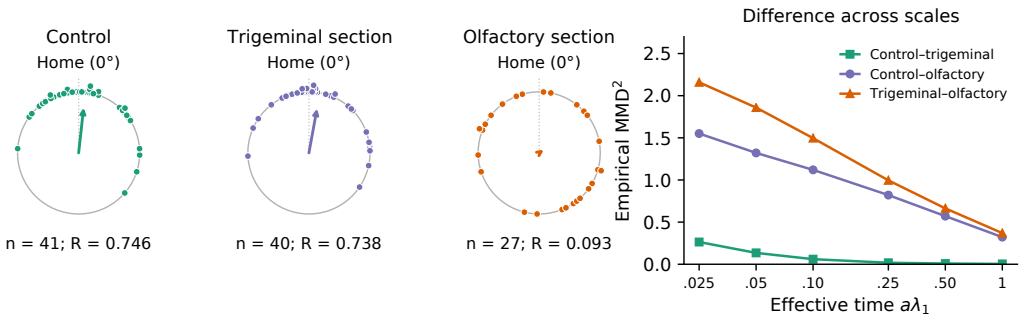


Figure 4: Archived vanishing bearings by group. Each dot is one bird; repeated angles are separated radially. Home is at 0°, and angles increase clockwise. Arrows show mean unit vectors, with length R measuring concentration. The olfactory-section bearings are much less concentrated around home. Right: squared distances between empirical mean feature vectors on the paired subsamples from Table 2, over the same diffusion grid.

Table 13: Larger-sample torus rejection rates at level 0.05, with 95% Wilson intervals from 300 replications per cell. Both procedures use the same finite diffusion kernels; RA denotes chi-square rank aggregation.

n	Frequency	RA chi-square	MMD-FUSE-N
1000	Null	0.060 [0.038, 0.093]	0.050 [0.031, 0.081]
1000	1	0.593 [0.537, 0.647]	0.933 [0.899, 0.956]
1000	3	0.177 [0.138, 0.224]	0.440 [0.385, 0.497]
1000	7	0.080 [0.054, 0.116]	0.107 [0.077, 0.147]
5000	Null	0.040 [0.023, 0.069]	0.063 [0.041, 0.097]
5000	1	1.000 [0.987, 1.000]	1.000 [0.987, 1.000]
5000	3	0.987 [0.966, 0.995]	1.000 [0.987, 1.000]
5000	7	0.613 [0.557, 0.667]	0.767 [0.716, 0.811]
20000	Null	0.027 [0.014, 0.052]	0.057 [0.036, 0.089]
20000	7	1.000 [0.987, 1.000]	1.000 [0.987, 1.000]

All six null intervals include 0.05. At $n = 1000$, FUSE's paired power gains over rank aggregation are 0.340 and 0.263 for frequencies 1 and 3. At $n = 5000$, both procedures detect every low-frequency sample, and rank aggregation misses four middle-frequency samples. The high-frequency gain from FUSE remains 0.153, with a paired 95% normal interval $[0.088, 0.218]$ based on differences of rejection indicators. At $n = 20000$, both methods detect all 300 high-frequency samples. The nonsaturated comparison below examines how computation changes the sample size each method can process and the resulting detection probability.

C.2 Exact finite-feature computation

The finite spectral representation permits an exact FUSE implementation without storing dense kernel matrices. For scale ℓ, let U_ℓ be the matrix whose ith row is $\psi_{a_\ell,J}(z_i)^\top$, including the constant coordinate, and let $s \in \{-1, 1\}^{2n}$ specify a balanced split. Write $\|\cdot\|_{\mathrm{F}}$ for the Frobenius norm. Expanding the two within-sample sums and the cross-sample sum gives

$$\widehat{\mathrm{MMD}}^2_{\ell,u} = \frac{(2n-1)\left\|U_\ell^\top s\right\|^2 + \left\|U_\ell^\top \mathbf{1}\right\|^2}{2n^2(n-1)} - \frac{\|U_\ell\|_{\mathrm{F}}^2}{n(n-1)}.$$

Since the kernel matrix is $U_\ell U_\ell^\top$, the normalizer from Section B.5 equals

$$\widehat{N}_\ell = \frac{\left\|U_\ell^\top U_\ell\right\|_{\mathrm{F}}^2 - \sum_{i=1}^{2n} \|\psi_{a_\ell,J}(z_i)\|^4}{n(n-1)}.$$

The first identity follows by writing the two feature sums as $(U_\ell^\top \mathbf{1} \pm U_\ell^\top s)/2$. The second uses equality of the squared Frobenius norms of the two Gram matrices and subtracts their observation-level diagonal squares. The constant coordinate cancels from the unbiased MMD but remains in this normalizer.

All scales share the unweighted feature cross-products. With a fixed diffusion grid and B permutations, this gives $O(nJ^2 + BnJ)$ work and $O(nJ + Bn + J^2)$ storage when caching all features. In the cost probe at $n = 2000$, the feature implementation took 0.17 seconds versus 5.46 seconds for six cached dense kernels. FUSE can also accumulate feature sums and cross-products in blocks. Our cost comparison uses 512-row blocks, storing balanced permutation labels as bytes; this reduces working storage to $O(Bn + BJ + J^2)$ and preserves every permutation statistic. Both methods therefore receive a bounded-block implementation.

For rank aggregation, Algorithm 1 can discard each feature block after updating the past sum, $\boldsymbol{S}_n$ and $\widehat{\boldsymbol{\Sigma}}_n$. We implement blocks of 256 pairs, retaining the same basis evaluator and pair order. Block boundaries carry all three accumulators, so this changes storage without changing the statistic. The supplied raw observations are included in the process-memory measurements below.

C.3 Complete cost and its statistical tradeoff

Timers start from raw observations and finish at the final p-value, including basis evaluation, covariance calculations, permutation generation and normalization where used. Data generation and compilation warmup are excluded. Each memory measurement runs in a fresh Python process, with one BLAS thread on the same AMD Ryzen Threadripper 3990X; peak resident memory includes the interpreter, imported libraries and raw data. Table 14 gives three process repetitions on a fixed independent uniform-null sample per size, separating runtime variation from the 300-sample power experiment.

Table 14: Complete time in seconds: median [range] across three fresh processes per entry. Peak resident memory is the median in MiB. RA uses 256-pair blocks and chi-square calibration; FUSE uses 512-row feature blocks and 199 permutations.

	Time (seconds)		Peak memory (MiB)	
n	RA, blocks	MMD-FUSE-N	RA, blocks	MMD-FUSE-N
1000	0.061 [0.059, 0.065]	0.109 [0.102, 0.109]	154	161
5000	0.287 [0.260, 0.298]	0.607 [0.528, 0.613]	155	163
20000	1.076 [1.037, 1.080]	1.953 [1.797, 2.082]	155	169

At $n = 20000$, blocking reduces FUSE peak memory from 525 to 169 MiB, close to the 155 MiB used by rank aggregation. The main computational gain is elapsed time. At fixed $n = 5000$, FUSE buys a 15-point high-frequency power gain for roughly twice the computation. The following comparison examines using rank aggregation's time saving to process more observations.

C.4 Power at similar computational cost

Suppose 5000 pairs are already available and computation limits how many can be processed. An independent timing study on eight uniform-null samples measured median complete times of 0.252 seconds for rank aggregation at $n = 5000$ and 0.239 seconds for FUSE at $n = 2500$. This fixed the smaller sample size before evaluating its power. Both methods use the bounded-block implementations above. FUSE retains 199 permutations and temperature $n(n-1)$; the kernel grid and rank cutoff are unchanged.

We reuse the 300 null and 300 frequency-7 samples at $n = 5000$ from Table 13, evaluating FUSE on the first 2500 observations of each group. Its null rejection rate is 0.060, with a 95% Wilson interval $[0.038, 0.093]$, compared with 0.040 $[0.023, 0.069]$ for rank aggregation. At signal strength 0.3 and frequency 7, rank aggregation has power 0.613 $[0.557, 0.667]$, whereas FUSE on the smaller sample has power 0.280 $[0.232, 0.333]$. The paired difference is 0.333 with a 95% normal interval $[0.268, 0.398]$, computed from the 300 differences of rejection indicators. Sequential timing on the first 30 null samples, alternating method order, gives median complete times of 0.251 seconds

for rank aggregation and 0.237 seconds for FUSE; the interquartile ranges are $[0.246, 0.260]$ and $[0.229, 0.245]$ seconds.

For these fixed procedures, processing twice as many available pairs more than compensates for rank aggregation's lower power at equal sample size. This supports its use when computation limits the number of observations processed. The comparison holds the permutation count and kernel family fixed and excludes data acquisition from the time budget.

D Reductions and special cases

Proposition D.1 (Reductions to familiar procedures)**.** *The methodology contains the following exact or asymptotic reductions.*

(i) At $t = 0$, the single-scale statistic is the row-average version of mMMD with the original base kernel k; by Proposition G.2, it has the same first-order null law as the normalization in Chatterjee and Ramdas (2026).

(ii) If $k = p_\tau$, the test at diffusion time t is exactly mMMD with kernel $p_{\tau+2t}$.

(iii) If $L = 1$, $Q_n^{\mathrm{RA}} = T_n(t_1)^2$ whenever the single estimated eigenvalue exceeds κ_n.

(iv) On the flat torus $\mathbb{T}^d$, the diffusion-kernel signal is a Fourier-weighted MMD with weights $e^{-4\pi^2 a\|m\|^2}$, $m \in \mathbb{Z}^d \setminus \{0\}$.

(v) For a fixed spectral truncation J, Algorithm 1 is an ordinary finite-dimensional self-normalized martingale test applied to the feature vectors in (4)*.*

Proof of Proposition D.1. For part (i), setting $t = 0$ in (1) gives $k_0 = k$, and (2) is the row-average mMMD increment. The first-order equivalence with the i^{-1} normalization is exactly Proposition G.2.

Part (ii) follows from the semigroup calculation in Proposition 3.1: $P_t^x P_t^y p_\tau = p_{\tau+2t}$ pointwise.

For part (iii), when $L = 1$,

$$\widehat{\boldsymbol{G}}_n = V_n(t_1)/\log n, \qquad \boldsymbol{S}_n/\sqrt{\log n} = S_n(t_1)/\sqrt{\log n}.$$

If the single eigenvalue is retained, substitution into (3) gives

$$Q_n^{\mathrm{RA}} = \frac{S_n(t_1)^2}{V_n(t_1)} = T_n(t_1)^2.$$

For part (iv), on the unit flat torus the complex Laplace eigenfunctions are $x \mapsto e^{2\pi i m^\top x}$ with eigenvalues $4\pi^2\|m\|^2$, $m \in \mathbb{Z}^d$. The zero mode cancels between two probability measures. Write $\widehat{P}(m) = \int e^{2\pi i m^\top x} P(\mathrm{d}x)$ and define $\widehat{Q}(m)$ analogously. Applying (5) in this basis gives

$$\Delta_a^2(P, Q) = \sum_{m \in \mathbb{Z}^d \setminus \{0\}} e^{-4\pi^2 a\|m\|^2} \left|\widehat{P}(m) - \widehat{Q}(m)\right|^2.$$

The equivalent real sine–cosine basis gives the same expression after combining conjugate modes.

Finally, the truncated feature map in (4) satisfies

$$p_{a,J}(x, y) = \langle \psi_{a,J}(x), \psi_{a,J}(y) \rangle_{\mathbb{R}^{J+1}}.$$

Thus every contrast and every predictable witness in Algorithm 1 is exactly an ordinary Euclidean inner product of the truncated feature differences. This proves part (v). □

E Preliminary bounds and martingale tools

All constants in the appendices are finite and may change from line to line. Since M is compact and every kernel used below is continuous, for each fixed finite kernel family there is $K < \infty$ such that

$$\sup_{\ell \le L} \sup_{x \in M} k_{t_\ell}(x, x) \le K.$$

Hence

$$\|\Phi_{t_\ell}(x)\| \le \sqrt{K}, \qquad \|\zeta_i(t_\ell)\| \le 2\sqrt{K}, \qquad |D_i(t_\ell)| \le 4K. \tag{17}$$

Lemma E.1 (Fourth moment of a bounded Hilbert sum)**.** *Let $\xi_1, \dots, \xi_m$ be independent, mean-zero random elements in a real Hilbert space and suppose $\|\xi_j\| \le B$ almost surely. Then*

$$\mathbb{E}\left\|\sum_{j=1}^m \xi_j\right\|^4 \le 6B^4 m^2.$$

Proof. Set $R_m = \sum_{j=1}^m \xi_j$ and write

$$\|R_m\|^2 = A_m + 2H_m, \quad A_m = \sum_{j=1}^m \|\xi_j\|^2, \quad H_m = \sum_{1\le r<s\le m} \langle \xi_r, \xi_s\rangle .$$

The elementary inequality $(a+2b)^2 \le 2a^2 + 8b^2$ gives

$$\mathbb{E}\,\|R_m\|^4 \le 2\mathbb{E}A_m^2 + 8\mathbb{E}H_m^2.$$

Because $A_m \le mB^2$, the first term is at most $2m^2B^4$. Expanding H_m^2, any product $\langle \xi_r, \xi_s\rangle \langle \xi_u, \xi_v\rangle$ with distinct unordered pairs $\{r,s\} \ne \{u,v\}$ contains an index appearing in only one inner product. Conditioning on the remaining variables and using $\mathbb{E}\xi_j = 0$ makes its expectation zero. Therefore

$$\mathbb{E}H_m^2 = \sum_{r<s} \mathbb{E}\,\langle \xi_r, \xi_s\rangle^2 \le \sum_{r<s} \mathbb{E}\{\|\xi_r\|^2 \|\xi_s\|^2\} \le \frac{m(m-1)}{2} B^4.$$

Combining the two bounds yields

$$\mathbb{E}\,\|R_m\|^4 \le 2m^2B^4 + 4m(m-1)B^4 \le 6m^2B^4.$$

□

Lemma E.2 (A harmonic-weight bound)**.** *For*

$$r_{n,j} = \sum_{m=j}^{n-1} \frac{1}{m}, \qquad 1 \le j \le n-1,$$

there is a universal constant C such that

$$\sum_{j=1}^{n-1} r_{n,j}^2 \le Cn.$$

Proof. For $j \ge 2$,

$$r_{n,j} \le \int_{j-1}^{n-1} \frac{\mathrm{d}x}{x} = \log \frac{n-1}{j-1},$$

and $r_{n,1} \le 1 + \log n$. Thus, after increasing a constant,

$$r_{n,j} \le 2 + \log(n/j), \qquad 1 \le j \le n-1.$$

The function $x \mapsto (2 + \log(1/x))^2$ is integrable on $(0,1]$, so a Riemann-sum comparison gives

$$\frac{1}{n}\sum_{j=1}^{n-1} (2 + \log(n/j))^2 \le C$$

for a universal C. Multiplying by n proves the result. □

Lemma E.3 (Logarithmic quadratic variation)**.** *For each n, let $\varepsilon_{n,1}, \dots, \varepsilon_{n,n}$ be i.i.d. mean-zero Hilbert-valued random elements satisfying $\|\varepsilon_{n,i}\| \le B$ almost surely. Let*

$$C_n = \mathbb{E}(\varepsilon_{n,1} \otimes \varepsilon_{n,1}), \quad \alpha_n = \mathrm{tr}(C_n^2), \quad U_{n,i} = \langle \bar{\varepsilon}_{n,i-1}, \varepsilon_{n,i}\rangle ,$$

where $\bar{\varepsilon}_{n,i-1} = (i-1)^{-1}\sum_{j<i}\varepsilon_{n,j}$ *and* $\mathcal{F}_{n,i} = \sigma(\varepsilon_{n,1},\dots,\varepsilon_{n,i})$. *If* $\alpha_n \to \alpha_0 \in (0,\infty)$, *then*

$$\frac{1}{\alpha_n \log n}\sum_{i=2}^{n}\mathbb{E}(U_{n,i}^2 \mid \mathcal{F}_{n,i-1}) \to 1, \tag{18}$$

$$\frac{1}{\alpha_n \log n}\sum_{i=2}^{n}U_{n,i}^2 \to 1, \tag{19}$$

$$\frac{\sum_{i=2}^{n}U_{n,i}}{\sqrt{\alpha_n \log n}} \Rightarrow N(0,1), \tag{20}$$

where the first two convergences are in probability.

Proof. Write $R_{n,m} = \sum_{j=1}^{m}\varepsilon_{n,j}$ and $B_{n,m} = \langle C_n R_{n,m}, R_{n,m}\rangle$. Conditional independence gives

$$q_{n,m+1} := \mathbb{E}(U_{n,m+1}^2 \mid \mathcal{F}_{n,m}) = \frac{B_{n,m}}{m^2}.$$

Because $\mathbb{E}(R_{n,m}\otimes R_{n,m}) = mC_n$,

$$\mathbb{E}B_{n,m} = m\,\mathrm{tr}(C_n^2) = m\alpha_n.$$

Consequently,

$$\mathbb{E}\sum_{i=2}^{n}q_{n,i} = \alpha_n\sum_{m=1}^{n-1}\frac{1}{m} = \alpha_n\log n + O(1). \tag{21}$$

By Lemma E.1, $\mathbb{E}\left\|R_{n,m}\right\|^4 \le 6B^4m^2$. Hence

$$\mathrm{Var}(B_{n,m}) \le \mathbb{E}B_{n,m}^2 \le \|C_n\|_{\mathrm{op}}^2\,\mathbb{E}\left\|R_{n,m}\right\|^4 \le Cm^2,$$

where the constant is uniform in n because $\|C_n\|_{\mathrm{op}} \le \mathbb{E}\left\|\varepsilon_{n,1}\right\|^2 \le B^2$. For $m \le \ell$, the difference $R_{n,\ell} - R_{n,m}$ is independent of $\mathcal{F}_{n,m}$ and has mean zero. Expanding the quadratic form gives

$$\mathbb{E}(B_{n,\ell} \mid \mathcal{F}_{n,m}) = B_{n,m} + (\ell - m)\alpha_n.$$

It follows that

$$\mathrm{Cov}(B_{n,m}, B_{n,\ell}) = \mathrm{Var}(B_{n,m}), \qquad |\mathrm{Cov}(B_{n,m}, B_{n,\ell})| \le Cm^2.$$

Therefore

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=2}^{n}q_{n,i}\right) &= \mathrm{Var}\left(\sum_{m=1}^{n-1}\frac{B_{n,m}}{m^2}\right)\\ &\le 2C\sum_{1\le m\le \ell\le n-1}\frac{m^2}{m^2\ell^2} \le C'\sum_{\ell=1}^{n-1}\frac{1}{\ell} = O(\log n).\end{aligned}$$

Together with (21) and $\alpha_n \to \alpha_0 > 0$, Chebyshev's inequality proves (18).

Next set

$$M_n^{(2)} = \sum_{i=2}^{n}\{U_{n,i}^2 - q_{n,i}\}.$$

This is a martingale. Since $|U_{n,i}| \le B^2$,

$$\mathbb{E}\{(U_{n,i}^2 - q_{n,i})^2 \mid \mathcal{F}_{n,i-1}\} \le \mathbb{E}(U_{n,i}^4 \mid \mathcal{F}_{n,i-1}) \le B^4 q_{n,i}.$$

Thus, by orthogonality of martingale increments and (21),

$$\mathbb{E}\{M_n^{(2)}\}^2 \le B^4\mathbb{E}\sum_{i=2}^{n}q_{n,i} = O(\log n).$$

Hence $M_n^{(2)}/\log n \to 0$ in probability. Combining this with (18) proves (19).

Finally, the predictable quadratic variation of $\sum_{i=2}^{n}U_{n,i}/\sqrt{\alpha_n\log n}$ converges to one by (18). The increments are bounded by $B^2/\sqrt{\alpha_n\log n} \to 0$, so the conditional Lindeberg condition holds. The martingale central limit theorem (Hall and Heyde, 1980) gives (20). □

Lemma E.4 (Diffusion covariance under total-variation convergence)**.** *Let $K \subset (0,\infty)$ be compact. For $a \in K$, the canonical diffusion feature map is $\Phi_a(x) = (e^{-a\lambda_j/2}\varphi_j(x))_{j\geq 0} \in \ell^2$. Set $\zeta_a = \Phi_a(X) - \Phi_a(Y)$ and $R_n = P_n \otimes Q_n$, and let $C_{n,a}$ be the covariance operator of ζ_a under R_n. If*

$$\|P_n - P_0\|_{\mathrm{TV}} + \|Q_n - P_0\|_{\mathrm{TV}} \to 0,$$

then, with $C_{0,a}$ defined under $P_0 \otimes P_0$,

$$\sup_{a\in K} \|C_{n,a} - C_{0,a}\|_{\mathrm{HS}} \to 0.$$

Proof. Since K is bounded away from zero and the diffusion kernel is smooth,

$$K_0 := \sup_{a\in K}\sup_{x\in M} p_a(x,x) < \infty.$$

Hence, uniformly over $a \in K$,

$$\|\Phi_a(x)\|_{\ell^2} \leq \sqrt{K_0}, \qquad \|\Phi_a(x) - \Phi_a(y)\|_{\ell^2} \leq 2\sqrt{K_0}.$$

The product measures satisfy

$$\|R_n - P_0 \otimes P_0\|_{\mathrm{TV}} \leq \|P_n - P_0\|_{\mathrm{TV}} + \|Q_n - P_0\|_{\mathrm{TV}} \to 0.$$

Therefore the defining dual bound for total variation, applied to bounded Bochner-valued functions, gives uniformly in $a \in K$,

$$\|\mathbb{E}_{R_n}\zeta_a - \mathbb{E}_{P_0\otimes P_0}\zeta_a\|_{\ell^2} \to 0,$$
$$\|\mathbb{E}_{R_n}(\zeta_a \otimes \zeta_a) - \mathbb{E}_{P_0\otimes P_0}(\zeta_a \otimes \zeta_a)\|_{\mathrm{HS}} \to 0.$$

Here $\|u \otimes u\|_{\mathrm{HS}} = \|u\|^2$. Write $m_{n,a} = \mathbb{E}_{R_n}\zeta_a$. Since

$$C_{n,a} = \mathbb{E}_{R_n}(\zeta_a \otimes \zeta_a) - m_{n,a} \otimes m_{n,a}, \qquad C_{0,a} = \mathbb{E}_{P_0\otimes P_0}(\zeta_a \otimes \zeta_a),$$

and $\|m_{n,a} \otimes m_{n,a}\|_{\mathrm{HS}} = \|m_{n,a}\|^2$, the two uniform convergences prove the claim. In particular, uniform boundedness of the Hilbert–Schmidt norms and the inequality

$$\left|\mathrm{tr}(C_{n,a}^2) - \mathrm{tr}(C_{0,a}^2)\right| \leq \|C_{n,a} - C_{0,a}\|_{\mathrm{HS}} \left\{\|C_{n,a}\|_{\mathrm{HS}} + \|C_{0,a}\|_{\mathrm{HS}}\right\}$$

also give locally uniform convergence of the corresponding null scales. □

F Proofs for diffusion pullback and spectral representations

Theorem F.1 (Diffusion-pullback identity and preservation of characteristicness)**.** *Let k be continuous, symmetric, and positive definite on M. For every $t > 0$:*

(i) k_t in (1) *is continuous, bounded, symmetric, and positive definite;*

(ii) for every finite signed Borel measure η,

$$\iint k_t(x,y)\,\eta(\mathrm{d}x)\eta(\mathrm{d}y) = \iint k(u,v)\,(P_t^*\eta)(\mathrm{d}u)(P_t^*\eta)(\mathrm{d}v); \tag{22}$$

(iii) P_t^* is injective on finite signed Borel measures;

(iv) if k is characteristic, then k_t is characteristic.

In particular,

$$\mathrm{MMD}_{k_t}(P,Q) = \mathrm{MMD}_k(P_t^*P, P_t^*Q).$$

Proof of Theorem F.1. We prove the four claims in order.

Step 1: construct a feature map for k_t. Let $\Phi(x) = k(x,\cdot) \in \mathcal{H}_k$. Since k is continuous on compact $M \times M$,

$$K_0 := \sup_{x\in M} k(x,x) < \infty, \qquad \|\Phi(x)\|_{\mathcal{H}_k} \leq \sqrt{K_0}.$$

For $t > 0$, define the Bochner integral

$$\Psi_t(x) = \int_M p_t(x,u)\Phi(u)\,\mathfrak{m}(\,\mathrm{d}u).$$

It exists because

$$\int_M p_t(x,u)\,\|\Phi(u)\|\,\mathfrak{m}(\,\mathrm{d}u) \le \sqrt{K_0}\int_M p_t(x,u)\,\mathfrak{m}(\,\mathrm{d}u) = \sqrt{K_0}.$$

Fubini's theorem and the reproducing property give

$$\begin{aligned}\langle \Psi_t(x), \Psi_t(y)\rangle_{\mathcal{H}_k} &= \iint p_t(x,u)p_t(y,v)\,\langle \Phi(u),\Phi(v)\rangle_{\mathcal{H}_k}\,\mathfrak{m}(\,\mathrm{d}u)\mathfrak{m}(\,\mathrm{d}v)\\ &= \iint p_t(x,u)p_t(y,v)k(u,v)\,\mathfrak{m}(\,\mathrm{d}u)\mathfrak{m}(\,\mathrm{d}v)\\ &= k_t(x,y).\end{aligned}$$

Hence, for arbitrary points $x_1,\dots,x_m$ and scalars $a_1,\dots,a_m$,

$$\sum_{r,s=1}^m a_r a_s k_t(x_r,x_s) = \left\|\sum_{r=1}^m a_r \Psi_t(x_r)\right\|_{\mathcal{H}_k}^2 \ge 0.$$

Thus k_t is positive definite. Symmetry is immediate. The bound

$$|k_t(x,y)| \le \iint p_t(x,u)p_t(y,v)\,|k(u,v)|\,\mathfrak{m}(\,\mathrm{d}u)\mathfrak{m}(\,\mathrm{d}v) \le \|k\|_\infty$$

proves boundedness. The diffusion kernel is smooth for $t > 0$, and dominated convergence on the compact domain proves continuity of k_t.

Step 2: prove the energy identity. Let η be a finite signed measure. Absolute integrability follows from boundedness of k and finiteness of $|\eta|\,(M)$. Therefore Fubini's theorem yields

$$\begin{aligned}&\iint k_t(x,y)\eta(\,\mathrm{d}x)\eta(\,\mathrm{d}y)\\ &\quad= \iint k(u,v)\left\{\int p_t(x,u)\eta(\,\mathrm{d}x)\right\}\left\{\int p_t(y,v)\eta(\,\mathrm{d}y)\right\}\mathfrak{m}(\,\mathrm{d}u)\mathfrak{m}(\,\mathrm{d}v).\end{aligned}$$

Because the diffusion kernel is symmetric, the function

$$u \longmapsto \int p_t(x,u)\eta(\,\mathrm{d}x)$$

is the density of $P_t^*\eta$ with respect to $\mathfrak{m}$. The last display is exactly the right-hand side of (22). Taking $\eta = P - Q$ proves the MMD identity.

Step 3: prove injectivity of P_t^*. Suppose $P_t^*\eta = 0$. For every eigenfunction φ_j,

$$0 = \int_M \varphi_j\;\mathrm{d}(P_t^*\eta) = \int_M P_t\varphi_j\;\mathrm{d}\eta = e^{-t\lambda_j}\int_M \varphi_j\;\mathrm{d}\eta.$$

Since $e^{-t\lambda_j} > 0$,

$$\int_M \varphi_j\;\mathrm{d}\eta = 0 \qquad \text{for every } j \ge 0. \tag{23}$$

Fix $f \in C(M)$ and $s > 0$. The smoothed function $P_s f$ has a uniformly convergent eigenfunction expansion on M; hence (23) implies

$$\int_M P_s f\;\mathrm{d}\eta = 0.$$

The diffusion semigroup is strongly continuous on $C(M)$, so $\|P_s f - f\|_\infty \to 0$ as $s \downarrow 0$. Therefore

$$\left|\int_M f\;\mathrm{d}\eta\right| = \left|\int_M (f - P_s f)\;\mathrm{d}\eta\right| \le \|f - P_s f\|_\infty\,|\eta|\,(M) \to 0.$$

Thus η annihilates every continuous function. The Riesz representation theorem gives $\eta = 0$.

Step 4: transfer characteristicness. Suppose k is characteristic and $\mathrm{MMD}_{k_t}(P, Q) = 0$. By the energy identity,

$$\mathrm{MMD}_k(P_t^* P, P_t^* Q) = 0.$$

Characteristicness of k gives $P_t^* P = P_t^* Q$. Injectivity applied to $P - Q$ then gives $P = Q$. Hence k_t is characteristic. □

For the truncated kernel in (4),

$$\Delta_{a,J}^2(P, Q) = \sum_{j=1}^{J} e^{-a\lambda_j} (P\varphi_j - Q\varphi_j)^2. \tag{24}$$

Moreover,

$$0 \le \Delta_a^2(P, Q) - \Delta_{a,J}^2(P, Q) \le 4 \sup_{x,y \in M} |p_a(x, y) - p_{a,J}(x, y)|,$$

which converges to zero as $J \to \infty$ for each fixed $a > 0$.

Proof of Proposition 3.1. The semigroup identity and symmetry of the diffusion kernel imply

$$\begin{aligned} k_t(x, y) &= \iint p_t(x, u) p_t(y, v) p_\tau(u, v)\, \mathfrak{m}(\mathrm{d}u) \mathfrak{m}(\mathrm{d}v) \\ &= \int p_t(x, u) p_{\tau+t}(u, y)\, \mathfrak{m}(\mathrm{d}u) \\ &= p_{\tau+2t}(x, y) = p_a(x, y). \end{aligned}$$

For $a > 0$, the spectral expansion of the diffusion kernel converges smoothly:

$$p_a(x, y) = \sum_{j \ge 0} e^{-a\lambda_j} \varphi_j(x) \varphi_j(y).$$

Let $\eta = P - Q$. Since $\eta(M) = 0$, the $j = 0$ term vanishes and

$$\begin{aligned} \Delta_a^2(P, Q) &= \iint p_a(x, y) \eta(\mathrm{d}x) \eta(\mathrm{d}y) \\ &= \sum_{j \ge 1} e^{-a\lambda_j} \left(\int \varphi_j \, \mathrm{d}\eta \right)^2, \end{aligned}$$

which is (5).

If this energy is zero, every coefficient $P\varphi_j - Q\varphi_j$ is zero. The coefficient for $j = 0$ is also zero because both measures have mass one. The argument in Step 3 of the preceding proof then shows $P = Q$. Thus p_a is characteristic.

For the truncated kernel, the feature map in (4) gives positive definiteness, and direct expansion gives (24). The difference kernel

$$r_{a,J}(x, y) = p_a(x, y) - p_{a,J}(x, y)$$

is positive semidefinite, so its MMD energy is nonnegative. Moreover,

$$\begin{aligned} \Delta_a^2(P, Q) - \Delta_{a,J}^2(P, Q) &= \iint r_{a,J}(x, y) \eta(\mathrm{d}x) \eta(\mathrm{d}y) \\ &\le \|r_{a,J}\|_\infty |\eta|(M)^2 \le 4 \|r_{a,J}\|_\infty. \end{aligned}$$

Smooth convergence of the eigenfunction expansion for fixed $a > 0$ implies $\|r_{a,J}\|_\infty \to 0$. □

G PROOFS FOR NULL SCALE, NORMALIZATION, AND FIXED ALTERNATIVES

G.1 NULL SCALE AND ROW NORMALIZATION

Fix a scale t and suppose $H_0 : P = Q$. Let

$$m_t = \mathbb{E}_P \Phi_t(X), \qquad \Sigma_t = \mathbb{E}_P[(\Phi_t(X) - m_t) \otimes (\Phi_t(X) - m_t)],$$

and let

$$C_t = \mathbb{E}[\zeta_1(t) \otimes \zeta_1(t)], \qquad \alpha_t = \operatorname{tr}(C_t^2).$$

Here $(u \otimes v)h = u \langle v, h \rangle$. Let Z, Z' be independent copies of (X, Y) under H_0.

Proposition G.1 (Equivalent forms of the null scale)**.** *Under H_0,*

$$C_t = 2\Sigma_t, \qquad \alpha_t = 4 \operatorname{tr}(\Sigma_t^2) = \mathbb{E}\{h_t(Z, Z')^2\}.$$

Consequently, $\alpha_t = 0$ if and only if $\Phi_t(X)$ is almost surely constant. If k_t is characteristic, then

$$\alpha_t > 0 \quad \Longleftrightarrow \quad P \text{ is not a point mass}. \tag{25}$$

The bounded U-statistic

$$\widehat{\alpha}_{t,n} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} h_t(Z_i, Z_j)^2 \tag{26}$$

converges almost surely to α_t.

The nondegeneracy condition $\alpha_t > 0$ is therefore exactly characterized in the characteristic-kernel case and can be estimated directly. When k_t is characteristic, it excludes only the case in which every observation equals one fixed point under the null. Near-zero values of (26) should nevertheless be reported because the asymptotic approximation may be poor when the effective feature variation is very small.

Proof of Proposition G.1. Let X, Y be independent with law P. Since

$$\zeta = \Phi_t(X) - \Phi_t(Y) = \{\Phi_t(X) - m_t\} - \{\Phi_t(Y) - m_t\},$$

independence and centering give

$$C_t = \mathbb{E}(\zeta \otimes \zeta) = \Sigma_t + \Sigma_t = 2\Sigma_t.$$

Therefore

$$\alpha_t = \operatorname{tr}(C_t^2) = 4 \operatorname{tr}(\Sigma_t^2).$$

For independent paired observations Z, Z', the contrast identity gives $h_t(Z, Z') = \langle \zeta, \zeta' \rangle$. Hence

$$\begin{aligned} \mathbb{E} h_t(Z, Z')^2 &= \mathbb{E} \langle \zeta, \zeta' \rangle^2 \\ &= \mathbb{E} \langle \zeta, C_t \zeta \rangle \\ &= \operatorname{tr}\{C_t \mathbb{E}(\zeta \otimes \zeta)\} = \operatorname{tr}(C_t^2) = \alpha_t. \end{aligned}$$

A positive trace-class operator has $\operatorname{tr}(\Sigma_t^2) = 0$ if and only if $\Sigma_t = 0$. Also

$$\operatorname{tr}(\Sigma_t) = \mathbb{E} \left\| \Phi_t(X) - m_t \right\|^2.$$

Thus $\Sigma_t = 0$ if and only if $\Phi_t(X) = m_t$ almost surely.

Suppose now that k_t is characteristic. Then the canonical feature map is injective: if $\Phi_t(x) = \Phi_t(y)$, the mean embeddings of δ_x and δ_y coincide, so characteristicness gives $x = y$. If $\Phi_t(X)$ is almost surely constant, injectivity implies that X is almost surely equal to one point. Conversely, a point mass clearly gives a constant feature and $\alpha_t = 0$. This proves (25).

Finally, the kernel

$$(z, z') \longmapsto h_t(z, z')^2$$

is bounded and symmetric. The strong law for U-statistics therefore gives $\widehat{\alpha}_{t,n} \to \mathbb{E} h_t(Z, Z')^2 = \alpha_t$ almost surely. □

Define the Chatterjee–Ramdas increment

$$\check{D}_i(t) = \frac{1}{i}\sum_{j=1}^{i-1} h_t(Z_i, Z_j) = \frac{i-1}{i} D_i(t),$$

and let $\check{T}_n(t)$ be its self-normalized sum.

Proposition G.2 (Equivalence of the two row normalizations)**.** *Under H_0 and $\alpha_t > 0$,*

$$T_n(t) - \check{T}_n(t) \to 0 \qquad \textit{in probability}.$$

The same conclusion holds under the local alternatives in Theorem 3.2.

Proof of Proposition G.2. Suppress the scale t and write

$$\check{D}_i = \frac{i-1}{i} D_i, \quad S_n = \sum_{i=2}^{n} D_i, \quad \check{S}_n = \sum_{i=2}^{n} \check{D}_i, \quad V_n = \sum_{i=2}^{n} D_i^2, \quad \check{V}_n = \sum_{i=2}^{n} \check{D}_i^2.$$

Under H_0, D_i is a martingale difference. With $C_t = \mathbb{E}(\zeta_1 \otimes \zeta_1)$ and $\alpha = \alpha_t = \operatorname{tr}(C_t^2)$,

$$\mathbb{E} D_i^2 = \frac{\alpha}{i-1}.$$

Indeed, conditional on the past,

$$\mathbb{E}(D_i^2 \mid \mathcal{F}_{i-1}) = \left\langle C_t \bar{\zeta}_{i-1}, \bar{\zeta}_{i-1} \right\rangle,$$

and $\mathbb{E}(\bar{\zeta}_{i-1} \otimes \bar{\zeta}_{i-1}) = C_t/(i-1)$.

The numerator difference is

$$S_n - \check{S}_n = \sum_{i=2}^{n} \frac{D_i}{i}.$$

Its summands are martingale differences and

$$\mathbb{E}(S_n - \check{S}_n)^2 = \sum_{i=2}^{n} \frac{\mathbb{E} D_i^2}{i^2} = \alpha \sum_{i=2}^{n} \frac{1}{i^2(i-1)} \le C.$$

Hence $S_n - \check{S}_n = O_{\mathbb{P}}(1)$. The denominator difference is nonnegative and satisfies

$$\begin{aligned} 0 \le V_n - \check{V}_n &= \sum_{i=2}^{n} \left\{ 1 - \left(\frac{i-1}{i} \right)^2 \right\} D_i^2 \\ &\le 2 \sum_{i=2}^{n} \frac{D_i^2}{i}. \end{aligned}$$

Its expectation is bounded by

$$2\alpha \sum_{i=2}^{n} \frac{1}{i(i-1)} \le 2\alpha.$$

Thus $V_n - \check{V}_n = O_{\mathbb{P}}(1)$.

Applying Lemma E.3 with $\varepsilon_i = \zeta_i$ gives $V_n/(\alpha \log n) \to 1$; the same holds for $\check{V}_n$. Also $S_n = O_{\mathbb{P}}(\sqrt{\log n})$. Therefore

$$\begin{aligned} \left| \frac{S_n}{\sqrt{V_n}} - \frac{\check{S}_n}{\sqrt{\check{V}_n}} \right| &\le \frac{|S_n - \check{S}_n|}{\sqrt{V_n}} + |\check{S}_n| \left| \frac{1}{\sqrt{V_n}} - \frac{1}{\sqrt{\check{V}_n}} \right| \\ &= o_{\mathbb{P}}(1). \end{aligned}$$

For the second term, use

$$\left| V_n^{-1/2} - \check{V}_n^{-1/2} \right| = \frac{|V_n - \check{V}_n|}{\sqrt{V_n \check{V}_n}(\sqrt{V_n} + \sqrt{\check{V}_n})} = O_{\mathbb{P}}((\log n)^{-3/2}).$$

Under the local alternatives of Theorem 3.2, write $\varepsilon_i = \varepsilon_{n,i}$, $w = w_n$ and $\bar{\varepsilon}_{i-1} = (i-1)^{-1}\sum_{j<i}\varepsilon_j$. Set $A_i = \langle w, \varepsilon_i\rangle$, $B_i = \langle \bar{\varepsilon}_{i-1}, w\rangle$ and $U_i = \langle \bar{\varepsilon}_{i-1}, \varepsilon_i\rangle$. The decomposition

$$D_i = d_n + A_i + B_i + U_i$$

in (27) gives, uniformly in i and n,

$$\mathbb{E}D_i^2 \le C\left\{d_n + \frac{1}{i-1}\right\}.$$

Consequently,

$$\mathbb{E}\sum_{i=2}^{n}\frac{D_i^2}{i} \le C\{d_n \log n + 1\} = O(1),$$

because (6) implies $d_n \log n \to 0$. Hence $V_n - \check{V}_n = O_{\mathbb{P}}(1)$.

For the numerator difference, write

$$S_n - \check{S}_n = d_n\sum_{i=2}^{n}\frac{1}{i} + \sum_{i=2}^{n}\frac{A_i}{i} + \sum_{i=2}^{n}\frac{B_i}{i} + \sum_{i=2}^{n}\frac{U_i}{i}.$$

The deterministic term is $o(1)$. Independence gives

$$\mathbb{E}\left(\sum_{i=2}^{n}\frac{A_i}{i}\right)^2 \le Cd_n\sum_{i=2}^{n} i^{-2} = o(1).$$

If $e_j = \langle \varepsilon_j, w_n\rangle$, then the telescoping identity

$$\sum_{i=2}^{n}\frac{B_i}{i} = \sum_{j=1}^{n-1}\left(\frac{1}{j} - \frac{1}{n}\right)e_j$$

and $\mathbb{E}e_j^2 \le Cd_n$ imply

$$\mathbb{E}\left(\sum_{i=2}^{n}\frac{B_i}{i}\right)^2 \le Cd_n\sum_{j=1}^{n-1}\left(\frac{1}{j} - \frac{1}{n}\right)^2 = o(1).$$

Finally, (U_i/i) is a martingale difference sequence and

$$\mathbb{E}\left(\sum_{i=2}^{n}\frac{U_i}{i}\right)^2 = \sum_{i=2}^{n}\frac{\alpha_n}{i^2(i-1)} = O(1).$$

Thus $S_n - \check{S}_n = O_{\mathbb{P}}(1)$. The proof of Theorem 3.2 gives $V_n \asymp_{\mathbb{P}} \log n$ and $S_n = O_{\mathbb{P}}(\sqrt{\log n})$; the same ratio argument completes the local case. □

Corollary G.3 (Fixed-scale null limit and consistency)**.** *Let k_t be a continuous positive definite kernel on M.*

(i) Under H_0 and $\alpha_t > 0$,

$$T_n(t) \Rightarrow N(0,1).$$

Thus $\mathbf{1}\{T_n(t) > z_{1-\alpha}\}$ has asymptotic level α.

(ii) If $\Delta_t(P,Q) > 0$, then

$$\frac{S_n(t)}{n} \to \Delta_t^2(P,Q) \qquad \textit{and} \qquad T_n(t) \to +\infty$$

in probability.

(iii) If k_t is characteristic, the test is consistent against every fixed $P \neq Q$.

Proof of Corollary G.3. Under H_0, apply Lemma E.3 with $\varepsilon_i = \zeta_i(t)$, $C = C_t$, and $\alpha = \alpha_t$. It gives

$$\frac{S_n(t)}{\sqrt{\alpha_t \log n}} \Rightarrow N(0,1), \qquad \frac{V_n(t)}{\alpha_t \log n} \to 1$$

in probability. Slutsky's theorem therefore yields $T_n(t) \Rightarrow N(0,1)$. This is the bounded fixed-kernel, row-average specialization of the mMMD null theorem of Chatterjee and Ramdas (2026); Proposition G.2 shows that it agrees to first order with their i^{-1} normalization.

For completeness, we verify the row-average consistency in part (ii), which also shows that the small normalization change does not affect fixed alternatives. Write

$$w = \mathbb{E}\zeta_1(t), \qquad d = \|w\|^2 = \Delta_t^2(P,Q) > 0.$$

The conditional mean of D_i is

$$m_i = \mathbb{E}(D_i \mid \mathcal{F}_{i-1}) = \left\langle \bar{\zeta}_{i-1}, w \right\rangle.$$

The Hilbert-space strong law gives $\bar{\zeta}_{i-1} \to w$ almost surely, hence $m_i \to d$ almost surely. Cesàro's lemma yields

$$\frac{1}{n}\sum_{i=2}^{n} m_i \to d \qquad \text{almost surely.}$$

The differences $D_i - m_i$ form a martingale difference sequence and are uniformly bounded. Therefore

$$\mathbb{E}\left\{\frac{1}{n}\sum_{i=2}^{n}(D_i - m_i)\right\}^2 = \frac{1}{n^2}\sum_{i=2}^{n}\mathbb{E}(D_i - m_i)^2 \le \frac{C}{n} \to 0.$$

It follows that $S_n/n \to d$ in probability.

By (17), $V_n \le 16K^2 n$. On the event $\{S_n \ge nd/2\}$, whose probability tends to one,

$$T_n = \frac{S_n}{\sqrt{V_n}} \ge \frac{nd/2}{4K\sqrt{n}} = \frac{d}{8K}\sqrt{n} \to \infty.$$

This proves part (ii). If k_t is characteristic, then $P \neq Q$ implies $d > 0$, proving part (iii). □

H Proof of the fixed-scale local alternative theorem

Proof of Theorem 3.2. All quantities in this proof may depend on n. We suppress that index when no confusion can arise. Let

$$\bar{\varepsilon}_{i-1} = \frac{1}{i-1}\sum_{j=1}^{i-1}\varepsilon_j, \qquad U_i = \left\langle \bar{\varepsilon}_{i-1}, \varepsilon_i \right\rangle.$$

Since $\zeta_i = w + \varepsilon_i$ and $\bar{\zeta}_{i-1} = w + \bar{\varepsilon}_{i-1}$,

$$D_i = d + A_i + B_i + U_i, \quad A_i = \left\langle w, \varepsilon_i \right\rangle, \quad B_i = \left\langle \bar{\varepsilon}_{i-1}, w \right\rangle. \tag{27}$$

We analyze the numerator and denominator separately.

Step 1: the linear terms are negligible on the $\sqrt{\log n}$ scale. Independence and centering give

$$\mathbb{E}\left(\sum_{i=2}^{n} A_i\right)^2 = (n-1)\mathbb{E}\left\langle w, \varepsilon_1 \right\rangle^2 \le Cnd.$$

For $e_j = \langle \varepsilon_j, w \rangle$,

$$\sum_{i=2}^{n} B_i = \sum_{j=1}^{n-1} r_{n,j} e_j, \qquad r_{n,j} = \sum_{m=j}^{n-1}\frac{1}{m}.$$

The e_j are independent, mean zero, and $\mathbb{E}e_j^2 \le Cd$. By Lemma E.2,

$$\mathbb{E}\left(\sum_{i=2}^{n} B_i\right)^2 \le Cd\sum_{j=1}^{n-1} r_{n,j}^2 \le Cnd.$$

Thus

$$\sum_{i=2}^{n}(A_i + B_i) = O_{\mathbb{P}}(\sqrt{nd}) = o_{\mathbb{P}}(\sqrt{\log n}) \tag{28}$$

because the second condition in (6) implies $nd/\log n \to 0$.

Step 2: the centered martingale has a Gaussian limit. The centered variables ε_i are uniformly bounded because the original feature differences are uniformly bounded. By Lemma E.3,

$$\frac{\sum_{i=2}^{n} U_i}{\sqrt{\alpha_n \log n}} \Rightarrow N(0,1), \qquad \frac{\sum_{i=2}^{n} U_i^2}{\alpha_n \log n} \to 1. \tag{29}$$

Combining (27), (28), and the first relation in (29),

$$\frac{S_n - (n-1)d_n}{\sqrt{\alpha_n \log n}} \Rightarrow N(0,1).$$

Since $nd_n/\sqrt{\log n} \to \rho$ and $\alpha_n \to \alpha_0$,

$$\frac{S_n}{\sqrt{\alpha_n \log n}} \Rightarrow N\left(\frac{\rho}{\sqrt{\alpha_0}}, 1\right). \tag{30}$$

Step 3: the observed denominator is asymptotically unchanged by the local mean. Let

$$R_i = d + A_i + B_i, \qquad D_i = U_i + R_i.$$

We first show

$$\sum_{i=2}^{n} R_i^2 = o_{\mathbb{P}}(\log n). \tag{31}$$

Using $(a+b+c)^2 \le 3(a^2+b^2+c^2)$,

$$\mathbb{E}\sum_{i=2}^{n} R_i^2 \le 3(n-1)d^2 + 3\sum_{i=2}^{n}\mathbb{E}A_i^2 + 3\sum_{i=2}^{n}\mathbb{E}B_i^2.$$

The first two terms are bounded by Cnd because d is uniformly bounded. For the last term,

$$\mathbb{E}B_i^2 = \frac{1}{(i-1)^2}\sum_{j=1}^{i-1}\mathbb{E}\langle \varepsilon_j, w\rangle^2 \le \frac{Cd}{i-1},$$

so

$$\sum_{i=2}^{n}\mathbb{E}B_i^2 \le Cd\log n.$$

Now $nd = o(\log n)$ and $d \to 0$, hence

$$\mathbb{E}\sum_{i=2}^{n} R_i^2 = o(\log n).$$

Markov's inequality proves (31).

By Cauchy–Schwarz and (29),

$$\left|\sum_{i=2}^{n} U_i R_i\right| \le \left(\sum_{i=2}^{n} U_i^2\right)^{1/2}\left(\sum_{i=2}^{n} R_i^2\right)^{1/2} = o_{\mathbb{P}}(\log n).$$

Therefore

$$\begin{aligned} V_n &= \sum_{i=2}^{n}(U_i + R_i)^2 \\ &= \sum_{i=2}^{n} U_i^2 + 2\sum_{i=2}^{n} U_i R_i + \sum_{i=2}^{n} R_i^2 = \alpha_n \log n\{1 + o_{\mathbb{P}}(1)\}. \end{aligned}$$

Combining this with (30) and Slutsky's theorem proves (7). The power expression follows from convergence of the rejection probability at the continuity point $z_{1-\alpha}$.

If $nd_n/\sqrt{\log n} \to \infty$ while $nd_n/\log n \to 0$, the same decomposition gives

$$T_n = \frac{nd_n}{\sqrt{\alpha_n \log n}}\{1 + o_{\mathbb{P}}(1)\} + O_{\mathbb{P}}(1) \to \infty.$$

□

I PROOF OF THE DIFFUSION-TIME EFFICIENCY CRITERION

For the canonical ℓ^2 diffusion feature map, let $C_{n,a}$ be the covariance operator of the centered paired feature difference at effective time a, and let $C_{0,a}$ be its counterpart when both coordinates have law P_0. Assume that, for every compact $K \subset (0, \infty)$,

$$\sup_{a \in K} \|C_{n,a} - C_{0,a}\|_{\mathrm{HS}} \to 0.$$

This condition is implied, for example, by convergence of both P_n and Q_n to P_0 in total variation; see Lemma E.4. It ensures locally uniform convergence of the null variance. To justify differentiation at every positive diffusion time, suppose additionally that $\theta_j = \int_M \varphi_j \, \mathrm{d}\nu$ for a finite signed measure ν with $\nu(M) = 0$; alternatively, one may directly assume local uniform convergence of the series in (8) and their first derivatives.

Proof of Theorem 3.3. For the diffusion kernel p_a, a canonical feature representation is

$$\Phi_a(x) = \left(e^{-a\lambda_j/2}\varphi_j(x)\right)_{j \geq 0} \in \ell^2.$$

Under P_0, the covariance operator of $\Phi_a(X)$ has matrix entries

$$(\Sigma_a)_{jk} = e^{-a(\lambda_j + \lambda_k)/2} c_{jk}.$$

For the paired feature difference, $C_{0,a} = 2\Sigma_a$ by Proposition G.1. Hence

$$\begin{aligned} \operatorname{tr}(C_{0,a}^2) &= 4 \operatorname{tr}(\Sigma_a^2) \\ &= 4 \sum_{j,k \geq 1} e^{-a(\lambda_j + \lambda_k)} c_{jk}^2 = 4N(a). \end{aligned}$$

Terms involving $j = 0$ or $k = 0$ vanish because φ_0 is constant.

The local spectral direction gives

$$\Delta_a^2(P_n, Q_n) = c_n^2 \sum_{j \geq 1} e^{-a\lambda_j} \theta_j^2 = c_n^2 A(a).$$

Since $nc_n^2/\sqrt{\log n} \to c$, Theorem 3.2 gives noncentrality

$$\frac{cA(a)}{\sqrt{4N(a)}} = c\mathcal{E}(a).$$

Thus maximizing local asymptotic power is equivalent to maximizing $\mathcal{E}(a)$.

Under the finite-signed-measure condition on θ, $A(a) = \iint p_a(x, y) \, \nu(\mathrm{d}x)\nu(\mathrm{d}y)$, and the uniform positive-time bound on $\partial_a p_a$ justifies differentiation on every compact subinterval of $(0, \infty)$. For the covariance series, write $v_j = \operatorname{Var}_{P_0}\{\varphi_j(X)\}$. Cauchy–Schwarz gives $c_{jk}^2 \leq v_j v_k$, so the absolute derivative series is bounded by

$$2 \left(\sum_{j \geq 1} \lambda_j e^{-a\lambda_j} v_j \right) \left(\sum_{k \geq 1} e^{-a\lambda_k} v_k \right),$$

which is locally uniformly finite by the differentiated and undifferentiated positive-time spectral trace bounds; see, for example, (Grigor'yan, 2009; Rosenberg, 1997). Under the alternative formulation in

the theorem, these convergence properties are assumed directly. Termwise differentiation is therefore valid in either case. We obtain

$$A'(a) = -\sum_{j\geq 1} \lambda_j e^{-a\lambda_j}\theta_j^2 = -A(a)\bar{\lambda}_{\mathrm{sig}}(a)$$

and

$$\begin{aligned} N'(a) &= -\sum_{j,k\geq 1} (\lambda_j+\lambda_k) e^{-a(\lambda_j+\lambda_k)} c_{jk}^2 \\ &= -2N(a)\bar{\lambda}_{\mathrm{noise}}(a). \end{aligned}$$

Consequently,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}a}\log \mathcal{E}(a) &= \frac{A'(a)}{A(a)} - \frac{1}{2}\frac{N'(a)}{N(a)} \\ &= -\bar{\lambda}_{\mathrm{sig}}(a) + \bar{\lambda}_{\mathrm{noise}}(a), \end{aligned}$$

which is (9). Since $a = \tau + 2t$, the chain rule gives the derivative with respect to t. Integrating the logarithmic derivative from a_1 to a_2 gives $\log\{\mathcal{E}(a_2)/\mathcal{E}(a_1)\}$, proving the finite-interval assertion. □

J Proofs for rank-adaptive multiscale inference

We first record two finite-dimensional facts used in both the null and local arguments.

Lemma J.1 (Matrix logarithmic quadratic variation)**.** *Fix $L < \infty$. For each n, let*

$$\{(\varepsilon_{n,i}^{(1)}, \dots, \varepsilon_{n,i}^{(L)}) : 1 \leq i \leq n\}$$

be i.i.d. centered random elements in the product of real Hilbert spaces $\mathcal{H}_1 \times \cdots \times \mathcal{H}_L$, and suppose $\max_{\ell\leq L} \left\|\varepsilon_{n,i}^{(\ell)}\right\| \leq B$ almost surely, uniformly in n. Put

$$U_{n,i}^{(\ell)} = \left\langle \bar{\varepsilon}_{n,i-1}^{(\ell)}, \varepsilon_{n,i}^{(\ell)} \right\rangle_{\mathcal{H}_\ell}, \qquad \boldsymbol{U}_{n,i} = (U_{n,i}^{(1)}, \dots, U_{n,i}^{(L)})^\top,$$

where $\bar{\varepsilon}_{n,i-1}^{(\ell)} = (i-1)^{-1}\sum_{j<i}\varepsilon_{n,j}^{(\ell)}$. Use the natural filtration $\mathcal{F}_{n,i} = \sigma(\varepsilon_{n,j}^{(\ell)} : j \leq i, \ell \leq L)$. Let

$$\Gamma_{n,ab} = \mathbb{E}\left[\left\langle \varepsilon_{n,1}^{(a)}, \varepsilon_{n,2}^{(a)} \right\rangle \left\langle \varepsilon_{n,1}^{(b)}, \varepsilon_{n,2}^{(b)} \right\rangle\right], \tag{32}$$

where indices 1 and 2 denote independent copies. Write $\boldsymbol{\Gamma}_n = (\Gamma_{n,ab})_{a,b=1}^L$. Then

$$\left\| \frac{1}{\log n}\sum_{i=2}^{n} \mathbb{E}(\boldsymbol{U}_{n,i}\boldsymbol{U}_{n,i}^\top \mid \mathcal{F}_{n,i-1}) - \boldsymbol{\Gamma}_n \right\|_{\mathrm{op}} = O_{\mathbb{P}}\{(\log n)^{-1/2}\}, \tag{33}$$

$$\left\| \frac{1}{\log n}\sum_{i=2}^{n} \boldsymbol{U}_{n,i}\boldsymbol{U}_{n,i}^\top - \boldsymbol{\Gamma}_n \right\|_{\mathrm{op}} = O_{\mathbb{P}}\{(\log n)^{-1/2}\}. \tag{34}$$

The constants implicit in these bounds are uniform over the triangular array.

Proof. Because L is fixed, it is enough to establish both assertions entry by entry. Fix $a, b \in \{1, \dots, L\}$. Define the cross-covariance operator

$$C_{n,ab} = \mathbb{E}\{\varepsilon_{n,1}^{(a)} \otimes \varepsilon_{n,1}^{(b)}\} : \mathcal{H}_b \longrightarrow \mathcal{H}_a$$

and the partial sums

$$R_{n,m}^{(\ell)} = \sum_{j=1}^{m} \varepsilon_{n,j}^{(\ell)}.$$

The operator is uniformly bounded because

$$\|C_{n,ab}\|_{\mathrm{op}} \leq \mathbb{E}\{\left\|\varepsilon_{n,1}^{(a)}\right\| \left\|\varepsilon_{n,1}^{(b)}\right\|\} \leq B^2.$$

Conditional independence of the $(m+1)$st observation gives

$$q_{n,m+1}^{ab} := \mathbb{E}\{U_{n,m+1}^{(a)} U_{n,m+1}^{(b)} \mid \mathcal{F}_{n,m}\} = \frac{1}{m^2} \left\langle R_{n,m}^{(a)}, C_{n,ab} R_{n,m}^{(b)} \right\rangle_{\mathcal{H}_a}. \tag{35}$$

Write

$$B_{n,m}^{ab} = \left\langle R_{n,m}^{(a)}, C_{n,ab} R_{n,m}^{(b)} \right\rangle_{\mathcal{H}_a}.$$

An independent-copy expansion shows

$$\mathbb{E} B_{n,m}^{ab} = m\,\mathbb{E} \left\langle \varepsilon_{n,1}^{(a)}, C_{n,ab} \varepsilon_{n,1}^{(b)} \right\rangle_{\mathcal{H}_a} = m\,\mathbb{E}\Big[\left\langle \varepsilon_{n,1}^{(a)}, \varepsilon_{n,2}^{(a)} \right\rangle \left\langle \varepsilon_{n,1}^{(b)}, \varepsilon_{n,2}^{(b)} \right\rangle \Big] = m \Gamma_{n,ab}.$$

Consequently,

$$\mathbb{E} \sum_{i=2}^{n} q_{n,i}^{ab} = \Gamma_{n,ab} \sum_{m=1}^{n-1} \frac{1}{m} = \Gamma_{n,ab} \log n + O(1). \tag{36}$$

We next control the fluctuation around this mean. By Cauchy–Schwarz and Lemma E.1,

$$\mathbb{E} \left| B_{n,m}^{ab} \right|^2 \le \|C_{n,ab}\|_{\mathrm{op}}^2 \left(\mathbb{E} \left\| R_{n,m}^{(a)} \right\|^4 \right)^{1/2} \left(\mathbb{E} \left\| R_{n,m}^{(b)} \right\|^4 \right)^{1/2} \le C m^2.$$

For $m \le s$, the vector of differences $(R_{n,s}^{(\ell)} - R_{n,m}^{(\ell)})_{\ell=1}^{L}$ is independent of $\mathcal{F}_{n,m}$ and centered. Expanding the bilinear form and taking the conditional expectation gives

$$\mathbb{E}(B_{n,s}^{ab} \mid \mathcal{F}_{n,m}) = B_{n,m}^{ab} + (s-m)\Gamma_{n,ab}.$$

It follows that

$$\mathrm{Cov}(B_{n,m}^{ab}, B_{n,s}^{ab}) = \mathrm{Var}(B_{n,m}^{ab}), \qquad \left| \mathrm{Cov}(B_{n,m}^{ab}, B_{n,s}^{ab}) \right| \le C m^2.$$

Using (35),

$$\mathrm{Var}\left(\sum_{i=2}^{n} q_{n,i}^{ab} \right) \le 2C \sum_{1 \le m \le s \le n-1} \frac{m^2}{m^2 s^2} \le C' \sum_{s=1}^{n-1} \frac{1}{s} = O(\log n).$$

Together with (36), this proves that

$$\sum_{i=2}^{n} q_{n,i}^{ab} = \Gamma_{n,ab} \log n + O_{\mathbb{P}}(\sqrt{\log n}). \tag{37}$$

It remains to replace the predictable products by observed products. Put

$$M_n^{ab} = \sum_{i=2}^{n} \{U_{n,i}^{(a)} U_{n,i}^{(b)} - q_{n,i}^{ab}\}.$$

This is a real-valued martingale. Since $\left| U_{n,i}^{(\ell)} \right| \le B^2$,

$$\mathbb{E}\Big[\{U_{n,i}^{(a)} U_{n,i}^{(b)} - q_{n,i}^{ab}\}^2 \mid \mathcal{F}_{n,i-1} \Big] \le \mathbb{E}\{(U_{n,i}^{(a)} U_{n,i}^{(b)})^2 \mid \mathcal{F}_{n,i-1}\} \le B^4 q_{n,i}^{aa}.$$

The expected sum of the last quantities is $O(\log n)$ by (36) with $a = b$. Orthogonality of martingale increments therefore gives

$$\mathbb{E} \left| M_n^{ab} \right|^2 = O(\log n), \qquad M_n^{ab} = O_{\mathbb{P}}(\sqrt{\log n}).$$

Combining this with (37) proves the two entrywise claims. Since L is fixed, the entrywise maximum and operator norm differ only by a constant factor, which proves (33) and (34). □

Lemma J.2 (Stability of a thresholded Moore–Penrose inverse)**.** *Let A be an $L \times L$ positive semidefinite matrix of rank $r \geq 1$, and let A_n be random positive semidefinite matrices. Suppose $\kappa_n \downarrow 0$ and*

$$\|A_n - A\|_{\mathrm{op}} = o_{\mathbb{P}}(\kappa_n).$$

Then

$$r_{\kappa_n}(A_n) \to r, \qquad \left\|(A_n)^+_{\kappa_n} - A^+\right\|_{\mathrm{op}} \to 0$$

in probability.

Proof. Let

$$\lambda_1(A) \geq \cdots \geq \lambda_r(A) > 0 = \lambda_{r+1}(A) = \cdots = \lambda_L(A)$$

and put $\lambda_\star = \lambda_r(A)$. Write $e_n = \|A_n - A\|_{\mathrm{op}}$. Weyl's eigenvalue inequality (Bhatia, 1997) gives

$$\max_{j \leq L} |\lambda_j(A_n) - \lambda_j(A)| \leq e_n.$$

On an event whose probability tends to one, $e_n < \kappa_n/2$ and $\kappa_n < \lambda_\star/4$. On this event,

$$\lambda_j(A_n) > \lambda_\star/2 > \kappa_n, \quad j \leq r, \qquad \lambda_j(A_n) < \kappa_n, \quad j > r.$$

Thus the threshold retains exactly r eigenvalues.

Let P be the orthogonal projector onto $\mathrm{Range}(A)$ and let P_n be the projector onto the retained eigenspace of A_n. The spectral gap between the positive and zero spectra of A is $\lambda_\star$, so the Davis–Kahan projector bound, another consequence of the spectral perturbation theory in Bhatia (1997), yields

$$\|P_n - P\|_{\mathrm{op}} \leq \frac{2e_n}{\lambda_\star} \to 0$$

in probability. Define the invertible matrices

$$B = A + (I - P), \qquad B_n = P_n A_n P_n + (I - P_n).$$

Then $B_n \to B$ in operator norm. Moreover, on the event above, every eigenvalue of B_n is at least $\min(\lambda_\star/2, 1)$, so $\left\|B_n^{-1}\right\|_{\mathrm{op}}$ is uniformly bounded. The resolvent identity gives

$$B_n^{-1} - B^{-1} = B_n^{-1}(B - B_n)B^{-1},$$

and therefore $B_n^{-1} \to B^{-1}$. Finally,

$$(A_n)^+_{\kappa_n} = B_n^{-1} - (I - P_n), \qquad A^+ = B^{-1} - (I - P).$$

The convergence of the inverse and of the projectors proves the result. □

Proposition J.3 (Rank of the contrast covariance)**.** *The matrix $\mathbf{\Gamma} = (\Gamma_{ab})$ is positive semidefinite and*

$$c^\top \mathbf{\Gamma} c = \mathbb{E}\left\{\sum_{\ell=1}^{L} c_\ell h_{t_\ell}(Z, Z')\right\}^2. \tag{38}$$

Hence $\mathbf{\Gamma}$ is positive definite if and only if the contrast functions $h_{t_1}, \ldots, h_{t_L}$ are linearly independent in $L^2(\mathcal{L}(Z) \otimes \mathcal{L}(Z))$. The estimator

$$\widetilde{\Gamma}_{ab,n} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} h_{t_a}(Z_i, Z_j) h_{t_b}(Z_i, Z_j)$$

converges almost surely to Γ_{ab}.

Proof of Proposition J.3. For $c = (c_1, \ldots, c_L)^\top$, linearity and Fubini's theorem give

$$\begin{aligned} c^\top \mathbf{\Gamma} c &= \sum_{a,b=1}^{L} c_a c_b \mathbb{E}\{h_{t_a}(Z, Z') h_{t_b}(Z, Z')\} \\ &= \mathbb{E}\left\{\sum_{\ell=1}^{L} c_\ell h_{t_\ell}(Z, Z')\right\}^2. \end{aligned}$$

The right-hand side is nonnegative, proving positive semidefiniteness and (38). It vanishes if and only if the displayed linear combination is zero in $L^2\{\mathcal{L}(Z) \otimes \mathcal{L}(Z)\}$. Thus the nullspace of $\mathbf{\Gamma}$ is exactly the set of L^2 linear relations among the contrast functions, and positive definiteness is equivalent to their linear independence.

For completeness, we verify the operator identity stated before the proposition. With $C_{ab} = \mathbb{E}\{\zeta(t_a) \otimes \zeta(t_b)\}$, boundedness of the feature vectors makes C_{ab} Hilbert–Schmidt and $C_{ba} = C_{ab}^*$. If (Z, Z') are independent, then

$$\begin{aligned}\operatorname{tr}(C_{ab}C_{ba}) &= \mathbb{E}[\langle \zeta(t_a), \zeta'(t_a)\rangle \langle \zeta(t_b), \zeta'(t_b)\rangle] \\ &= \mathbb{E}\{h_{t_a}(Z, Z')h_{t_b}(Z, Z')\} = \Gamma_{ab}.\end{aligned}$$

One may justify the first equality by expanding both operators in arbitrary orthonormal bases and using Fubini; absolute convergence follows from the uniform feature bound.

Finally, the kernel

$$(z, z') \longmapsto h_{t_a}(z, z')h_{t_b}(z, z')$$

is bounded and symmetric. The strong law for order-two U-statistics (Serfling, 1980) therefore gives $\widetilde{\Gamma}_{ab,n} \to \Gamma_{ab}$ almost surely. □

Proof of Theorem 3.4. Under H_0, every feature difference is centered. Apply Lemma J.1 with $\varepsilon_i^{(\ell)} = \zeta_i(t_\ell)$. The matrix in (32) is then the fixed matrix $\mathbf{\Gamma}$ in (10). Equation (34) gives

$$\left\|\widehat{\boldsymbol{G}}_n - \mathbf{\Gamma}\right\|_{\mathrm{op}} = O_{\mathbb{P}}\{(\log n)^{-1/2}\},$$

which is the second assertion in (11).

We next prove the vector central limit theorem. Fix $c \in \mathbb{R}^L$ and define

$$X_{n,i} = c^\top \boldsymbol{D}_i.$$

Under H_0, $(X_{n,i}, \mathcal{F}_i)$ is a bounded martingale difference array. By (33),

$$\frac{1}{\log n}\sum_{i=2}^{n} \mathbb{E}(X_{n,i}^2 \mid \mathcal{F}_{i-1}) \to c^\top \mathbf{\Gamma} c$$

in probability. If $c^\top \mathbf{\Gamma} c > 0$, division by $\sqrt{c^\top \mathbf{\Gamma} c \log n}$ yields predictable quadratic variation converging to one. The increments are uniformly bounded and hence are $o(\sqrt{\log n})$, so the conditional Lindeberg condition holds. The martingale central limit theorem (Hall and Heyde, 1980) gives

$$\frac{c^\top \boldsymbol{S}_n}{\sqrt{\log n}} \Rightarrow N(0, c^\top \mathbf{\Gamma} c).$$

If $c^\top \mathbf{\Gamma} c = 0$, (38) implies $\sum_\ell c_\ell h_{t_\ell}(Z, Z') = 0$ almost surely. Hence $c^\top \boldsymbol{D}_i = 0$ almost surely for every i, and the same degenerate limit holds. The Cramér–Wold device proves

$$\boldsymbol{S}_n / \sqrt{\log n} \Rightarrow N_L(\mathbf{0}, \mathbf{\Gamma}).$$

Condition (12) and the covariance rate imply

$$\left\|\widehat{\boldsymbol{G}}_n - \mathbf{\Gamma}\right\|_{\mathrm{op}} = o_{\mathbb{P}}(\kappa_n).$$

By Lemma J.2,

$$\widehat{r}_n \to r, \qquad \left\|(\widehat{\boldsymbol{G}}_n)_{\kappa_n}^{+} - \mathbf{\Gamma}^{+}\right\|_{\mathrm{op}} \to 0$$

in probability. Slutsky's theorem and the continuous mapping theorem now give

$$Q_n^{\mathrm{RA}} \Rightarrow X^\top \mathbf{\Gamma}^{+} X, \qquad X \sim N_L(\mathbf{0}, \mathbf{\Gamma}).$$

Let $\mathbf{\Gamma} = U_r \Lambda_r U_r^\top$ be its positive-eigenvalue decomposition, where $\Lambda_r = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$. We may write $X = U_r \Lambda_r^{1/2} Z$ with $Z \sim N_r(0, I_r)$. Therefore

$$X^\top \mathbf{\Gamma}^{+} X = Z^\top Z \sim \chi_r^2.$$

Because $\mathbb{P}(\widehat{r}_n = r) \to 1$ and the chi-square distribution is continuous at its $(1-\alpha)$ quantile, replacing the fixed critical value by $\chi^2_{\widehat{r}_n, 1-\alpha}$ does not change the limiting rejection probability. This proves asymptotic level α. □

Condition (14) is needed only for rank recovery under a triangular local alternative; mere convergence $\mathbf{\Gamma}_n \to \mathbf{\Gamma}$ does not specify whether a vanishing eigenvalue lies above or below the threshold. For example, convergence of the joint paired law to its limit at total-variation rate $o(\kappa_n)$ implies $\|\mathbf{\Gamma}_n - \mathbf{\Gamma}\|_{\mathrm{op}} = o(\kappa_n)$, because all contrast products are bounded. This verifies (14). Under the exact null, no such additional assumption is needed.

Proof of Theorem 3.5. Step 1: local decomposition. For scale ℓ, write

$$w_{n\ell} = \mathbb{E}\zeta_{n,1}(t_\ell), \qquad \varepsilon^{(\ell)}_{n,i} = \zeta_{n,i}(t_\ell) - w_{n\ell}, \qquad d_{n\ell} = \|w_{n\ell}\|^2 .$$

Then

$$D^{(\ell)}_{n,i} = d_{n\ell} + A^{(\ell)}_{n,i} + B^{(\ell)}_{n,i} + U^{(\ell)}_{n,i}, \tag{39}$$

where

$$A^{(\ell)}_{n,i} = \left\langle w_{n\ell}, \varepsilon^{(\ell)}_{n,i} \right\rangle, \quad B^{(\ell)}_{n,i} = \left\langle \bar{\varepsilon}^{(\ell)}_{n,i-1}, w_{n\ell} \right\rangle, \quad U^{(\ell)}_{n,i} = \left\langle \bar{\varepsilon}^{(\ell)}_{n,i-1}, \varepsilon^{(\ell)}_{n,i} \right\rangle .$$

Let $\boldsymbol{A}_{n,i}$ and $\boldsymbol{B}_{n,i}$ collect these components in scale order, and let $d_{n,\max} = \max_{\ell \le L} d_{n\ell}$. The scalar estimates in Step 1 of the proof of Theorem 3.2, summed over the fixed number of scales, yield

$$\left\| \sum_{i=2}^{n} (\boldsymbol{A}_{n,i} + \boldsymbol{B}_{n,i}) \right\| = O_{\mathbb{P}}(\sqrt{n d_{n,\max}}) = o_{\mathbb{P}}(\sqrt{\log n}). \tag{40}$$

The deterministic contribution is $(n-1)\boldsymbol{d}_n/\sqrt{\log n} \to \boldsymbol{b}$.

Step 2: multivariate centered martingale limit. Fix $c \in \mathbb{R}^L$. The scalar process $c^\top \boldsymbol{U}_{n,i}$ is a bounded martingale difference. By the predictable part of Lemma J.1,

$$\frac{1}{\log n} \sum_{i=2}^{n} \mathbb{E}\{(c^\top \boldsymbol{U}_{n,i})^2 \mid \mathcal{F}_{n,i-1}\} - c^\top \mathbf{\Gamma}_n c \to 0$$

in probability. If $c^\top \mathbf{\Gamma} c > 0$, then $c^\top \mathbf{\Gamma}_n c \to c^\top \mathbf{\Gamma} c$ and the martingale central limit theorem gives

$$\frac{c^\top \sum_{i=2}^{n} \boldsymbol{U}_{n,i}}{\sqrt{\log n}} \Rightarrow N(0, c^\top \mathbf{\Gamma} c).$$

If $c^\top \mathbf{\Gamma} c = 0$, orthogonality of martingale increments gives

$$\mathbb{E}\left(\frac{c^\top \sum_{i=2}^{n} \boldsymbol{U}_{n,i}}{\sqrt{\log n}} \right)^2 = \frac{\sum_{m=1}^{n-1} m^{-1}}{\log n} c^\top \mathbf{\Gamma}_n c \to 0.$$

Thus the same possibly degenerate Gaussian conclusion holds for every c. The Cramér–Wold device, (39), and (40) prove

$$\frac{\boldsymbol{S}_n}{\sqrt{\log n}} \Rightarrow N_L(\boldsymbol{b}, \mathbf{\Gamma}).$$

Step 3: covariance-estimator perturbation under local alternatives. Let

$$\boldsymbol{R}_{n,i} = \boldsymbol{d}_n + \boldsymbol{A}_{n,i} + \boldsymbol{B}_{n,i}, \qquad \boldsymbol{D}_{n,i} = \boldsymbol{U}_{n,i} + \boldsymbol{R}_{n,i}.$$

The bounds used in the single-scale proof imply

$$\mathbb{E} \sum_{i=2}^{n} \|\boldsymbol{R}_{n,i}\|^2 \le C\{n d_{n,\max} + d_{n,\max} \log n\}.$$

Under (13), $n d_{n,\max} = O(\sqrt{\log n})$ and $d_{n,\max} \to 0$, so

$$\sum_{i=2}^{n} \|\boldsymbol{R}_{n,i}\|^2 = O_{\mathbb{P}}(\sqrt{\log n}).$$

By Lemma J.1, $\sum_i \|\boldsymbol{U}_{n,i}\|^2 = O_{\mathbb{P}}(\log n)$. The elementary matrix inequality

$$\left\| \sum_{i=2}^{n} (\boldsymbol{D}_{n,i}\boldsymbol{D}_{n,i}^\top - \boldsymbol{U}_{n,i}\boldsymbol{U}_{n,i}^\top) \right\|_{\mathrm{op}}$$
$$\le 2 \left(\sum_{i=2}^{n} \|\boldsymbol{U}_{n,i}\|^2 \right)^{1/2} \left(\sum_{i=2}^{n} \|\boldsymbol{R}_{n,i}\|^2 \right)^{1/2} + \sum_{i=2}^{n} \|\boldsymbol{R}_{n,i}\|^2$$

therefore gives

$$\left\|\widehat{\boldsymbol{G}}_n - \boldsymbol{\Gamma}_n\right\|_{\mathrm{op}} = O_{\mathbb{P}}\{(\log n)^{-1/4}\}.$$

In particular, $\widehat{\boldsymbol{G}}_n \to \boldsymbol{\Gamma}$ in probability.

Under (14) and (15),

$$\left\|\widehat{\boldsymbol{G}}_n - \boldsymbol{\Gamma}\right\|_{\mathrm{op}} \le \left\|\widehat{\boldsymbol{G}}_n - \boldsymbol{\Gamma}_n\right\|_{\mathrm{op}} + \|\boldsymbol{\Gamma}_n - \boldsymbol{\Gamma}\|_{\mathrm{op}} = o_{\mathbb{P}}(\kappa_n).$$

The stability lemma now yields

$$\widehat{r}_n \to r, \qquad (\widehat{\boldsymbol{G}}_n)^+_{\kappa_n} \to \boldsymbol{\Gamma}^+$$

in operator norm and in probability. If $X \sim N_L(\boldsymbol{b}, \boldsymbol{\Gamma})$, Slutsky's theorem gives

$$Q_n^{\mathrm{RA}} \Rightarrow X^\top \boldsymbol{\Gamma}^+ X.$$

Using the positive-eigenvalue decomposition from the null proof,

$$X^\top \boldsymbol{\Gamma}^+ X = \left\|Z + \Lambda_r^{-1/2} U_r^\top \boldsymbol{b}\right\|^2, \qquad Z \sim N_r(0, I_r).$$

This is noncentral chi-square with r degrees of freedom and noncentrality

$$\left\|\Lambda_r^{-1/2} U_r^\top \boldsymbol{b}\right\|^2 = \boldsymbol{b}^\top \boldsymbol{\Gamma}^+ \boldsymbol{b}.$$

This proves the local assertion.

□

Proposition J.4 (Fixed-alternative consistency of multiscale aggregation)**.** *Let P, Q be fixed, let $L < \infty$, and let the selected kernels be bounded. If $\kappa_n \downarrow 0$ and $\kappa_n \log n/n \to 0$, then any alternative satisfying $\Delta^2_{t_\ell}(P, Q) > 0$ for at least one ℓ has*

$$Q_n^{\mathrm{RA}} \to \infty \qquad \textit{in probability}.$$

Consequently, the rank-adaptive test is consistent whenever one selected scale separates P and Q.

Proof of Proposition J.4. Now let P, Q be fixed and suppose at least one $d_\ell = \Delta^2_{t_\ell}(P, Q)$ is positive. Put

$$w_\ell = \mathbb{E}\zeta_1(t_\ell), \qquad g_i^{(\ell)} = \langle w_\ell, \zeta_i(t_\ell)\rangle, \qquad \boldsymbol{g}_i = (g_i^{(1)}, \dots, g_i^{(L)})^\top.$$

Then $\mathbb{E}\boldsymbol{g}_i = \boldsymbol{d} = (d_1, \dots, d_L)^\top \neq \boldsymbol{0}$. Also

$$D_i^{(\ell)} - g_i^{(\ell)} = \left\langle \bar{\zeta}_{i-1}(t_\ell) - w_\ell, \zeta_i(t_\ell)\right\rangle.$$

Independence of the current observation from the past and boundedness give

$$\mathbb{E}\left\|\boldsymbol{D}_i - \boldsymbol{g}_i\right\|^2 \le \frac{C}{i-1}.$$

Consequently,

$$\frac{1}{n}\sum_{i=2}^{n} \left\|\boldsymbol{D}_i - \boldsymbol{g}_i\right\|^2 \to 0 \tag{41}$$

in L^1 and in probability. By Cauchy–Schwarz, (41), and the ordinary law of large numbers,

$$\frac{\boldsymbol{S}_n}{n} \to \boldsymbol{d}, \qquad \frac{\widehat{\boldsymbol{\Sigma}}_n}{n} \to B := \mathbb{E}(\boldsymbol{g}_1 \boldsymbol{g}_1^\top)$$

in probability. To verify the second convergence explicitly, use

$$\begin{aligned}
&\left\|\frac{1}{n}\sum_i (\boldsymbol{D}_i \boldsymbol{D}_i^\top - \boldsymbol{g}_i \boldsymbol{g}_i^\top)\right\|_{\mathrm{op}} \\
&\quad \le \left(\frac{1}{n}\sum_i \left\|\boldsymbol{D}_i - \boldsymbol{g}_i\right\|^2\right)^{1/2} \left[\left(\frac{1}{n}\sum_i \left\|\boldsymbol{D}_i\right\|^2\right)^{1/2} + \left(\frac{1}{n}\sum_i \left\|\boldsymbol{g}_i\right\|^2\right)^{1/2}\right],
\end{aligned}$$

and note that the bracket is bounded in probability.

The matrix B is positive semidefinite. If $c \in \mathrm{Null}(B)$, then

$$0 = c^\top B c = \mathbb{E}(c^\top \boldsymbol{g}_1)^2,$$

so $c^\top \boldsymbol{g}_1 = 0$ almost surely and hence $c^\top \boldsymbol{d} = 0$. Therefore $\boldsymbol{d} \in \mathrm{Range}(B)$. Since $\boldsymbol{d} \neq \mathbf{0}$,

$$\boldsymbol{d}^\top B^+ \boldsymbol{d} > 0. \tag{42}$$

Set $A_n = \widehat{\boldsymbol{\Sigma}}_n / n$ and $\tau_n = \kappa_n \log n / n$. By assumption $\tau_n \to 0$. An eigenvalue of $\widehat{\boldsymbol{G}}_n$ exceeds κ_n exactly when the corresponding eigenvalue of A_n exceeds τ_n. Algebraically,

$$\frac{Q_n^{\mathrm{RA}}}{n} = \left(\frac{\boldsymbol{S}_n}{n}\right)^\top (A_n)^+_{\tau_n} \left(\frac{\boldsymbol{S}_n}{n}\right). \tag{43}$$

Let β be the smallest positive eigenvalue of B and let P_n^+ be the spectral projector of A_n onto eigenvalues above $\beta/2$. Because $\tau_n \to 0$ and $A_n \to B$, with probability tending to one this cluster is retained, $P_n^+ \to P_{\mathrm{Range}(B)}$, and the inverse of A_n restricted to this cluster converges to B^+. Every additional direction retained above τ_n contributes a nonnegative squared-coordinate term to (43). Taking the lower bound supplied by the positive cluster and using $\boldsymbol{S}_n / n \to \boldsymbol{d}$ therefore gives

$$\liminf_{n \to \infty} \frac{Q_n^{\mathrm{RA}}}{n} \geq \boldsymbol{d}^\top B^+ \boldsymbol{d}$$

in probability. The right-hand side is strictly positive by (42); hence $Q_n^{\mathrm{RA}} \to \infty$ in probability. Since $0 \leq \widehat{r}_n \leq L$, all possible chi-square critical values are fixed finite constants. The rejection probability therefore tends to one. □