Semiparametric Efficient Test for Interpretable Distributional Treatment Effects

Houssam Zenati **Arthur Gretton**
Gatsby Computational Neuroscience Unit
University College London

Abstract

Distributional treatment effects can be invisible to means: a treatment may preserve average outcomes while changing tails, modes, dispersion, or rare-event probabilities. Kernel tests can detect discrepancies between interventional outcome laws, but global tests do not reveal where the laws differ. We propose DR-ME, to our knowledge the first semiparametrically efficient finite-location test for interpretable distributional treatment effects. DR-ME learns a small set of reference outcomes. For each reference, it estimates the difference in average kernel similarity between outcomes under treatment and control. It returns the selected references, signed discrepancy estimates, and a joint p-value. From observational data, we derive orthogonal doubly robust kernel features whose centered oracle form is the canonical gradient of this finite witness. For fixed locations, we characterize the local testing limit: DR-ME is chi-square calibrated under the null, has noncentral chi-square local power, and uses the covariance whitening that optimizes local signal-to-noise for discrepancies visible through the selected coordinates. This efficient local-power geometry yields a principled location-learning criterion, with sample splitting preserving post-selection validity. Experiments assess calibration, power, and the benefit of covariance-whitened location learning. A semi-synthetic medical-imaging example illustrates how an image-valued reference can be reported together with its held-out witness estimate.

1 Introduction

Average treatment effects are often too coarse for causal questions involving risk, heterogeneity, or structured outcomes. A treatment may leave the mean outcome nearly unchanged while altering dispersion, tails, multimodality, or the probability of rare but consequential events. This is the motivation behind distributional treatment-effect analysis, including nonparametric policy effects [33] and inference on interventional distributions [3]. The issue is even sharper when outcomes are images, sequences, graphs, embeddings, or high-dimensional measurements [8]: reducing such outcomes to a small number of hand-chosen scalar summaries can obscure the effect of interest.

In retinal imaging, for example, an intervention may change how often a localized alteration of retinal structure occurs, or how pronounced it is (Figure 1). Pixelwise averaging across scans can blur alterations whose positions or shapes vary. A scalar summary such as mean intensity also discards how neighboring pixels are arranged into edges, contours, and textures—the local patterns that suitable representations detect. Beyond detecting a distributional effect, we would also like to understand which image patterns distinguish the outcomes under treatment and control.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

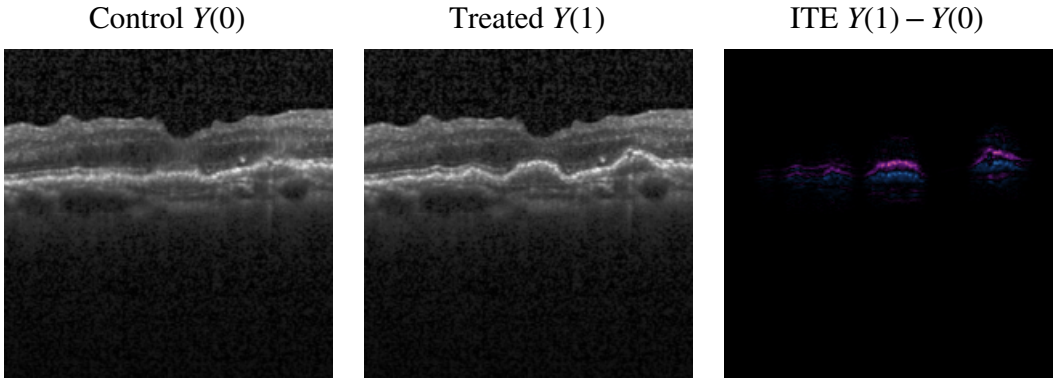


Figure 1: Semi-synthetic OCTMNIST example [18, 42]. The treated potential outcome retains a real OCT scan; the control is constructed by reducing its existing retinal elevations. The third panel shows their known individual treatment effect (ITE), with pink and blue indicating increased and decreased intensity. Appendix B.12 gives the construction.

We therefore study tests that compare the outcome distributions under treatment and control in the same target population and reveal which outcome patterns distinguish them. A useful representation should retain distributional information while allowing us to relate a detected difference back to particular images, sequences, or other outcomes.

Kernel mean embeddings make this connection through a positive definite kernel, which measures similarity between outcomes. For any reference outcome, the expected kernel similarity describes how closely outcomes drawn from a law resemble that reference. Collecting these expectations over all references represents the law as a function in a reproducing kernel Hilbert space [37]. With a characteristic kernel, equality of these embeddings is equivalent to equality of the underlying laws [38]. Their RKHS distance is the maximum mean discrepancy, used in kernel two-sample testing [9]. Applying this representation to outcome laws under treatment interventions gives interventional mean embeddings [24].[1] Policy mean embeddings extend it to outcome distributions induced by treatment policies, with doubly robust estimation [43].

Global kernel discrepancies are powerful omnibus tools, but they are often hard to interpret. A global rejection says that two interventional outcome laws differ, but not where the difference is expressed. Recent doubly robust kernel tests provide such global comparisons for observational causal studies [6, 23, 44]. In ordinary two-sample testing, finite-location mean-embedding tests instead evaluate a witness function at a small number of spatial or frequency locations [5, 16]. These locations indicate where the laws are distinguishable after accounting for sampling variability, but extending this comparison to observational causal data requires adjustment for confounding.

We develop DR-ME, a semiparametrically efficient finite-location test for distributional treatment effects. Each location compares the interventional laws through their kernel similarity to a selected outcome point. Under selection on observables and positivity [11, 15, 32, 34], these comparisons are identified through treatment probabilities and outcome regressions. To account for error in these estimates, we derive an orthogonal augmented inverse-propensity feature [1, 10, 31]. For fixed locations, its cross-fitted mean has a first-order expansion governed by the canonical gradient, while nuisance errors enter through second-order remainders [4].

Operationally, DR-ME separates nuisance fitting, location learning, and testing. We first estimate treatment probabilities and regress kernel similarities on the covariates. On a second split, we select locations whose estimated causal discrepancies are large relative to their sampling variability. We then freeze the selected locations and estimate their signed discrepancies on an independent test split. The resulting joint p-value tests for a

[1] Muandet et al. [24] introduced the term "counterfactual mean embeddings." Our targets $P_{Y(a)} = P(Y \mid \mathrm{do}(A = a))$ are interventional, on the second rung of Pearl's causal ladder [29].

distributional effect through these selected comparisons, while the references and witness estimates describe the direction of the discrepancies.

The choice of locations and the normalization used in the final test are guided by our local-efficiency analysis. For fixed locations, the canonical gradient determines the smallest asymptotic covariance available to regular estimators of the finite signal. The corresponding local-power calculation provides a criterion for learning locations that accounts for variability and dependence among their witness estimates.

Our contributions are fourfold. First, we formulate interpretable finite-location testing for distributional treatment effects by evaluating the interventional RKHS witness at selected outcome locations. Second, we derive the canonical gradient of this finite signal; its augmented inverse-propensity form yields an orthogonal doubly robust feature, whose cross-fitted version defines the DR-ME Hotelling statistic. We show that DR-ME attains the efficient finite-signal Gaussian experiment, giving chi-square calibration, noncentral local power, and canonical-gradient covariance whitening. Third, we use this geometry to learn locations and prove uniform consistency of the empirical criterion. Fourth, we evaluate calibration, power, and covariance-whitened location learning in simulations, including sensitivity to the number of locations and to difficult nuisance-estimation settings. A semi-synthetic medical-imaging study illustrates the construction and interpretation of an image-valued location.

2 Finite-location causal witnesses and observed-data features

We use the potential-outcomes notation [15, 34]. Let $A \in \{0,1\}$ be a binary treatment, let $Y(0), Y(1) \in \mathcal{Y}$ be potential outcomes, and let k_Y be a bounded positive definite kernel on $\mathcal{Y}$, with RKHS $\mathcal{H}_Y$ and feature map $\varphi_Y(y) = k_Y(\cdot, y)$. For each arm a, define the interventional mean embedding $\chi(a) := \mathbb{E}[\varphi_Y(Y(a))]$, following interventional and policy mean embedding constructions [24, 43]. The interventional embedding difference is $\Delta := \chi(1) - \chi(0)$, with witness function

$$w(y) := \langle \Delta, \varphi_Y(y) \rangle_{\mathcal{H}_Y} = \mathbb{E}[k_Y(y, Y(1))] - \mathbb{E}[k_Y(y, Y(0))]. \tag{1}$$

Thus $w(y)$ compares the expected kernel similarity to y under the two interventions. Its sign indicates which law has greater similarity to this reference, as measured by k_Y. Moreover, $\|\Delta\|^2_{\mathcal{H}_Y}$ is the squared maximum mean discrepancy between $P_{Y(1)}$ and $P_{Y(0)}$ [9]; if k_Y is characteristic, then $\Delta = 0$ is equivalent to equality of the two interventional outcome laws [38].

To obtain an interpretable finite signal, fix $J \geq 1$ outcome locations $V = (v_1, \ldots, v_J) \in \mathcal{Y}^J$, and write $k_V(y) := (k_Y(v_1, y), \ldots, k_Y(v_J, y))^\top \in \mathbb{R}^J$. The finite-location causal witness is

$$\mu_V := \big(w(v_1), \ldots, w(v_J)\big)^\top = \mathbb{E}[k_V(Y(1))] - \mathbb{E}[k_V(Y(0))] \in \mathbb{R}^J. \tag{2}$$

This is a finite-dimensional projection of the RKHS discrepancy Δ, with coordinates indexed by interpretable outcome locations. When Y is an image, each v_j is itself a complete image-space reference; when Y is a sequence, it can be a reference sequence. A location indexes a kernel comparison in outcome space, rather than a subgroup defined by X or an individual pixel. The target averages these comparisons over the same covariate population under both interventions.

These finite comparisons also determine which distributional differences the test can detect. The global null $H_0 : \Delta = 0$ implies the finite-location null $H_{0,V} : \mu_V = 0$ for every V. Conversely, a fixed finite V need not characterize every global alternative. The procedure is therefore calibrated under the global causal null, with power against alternatives visible through the selected witness coordinates. This follows the finite-location mean-embedding testing principle of Chwialkowski et al. [5] and Jitkrittum et al. [16], but for interventional rather than ordinary two-sample distributions.

If samples from $P_{Y(1)}$ and $P_{Y(0)}$ were directly available, the problem would reduce to a J-dimensional mean test based on the oracle contrasts $k_V(Y^{(1)}) - k_V(Y^{(0)})$. In observational data, these contrasts are unavailable because each unit reveals only one potential outcome. We therefore construct an observed-data pseudo-feature whose mean is μ_V and whose centered form is the canonical gradient of this finite-location signal.

Assumption A1 (Identification)**.** Let $Z = (X, A, Y) \sim P_0$, where X contains pre-treatment covariates. For each $a \in \{0, 1\}$, assume consistency, $Y = Y(a)$ when $A = a$; conditional exchangeability, $Y(a) \perp A \mid X$; and positivity, $\pi_0(a \mid X) := P_0(A = a \mid X) \geq \varepsilon > 0$ almost surely [11, 32].
For fixed V, define

$$m_a(x; V) := \mathbb{E}[k_V(Y) \mid A = a, X = x], \qquad \mu_{a,V} := \mathbb{E}[k_V(Y(a))]. \tag{3}$$

Each coordinate of m_a regresses a scalar kernel similarity on X within arm a, even when the outcome is an image or a sequence. Averaging these regressions over the covariate population identifies the interventional means: $\mu_{a,V} = \mathbb{E}[m_a(X; V)]$, and hence $\mu_V = \mathbb{E}\big[m_1(X; V) - m_0(X; V)\big]$. Thus μ_V is identified by a regression contrast, but the feasible plug-in contrast is not orthogonal to nuisance error: regression estimation error induces first-order bias. We therefore use an augmented feature whose moment is locally insensitive to nuisance perturbations, so that first-stage errors enter only through higher-order remainders [4].

Orthogonal observed-data feature. Let $\eta = (\pi, r_0, r_1)$ be a nuisance tuple, where $\pi(a \mid x) > 0$ is a propensity model and $r_a(x; V) \in \mathbb{R}^J$ is a candidate regression for $m_a(x; V)$. Define the arm-specific augmented feature

$$\phi_V^a(Z; \eta) := \frac{\mathbf{1}\{A = a\}}{\pi(a \mid X)}\big(k_V(Y) - r_a(X; V)\big) + r_a(X; V), \qquad a \in \{0, 1\},$$

and the doubly robust contrast $z_V^{\mathrm{dr}}(Z; \eta) := \phi_V^1(Z; \eta) - \phi_V^0(Z; \eta)$. This is the vector-valued analogue of augmented inverse-propensity features [1, 10, 31]. The regression term predicts the similarity for every unit. The weighted residual then corrects that prediction using units whose observed treatment matches arm a.

Proposition 2.1 (Doubly robust feature and canonical gradient)**.** *Under Assumption A1, for any nuisance tuple $\eta = (\pi, r_0, r_1)$ with $\pi(a \mid X) > 0$ almost surely,*

$$\mathbb{E}\{\phi_V^a(Z; \eta)\} - \mu_{a,V} = \mathbb{E}\left[\frac{\pi_0(a \mid X) - \pi(a \mid X)}{\pi(a \mid X)}\{m_a(X; V) - r_a(X; V)\}\right].$$

Consequently, $\mathbb{E}\{\phi_V^a(Z; \eta)\} = \mu_{a,V}$ if either $\pi(a \mid \cdot) = \pi_0(a \mid \cdot)$ or $r_a(\cdot; V) = m_a(\cdot; V)$. Let $\eta_0 = (\pi_0, m_0, m_1)$. Then $\mathbb{E}\{z_V^{\mathrm{dr}}(Z; \eta_0)\} = \mu_V$, and

$$\psi_V(Z) := z_V^{\mathrm{dr}}(Z; \eta_0) - \mu_V$$

is the canonical gradient of μ_V in the nonparametric model.

The proof is given in Appendix E. Related pathwise differentiability and efficient influence function results for the full RKHS-valued interventional mean embedding are established by Luedtke and Chung [21]. For the finite witness considered here, Proposition 2.1 has two roles. First, its bias identity explains how the augmentation controls bias: the target-moment error is a product of propensity and outcome-regression errors. Second, the canonical gradient determines the covariance governing estimation uncertainty. We use this representation to construct the statistic below.

DR-ME statistic. Let $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P_0$. For each i, let $\hat{\eta}_{-i}$ be a nuisance estimate trained on data independent of Z_i; this notation covers both K-fold cross-fitting and the independent nuisance split used after location learning. Define

$$\hat{z}_{i,V}^{\mathrm{dr}} := z_V^{\mathrm{dr}}(Z_i; \hat{\eta}_{-i}), \quad \bar{z}_{n,V}^{\mathrm{dr}} := \frac{1}{n}\sum_{i=1}^n \hat{z}_{i,V}^{\mathrm{dr}}, \quad S_{n,V}^{\mathrm{dr}} := \frac{1}{n-1}\sum_{i=1}^n (\hat{z}_{i,V}^{\mathrm{dr}} - \bar{z}_{n,V}^{\mathrm{dr}})(\hat{z}_{i,V}^{\mathrm{dr}} - \bar{z}_{n,V}^{\mathrm{dr}})^\top.$$

The fixed-location doubly robust mean-embedding statistic, abbreviated DR-ME, is

$$\hat{\lambda}_{n,V}^{\mathrm{dr}} := n\, \bar{z}_{n,V}^{\mathrm{dr}\top}(S_{n,V}^{\mathrm{dr}} + \gamma_n I_J)^{-1} \bar{z}_{n,V}^{\mathrm{dr}}, \qquad \gamma_n \downarrow 0. \tag{4}$$

This is the observed-data analogue of the finite-location Hotelling statistic [13]. Covariance normalization accounts for variability and dependence among the witness coordinates, while γ_n stabilizes the inverse.

Assumption A2 (Nuisance estimation at fixed locations)**.** For fixed $V \in \mathcal{Y}^J$, nuisance estimates use cross-fitting with a fixed number K of folds of nonvanishing relative size, or an independent nuisance sample. With probability tending to one, estimated propensities are uniformly bounded away from zero and fitted regressions are uniformly bounded. Define

$$\alpha_{a,k} := \|\hat{\pi}_k(a \mid \cdot) - \pi_0(a \mid \cdot)\|_{L_2(P_X)}, \qquad \beta_{a,k}(V) := \|\hat{m}_{a,k}(\cdot; V) - m_a(\cdot; V)\|_{L_2(P_X;\mathbb{R}^J)}.$$

The errors satisfy

$$\max_k \sum_{a=0}^{1} \{\alpha_{a,k} + \beta_{a,k}(V)\} = o_p(1), \qquad \max_k \sum_{a=0}^{1} \alpha_{a,k}\beta_{a,k}(V) = o_p(n^{-1/2}).$$

For an independent nuisance split, the maxima over k are omitted.
The product condition in Assumption A2 makes the bias from nuisance estimation negligible at the scale used by the test [4, 17]. Both errors may converge more slowly than $n^{-1/2}$; for example, $o_p(n^{-1/4})$ for each is sufficient.

Theorem 2.2 (Fixed-location first-order representation)**.** *Fix $V \in \mathcal{Y}^J$, and let $\Sigma_V := \mathrm{Var}_{P_0}\{\psi_V(Z)\}$. Under Assumptions A1–A2 and the bounded-kernel condition,*

$$\sqrt{n}\big(\bar{z}^{\mathrm{dr}}_{n,V} - \mu_V\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_V(Z_i) + o_p(1), \qquad S^{\mathrm{dr}}_{n,V} \to_p \Sigma_V.$$

If, in addition, Σ_V is positive definite and $\gamma_n \downarrow 0$, then $(S^{\mathrm{dr}}_{n,V} + \gamma_n I_J)^{-1} \to_p \Sigma_V^{-1}$.

The proof is given in Appendix E. The theorem reduces the feasible fixed-location statistic to the empirical mean of the canonical gradient. The covariance Σ_V therefore links uncertainty in the estimated witness to the efficient local testing experiment developed next.

3 Efficient local testing geometry

We now explain how to calibrate the statistic and use its power to guide location learning. The null law will give the final p-value, while the local-power calculation will determine the objective optimized on the learning split.

To compare the sensitivity of tests, we study the fixed-location test $H_{0,V} : \mu_V = 0$ against $H_{1,V} : \mu_V \neq 0$, in the local regime where alternatives approach the null at $n^{-1/2}$ rate. This is the regime in which asymptotic power is nondegenerate: fixed alternatives are detected with probability tending to one, while faster local alternatives are invisible to regular tests. Appendix D recalls the local-asymptotic testing tools used below, including local asymptotic normality (LAN), contiguity, regular procedures, and the efficiency interpretation.

To describe these local alternatives, let $\mathcal{P}$ be the nonparametric model for the observed-data law on $\mathcal{Z} = \mathcal{X} \times \{0, 1\} \times \mathcal{Y}$, and fix $P_0 \in \mathcal{P}$ with $\mu_V(P_0) = 0$. Let T_{P_0} denote the tangent space at P_0 [2, 39]. For a score direction $g \in T_{P_0}$, let $t \mapsto P_{t,g}$ be a quadratic-mean differentiable regular path through P_0 with score g, and define the contiguous local alternatives

$$P_{n,h,g} := P^{\otimes n}_{h/\sqrt{n},g}, \qquad h \in \mathbb{R}.$$

Quadratic-mean differentiability means that the square-root likelihood is differentiable in $L_2(P_0)$ with score g. It supplies the local Gaussian likelihood approximation and contiguity used below [39, Theorem 7.2]. We use the same path to measure how the finite witness moves under a local alternative. Since Proposition 2.1 identifies ψ_V as the canonical gradient of μ_V, define the local drift

$$\eta_V(g) := \frac{d}{dt}\bigg|_{t=0} \mu_V(P_{t,g}) = \mathbb{E}_{P_0}\{\psi_V(Z)g(Z)\}. \tag{5}$$

Its covariance-whitened magnitude is

$$\lambda_V(g) := \eta_V(g)^\top \Sigma_V^{-1} \eta_V(g), \qquad \Sigma_V := \mathrm{Var}_{P_0}\{\psi_V(Z)\}. \tag{6}$$

Thus $\eta_V(g)$ records how the selected witness coordinates move under the local alternative, while $\lambda_V(g)$ measures the size of that movement after normalization by the canonical-gradient

covariance. If $\eta_V(g) = 0$, the selected locations are locally blind to g; if $\eta_V(g) \neq 0$, the local alternative produces a first-order shift in the finite witness coordinates.

Theorem 3.1 (Efficient local testing geometry)**.** *Fix $V \in \mathcal{Y}^J$ and let $P_0 \in \mathcal{P}$ satisfy $\mu_V(P_0) = 0$. Let $t \mapsto P_{t,g}$ be a quadratic-mean differentiable regular path through P_0 with score $g \in T_{P_0}$. Assume $\Sigma_V \succ 0$, $\gamma_n \downarrow 0$, and that the first-order representation and covariance consistency in Theorem 2.2 hold under $P_0^{\otimes n}$. Then, under $P_{n,h,g}$,*

$$\sqrt{n}\,\bar{z}^{\mathrm{dr}}_{n,V} \xrightarrow{d} N\big(h\eta_V(g), \Sigma_V\big), \qquad \hat{\lambda}^{\mathrm{dr}}_{n,V} \xrightarrow{d} \chi^2_J\big(h^2\lambda_V(g)\big)\,.$$

The proof is given in Appendix F. The canonical-gradient covariance Σ_V is the semiparametric efficiency bound for estimating μ_V: any regular limit law is the convolution of $N(0, \Sigma_V)$ with a further noise distribution [39, Theorem 25.20]. DR-ME attains the bound.

In the resulting finite-signal Gaussian experiment, a linear contrast with coefficients b has local signal determined by $b^\top \eta_V(g)$ and variance $b^\top \Sigma_V b$. Maximizing the squared signal-to-noise ratio gives

$$\sup_{b \neq 0} \frac{\{b^\top \eta_V(g)\}^2}{b^\top \Sigma_V b} = \eta_V(g)^\top \Sigma_V^{-1} \eta_V(g) = \lambda_V(g).$$

Thus covariance whitening gives the largest squared signal-to-noise ratio available from the selected witness coordinates. The same quantity governs the local power of the Hotelling test.

The efficiency comparison is made along each regular local submodel; it does not assert uniformly most powerful testing over all distributional alternatives. Appendix G gives the directional and omnibus interpretation.

The noncentrality in Theorem 3.1 also explains how locations should be chosen. The relevant signal is not the raw Euclidean size of μ_V, but its size after whitening by the canonical-gradient covariance.

Proposition 3.2 (Local-power criterion for locations)**.** *For fixed V, define*

$$\Lambda_V(P) := \mu_V(P)^\top \Sigma_V(P)^{-1} \mu_V(P),$$

where $\Sigma_V(P)$ is the covariance of the canonical gradient of μ_V at P. Suppose $\Sigma_V(P)$ is nonsingular in a neighborhood of P_0 and continuous along $t \mapsto P_{t,g}$. Then, along $P_{h/\sqrt{n},g}$,

$$\mu_V(P_{h/\sqrt{n},g}) = \frac{h}{\sqrt{n}}\eta_V(g) + o(n^{-1/2}), \qquad n\,\Lambda_V(P_{h/\sqrt{n},g}) \to h^2\lambda_V(g).$$

The proof is given in Appendix F. Proposition 3.2 links the population criterion for choosing V to the noncentrality parameter in Theorem 3.1. It also explains why an unwhitened witness norm is not the right learning objective: a large discrepancy in a high-variance direction may be weak for testing, while a smaller discrepancy in a low-noise direction may be more informative.

Finally, the fixed-location theory gives a direct route to valid testing when the locations are learned on data independent of the final test split.

Corollary 3.3 (Post-selection null law for split-sample locations)**.** *Assume the global null $H_0 : \Delta = 0$. Let $\hat{V} \in \mathcal{V}$ be selected using data independent of the final testing split, and compute $\hat{\lambda}^{\mathrm{dr}}_{n_{\mathrm{te}},\hat{V}}$ on that final split with $\gamma_{n_{\mathrm{te}}} \downarrow 0$, using nuisance estimates independent of each test observation. Suppose that, conditionally on the learning split, the fixed-location first-order representation and covariance consistency of Theorem 2.2 hold at $V = \hat{V}$, with nondegenerate covariance; alternatively, suppose these conditions hold uniformly over $V \in \mathcal{V}$. Then, conditionally on the learning split,*

$$\hat{\lambda}^{\mathrm{dr}}_{n_{\mathrm{te}},\hat{V}} \xrightarrow{d} \chi^2_J.$$

Corollary 3.3 gives the calibration needed for the final test. After learning $\hat{V}$ independently of I_{te}, we compute

$$p^{\mathrm{dr}}_{\mathrm{te}} := 1 - F_{\chi^2_J}\left(\hat{\lambda}^{\mathrm{dr}}_{n_{\mathrm{te}},\hat{V}}\right),$$

where $F_{\chi^2_J}$ is the central chi-square distribution function, and reject $H_0 : \Delta = 0$ when $p^{\mathrm{dr}}_{\mathrm{te}} \le \alpha$. Since the global null implies $\mu_V = 0$ for every V, this asymptotic calibration remains valid after selection on independent data [7, 19]. The test assesses the selected coordinates jointly, so discrepancies outside these comparisons can remain undetected.

4 Learning locations by local power

We now turn the local-power analysis into a criterion for learning locations. Theorem 3.1 shows that local power is governed by the witness after whitening by the efficient observed-data covariance. We therefore maximize the ridge-stabilized criterion

$$\mathcal{P}_\tau(V) := \mu_V^\top (\Sigma_V + \tau I_J)^{-1} \mu_V, \qquad \tau > 0, \tag{7}$$

which approaches $\Lambda_V(P)$ as $\tau \downarrow 0$ when the covariance is nonsingular. This follows the ME-test feature-learning principle [16], using the covariance of the causal observed-data feature to account for variability and dependence among locations.

To estimate this criterion without using the final test data, we split the sample as $I_\eta \cup I_{\mathrm{tr}} \cup I_{\mathrm{te}}$. Nuisances $\hat{\eta}$ are fitted on I_η, locations are learned on I_{tr}, and the final DR-ME test is computed on I_{te}. For $i \in I_{\mathrm{tr}}$, let $\hat{z}^{\mathrm{dr}}_{i,V} := z^{\mathrm{dr}}_V(Z_i; \hat{\eta})$, with empirical mean and covariance $\bar{z}^{\mathrm{dr}}_{\mathrm{tr},V}$ and $S^{\mathrm{dr}}_{\mathrm{tr},V}$. The training objective is

$$\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V) := \bar{z}^{\mathrm{dr}\top}_{\mathrm{tr},V} (S^{\mathrm{dr}}_{\mathrm{tr},V} + \tau I_J)^{-1} \bar{z}^{\mathrm{dr}}_{\mathrm{tr},V}. \tag{8}$$

Because each candidate V defines different kernel responses, the fitted regressions must remain evaluable across the search class. Our ridge-regression implementation reuses the same covariate matrices for these responses, as detailed in Appendix A.

Figure 2 illustrates this objective in a representative two-location example. Since the optimization procedure need not find its global maximum, we measure the remaining gap for the learned output $\hat{V}$ by

$$\varepsilon_{\mathrm{tr}}(\hat{V}) := \sup_{V \in \mathcal{V}} \hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V) - \hat{\mathcal{P}}_{\tau,\mathrm{tr}}(\hat{V}).$$

This separates statistical error from numerical optimization error. Exhaustive search over a finite dictionary gives zero gap; greedy search and gradient ascent generally leave a nonzero gap, which we retain in the guarantee below. Finite dictionaries also allow locations to be actual structured outcomes, such as DNA or protein sequences [35, 40, 41].

For a compact Euclidean search class, the following result bounds the loss in the population criterion in terms of estimation and optimization error.

Assumption A3 (Regularity for location learning)**.** Assume $\mathcal{Y} \subset \mathbb{R}^d$ and let $\mathcal{V} \subset [-R, R]^{Jd}$ be compact, with J and $\tau > 0$ fixed. Let $\Pi(\mathcal{V}) := \{v_j : V = (v_1, \ldots, v_J) \in \mathcal{V},\ 1 \le j \le J\}$ denote the candidate reference locations. Suppose $k_Y(v, y)$ is uniformly bounded and Lipschitz in v over $\Pi(\mathcal{V})$. Assume that, with probability tending to one, estimated propensities are uniformly bounded away from zero and fitted regressions $\hat{m}_a(x; v)$ are uniformly bounded and Lipschitz in v, with constants independent of n_η. Define

$$\rho_{n_\eta} := \sum_{a=0}^{1} \left[\|\hat{\pi}(a \mid \cdot) - \pi_0(a \mid \cdot)\|_{L_2(P_X)} + \sup_{v \in \Pi(\mathcal{V})} \|\hat{m}_a(\cdot; v) - m_a(\cdot; v)\|_{L_2(P_X)} \right].$$

Theorem 4.1 (Uniform consistency of the learned criterion)**.** *Under Assumptions A1 and A3, if $n_\eta, n_{\mathrm{tr}} \to \infty$ and $\rho_{n_\eta} = o_p(1)$, then*

$$\Delta_{\mathrm{tr}} := \sup_{V \in \mathcal{V}} \left| \hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V) - \mathcal{P}_\tau(V) \right| = O_p\left(\sqrt{\frac{Jd \log n_{\mathrm{tr}}}{n_{\mathrm{tr}}}} + \rho_{n_\eta} \right).$$

Consequently, for any output $\hat{V}$ for which $\varepsilon_{\mathrm{tr}}(\hat{V}) = O_p(e_{\mathrm{tr}})$ and any $V^\star_\tau \in \arg\max_{V \in \mathcal{V}} \mathcal{P}_\tau(V)$,

$$\mathcal{P}_\tau(V^\star_\tau) - \mathcal{P}_\tau(\hat{V}) = O_p\left(\sqrt{\frac{Jd \log n_{\mathrm{tr}}}{n_{\mathrm{tr}}}} + \rho_{n_\eta} + e_{\mathrm{tr}} \right).$$

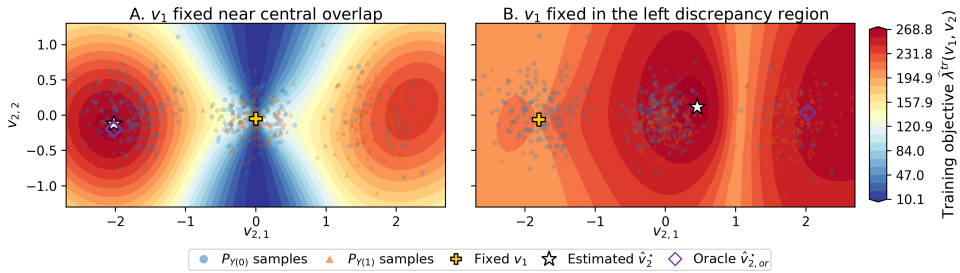


Figure 2: Training objective for learned causal locations. Each panel fixes v_1 and plots $v_2 \mapsto \hat{\mathcal{P}}_{\tau,\mathrm{tr}}(v_1, v_2)$. Blue/orange points show samples from $P_{Y(0)}/P_{Y(1)}$; the gold marker is v_1, the white star is the feasible maximizer, and the purple diamond is the oracle-nuisance maximizer on the same grid.

The proof is given in Appendix H. The first term is the statistical price of optimizing over a Jd-dimensional Euclidean class, the second is the nuisance-induced discrepancy, and the third is the empirical optimization gap. Learning this criterion requires only $\rho_{n_\eta} = o_p(1)$; the product-rate condition in Assumption A2 is needed for calibration of the final test.

The theorem gives near-optimality within the chosen search class and for the ridge-stabilized criterion; it does not claim recovery of globally optimal locations over all of $\mathcal{Y}^J$. Here the data-generating law remains fixed as the sample size grows; whether learned locations attain oracle local power when the alternative also approaches the null requires a separate analysis.

For structured outcomes where Euclidean optimization is inappropriate, Appendix I gives a finite-dictionary analogue. If $\mathcal{C} = \{c_1, \ldots, c_M\} \subset \mathcal{Y}$ and $\mathcal{V}_J(\mathcal{C}) \subset \mathcal{C}^J$, the Euclidean complexity term is replaced by $\sqrt{\frac{\log |\mathcal{V}_J(\mathcal{C})|}{n_{\mathrm{tr}}}} \le \sqrt{\frac{J \log M}{n_{\mathrm{tr}}}}$. With exhaustive dictionary search, $\varepsilon_{\mathrm{tr}}(\hat{V}) = 0$. Matrix formulas and implementation details are given in Appendix A.

Once locations have been learned, we return to the independent test split. Conditional on $(I_\eta, I_{\mathrm{tr}})$, the selected $\hat{V}$ is fixed relative to I_{te}. Hence the final statistic computed on I_{te} is a fixed-location statistic at $V = \hat{V}$, and Corollary 3.3 gives post-selection calibration under the global null. Under its conditional or uniform assumptions, the learning step affects power while preserving asymptotic null calibration. Algorithm 1 summarizes the resulting procedure.

Algorithm 1 DR-ME: from observations to a distributional comparison

Require: Pre-treatment covariates X, treatment A, outcomes Y, an outcome kernel, a search class, a number of locations J, learning and test ridges, and a level α.

1: Split the observations into nuisance, learning, and test samples.
2: On the nuisance sample, fit treatment probabilities and regress kernel similarities on X separately in each arm.
3: On the learning sample, select J reference outcomes with large adjusted witness values relative to their covariance, using the fitted nuisances.
4: Freeze the locations and nuisance fits. On the test sample, estimate their signed witness coordinates and covariance, and compute the quadratic statistic T in Eq. (4).
5: Report the references, witness estimates, standard errors, and joint p-value $p = 1 - F_{\chi^2_J}(T)$. Reject when $p \le \alpha$.

5 Experiments

We evaluate calibration and power under observational confounding, examine the role of covariance whitening, and illustrate learned locations on image outcomes.[2]

The main synthetic experiments use a confounded observational design with nonlinear outcome structure, three equally sized splits, and 200 Monte Carlo replications at nominal

[2]Code is available at `https://github.com/houssamzenati/interpretable-dr-me-test`.

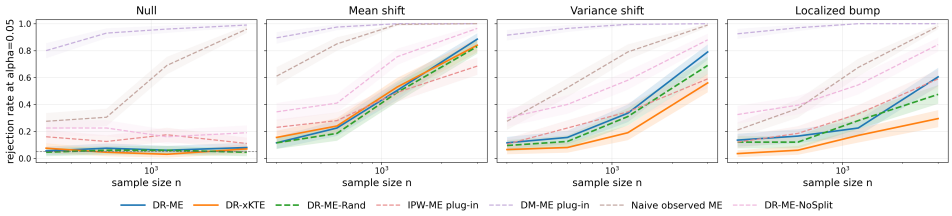


Figure 3: Calibration and power under observational confounding. Left: null rejection rates; the dashed line is the nominal level 0.05. Right: rejection rates under mean, variance, and localized-mixture shifts. Shading shows ±1.96 Monte Carlo standard errors over 200 repetitions. Among calibrated methods, learned locations improve power on the variance and localized-mixture alternatives in this design.

level $\alpha = 0.05$. Nuisances are estimated from I_η, locations are learned from I_{tr}, and the split-sample tests are evaluated on I_{te}; the no-split diagnostic deliberately reuses data for learning and testing. Unless stated otherwise, DR-ME uses a Gaussian outcome kernel, $M = 80$ candidate locations, and $J = 2$ selected locations. Full data-generating processes, nuisance models, bandwidth choices, and implementation details are given in Appendix B.

Calibration and power. Figure 3 compares DR-ME with finite-location diagnostics and the global kernel baseline. DR-ME-Rand uses the same statistic with random locations, IPW-ME uses only inverse-propensity weighting, DM-ME uses the fitted regression contrast, Naive observed ME ignores X, and DR-ME-NoSplit learns and tests on the same data. DR-xKTE [23] is the global doubly robust kernel competitor. The left panel illustrates the roles of orthogonalization and sample splitting: naive, plug-in, and no-split variants over-reject under the null in this confounded setting. The right panels show that, among calibrated methods, learned locations improve over random locations and outperform DR-xKTE on the variance and localized-mixture alternatives at larger sample sizes. Figure 7 gives a complementary summary of this comparison.

Additional results assess calibration across nominal levels, sensitivity to J, and regional stability (Appendices B.5–B.8). Appendix B.9 examines high-dimensional covariates, where calibration can deteriorate, and reports an independent size-adjusted comparison.

Covariance geometry. We next isolate the role of covariance whitening in location learning. The final test is fixed across variants: all use the same split-sample DR-ME Hotelling statistic on I_{te}; only the training-split location criterion changes. We compare full whitening, $\hat{\mu}_V^\top(\hat{\Sigma}_V + \tau I)^{-1}\hat{\mu}_V$, with raw witness maximization, $\|\hat{\mu}_V\|^2$, and random locations. Outcomes are non-scalar, $Y \in \mathbb{R}^{d_Y}$, and the alternative places rare localized mass in two sparse regions, so most dictionary locations are uninformative. This tests whether whitening selects high signal-to-noise, nonredundant witness coordinates rather than large or random witness values. Table 1 shows that the difference is power, not calibration: all three rules are close to nominal level under the null, but full whitening is strongest at every dimension, with large gains when informative locations are rare. This supports the criterion $\mu_V^\top \Sigma_V^{-1} \mu_V$: covariance whitening selects finite-location coordinates with favorable observed-data signal-to-noise. Additional selection diagnostics are deferred to Appendix B.

Interpretable image outcomes on OCTMNIST. To examine how locations can be interpreted when outcomes are images, we use a semi-synthetic OCTMNIST experiment [42]. Synthetic pre-treatment covariates drive treatment assignment and the frequency of a fluid-like pattern, while normal OCT images provide baseline anatomy. Potential outcomes are full 28×28 images $Y(a) = B + R(a)$. The residuals have equal conditional means but different distributions: one arm has mild fluid-like changes, while the other has rare severe changes. A single reference image is selected using a training-only search with multiple initializations, smoothing, and a support prior derived from the synthetic fluid template. The displayed reference is thresholded for visualization; these search choices shape its appearance.

Table 1: Non-scalar outcome two-bump ablation with $n = 3000$, $M = 300$, and $J = 3$. Entries are rejection rates at level $\alpha = 0.05$ over 200 repetitions. All methods use the same final DR-ME statistic; only the location-learning rule differs. Full covariance whitening remains calibrated under the null and gives the strongest power.

	Null				Two-bump alternative			
Method	$d_Y = 5$	10	25	50	$d_Y = 5$	10	25	50
Full whitening	0.070	0.025	0.050	0.070	0.940	0.945	0.760	0.660
Raw witness	0.055	0.035	0.070	0.055	0.135	0.210	0.380	0.500
Random	0.070	0.035	0.060	0.025	0.100	0.190	0.295	0.280

Figure 4 illustrates the resulting reference and mean contrasts. Its held-out witness estimate is 0.00745 with estimated standard error 0.02265, giving the nominal DR-ME p-value 0.742 on 375 test observations. Thus this run does not reject at level 0.05. The example demonstrates how to represent and inspect a structured location, but does not establish detection or pixelwise causal localization. Appendix B.12 reports the unthresholded reference, the complete search procedure, and the dependence introduced by sample-level mean matching.

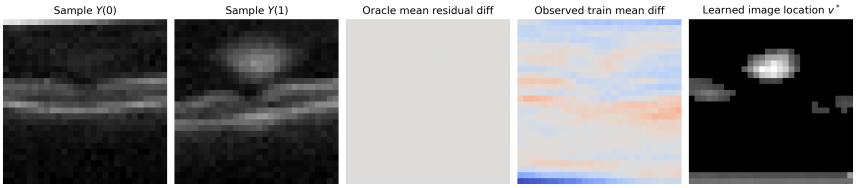


Figure 4: OCTMNIST image-space reference. From left to right: sampled potential images $Y(0)$ and $Y(1)$, oracle residual mean difference, observed training mean difference, and the displayed reference $v^\star$. The oracle contrast is zero after sample-level recentering; the observed contrast can remain nonzero under confounding. The last panel is a thresholded display of a reference selected with a fluid-support prior, rather than a map of significant pixels. Its nominal held-out p-value is 0.742.

6 Discussion

This paper constructs a causal DR-ME Hotelling test for finite-location projections of interventional distributional discrepancies. For fixed locations, the target is a regular observed-data parameter whose canonical-gradient covariance determines local power through $\eta_V(g)^\top \Sigma_V^{-1} \eta_V(g)$, motivating the location-learning criterion. DR-ME complements global kernel tests by targeting interpretable, locally efficient witness coordinates, with power depending on whether the learned locations capture the discrepancy. Applied use requires a meaningful outcome kernel, pre-treatment covariates sufficient for exchangeability, and adequate treatment overlap. Our imaging example uses real images within a synthetic causal design; an observational clinical study would require these assumptions to be assessed in its own setting. Beyond point treatments under observed confounding, longitudinal, adaptive, missing-data, instrumental-variable, or proximal settings would require new canonical gradients and local experiments. Other extensions include deep kernels or learned feature maps, with representation learning separated from final testing to preserve calibration, and location searches that exploit outcome structure through finite dictionaries for DNA sequences or constrained search classes for images.

Acknowledgments and Disclosure of Funding

Houssam Zenati and Arthur Gretton are supported by the Gatsby Charitable Foundation.

References

[1] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005. doi: 10.1111/j.1541-0420.2005.00377.x.

[2] Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

[3] Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

[4] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

[5] Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[6] Jake Fawkes, Robert Hu, Robin J. Evans, and Dino Sejdinovic. Doubly robust kernel statistics for testing distributional treatment effects. *Transactions on Machine Learning Research*, 2024.

[7] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection, 2014.

[8] Thomas Gärtner. A survey of kernels for structured data. *ACM SIGKDD explorations newsletter*, 5(1):49–58, 2003.

[9] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[10] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998. doi: 10.2307/2998560.

[11] Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

[12] Reiner Horst and Hoang Tuy. *Global Optimization: Deterministic Approaches*. Springer, Berlin, 3 edition, 1996.

[13] Harold Hotelling. The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, 1931. doi: 10.1214/aoms/1177732979.

[14] Aaron Hudson, Marco Carone, and Ali Shojaie. Inference on function-valued parameters using a restricted score test, 2021.

[15] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

[16] Wittawat Jitkrittum, Zoltán Szabó, Kacper P. Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems 29*, pages 181–189, 2016.

[17] Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: A review, 2022.

[18] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y. L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xiaogang Fu, Yaou Duan, Viet Anh Nguyen Huu, Cindy Wen, Edward D. Zhang, Chenyang L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, 2018.

[19] Arun Kumar Kuchibhotla, John E. Kolassa, and Todd A. Kuffner. Post-selection inference. *Annual Review of Statistics and Its Application*, 9:505–527, 2022. doi: 10.1146/annurev-statistics-100421-044639.

[20] Lucien M. Le Cam and Grace Lo Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer Series in Statistics. Springer, 2 edition, 2000. doi: 10.1007/978-1-4612-1166-2.

[21] Alex Luedtke and Incheoul Chung. One-step estimation of differentiable hilbert-valued parameters. *The Annals of Statistics*, 52(4):1534–1563, 2024.

[22] Alexander R. Luedtke, Marco Carone, and Mark J. van der Laan. An omnibus nonparametric test of equality in distribution for unknown functions. *Journal of the Royal Statistical Society: Series B*, 81(1):75–99, 2019.

[23] Diego Martinez Taboada, Aaditya Ramdas, and Edward Kennedy. An efficient doubly-robust test for the kernel treatment effect. In *Advances in Neural Information Processing Systems*, volume 36, pages 59924–59952, 2023.

[24] Krikamol Muandet, Motonobu Kanagawa, Sorawit Saengkyongam, and Sanparith Marukatat. Counterfactual mean embeddings. *Journal of Machine Learning Research*, 22(162):1–71, 2021.

[25] Jeffrey Näf, Junhyung Park, and Herbert Susmann. Causal-DRF: Conditional kernel treatment effect estimation using distributional random forest. *arXiv preprint arXiv:2411.08778*, 2026. URL `https://arxiv.org/abs/2411.08778v2`. Version 2.

[26] Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694–706):289–337, 1933.

[27] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, 2 edition, 2006.

[28] Junhyung Park, Uri Shalit, Bernhard Schölkopf, and Krikamol Muandet. Conditional distributional treatment effect with kernel conditional mean embeddings and u-statistic regression. In *International conference on machine learning*, pages 8401–8412, 2021.

[29] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[30] S. A. Piyavskii. An algorithm for finding the absolute extremum of a function. *USSR Computational Mathematics and Mathematical Physics*, 12(4):57–67, 1972.

[31] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. doi: 10.1080/01621459.1994.10476818.

[32] Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41.

[33] Christoph Rothe. Nonparametric estimation of distributional policy effects. *Journal of Econometrics*, 155(1):56–70, 2010.

[34] Donald B. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. doi: 10.1037/h0037350.

[35] Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert, editors. *Kernel Methods in Computational Biology*. MIT Press, Cambridge, MA, 2004. ISBN 9780262195096.

[36] Bruno O. Shubert. A sequential method seeking the global maximum of a function. *SIAM Journal on Numerical Analysis*, 9(3):379–388, 1972.

[37] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In *International conference on algorithmic learning theory*, pages 13–31. Springer, 2007.

[38] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

[39] A. W. van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

[40] Jean-Philippe Vert. Classification of biological sequences with kernel methods. In Yasubumi Sakakibara, Satoshi Kobayashi, Kengo Sato, Tetsuro Nishino, and Etsuji Tomita, editors, *Grammatical Inference: Algorithms and Applications*, volume 4201 of *Lecture Notes in Computer Science*, pages 7–18, Berlin, Heidelberg, 2006. Springer. doi: 10.1007/11872436_2.

[41] Jean-Philippe Vert, Robert Thurman, and William S. Noble. Kernels for gene regulatory regions. In Yair Weiss, Bernhard Schölkopf, and John Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1401–1408, Cambridge, MA, 2005. MIT Press.

[42] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[43] Houssam Zenati, Bariscan Bozkurt, and Arthur Gretton. Doubly-robust estimation of counterfactual policy mean embeddings. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=0GDlX9JFf2`.

[44] Houssam Zenati, Bariscan Bozkurt, and Arthur Gretton. Kernel treatment effects with adaptively collected data. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL `https://openreview.net/forum?id=di2b5POqcE`.

Appendix

Appendix organization. Readers implementing DR-ME can begin with Appendix A, which gives the full procedure, reusable regression matrices, and numerical choices. Appendix B collects the experimental settings and results, including calibration across nominal levels, sensitivity to J, location stability, high-dimensional covariate stress tests, and the image example. The remaining appendices develop the theory: Appendix C states the full assumptions, Appendix D introduces local asymptotic testing and related inferential targets, and Appendices E–I provide proofs and further results. An implementation accompanies the supplementary material.

A Implementation details

This appendix records the matrix implementation of the split-sample DR-ME test. Algorithm 1 summarizes the procedure in the main text; Algorithm 2 gives its detailed matrix implementation. The remaining subsections define the matrices used to evaluate the training objective and the final statistic, and give the closed-form regression formulas used for the nuisance estimates.

Algorithm 2 Split-sample DR-ME test with learned locations

Require: Data $Z_{1:n} = \{(X_i, A_i, Y_i)\}_{i=1}^n$, search class $\mathcal{V}$, number of locations J, outcome kernel k_Y, learning ridge $\tau > 0$, test ridge $\gamma_{n_{\mathrm{te}}} > 0$, level α

Ensure: Learned locations $\hat{V}$, statistic $\hat{\lambda}^{\mathrm{dr}}_{n_{\mathrm{te}},\hat{V}}$, p-value $p^{\mathrm{dr}}_{\mathrm{te}}$

Split the sample into disjoint sets I_η, I_{tr}, and I_{te}

Fit $\hat{\eta} = (\hat{\pi}, \hat{m}_0, \hat{m}_1)$ on I_η

for each candidate or optimizer evaluation $V \in \mathcal{V}$ **do**

  Compute $\hat{z}^{\mathrm{dr}}_{i,V} = z^{\mathrm{dr}}_V(Z_i; \hat{\eta})$ for all $i \in I_{\mathrm{tr}}$

  Compute $\bar{z}^{\mathrm{dr}}_{\mathrm{tr},V}$ and $S^{\mathrm{dr}}_{\mathrm{tr},V}$

  Compute

$$\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V) = \bar{z}^{\mathrm{dr}\top}_{\mathrm{tr},V}(S^{\mathrm{dr}}_{\mathrm{tr},V} + \tau I_J)^{-1}\bar{z}^{\mathrm{dr}}_{\mathrm{tr},V}$$

Select $\hat{V} \in \mathcal{V}$ with small empirical gap

$$\sup_{V\in\mathcal{V}} \hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V) - \hat{\mathcal{P}}_{\tau,\mathrm{tr}}(\hat{V}) \le \varepsilon_{\mathrm{tr}}$$

Compute $\hat{z}^{\mathrm{dr}}_{i,\hat{V}} = z^{\mathrm{dr}}_{\hat{V}}(Z_i; \hat{\eta})$ for all $i \in I_{\mathrm{te}}$

Compute $\bar{z}^{\mathrm{dr}}_{\mathrm{te},\hat{V}}$ and $S^{\mathrm{dr}}_{\mathrm{te},\hat{V}}$

Compute

$$\hat{\lambda}^{\mathrm{dr}}_{n_{\mathrm{te}},\hat{V}} = n_{\mathrm{te}}\,\bar{z}^{\mathrm{dr}\top}_{\mathrm{te},\hat{V}}(S^{\mathrm{dr}}_{\mathrm{te},\hat{V}} + \gamma_{n_{\mathrm{te}}} I_J)^{-1}\bar{z}^{\mathrm{dr}}_{\mathrm{te},\hat{V}}$$

Set

$$p^{\mathrm{dr}}_{\mathrm{te}} = 1 - F_{\chi^2_J}\left(\hat{\lambda}^{\mathrm{dr}}_{n_{\mathrm{te}},\hat{V}}\right)$$

Reject if $p^{\mathrm{dr}}_{\mathrm{te}} \le \alpha$

A.1 Three-way split and matrix notation

Let

$$\{1, \dots, n\} = I_\eta \cup I_{\mathrm{tr}} \cup I_{\mathrm{te}}$$

be the three-way split used for nuisance fitting, location learning, and final testing. For any evaluation set $I = \{i_1, \dots, i_m\}$, let $m = |I|$, let $1_m \in \mathbb{R}^m$ be the all-ones vector, and define

$$H_m := I_m - \frac{1}{m} 1_m 1_m^\top.$$

For $V = (v_1, \dots, v_J) \in \mathcal{Y}^J$, define

$$K_Y(Y_I, V) \in \mathbb{R}^{m\times J}, \qquad [K_Y(Y_I, V)]_{rj} := k_Y(v_j, Y_{i_r}).$$

Thus the r-th row of $K_Y(Y_I, V)$ is $k_V(Y_{i_r})^\top$.

For $a \in \{0,1\}$, define

$$D_{a,I} := \mathrm{diag}\left(\frac{\mathbf{1}\{A_{i_1} = a\}}{\hat{\pi}(a \mid X_{i_1})}, \ldots, \frac{\mathbf{1}\{A_{i_m} = a\}}{\hat{\pi}(a \mid X_{i_m})} \right).$$

Let $\hat{M}_{a,I}(V) \in \mathbb{R}^{m \times J}$ be the matrix of fitted arm-a regressions on I, with

$$[\hat{M}_{a,I}(V)]_{rj} = \hat{m}_a(X_{i_r}; v_j).$$

A.2 Matrix form of the pseudo-features

The arm-specific pseudo-feature matrices are

$$\hat{\Phi}^a_I(V) := D_{a,I}\big(K_Y(Y_I, V) - \hat{M}_{a,I}(V)\big) + \hat{M}_{a,I}(V), \qquad a \in \{0,1\}.$$

The doubly robust contrast matrix is

$$\hat{Z}_I(V) := \hat{\Phi}^1_I(V) - \hat{\Phi}^0_I(V) \in \mathbb{R}^{m \times J}.$$

Its r-th row is $\hat{z}^{\mathrm{dr}\top}_{i_r,V}$. Therefore

$$\bar{z}^{\mathrm{dr}}_I(V) = \frac{1}{m}\hat{Z}_I(V)^\top 1_m, \qquad S^{\mathrm{dr}}_I(V) = \frac{1}{m-1}\hat{Z}_I(V)^\top H_m \hat{Z}_I(V).$$

The empirical learning criterion on I is

$$\hat{\mathcal{P}}_{\tau,I}(V) = \bar{z}^{\mathrm{dr}}_I(V)^\top \big(S^{\mathrm{dr}}_I(V) + \tau I_J\big)^{-1} \bar{z}^{\mathrm{dr}}_I(V),$$

and the final test statistic on I_{te} is

$$\hat{\lambda}^{\mathrm{dr}}_{n_{\mathrm{te}},V} = n_{\mathrm{te}}\, \bar{z}^{\mathrm{dr}}_{I_{\mathrm{te}}}(V)^\top \big(S^{\mathrm{dr}}_{I_{\mathrm{te}}}(V) + \gamma_{n_{\mathrm{te}}} I_J\big)^{-1} \bar{z}^{\mathrm{dr}}_{I_{\mathrm{te}}}(V).$$

A.3 Outcome-regression matrices

Suppose the nuisance regressions $m_a(\cdot; V)$ are estimated on I_η by kernel ridge regression with a covariate kernel k_X. The finite-feature ridge regressions used in the experiments are obtained as the corresponding linear-kernel case on the chosen feature basis.

For each arm a, let

$$I_{\eta,a} := \{i \in I_\eta : \ A_i = a\}, \qquad n_{\eta,a} := |I_{\eta,a}|.$$

Define

$$K^X_{aa} := \big[k_X(X_i, X_j)\big]_{i,j \in I_{\eta,a}}, \qquad G^X_{I,a} := \big[k_X(X_i, X_j)\big]_{i \in I,\ j \in I_{\eta,a}}.$$

For a location set V, define the arm-a target matrix

$$U_a(V) := K_Y(Y_{I_{\eta,a}}, V) \in \mathbb{R}^{n_{\eta,a} \times J}.$$

With ridge parameter $\lambda_a > 0$,

$$\hat{M}_{a,I}(V) = G^X_{I,a}\big(K^X_{aa} + \lambda_a n_{\eta,a} I_{n_{\eta,a}}\big)^{-1} U_a(V).$$

Thus all J regression coordinates are fitted simultaneously. The matrices K^X_{aa}, $G^X_{I_{\mathrm{tr}},a}$, and $G^X_{I_{\mathrm{te}},a}$ do not depend on V; only $U_a(V)$ changes with the candidate locations.

A.4 Training and testing criteria

On the training split,

$$\hat{Z}_{\mathrm{tr}}(V) = \hat{\Phi}^1_{\mathrm{tr}}(V) - \hat{\Phi}^0_{\mathrm{tr}}(V),$$

$$\bar{z}^{\mathrm{dr}}_{\mathrm{tr}}(V) = \frac{1}{n_{\mathrm{tr}}}\hat{Z}_{\mathrm{tr}}(V)^\top 1_{n_{\mathrm{tr}}}, \qquad S^{\mathrm{dr}}_{\mathrm{tr}}(V) = \frac{1}{n_{\mathrm{tr}}-1}\hat{Z}_{\mathrm{tr}}(V)^\top H_{n_{\mathrm{tr}}} \hat{Z}_{\mathrm{tr}}(V),$$

and

$$\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V) = \bar{z}^{\mathrm{dr}}_{\mathrm{tr}}(V)^\top \big(S^{\mathrm{dr}}_{\mathrm{tr}}(V) + \tau I_J\big)^{-1} \bar{z}^{\mathrm{dr}}_{\mathrm{tr}}(V).$$

After selecting $\hat{V}$, the final statistic is computed by replacing I_{tr} with I_{te} and τ with $\gamma_{n_{\mathrm{te}}}$.

A.5 Ridges, covariance conditioning, and reported coordinates

The learning ridge τ and test ridge γ_n serve different purposes. The former defines the regularized objective being optimized; the latter stabilizes the covariance inversion in the final statistic. Our fixed-J null limit assumes a positive-definite population covariance and a deterministic $\gamma_n \to 0$, with no additional rate requirement. For example, $\gamma_n = n^{-1/2}$ satisfies this condition. A fixed positive limiting ridge instead yields a weighted chi-square limit. The simulations use the small fixed numerical values recorded below.

Repeated or nearly redundant locations can produce an ill-conditioned covariance. Such choices can be excluded or penalized during learning using I_{tr}; the selected set must still satisfy the covariance conditions of Corollary 3.3. Adding a ridge makes a matrix invertible numerically but does not restore a J-dimensional chi-square reference when the population covariance has smaller rank. A rank-adaptive extension would require its own calibration argument.

For each frozen location $\hat{v}_j$, report the held-out estimate $\bar{z}^{\mathrm{dr}}_{\mathrm{te},\hat{V},j}$ and estimated standard error $\{S^{\mathrm{dr}}_{\mathrm{te},\hat{V},jj}/n_{\mathrm{te}}\}^{1/2}$. The p-value from the quadratic statistic concerns the vector jointly. Coordinatewise significance statements require their own uncertainty assessment, with multiplicity accounted for when several coordinates are inspected.

A.6 Gradient-based Euclidean optimization

When $\mathcal{Y} \subset \mathbb{R}^d$ and $k_Y(v, y)$ is differentiable in v, the criterion $V \mapsto \hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V)$ is differentiable. Let

$$R_a := \left(K^X_{aa} + \lambda_a n_{\eta,a} I_{n_{\eta,a}}\right)^{-1}.$$

Since $\hat{M}_{a,I}(V) = G^X_{I,a} R_a U_a(V)$, for each location v_j,

$$\partial_{v_j} \hat{M}_{a,I}(V) = G^X_{I,a} R_a \, \partial_{v_j} U_a(V).$$

The matrix $\partial_{v_j} U_a(V)$ has only one nonzero column, with entries

$$[\partial_{v_j} U_a(V)]_{rj} = \nabla_1 k_Y(v_j, Y_{i_r}), \qquad i_r \in I_{\eta,a}.$$

The same structure holds for $\partial_{v_j} K_Y(Y_I, V)$.

Writing

$$\mu_I(V) := \bar{z}^{\mathrm{dr}}_I(V), \qquad A_I(V) := S^{\mathrm{dr}}_I(V) + \tau I_J,$$

the differential of the criterion is

$$d\hat{\mathcal{P}}_{\tau,I}(V) = 2\,(d\mu_I(V))^\top A_I(V)^{-1}\mu_I(V) - \mu_I(V)^\top A_I(V)^{-1}(dA_I(V))A_I(V)^{-1}\mu_I(V).$$

These formulas give analytic gradients. Automatic differentiation can also be applied directly to the matrix expressions above.

A.7 Finite-dictionary implementation

For a finite dictionary

$$\mathcal{C} = \{c_1, \ldots, c_M\} \subset \mathcal{Y},$$

precompute

$$K_Y(Y_{I_{\eta,a}}, \mathcal{C}), \qquad K_Y(Y_{I_{\mathrm{tr}}}, \mathcal{C}), \qquad K_Y(Y_{I_{\mathrm{te}}}, \mathcal{C}).$$

A candidate $V = (c_{j_1}, \ldots, c_{j_J}) \in \mathcal{C}^J$ is then evaluated by selecting columns $j_1, \ldots, j_J$ from these matrices. The formulas above apply without modification.

A.8 Numerical precomputation

The implementation precomputes

$$R_a, \qquad G^X_{I_{\mathrm{tr}},a}, \qquad G^X_{I_{\mathrm{te}},a}, \qquad D_{a,\mathrm{tr}}, \qquad D_{a,\mathrm{te}}, \qquad a \in \{0, 1\}.$$

For each new V, the remaining operations are the construction or column selection of $K_Y(Y_I, V)$ and $U_a(V)$, followed by multiplications involving $|I| \times n_{\eta,a}$ and $n_{\eta,a} \times J$ matrices. In the finite-location regime, J is small, so the dominant cost is typically location search rather than the final Hotelling statistic.

A.9 Search classes and optimization error

The empirical optimization gap $\varepsilon_{\mathrm{tr}}(\hat{V})$ separates the quality of numerical search from statistical estimation error. Exhaustive search over a finite dictionary gives zero gap, whereas greedy selection generally does not. For Euclidean classes, the boundedness and Lipschitz assumptions make the empirical criterion Lipschitz in V. A δ-net therefore gives $\varepsilon_{\mathrm{tr}}(\hat{V}) \leq \mathrm{Lip}(\hat{\mathcal{P}}_{\tau,\mathrm{tr}})\delta$; certified Lipschitz and branch-and-bound methods can also control the gap [12, 30, 36]. Plain gradient ascent generally certifies stationarity rather than global optimality [27]. Figure 2 illustrates the objective for two locations.

B Additional experiments

Common synthetic setup. Unless stated otherwise, we generate $X_i \sim N(0, I_5)$ and use the confounded treatment mechanism

$$\pi_0(1 \mid X_i) = \mathrm{clip}\{\sigma(0.90X_{i1} - 0.75X_{i2} + 0.55X_{i3} - 0.40X_{i4}), 0.06, 0.94\},$$

with prognostic component

$$g(x) = 0.90x_1 + 0.60\sin(x_2) + 0.35(x_3^2 - 1) + 0.25x_1x_4 - 0.20\cos(x_5).$$

Each sample is split into I_η, I_{tr}, and I_{te}, used respectively for nuisance fitting, location learning, and final testing. Propensities are estimated by ridge-regularized logistic regression, and the conditional kernel regressions $m_a(x; c) = E[k_Y(c, Y) \mid A = a, X = x]$ are estimated by multi-output ridge regression on nonlinear features of X. We use a Gaussian outcome kernel with bandwidth chosen on $I_\eta \cup I_{\mathrm{tr}}$, a dictionary of $M = 80$ candidate outcome locations, $J = 2$ selected locations, and nominal level $\alpha = 0.05$. The main synthetic studies use 200 Monte Carlo replications; the additional J, stability, and high-dimensional covariate studies below use 300. The Monte Carlo standard error near level 0.05 is approximately 0.015 and 0.013, respectively.

Numerical settings. For the scalar calibration and power experiments, the three splits contain $\lfloor n/3 \rfloor$, $\lfloor n/3 \rfloor$, and the remaining observations. The outcome bandwidth is the median nonzero pairwise distance on at most 512 observations from $I_\eta \cup I_{\mathrm{tr}}$, with a floor of 10^{-3}. The scalar dictionary consists of 80 equally spaced points over the learning-sample 2nd–98th percentile range, extended by 5% at each end. Propensities use logistic regression on X with penalty 10^{-2} and are clipped to $[0.05, 0.95]$. Outcome regressions use an intercept, X_j, $X_j^2 - 1$, $\sin(X_j)$, $\cos(X_j)$, and all pairwise products X_jX_k, standardized on I_η. Their ridge penalty is 1, with the intercept effectively unpenalized, and predictions are clipped to $[-0.05, 1.05]$. The learning ridge is $\tau = 10^{-4}$ and the test ridge is $\gamma = 10^{-8}$. The high-dimensional covariate study uses the different nuisance basis and penalties stated in Appendix B.9.

B.1 Computation infrastructure

The runtime study uses the local CPU setup listed below. The additional 300-repetition studies have saved seeds, scripts, and raw per-repetition outputs in their execution bundles. The high-dimensional covariate bundle was generated in a hosted Python environment; its original hardware and package versions were not recorded. Consequently, the machine specification below should not be interpreted as the provenance of every additional simulation.

- **Machine:** Apple MacBook Pro
- **Model Identifier:** Mac16,1
- **Chip:** Apple M4
- **CPU:** 10 cores (4 performance cores, 6 efficiency cores)
- **Memory:** 24 GB
- **Operating System:** macOS 15.0
- **Kernel / Architecture:** Darwin 24.0.0, `arm64`
- **Compute:** CPU-only execution for the runtime study

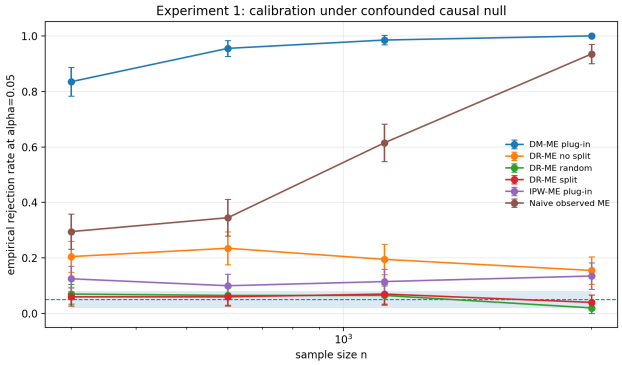


Figure 5: Sharp-null calibration under observational confounding. The dashed line is the nominal level $\alpha = 0.05$. Split DR-ME and random-location DR-ME remain calibrated; naive observed testing, no-split selection, and non-orthogonal plug-in baselines over-reject.

Table 2: Empirical rejection rates under the sharp confounded null $Y(0) = Y(1)$. The interventional laws are equal, but observed treated and control outcome distributions differ because of confounding.

Method	$n = 300$	$n = 600$	$n = 1200$	$n = 3000$
DR-ME split	0.060	0.060	0.070	0.040
DR-ME random	0.070	0.065	0.065	0.020
IPW-ME plug-in	0.125	0.100	0.115	0.135
DM-ME plug-in	0.835	0.955	0.985	1.000
Naive observed ME	0.295	0.345	0.615	0.935
DR-ME no split	0.205	0.235	0.195	0.155

B.2 Sharp-null calibration under observational confounding.

We first isolate type-I error under a sharp interventional null:

$$Y_i(0) = Y_i(1) = g(X_i) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, 1).$$

Thus $P_{Y(0)} = P_{Y(1)}$, but the observed treated and control outcome distributions differ because treatment depends on prognostic covariates. Figure 5 and Table 2 show that split-sample DR-ME remains close to nominal level across sample sizes, as does the random-location DR variant. By contrast, the naive observed ME test strongly over-rejects because it ignores confounding, the no-split DR-ME ablation over-rejects because locations are learned and tested on the same data, and the plug-in DM/IPW variants show the instability expected from non-orthogonal or non-augmented procedures. This experiment confirms that calibration comes from the orthogonal feature together with an independent final test split.

B.3 Power under confounded alternatives.

We next use the same confounded observational design, nuisance estimators, and three-way split, but replace the sharp null by three alternatives. With $Y_i(0) = g(X_i) + \varepsilon_{i0}$, we consider a broad mean shift,

$$Y_i(1) = g(X_i) + \varepsilon_{i1} + 0.35,$$

a variance shift,

$$Y_i(1) = g(X_i) + 1.45\,\varepsilon_{i1},$$

and a mean-zero mixture perturbation,

$$Y_i(1) = g(X_i) + \varepsilon_{i1} + \frac{B_i - q}{\sqrt{q(1-q)}}, \qquad B_i \sim \text{Bernoulli}(q), \quad q = 0.12.$$

The last alternative changes the shape and tail behavior of the interventional outcome law without changing the mean of the added component. We compare DR-ME with DR-xKTE,

Table 3: Empirical rejection rates for DR-ME and DR-xKTE. Under the null, the entries estimate type-I error. Under the alternatives, they estimate power.

Setting	Method	$n = 300$	$n = 600$	$n = 1200$	$n = 3000$
Null	DR-ME	0.055	0.075	0.060	0.080
	DR-xKTE	0.075	0.045	0.030	0.070
Mean shift	DR-ME	0.115	0.225	0.505	0.885
	DR-xKTE	0.155	0.240	0.530	0.840
Variance shift	DR-ME	0.115	0.155	0.335	0.790
	DR-xKTE	0.065	0.080	0.190	0.560
Localized bump	DR-ME	0.135	0.165	0.225	0.605
	DR-xKTE	0.035	0.060	0.170	0.295

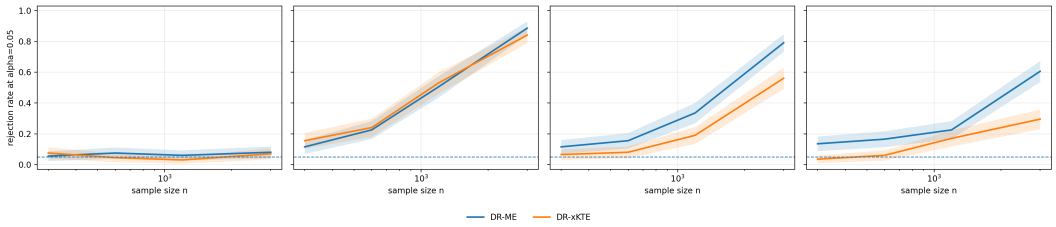


Figure 6: DR-ME versus the global DR-xKTE baseline. Both methods are close to nominal level under the confounded null. DR-ME is comparable on the mean shift and stronger on the variance-shift and localized-bump alternatives in this design.

a global doubly robust kernel test, using the same held-out test split. We also retain finite-location diagnostics from the calibration experiment: random-location DR-ME, IPW-ME, DM-ME, naive observed ME, and the no-split DR-ME ablation.

Table 3 and Figure 6 show that DR-ME and DR-xKTE have comparable calibration under the confounded null. On the broad mean-shift alternative, the two methods are similar: DR-xKTE is slightly stronger at smaller sample sizes, while DR-ME is slightly stronger at $n = 3000$. On the variance-shift and localized-bump alternatives, DR-ME is stronger at larger sample sizes. At $n = 3000$, DR-ME rejects with probabilities 0.790 and 0.605, compared with 0.560 and 0.295 for DR-xKTE. This is consistent with the finite-location objective: when a small number of whitened witness evaluations captures the discrepancy, location learning can concentrate power on informative parts of the outcome distribution rather than averaging globally.

Table 4 clarifies which rejection rates can be interpreted as calibrated power. The naive observed ME test rejects the null with probability 0.960, because it tests observed treated-versus-control outcome distributions rather than interventional laws. The no-split ablation also over-rejects, showing the selection bias induced by learning and testing locations on the same data. The DM plug-in baseline is not a valid calibrated comparator in this experiment, since it has near-unit null rejection. The calibrated comparison is therefore mainly among DR-ME, DR-ME-Rand, and DR-xKTE. Within that group, learned locations improve over random locations, and DR-ME gives the largest rejection rates on the variance and localized-bump alternatives.

B.4 High-dimensional location-learning ablation

This experiment isolates the role of the covariance-whitened learning criterion. The final test is held fixed across methods: all variants compute the same split-sample DR-ME Hotelling statistic on I_{te}. Only the location-learning rule on I_{tr} changes. We compare full whitening,

$$\hat{\mu}_V^\top (\hat{\Sigma}_V + \tau I)^{-1} \hat{\mu}_V,$$

raw witness maximization, $\|\hat{\mu}_V\|^2$, and random dictionary locations.

Table 4: Diagnostic baselines at $n = 3000$. These baselines are included to interpret the power curves, not all as valid calibrated competitors.

Method	Null	Mean shift	Variance shift	Localized bump
DR-ME	0.080	0.885	0.790	0.605
DR-xKTE	0.070	0.840	0.560	0.295
DR-ME-Rand	0.045	0.830	0.690	0.475
IPW-ME plug-in	0.110	0.685	0.590	0.590
DM-ME plug-in	0.990	1.000	1.000	1.000
Naive observed ME	0.960	1.000	0.990	0.980
DR-ME-NoSplit	0.190	0.965	0.880	0.845

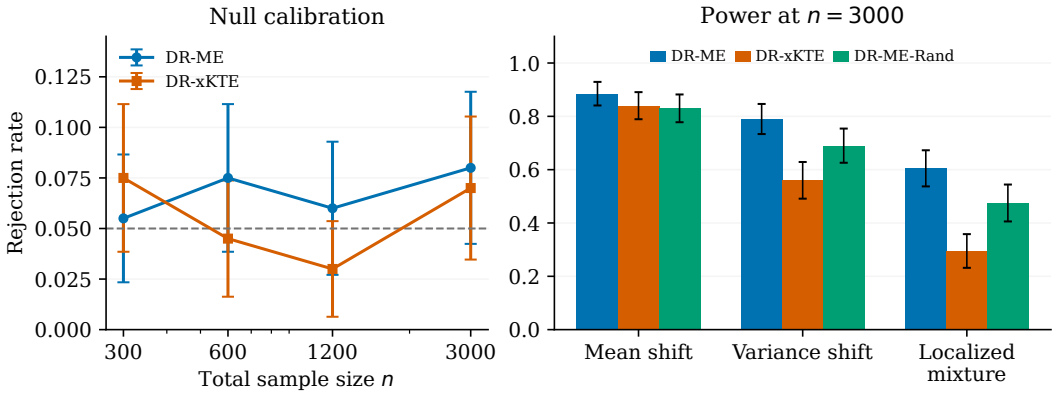


Figure 7: Complementary summary of the calibration and power results in Figure 3. Left: null rejection rates for DR-ME and DR-xKTE across sample sizes. Right: power at $n = 3000$, including random-location DR-ME. Error bars show ± 1.96 Monte Carlo standard errors over 200 repetitions.

Outcomes are vector-valued, $Y \in \mathbb{R}^{d_Y}$, with $d_Y \in \{5, 10, 25, 50\}$. Under the null,

$$Y(a) = \mu(X) + \varepsilon_a, \qquad a \in \{0, 1\}, \quad \varepsilon_0, \varepsilon_1 \text{ independent with the same Gaussian law.}$$

Under the alternative, rare mass is shifted toward two sparse regions:

$$Y(0) = \mu(X) + \varepsilon_0 - \Delta v_1 \mathbf{1}\{B_0 = 1\} - \Delta v_2 \mathbf{1}\{B_0 = 2\},$$
$$Y(1) = \mu(X) + \varepsilon_1 + \Delta v_1 \mathbf{1}\{B_1 = 1\} + \Delta v_2 \mathbf{1}\{B_1 = 2\},$$

where $\Pr(B_a = 1) = \Pr(B_a = 2) = 0.04$ and $\Delta = 4$. Candidate locations are observed outcomes from the training split, so most dictionary points are background points rather than bump locations. We use $n = 3000$, $M = 300$ dictionary candidates, $J = 3$ selected locations, and 200 Monte Carlo replications.

The main text reports null rejection rates and power. Here we report selection diagnostics explaining the power gap.

The simulation labels dictionary points by their generating mixture component. The number of distinct bump regions counts signed labels in $\{-2, -1, 1, 2\}$, so it can exceed two; the two bump types each have a treatment-specific sign. These labels are used only to evaluate selection, and are not available to the learning rule.

Table 5 shows that full whitening selects informative and less redundant coordinates. It hits at least one true bump region much more often than random locations, and it selects more distinct bump regions than the raw witness objective. It also has lower average feature correlation than raw witness maximization. Thus the gain in the main power table is not merely from finding large empirical witness values: the covariance term changes the selected finite-dimensional projection by penalizing redundant or high-noise coordinates, as predicted by the criterion $\mu_V^\top \Sigma_V^{-1} \mu_V$.

Table 5: Selection diagnostics under the high-dimensional two-bump alternative. "Any bump" is the probability that at least one selected location lies in a true bump region. "Distinct regions" is the average number of distinct bump regions selected. "Avg. corr." is the average absolute correlation among the selected DR features on the training split.

d_Y	Method	Any bump	Distinct regions	Avg. corr.
5	DR-ME	1.000	2.495	0.173
	Raw witness	0.180	0.205	0.765
	Random locations	0.200	0.230	0.384
10	DR-ME	1.000	2.240	0.170
	Raw witness	0.295	0.315	0.646
	Random locations	0.245	0.270	0.383
25	DR-ME	0.870	1.470	0.248
	Raw witness	0.420	0.500	0.529
	Random locations	0.205	0.210	0.378
50	DR-ME	0.890	1.300	0.274
	Raw witness	0.750	1.010	0.434
	Random locations	0.210	0.220	0.360

B.5 Calibration across nominal levels

We evaluate rejection rates at several nominal levels using the saved equal-law null p-values from the scalar J study below. Each sample size uses 300 repetitions with $J = 2$. Figure 8 and Table 6 compare empirical rejection with the nominal level. This complements the fixed-level calibration experiment and the full null-distribution diagnostics in the local-limit study.

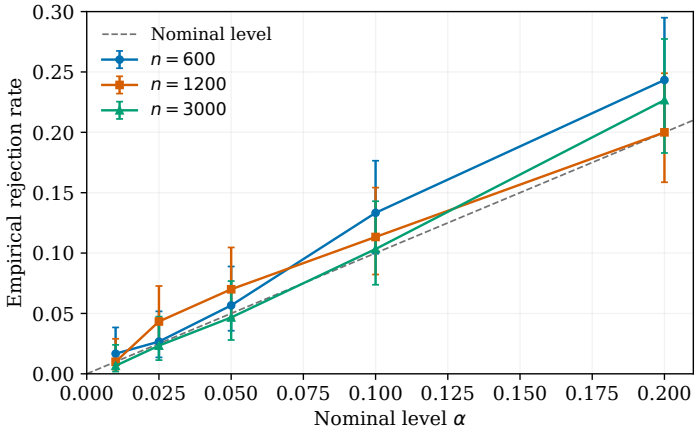


Figure 8: Rejection rates across nominal levels for the scalar equal-law null. Each curve uses 300 repetitions; vertical bars are 95% Wilson binomial intervals. The diagonal is exact calibration.

B.6 Sensitivity to the number of locations

We vary $J \in \{1, 2, 3, 5\}$ in the scalar design, with sample sizes $n \in \{600, 1200, 3000\}$ and 300 repetitions per setting. The nuisance fits, bandwidth rule, 80-point candidate dictionary, and three-way split follow the common setup. The final test is evaluated under the equal-law null and the mean-zero localized-mixture alternative. Locations are selected by greedy maximization of the full covariance-whitened criterion.

The null rejection rates remain close to 0.05 in these settings, while the effect of J on power depends on sample size. At $n = 1200$, power increases up to $J = 3$ and decreases at $J = 5$;

Table 6: Empirical rejection rates for $J = 2$ under the scalar equal-law null, over 300 repetitions.

Nominal α	$n = 600$	$n = 1200$	$n = 3000$
0.010	0.017	0.010	0.007
0.025	0.027	0.043	0.023
0.050	0.057	0.070	0.047
0.100	0.133	0.113	0.103
0.200	0.243	0.200	0.227

Table 7: Effect of the number of learned locations on type-I error and power under the mean-zero localized-mixture alternative. Each entry is an empirical rejection rate at nominal level $\alpha = 0.05$ over 300 repetitions.

	Type-I error			Power		
J	$n = 600$	$n = 1200$	$n = 3000$	$n = 600$	$n = 1200$	$n = 3000$
1	0.053	0.047	0.060	0.143	0.297	0.697
2	0.057	0.070	0.047	0.150	0.320	0.700
3	0.047	0.070	0.047	0.163	0.330	0.747
5	0.067	0.060	0.057	0.160	0.280	0.757

at $n = 3000$, the larger sets capture more of the discrepancy. Increasing J can improve representation of the effect, but also increases the chi-square critical value and the dimension of covariance estimation. These results support $J = 2$ as a compact default, with larger choices assessed on learning data when the application warrants them.

B.7 An exploratory adaptive choice of J

We also examine selection over the same finite grid using only I_{tr}. For each candidate J, two-fold internal validation learns locations on one half of I_{tr} and evaluates the quadratic criterion on the other, then exchanges the roles of the halves. The validation criterion is corrected by subtracting its first-order J/m null noise floor, where m is the validation-fold size, and truncating at zero. Multiplying the averaged corrected criterion by n_{te} gives an estimated noncentrality parameter. We choose J to maximize the corresponding noncentral-chi-square power at level 0.05 and relearn its locations on the full I_{tr}. The final test data are used only after this selection.

Table 8 reports the archived experiment at $n = 3000$. The adaptive rule increases observed power from 0.700 to 0.750, with null rejection 0.033. The selected J is dispersed across the grid, consistent with the similar fixed-J powers in Table 7.

There is a numerical qualification to this exploratory rule. When estimated noncentrality is zero, candidate powers are theoretically equal to 0.05. The archived implementation compares floating-point values directly, so numerical rounding can determine the selected J. Recomputing these powers with a different numerical library, while retaining the saved candidate statistics, changes the $n = 3000$ null rejection rate to 0.037 and alternative rejection rate to 0.747. We retain the original table for transparency. Reuse of this adaptive rule should specify a numerical tie tolerance and a deterministic tie convention; fixed $J = 2$ remains the default procedure evaluated in the main text.

B.8 Regional stability of selected locations

Coordinatewise agreement between learned locations can be misleading when several locations represent the same distributional feature. We therefore evaluate whether the selected set repeatedly includes dictionary points from known informative mixture components. This experiment uses the two-bump outcome design with $d_Y = 5$, $n = 1200$, 300 repetitions, $M = 200$, effect amplitude 5, noise standard deviation 0.70, and bump probability 0.06 for

Table 8: Data-adaptive selection of J at $n = 3000$. The first two numerical columns report rejection rates; the remaining columns report the empirical distribution of the selected number of locations.

Setting	Fixed $J = 2$	Adaptive J	$\Pr(\widehat{J} = 1)$	$\Pr(\widehat{J} = 2)$	$\Pr(\widehat{J} = 3)$	$\Pr(\widehat{J} = 5)$
Null	0.047	0.033	0.257	0.343	0.163	0.237
Localized mixture	0.700	0.750	0.327	0.183	0.207	0.283

Table 9: Regional stability of covariance-whitened learned locations under the two-bump alternative ($d_Y = 5$, $n = 1200$).

J	Power	Any region	Both regions	Informative fraction	Avg. |corr|
1	0.550	0.910	0.000	0.910	0.000
2	0.790	0.997	0.667	0.867	0.236
3	0.867	1.000	0.930	0.800	0.228
5	0.933	1.000	0.987	0.703	0.260

each type. The bandwidth is 0.70 times the median-distance rule. All methods use the common nuisance fits and a final DR-ME test; only J varies.

"Any region" records at least one selected point from a non-background component. "Both regions" records representation of both bump types, irrespective of sign; "informative fraction" is the fraction of selected points from either type. These are component-label diagnostics in a controlled simulation, not confidence statements about spatial boundaries. Labels are unavailable to the location learner.

With $J = 1$, selection hits an informative component in 91% of repetitions. With $J = 2$, both types are represented in 66.7% of repetitions, increasing to 93.0% with $J = 3$. Larger sets cover the informative components more consistently, but the informative fraction falls as background points are added. The results distinguish stable recovery of a distributional feature from recovery of one unique coordinate.

B.9 High-dimensional covariate stress test

We stress nuisance estimation by varying $d_X \in \{20, 50, 100\}$ and $n \in \{300, 600, 1200\}$. Covariates are centered Gaussian with $\mathrm{Cov}(X_j, X_k) = 0.3^{|j-k|}$. In the sparse regime, only the first five coordinates enter the propensity and prognostic function. In the dense regime, all coordinates enter through normalized coefficient vectors. Unlike the outcome-dimension ablation, this experiment increases the dimension of the confounders while keeping the outcome scalar.

In the sparse regime, let

$$s(X) = 0.90X_1 - 0.75X_2 + 0.55X_3 - 0.40X_4 + 0.30X_5,$$
$$g(X) = 0.80X_1 + 0.55\sin(X_2) + 0.30(X_3^2 - 1) + 0.22X_4X_5.$$

In the dense regime, define $u_j(\varphi) = \cos(0.37j + \varphi) + 0.60\sin(0.19j - 0.50\varphi)$ for $j = 1, \ldots, d_X$, and $w(\varphi) = u(\varphi)/\|u(\varphi)\|$. Then

$$s(X) = X^\top w(0), \qquad g(X) = 0.80X^\top w(1.1) + 0.45\sin\{X^\top w(2.3)\} + 0.18\big[\{X^\top w(3.7)\}^2 - 1\big].$$

Treatment probabilities are $\mathrm{clip}\{\sigma(0.85s(X)), 0.06, 0.94\}$. Under the sharp null, $Y(0) = Y(1) = g(X) + \varepsilon$, with standard Gaussian noise. Under alternatives, $Y(0) = g(X) + \varepsilon_0$ and

$$Y(1) = g(X) + \varepsilon_1 + 1.20\frac{B - q}{\sqrt{q(1-q)}} \quad \text{or} \quad Y(1) = g(X) + 1.55\varepsilon_1,$$

where $B \sim \mathrm{Bernoulli}(q)$, $q = 0.12$, and the two Gaussian noises are independent. The first perturbation has mean zero and changes the mixture shape; the second changes variance more diffusely.

Table 10: Type-I error in the high-dimensional covariate stress test at nominal level $\alpha = 0.05$. Each cell uses 300 Monte Carlo replications.

Regime / d_X	Method	$n = 300$	$n = 600$	$n = 1200$
Sparse / 20	DR-ME	0.077	0.067	0.033
	DR-xKTE	0.127	0.147	0.173
	DR-ME-Rand	0.080	0.053	0.020
Sparse / 50	DR-ME	0.063	0.067	0.090
	DR-xKTE	0.120	0.140	0.110
	DR-ME-Rand	0.070	0.090	0.080
Sparse / 100	DR-ME	0.050	0.087	0.133
	DR-xKTE	0.163	0.127	0.090
	DR-ME-Rand	0.050	0.063	0.120
Dense / 20	DR-ME	0.050	0.080	0.080
	DR-xKTE	0.063	0.127	0.073
	DR-ME-Rand	0.053	0.067	0.083
Dense / 50	DR-ME	0.103	0.057	0.077
	DR-xKTE	0.150	0.097	0.083
	DR-ME-Rand	0.080	0.047	0.073
Dense / 100	DR-ME	0.070	0.073	0.110
	DR-xKTE	0.160	0.133	0.120
	DR-ME-Rand	0.067	0.073	0.090

Each of the 54 cells uses 300 repetitions. All three splits have equal sizes. The propensity model uses linear covariates with penalty 20 and clipping to $[0.05, 0.95]$; kernel-similarity regressions also use a linear basis with ridge penalty 20. The outcome bandwidth follows the median rule, $M = 80$, $J = 2$, the learning ridge is 0.05, and the test ridge is 10^{-8}. DR-ME and DR-ME-Rand share the nuisance fits and final pseudo-feature matrix, differing only in selected locations. DR-xKTE uses the same held-out observations and estimated propensity, with an internal split for its cross-statistic. The design therefore probes the effect of difficult and potentially misspecified nuisance estimation on calibration as well as power.

Across the 18 null settings, median rejection is 0.075 for DR-ME, 0.072 for DR-ME-Rand, and 0.127 for DR-xKTE. DR-ME is closer to nominal level overall, but its range of 0.033–0.133 also shows substantial finite-sample distortion in some cells. Alternative rejection rates must be interpreted together with this null behavior.

Optimized locations improve the raw localized-mixture rejection rate over random locations in 17 of 18 cells, with an average gain of 4.7 percentage points. The improvement is largest when enough data are available to identify informative locations. For the variance shift, the mean gain is smaller, at 1.6 percentage points.

Independent size adjustment. To separate power from null distortion, we partition the 300 null repetitions in each (d_X, n, regime) cell by seed parity. For each method, the 150 even-seed repetitions define an empirical 5% lower-tail p-value threshold using the lower order-statistic quantile. The 150 odd-seed null repetitions assess its achieved size, and all 300 alternative repetitions are evaluated at that same threshold. This is a simulation diagnostic using separate calibration repetitions, not the nominal chi-square procedure deployed in Algorithm 1. Averaged over the 18 cells, adjusted power is 0.254, 0.211, and 0.101 for DR-ME, DR-ME-Rand, and DR-xKTE under the localized mixture, and 0.225, 0.219, and 0.119 under the variance shift. The finite calibration sample leaves uncertainty in each estimated threshold; the saved per-cell results include the independent achieved-size estimates.

B.10 Local noncentral chi-square validation

This experiment validates the local asymptotic law of the fixed-location orthogonalized ME statistic. Unlike the previous power experiments, the goal is not to compare against external baselines. The goal is to check whether, once the locations are fixed independently of the

Table 11: Rejection rates under the mean-zero localized-mixture alternative. Each cell uses 300 Monte Carlo replications.

Regime / d_X	Method	$n = 300$	$n = 600$	$n = 1200$
Sparse / 20	DR-ME	0.220	0.433	0.783
	DR-xKTE	0.130	0.210	0.327
	DR-ME-Rand	0.157	0.387	0.637
Sparse / 50	DR-ME	0.157	0.250	0.600
	DR-xKTE	0.133	0.197	0.243
	DR-ME-Rand	0.127	0.233	0.483
Sparse / 100	DR-ME	0.120	0.173	0.397
	DR-xKTE	0.183	0.217	0.227
	DR-ME-Rand	0.093	0.153	0.340
Dense / 20	DR-ME	0.180	0.270	0.473
	DR-xKTE	0.153	0.160	0.243
	DR-ME-Rand	0.140	0.253	0.383
Dense / 50	DR-ME	0.143	0.243	0.397
	DR-xKTE	0.150	0.150	0.213
	DR-ME-Rand	0.123	0.213	0.333
Dense / 100	DR-ME	0.123	0.150	0.277
	DR-xKTE	0.150	0.150	0.163
	DR-ME-Rand	0.100	0.153	0.233

Table 12: Rejection rates under the variance-shift alternative. Each cell uses 300 Monte Carlo replications.

Regime / d_X	Method	$n = 300$	$n = 600$	$n = 1200$
Sparse / 20	DR-ME	0.213	0.367	0.677
	DR-xKTE	0.177	0.263	0.433
	DR-ME-Rand	0.207	0.333	0.607
Sparse / 50	DR-ME	0.120	0.260	0.570
	DR-xKTE	0.190	0.250	0.350
	DR-ME-Rand	0.120	0.277	0.527
Sparse / 100	DR-ME	0.147	0.207	0.350
	DR-xKTE	0.210	0.217	0.243
	DR-ME-Rand	0.147	0.177	0.313
Dense / 20	DR-ME	0.157	0.253	0.417
	DR-xKTE	0.120	0.157	0.273
	DR-ME-Rand	0.133	0.267	0.393
Dense / 50	DR-ME	0.130	0.223	0.343
	DR-xKTE	0.190	0.167	0.210
	DR-ME-Rand	0.137	0.227	0.330
Dense / 100	DR-ME	0.093	0.187	0.240
	DR-xKTE	0.173	0.193	0.223
	DR-ME-Rand	0.123	0.120	0.220

test sample, the empirical Hotelling statistic follows the predicted central and noncentral chi-square laws.

The theoretical prediction is the following. For fixed locations V and local alternatives of size $n^{-1/2}$,

$$\sqrt{n}\,\bar{z}^{\mathrm{dr}}_{n,V} \Rightarrow N\left(h\eta_V(g), \Sigma_V\right), \tag{9}$$

and therefore

$$\widehat{\lambda}^{\mathrm{dr}}_{n,V} = n\bar{z}^{\mathrm{dr}\top}_{n,V}\left(S^{\mathrm{dr}}_{n,V} + \gamma_n I_J\right)^{-1}\bar{z}^{\mathrm{dr}}_{n,V} \Rightarrow \chi^2_J\left(h^2\lambda_V(g)\right), \tag{10}$$

where

$$\lambda_V(g) = \eta_V(g)^\top \Sigma_V^{-1}\eta_V(g). \tag{11}$$

Thus, under the null $h = 0$, the statistic should be approximately central χ^2_J. Under local alternatives, the rejection probability should follow the power curve of a noncentral χ^2_J distribution with noncentrality parameter $h^2\lambda_V(g)$.

We use the same observational structure as in the previous experiments:

$$X_i \sim N(0, I_5), \qquad A_i \mid X_i \sim \text{Bernoulli}(\pi_0(X_i)), \tag{12}$$

with

$$\pi_0(1 \mid X_i) = \text{clip}\left\{\sigma(0.90X_{i1} - 0.75X_{i2} + 0.55X_{i3} - 0.40X_{i4}), 0.06, 0.94\right\}. \tag{13}$$

The prognostic component is

$$g(x) = 0.90x_1 + 0.60\sin(x_2) + 0.35(x_3^2 - 1) + 0.25x_1x_4 - 0.20\cos(x_5). \tag{14}$$

The local path is a treatment-arm mean shift of order $n^{-1/2}$:

$$Y_i(0) = g(X_i) + \varepsilon_{i0}, \qquad Y_i(1) = g(X_i) + \varepsilon_{i1} + \delta_n, \qquad \delta_n = \frac{h}{\sqrt{n}}, \tag{15}$$

with $\varepsilon_{i0}, \varepsilon_{i1} \sim N(0, 1)$. The treatment mechanism and the covariate distribution are unchanged along the path. Hence the only local perturbation is in the interventional outcome law under treatment.

We use a Gaussian kernel on the outcome. Since the conditional residual law is Gaussian, the nuisance regressions

$$m_a(x; v) = E[k_Y(v, Y) \mid A = a, X = x]$$

are available in closed form. For $Y = g(x) + \delta + \varepsilon$, $\varepsilon \sim N(0, \sigma_Y^2)$, and $k_Y(v, y) = \exp\{-(v - y)^2/(2\ell_Y^2)\}$,

$$m_\delta(x; v) = \left(\frac{\ell_Y^2}{\ell_Y^2 + \sigma_Y^2}\right)^{1/2} \exp\left\{-\frac{(v - g(x) - \delta)^2}{2(\ell_Y^2 + \sigma_Y^2)}\right\}. \tag{16}$$

We use these true nuisance functions in this experiment to isolate the local chi-square approximation from nuisance-estimation error. This is a deliberate theory-validation experiment: first-stage estimation is not the object being tested here.

The locations are also fixed independently of the Monte Carlo testing samples. A large independent pilot sample under the null is used to select $J = 2$ locations by maximizing the local noncentrality proxy

$$\eta_V^\top(\Sigma_V + \tau I_J)^{-1}\eta_V. \tag{17}$$

The selected locations are then frozen for all repetitions, all sample sizes, and all local-alternative strengths. This avoids any leakage from the test sample into the location choice.

The pilot configuration is:

$$n_{\text{pilot}} = 100000, \tag{18}$$
$$M = 100, \tag{19}$$
$$J = 2, \tag{20}$$
$$\widehat{V} = (-2.6752,\ 3.9031), \tag{21}$$
$$\ell_Y = 1.4289, \tag{22}$$
$$\widehat{\lambda}_V = 0.1383. \tag{23}$$

We then run 2000 Monte Carlo repetitions for each $n \in \{500, 1000, 3000\}$ and $h \in \{0, 1, 2, 3, 4, 6, 8\}$. The rejection threshold is the 0.95 quantile of χ^2_2.

Table 13 shows that the empirical rejection probabilities closely follow the noncentral chi-square prediction. Under the null, corresponding to $h = 0$, the rejection rate is approximately 5% at all sample sizes. This verifies the central chi-square calibration of the fixed-location statistic. Under local alternatives, the rejection probability increases with h and tracks the theoretical curve well. The agreement is especially good at $n = 1000$ and $n = 3000$, as expected from an asymptotic local theory.

Table 13: Empirical rejection rates under local alternatives, compared with the noncentral chi-square prediction. The theoretical power is computed from $\chi^2_2(h^2\widehat{\lambda}_V)$ with $\widehat{\lambda}_V = 0.1383$.

h	$h^2\widehat{\lambda}_V$	Theory	$n = 500$	$n = 1000$	$n = 3000$
0	0.000	0.050	0.053	0.044	0.054
1	0.138	0.061	0.063	0.059	0.049
2	0.553	0.094	0.107	0.101	0.095
3	1.245	0.155	0.164	0.166	0.167
4	2.213	0.246	0.260	0.269	0.275
6	4.979	0.502	0.555	0.519	0.524
8	8.851	0.763	0.810	0.786	0.770

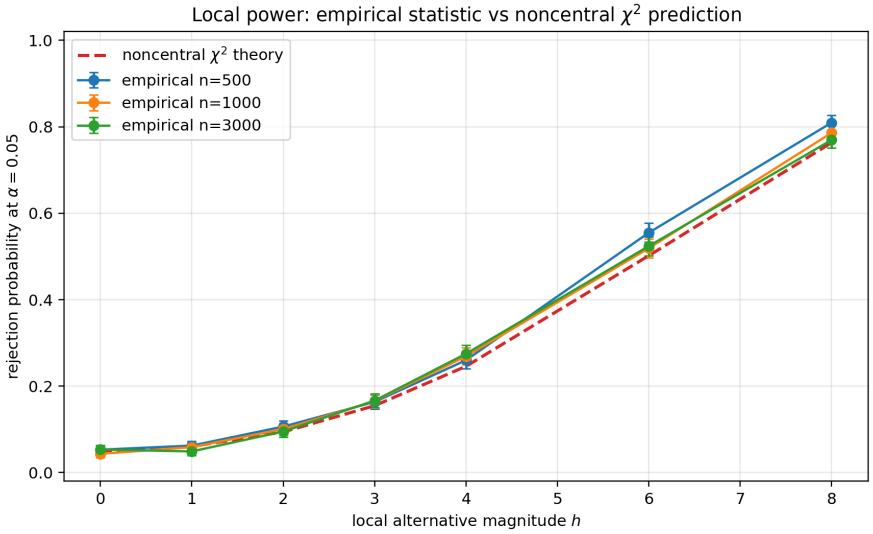


Figure 9: Empirical rejection rates against the noncentral chi-square prediction. The black theoretical curve is computed from $\chi^2_2(h^2\widehat{\lambda}_V)$. The empirical curves for $n = 500, 1000, 3000$ closely follow the predicted local-power curve, with the largest-sample curve giving the closest agreement.

At the smaller sample size $n = 500$, the empirical power is slightly above the asymptotic prediction for larger values of h, especially at $h = 6$ and $h = 8$. This is not a failure of the theory. For large h, the finite-sample shift $\delta_n = h/\sqrt{n}$ is no longer extremely small, so the experiment is moving away from the strictly local regime. The asymptotic prediction should be most accurate when n is larger and the same h corresponds to a smaller actual shift.

Figure 9 is the main figure for the experiment. It directly validates the local-power formula. The horizontal axis is the local strength h, not the raw shift δ_n. The raw shift changes with n as $\delta_n = h/\sqrt{n}$, while the theory predicts that the asymptotic power should depend on h through $h^2\lambda_V$. The empirical curves are close to the theoretical curve, which supports the claim that λ_V is the relevant whitened local signal-to-noise ratio.

Figure 10 checks the full null distribution, not only the rejection probability at level 0.05. The empirical histogram aligns well with the central χ^2_2 reference law. This supports the fixed-location null calibration of the statistic and confirms that the covariance normalization is working as intended.

Figure 11 gives a stronger distributional diagnostic. A power curve only checks one tail probability for each h, whereas the QQ plot checks the entire distribution. The approximate alignment with the diagonal indicates that the statistic is not merely producing the correct

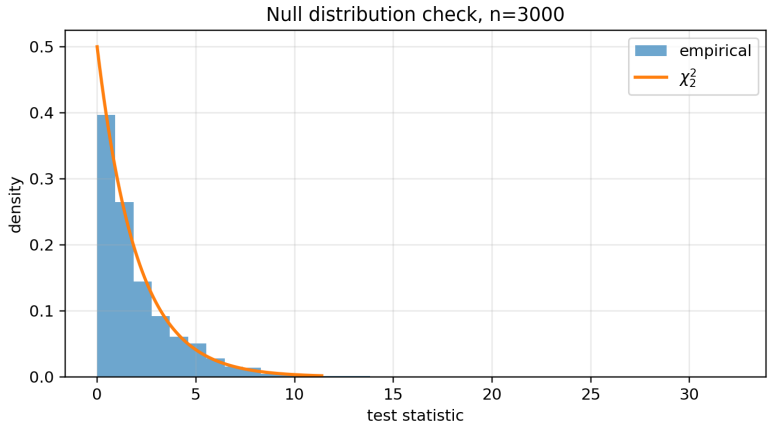


Figure 10: Null distribution of the statistic at the largest sample size. The histogram shows the Monte Carlo distribution of $\widehat{\lambda}^{\mathrm{dr}}_{n,V}$ under $h = 0$, while the curve shows the central χ^2_2 density.

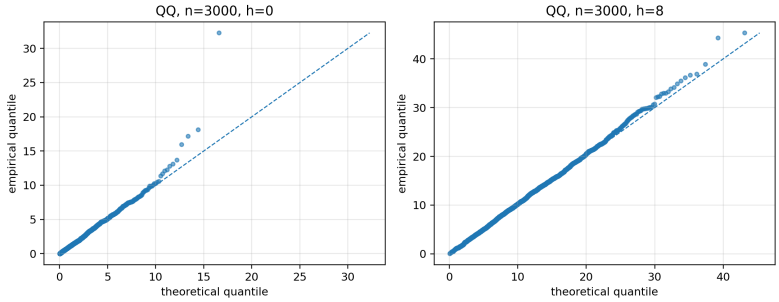


Figure 11: QQ plot at the largest sample size. Empirical quantiles of the statistic are compared with the corresponding noncentral chi-square quantiles.

rejection probability, but is close to the predicted noncentral chi-square distribution over a broad range of quantiles.

Figure 12 compares the empirical first moment of the statistic with the theoretical mean of the limiting noncentral chi-square law:

$$E\left[\chi^2_J(h^2\lambda_V)\right] = J + h^2\lambda_V. \tag{24}$$

The empirical means follow this prediction closely. The mild upward deviation for larger h at smaller n is consistent with the finite-sample behavior already visible in the power curves: for large local strengths and moderate n, the alternatives are no longer perfectly infinitesimal.

Overall, this experiment supports the local asymptotic theory in two ways. First, under $h = 0$, the statistic has the predicted central chi-square behavior. Second, under $n^{-1/2}$ local alternatives, its rejection probability and distributional shape are well described by the noncentral chi-square law with noncentrality parameter $h^2\lambda_V$. This validates the main theoretical interpretation of the statistic and justifies using the whitened quantity λ_V as the target criterion for location learning.

Once V is fixed independently of the test sample, the statistic behaves exactly as the local Gaussian and noncentral chi-square theory predicts.

B.11 Runtime comparison with a global DR kernel test

We finally compare the computational cost of the proposed finite-location test against a global doubly robust kernel treatment-effect test. The purpose of this experiment is not to

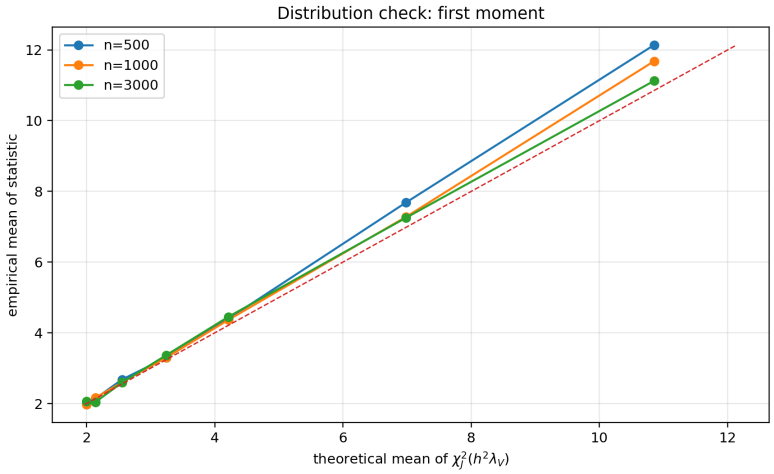


Figure 12: Moment diagnostic. The empirical mean of $\widehat{\lambda}^{\mathrm{dr}}_{n,V}$ is compared with the theoretical mean $J + h^2\widehat{\lambda}_V$ of the noncentral χ^2_J distribution.

study power, but to quantify the practical cost of the two ways of implementing DR-ME and to compare them with DR-xKTE. The comparison is important because DR-ME ultimately tests a low-dimensional statistic indexed by J learned locations, whereas DR-xKTE remains a global kernel statistic over the evaluation sample.

We use the same observational data-generating mechanism as in the previous experiments and the same three-way split

$$\{1, \dots, n\} = I_\eta \cup I_{\mathrm{tr}} \cup I_{\mathrm{te}}.$$

The nuisance split I_η is used to fit the propensity and, for DR-ME, the conditional kernel regressions. The training split I_{tr} is used to learn locations for DR-ME, and the final test is evaluated on I_{te}. This split is kept fixed across implementations so that the measured difference is due to the testing procedure rather than to different data usage.

We compare three methods:

- **DR-ME-Dict**: the finite-dictionary implementation of DR-ME. A dictionary $C = \{c_1, \dots, c_M\}$ of candidate outcome locations is formed, all outcome-kernel evaluations $K_Y(Y, C)$ are precomputed, and locations are selected by greedy maximization of the ridge-stabilized local-power criterion. We use $J = 2$ selected locations and $M = 80$ dictionary candidates.
- **DR-ME-Grad**: the continuous-location implementation of DR-ME. Locations are optimized by gradient ascent in the Euclidean outcome space, using the differentiable version of the same ridge-stabilized criterion. We use three gradient steps in this runtime experiment.
- **DR-xKTE**: a global doubly robust kernel test for the kernel treatment effect. The propensity is fit on I_η and evaluated on I_{te}, as in DR-ME. The statistic is then computed on the final test split. We use a Cholesky-based linear solver for the internal kernel ridge systems, so this is not an intentionally slow implementation.

For each method, we decompose runtime into nuisance fitting, location learning when applicable, and final testing. For DR-ME-Dict, the learning time includes precomputing the dictionary kernel matrices and greedy location search. For DR-ME-Grad, the learning time is the continuous gradient-ascent optimization. For DR-xKTE, there is no location-learning phase; the method-specific cost is the global kernel statistic computed on I_{te}. We report averages over five random seeds for each sample size

$$n \in \{300, 600, 1200, 3000, 5000, 10000\}.$$

Table 14: Total runtime in seconds. Total runtime includes nuisance fitting, location learning when applicable, and final testing. Results are averaged over five random seeds with $J = 2$, $M = 80$, and three gradient steps for DR-ME-Grad.

n	DR-ME-Dict	DR-ME-Grad	DR-xKTE
300	0.0048	0.0019	0.0006
600	0.0059	0.0032	0.0012
1200	0.0132	0.0107	0.0046
3000	0.0223	0.0282	0.0366
5000	0.0395	0.0508	0.1252
10000	0.2080	0.3631	0.9982

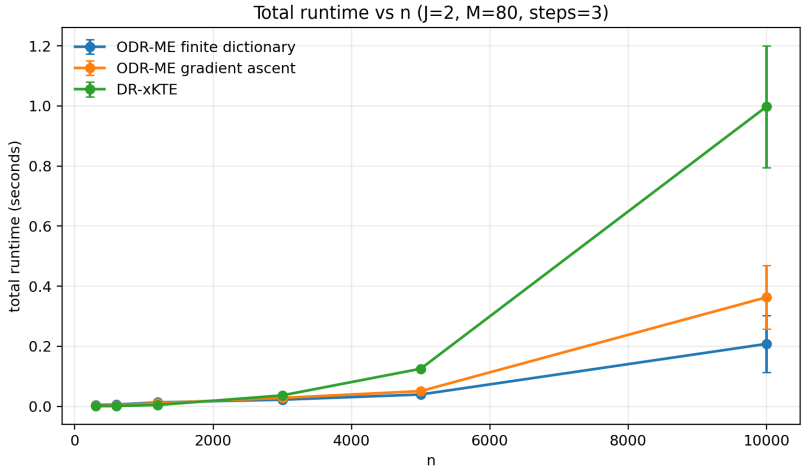


Figure 13: Total runtime versus sample size. DR-xKTE is competitive at small sample sizes, where the global kernel computation is still cheap. At larger sample sizes, the finite-dictionary DR-ME implementation becomes faster. With three gradient steps, continuous DR-ME is slower than the dictionary variant at larger n, while DR-xKTE is the most expensive of the three at $n = 10000$.

Table 14 and Figure 13 show that the relative cost changes with sample size. For $n \leq 1200$, all methods are cheap, and DR-xKTE is often the fastest method. This is expected: at small sample sizes, the overhead of precomputing dictionary features and running location search is not yet offset by DR-ME's low-dimensional final statistic. From $n = 3000$ onward, DR-ME-Dict is faster than DR-xKTE in the reported table; at $n = 10000$, total times are 0.2080 and 0.9982 seconds, respectively.

This scaling is the relevant practical point. DR-xKTE is a global kernel test and therefore continues to manipulate kernel matrices on the evaluation sample. In contrast, after location learning, DR-ME computes a Hotelling statistic in only $J = 2$ dimensions. The finite-dictionary implementation also exploits the fact that all candidate outcome-kernel evaluations can be precomputed once and reused during greedy search.

The core-runtime comparison in Table 15 and Figure 14 gives the cleanest view of method-specific scaling. Nuisance fitting is shared in spirit across the methods and is not the main object of the comparison. The core runtime shows that DR-ME-Dict is not uniformly faster than DR-xKTE at every sample size, but it becomes faster in the large-n regime. This is the regime where using a small number of learned locations matters computationally.

The continuous gradient implementation has a different profile. It avoids a finite dictionary and therefore its cost does not depend on M, but each gradient step requires evaluating the full criterion and its derivative over the training split. With three gradient steps, it is already slower than DR-ME-Dict at $n \geq 3000$. Increasing the number of gradient steps would

Table 15: Core method runtime in seconds. Core runtime removes nuisance fitting and keeps only location learning plus the final test for DR-ME, or the global kernel statistic for DR-xKTE. Results are averaged over five random seeds with $J = 2$, $M = 80$, and three gradient steps for DR-ME-Grad.

n	DR-ME-Dict	DR-ME-Grad	DR-xKTE
300	0.0037	0.0007	0.0004
600	0.0038	0.0011	0.0010
1200	0.0054	0.0030	0.0043
3000	0.0140	0.0200	0.0359
5000	0.0296	0.0409	0.1236
10000	0.1924	0.3475	0.9964

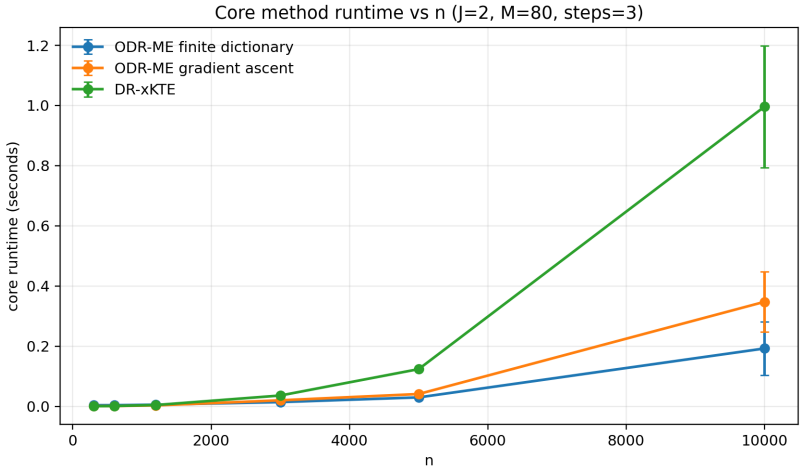


Figure 14: Core method runtime versus sample size. This plot removes nuisance fitting and focuses on the computational cost specific to the test. The finite -dictionary implementation of DR-ME overtakes DR-xKTE at large n, while the gradient implementation becomes substantially slower.

further increase this gap. Therefore, continuous gradient ascent is useful as an alternative for low-dimensional Euclidean outcomes, but it should not be the default scalable implementation in the current empirical suite.

Figure 15 shows the breakdown at the largest sample size. The finite-dictionary implementation is the most attractive option: it is faster than DR-xKTE at $n = 10000$ and still returns learned interpretable locations. DR-xKTE remains a strong global baseline, but its runtime is tied to a global kernel computation over the evaluation sample. DR-ME-Grad is dominated by its repeated gradient evaluations and is therefore less attractive for large n.

Overall, this experiment supports a task-dependent implementation choice. For hypothesis testing, as in our main experiments, DR-ME is best implemented with a finite dictionary and small J. This yields a practical low-dimensional test, scales favorably at larger sample sizes, and preserves interpretability through selected outcome locations. Continuous gradient learning is better viewed as an optional Euclidean-location variant: it is natural when the goal is to optimize interpretable locations in structured Euclidean outcome spaces, such as images, but it is not the primary scalable implementation for testing. DR-xKTE remains the main global-kernel baseline for power comparisons, while the runtime results show that finite-location testing can provide a computational advantage in the larger-n regime.

Table 16: Runtime breakdown in seconds at $n = 10000$, with $J = 2$, $M = 80$, and three gradient steps for DR-ME-Grad. Entries are averages over five random seeds.

Method	Nuisance	Learning	Test	Total
DR-ME-Dict	0.0156	0.1236	0.0688	0.2080
DR-ME-Grad	0.0156	0.3033	0.0441	0.3631
DR-xKTE	0.0017	0.0000	0.9964	0.9982

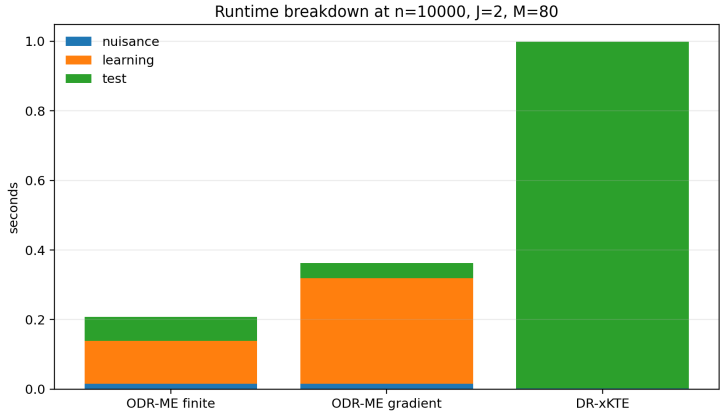


Figure 15: Runtime breakdown at $n = 10000$, $J = 2$, and $M = 80$. The finite-dictionary DR-ME implementation keeps the final test low-dimensional and is faster than DR-xKTE, whose cost is dominated by the global kernel test. Gradient-based DR-ME spends most of its time in location learning.

B.12 OCTMNIST mean-matched image-location experiment

Introductory illustration. Figure 1 uses a single OCTMNIST scan at resolution 224×224: drusen-labelled test image 127 (zero-based indexing). Denote this image by I. It supplies the displayed treated response, retaining the scan's existing contour and texture. We construct a reference image B with smaller retinal elevations by deforming the same scan locally. The pair therefore shares the anatomy of one source image, while its treatment mechanism is introduced by construction.

We estimate the bright retinal band by smoothing I with Gaussian widths $(1.5, 1.2)$ pixels in the row and column directions, locating the brightest pixel in rows 55–154 of each column, and smoothing this trace with width 2. A morphological closing of width 51, followed by smoothing of width 8, gives a smooth reference contour. We retain the positive difference between this contour and the estimated band, smooth it with width 2.5, and scale it by 0.8. A cubic taper reduces the displacement to zero over columns 12–30 and 193–211, fixing the outer columns. Writing r for the band position and d for this displacement in a given column, the inverse deformation maps $(r - 22, r + d, r + 30)$ to $(r - 22, r, r + 30)$ by cubic Hermite interpolation, with derivative one at each knot and the identity map outside this interval. Bilinear sampling gives B. The largest displacement is 6.93 pixels, and the inverse map has positive derivative throughout, so it preserves the ordering of tissue along each column.

With $R = I - B$, $q = 0.2$, and $S \sim \text{Bernoulli}(q)$, set $Y(0) = B + qR$ and $Y(1) = B + SR$. Thus $\mathbb{E}[Y(1) - Y(0)] = 0$, although the image distributions differ. The figure shows the response $S = 1$, for which $Y(1) = I$ and the individual effect is $(1 - q)R$. Both potential outcomes are shown to explain the construction; an observed unit would contribute only $Y = Y(A)$. We crop eight boundary pixels from each side of all panels. The outcome images share the display range $[0, 1]$, without clipping or separate normalization, and the signed effect uses a symmetric range $[-0.6, 0.6]$, with black at zero. This illustration is separate from the experiment below, and the known effect map is not a fitted DR-ME reference.

Table 17: Total runtime in seconds as the number of locations J varies, with $n = 1200$, $M = 80$, and three gradient steps for DR-ME-Grad. Results are averaged over five random seeds.

J	DR-ME-Dict	DR-ME-Grad	DR-xKTE
1	0.0133	0.0118	0.0043
2	0.0158	0.0128	0.0058
5	0.0187	0.0108	0.0044
10	0.0320	0.0128	0.0046

Table 18: Total runtime in seconds as the dictionary size M varies, with $n = 1200$, $J = 2$, and three gradient steps for DR-ME-Grad. Results are averaged over five random seeds.

M	DR-ME-Dict	DR-ME-Grad	DR-xKTE
40	0.0114	0.0109	0.0052
80	0.0139	0.0113	0.0039
160	0.0184	0.0111	0.0043
320	0.0260	0.0106	0.0045

Experimental construction. We use OCTMNIST from MedMNIST v2 [42], [18]. MedMNIST v2 is released under the Creative Commons Attribution 4.0 International license (CC BY 4.0). OCTMNIST is used to construct a semi-synthetic observational experiment with full image outcomes. The goal is to illustrate how an image-space reference is constructed and how its visual appearance relates to the held-out witness estimate. Images serve only as outcomes, while pre-treatment covariates are generated separately. We therefore generate low-dimensional pre-treatment covariates synthetically, use them to induce confounding and heterogeneity, and use OCTMNIST images only on the outcome side.

For each unit, we first sample synthetic covariates

$$X_i \in \mathbb{R}^{d_X}, \qquad X_i \sim N(0, I_{d_X}),$$

with $d_X = 6$ in our implementation. Treatment is assigned observationally by a clipped logistic propensity,

$$A_i \mid X_i \sim \text{Bernoulli}\{\pi_0(1 \mid X_i)\}, \qquad \pi_0(1 \mid X_i) = \text{clip}\big\{\sigma(\alpha^\top X_i),\, 0.07,\, 0.93\big\},$$

so confounding is driven by the synthetic covariates rather than by the OCT images themselves.

The outcome construction begins with baseline images

$$B_i \in [0,1]^{28\times 28}$$

sampled from the *normal* OCTMNIST class. These images play the role of untreated anatomy. To define a clinically interpretable treatment pattern, we construct a fixed fluid-like template

$$M \in [0,1]^{28\times 28}$$

from OCTMNIST using the positive part of the smoothed mean difference between the DME and normal classes, together with a mild central spatial prior. The DME images are therefore used only to define this template; the causal sample itself is built from normal baseline images.

Potential outcomes are generated by adding treatment-specific residual images to the same baseline anatomy:

$$Y_i(a) = B_i + R_i(a), \qquad a \in \{0,1\}.$$

The residuals are designed so that the two treatment arms have the same average residual signal, but different residual distributions. Let

$$q_i = q_{\min} + (q_{\max} - q_{\min})\,\sigma(\beta^\top X_i),$$

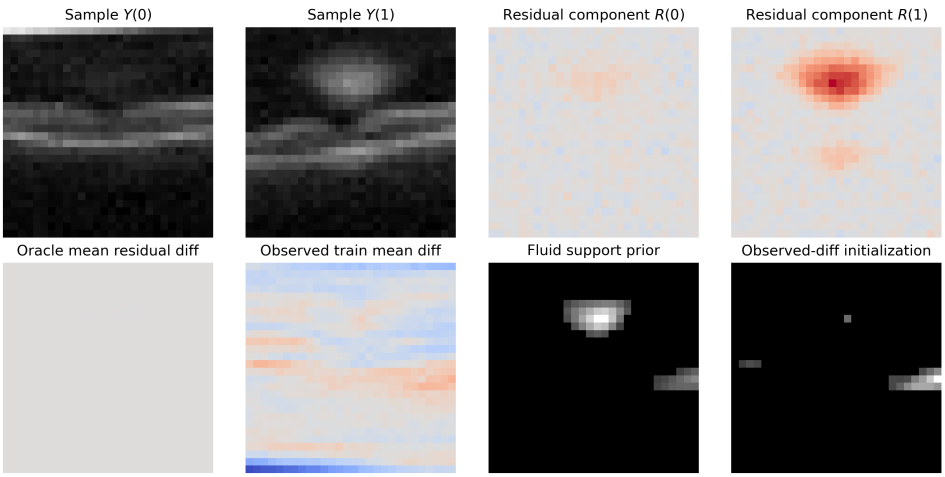


Figure 16: Additional diagnostics for the OCTMNIST mean-matched image-location experiment. Top row: sampled control and treated images $Y(0)$, $Y(1)$, together with their residual components $R(0)$, $R(1)$. Bottom row: oracle mean residual difference, observed treated-versus-control mean image difference on the training split, the fluid support prior, and the observed-difference initialization. The observed-difference image is one candidate initialization in the search. The selected reference in the saved run instead comes from a projected training image; it is reported together with its held-out estimate below. After sample-level recentering, the oracle mean residual difference is zero up to numerical precision. This removes that mean contrast but does not itself establish evidence at the selected location.

where $q_i \in (q_{\min}, q_{\max})$ is a covariate-dependent probability of a severe fluid event. We then sample

$$S_i \mid X_i \sim \text{Bernoulli}(q_i),$$

and define mean-one multiplicative amplitude jitters

$$J_{ia} = \exp(\tau U_{ia} - \tau^2/2), \qquad U_{ia} \sim N(0,1).$$

The two residual laws are

$$R_i(0) = q_i \theta J_{i0} M + \varepsilon_{i0}, \qquad R_i(1) = S_i \theta J_{i1} M + \varepsilon_{i1},$$

with independent Gaussian pixel noise

$$\varepsilon_{i0}, \varepsilon_{i1} \sim N(0, \sigma^2 I).$$

Since $\mathbb{E}[J_{ia}] = 1$ and $\mathbb{E}[S_i \mid X_i] = q_i$, the construction satisfies

$$\mathbb{E}\{R_i(1) - R_i(0) \mid X_i\} = 0.$$

Thus the treatment effect is distributional rather than mean-based: the control arm carries a mild diffuse fluid pattern, while the treated arm exhibits rare but much more severe localized deviations.

In finite samples, the equality above holds exactly only in expectation. To make the oracle mean residual difference visually and numerically null in the realized sample, we additionally recenter the sampled residuals. Writing

$$\bar{\Delta}_R := \frac{1}{n} \sum_{i=1}^{n} \{R_i(1) - R_i(0)\},$$

we replace

$$R_i(0) \leftarrow R_i(0) + \frac{1}{2}\bar{\Delta}_R, \qquad R_i(1) \leftarrow R_i(1) - \frac{1}{2}\bar{\Delta}_R.$$

After this recentering, the empirical oracle residual mean difference is zero up to numerical precision, while the residual distributions remain different. Because the same sample-wide

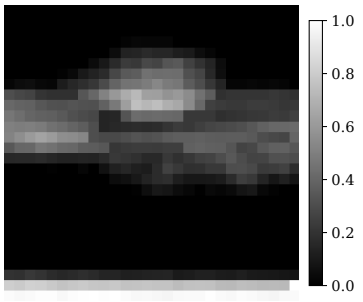


Figure 17: Unthresholded reference used for kernel evaluation in the OCTMNIST run. The main-text image includes additional display processing.

correction enters multiple observations, this recentering introduces dependence across units and across the subsequently formed splits. The example is therefore a descriptive construction; the i.i.d. calibration theorem does not directly justify its nominal chi-square p-value. It should not be treated as a calibration experiment.

The observed outcome is

$$Y_i = A_i Y_i(1) + (1 - A_i) Y_i(0).$$

For inference we work directly with the full image outcomes, not with residuals. We use a Gaussian image kernel

$$k_Y(v, y) = \exp\left\{ -\frac{\|v - y\|_2^2}{2 d_Y \ell_Y^2} \right\}, \qquad d_Y = 28 \cdot 28,$$

where the factor d_Y makes ℓ_Y interpretable as a per-pixel scale. The propensity is estimated on I_η by logistic regression, and the nuisance regressions

$$m_a(x; v) = \mathbb{E}\{k_Y(v, Y) \mid A = a, X = x\}$$

are estimated on I_η by ridge regression over a fixed nonlinear feature basis of X.

We learn a single image location v on the training split I_{tr}, so $J = 1$. In this scalar case, the DR-ME location-learning criterion reduces to the squared standardized doubly robust moment,

$$v \mapsto \hat{\mathcal{P}}_{\tau,\mathrm{tr}}(v) = \frac{\bar{z}_{\mathrm{tr}}^{\mathrm{dr}}(v)^2}{\widehat{\mathrm{Var}}_{\mathrm{tr}}\{z^{\mathrm{dr}}(v)\} + \tau}.$$

The final one-location DR-ME statistic is then computed only on the held-out test split I_{te}.

Because v lives in a continuous 28×28 image space, the search uses several initializations and explicit image constraints. The candidate starts include the observed training mean difference, a fluid-support template, the training mean image, and a dictionary of training images. Candidates are ranked by the training objective, and a subset is refined by gradient steps. Projection uses smoothing, nonnegativity, support weighting, soft thresholding, normalization, and clipping to $[0, 1]$, with light anchoring toward the initialization. The support prior is derived partly from the same fluid template used to construct the synthetic effect, so overlap between the displayed reference and that template cannot be attributed solely to the learning criterion.

The saved run contains $n = 1500$ observations, split into 525 nuisance, 600 learning, and 375 test observations. Its construction uses $d_X = 6$, $\theta = 0.55$, pixel-noise standard deviation 0.02, amplitude-jitter scale 0.20, and $(q_{\min}, q_{\max}) = (0.06, 0.24)$. The image-kernel scale is $\ell_Y = 0.22$, and both learning and test ridges are 10^{-6}. The recorded winning start is `train_image_503`; its objective equals the best recorded initialization objective, 0.0042942, so the saved result does not establish an improvement from gradient refinement over that starting reference. The figure's observed-difference initialization is shown as a diagnostic, not as the winning start.

The raw selected array is used for kernel evaluation. Display processing applies additional support weighting and thresholding; Figure 17 shows the unthresholded array for comparison. The nominal test output is

$$\bar{z}^{\mathrm{dr}}_{\mathrm{te}} = 0.0074543, \quad \widehat{\mathrm{se}} = \frac{0.4385637}{\sqrt{375}} = 0.0226473,$$
$$\hat{\lambda}^{\mathrm{dr}}_{\mathrm{te}} = 0.1083374, \quad p = 0.7420449.$$

The selected reference therefore yields no rejection at level 0.05, even under the nominal reference law. Its image structure should be interpreted alongside this weak held-out signal and the sample-level dependence described above.

The example illustrates the objects returned by DR-ME, while also showing why visual concentration alone is insufficient evidence of a distributional effect. A stronger imaging evaluation would separate the support prior from the planted effect and compare prespecified search classes under an i.i.d. construction. Finite dictionaries, sparse or wavelet-based parameterizations, and shape-constrained search are possible ways to use image structure while retaining a final testing sample separate from all fitting and selection.

C Notation and full assumptions

Observed data and potential outcomes. We observe $Z = (X, A, Y) \sim P_0$, with binary treatment $A \in \{0, 1\}$. The potential outcomes are $Y(0), Y(1) \in \mathcal{Y}$. Unless a split-sample construction is explicitly used, $Z_1, \ldots, Z_n \overset{\text{i.i.d.}}{\sim} P_0$.

Kernel and finite-location notation. The outcome kernel $k_Y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ has RKHS $\mathcal{H}_Y$ and feature map $\varphi_Y(y) = k_Y(\cdot, y)$. Throughout, k_Y is measurable, positive definite, and bounded: $\sup_y k_Y(y, y) \le \kappa^2 < \infty$. For $V = (v_1, \ldots, v_J) \in \mathcal{Y}^J$, write

$$k_V(y) := k_Y(V, y) := (k_Y(v_1, y), \ldots, k_Y(v_J, y))^\top \in \mathbb{R}^J.$$

Interventional embeddings and witness coordinates. For $a \in \{0, 1\}$, let $\chi(a) := \mathbb{E}[\varphi_Y(Y(a))] \in \mathcal{H}_Y$, $\Delta := \chi(1) - \chi(0)$, and

$$w(y) := \langle \Delta, \varphi_Y(y) \rangle_{\mathcal{H}_Y} = \mathbb{E}[k_Y(y, Y(1))] - \mathbb{E}[k_Y(y, Y(0))].$$

The finite-location arm means and contrast are

$$\mu_{a,V} := \mathbb{E}[k_V(Y(a))], \qquad \mu_V := \mu_{1,V} - \mu_{0,V} = (w(v_1), \ldots, w(v_J))^\top \in \mathbb{R}^J.$$

Observed-data nuisance functions. For $a \in \{0, 1\}$, define

$$\pi_0(a \mid x) := P_0(A = a \mid X = x), \qquad m_a(x; V) := \mathbb{E}[k_V(Y) \mid A = a, X = x].$$

A generic nuisance tuple is $\eta = (\pi, r_0, r_1)$, and the true tuple is $\eta_0 = (\pi_0, m_0, m_1)$.

Orthogonal pseudo-features. For $a \in \{0, 1\}$, define

$$\phi^a_V(Z; \eta) := \frac{\mathbf{1}\{A = a\}}{\pi(a \mid X)} \{k_V(Y) - r_a(X; V)\} + r_a(X; V), \qquad z^{\mathrm{dr}}_V(Z; \eta) := \phi^1_V(Z; \eta) - \phi^0_V(Z; \eta).$$

At the truth,

$$\psi_V(Z) := z^{\mathrm{dr}}_V(Z; \eta_0) - \mu_V, \qquad \Sigma_V := \mathrm{Var}_{P_0}\{\psi_V(Z)\}.$$

Proposition 2.1 shows that ψ_V is the canonical gradient of μ_V in the observed-data model.

Fixed-location statistic. For fixed V, write the feasible pseudo-feature as

$$\hat{z}^{\mathrm{dr}}_{i,V} := z^{\mathrm{dr}}_V(Z_i; \hat{\eta}_{-i}),$$

where $\hat{\eta}_{-i}$ is trained on data independent of Z_i. This covers both K-fold cross-fitting and the independent test split used after location learning. Define

$$\bar{z}^{\mathrm{dr}}_{n,V} := \frac{1}{n} \sum_{i=1}^n \hat{z}^{\mathrm{dr}}_{i,V}, \qquad S^{\mathrm{dr}}_{n,V} := \frac{1}{n-1} \sum_{i=1}^n (\hat{z}^{\mathrm{dr}}_{i,V} - \bar{z}^{\mathrm{dr}}_{n,V})(\hat{z}^{\mathrm{dr}}_{i,V} - \bar{z}^{\mathrm{dr}}_{n,V})^\top.$$

The fixed-location Hotelling statistic is

$$\hat{\lambda}^{\mathrm{dr}}_{n,V} := n\, \bar{z}^{\mathrm{dr}\top}_{n,V} (S^{\mathrm{dr}}_{n,V} + \gamma_n I_J)^{-1} \bar{z}^{\mathrm{dr}}_{n,V}, \qquad \gamma_n \downarrow 0.$$

Local asymptotic notation. Let $\mathcal{P}$ be a semiparametric model for the observed-data law. For a fixed V, let $P_0 \in \mathcal{P}$ satisfy $\mu_V(P_0) = 0$, and let $T_{P_0} \subset L_2^0(P_0)$ be the tangent space. For a score direction $g \in T_{P_0}$, let $t \mapsto P_{t,g}$ be a regular quadratic-mean differentiable path through P_0 with score g, and define $P_{n,h,g} := P_{h/\sqrt{n},g}^{\otimes n}$. The local drift and whitened local signal are

$$\eta_V(g) := \left.\frac{d}{dt}\right|_{t=0} \mu_V(P_{t,g}) = \mathbb{E}_{P_0}\{\psi_V(Z)g(Z)\}, \qquad \lambda_V(g) := \eta_V(g)^\top \Sigma_V^{-1} \eta_V(g).$$

Location learning notation. The full sample is split as $\{1, \dots, n\} = I_\eta \cup I_{\mathrm{tr}} \cup I_{\mathrm{te}}$, with $n_\eta = |I_\eta|$, $n_{\mathrm{tr}} = |I_{\mathrm{tr}}|$, and $n_{\mathrm{te}} = |I_{\mathrm{te}}|$. Nuisances are fitted on I_η, locations are learned on I_{tr}, and the final test is computed on I_{te}. The population and empirical ridge-stabilized criteria are

$$\mathcal{P}_\tau(V) := \mu_V^\top (\Sigma_V + \tau I_J)^{-1} \mu_V, \qquad \hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V) := \bar{z}_{\mathrm{tr},V}^{\mathrm{dr}\top} (S_{\mathrm{tr},V}^{\mathrm{dr}} + \tau I_J)^{-1} \bar{z}_{\mathrm{tr},V}^{\mathrm{dr}}.$$

For a learned location set $\hat{V}$, define its empirical optimization gap

$$\varepsilon_{\mathrm{tr}}(\hat{V}) := \sup_{V \in \mathcal{V}} \hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V) - \hat{\mathcal{P}}_{\tau,\mathrm{tr}}(\hat{V}),$$

and the uniform learning error

$$\Delta_{\mathrm{tr}} := \sup_{V \in \mathcal{V}} \left| \hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V) - \mathcal{P}_\tau(V) \right|.$$

Norms. Unless stated otherwise, $\|\cdot\|$ is the Euclidean norm, $\|\cdot\|_F$ is the Frobenius norm, and $\|\cdot\|_{L_2(P_X;\mathbb{R}^J)}$ is the $L_2(P_X)$ norm for $\mathbb{R}^J$-valued functions.

C.1 Observed-data identification

Assumption C.1 (Observed-data identification)**.** For each $a \in \{0, 1\}$, the following hold.

(i) *Consistency*: $Y = Y(a)$ whenever $A = a$.

(ii) *Conditional exchangeability*: $Y(a) \perp A \mid X$.

(iii) *Positivity*: there exists $\varepsilon > 0$ such that $\pi_0(a \mid X) \geq \varepsilon$ almost surely.

These are the identification conditions in Assumption A1. Under Assumption C.1, $\mu_{a,V} = \mathbb{E}[m_a(X; V)]$ and hence $\mu_V = \mathbb{E}[m_1(X; V) - m_0(X; V)]$.

C.2 Fixed-location nuisance and covariance conditions

The following conditions give the detailed form of Assumption A2 used in Theorem 2.2. They are stated for K-fold cross-fitting; for an independent nuisance split, remove the maxima over folds and use the single nuisance estimate $\hat{\eta}$.

Assumption C.2 (Fixed-location first-stage conditions)**.** Fix $V \in \mathcal{Y}^J$. The validation folds $\mathcal{I}_1, \dots, \mathcal{I}_K$ satisfy $\min_k |\mathcal{I}_k|/n \geq c_K > 0$, with fixed K. For $i \in \mathcal{I}_k$, the nuisance estimate $\hat{\eta}_k = (\hat{\pi}_k, \hat{m}_{0,k}, \hat{m}_{1,k})$ is trained outside fold k.

There exist constants $\underline{\varepsilon} > 0$ and $M < \infty$ such that, with probability tending to one,

$$\inf_{k,a,x} \hat{\pi}_k(a \mid x) \geq \underline{\varepsilon}, \qquad \sup_{k,a,x} \|\hat{m}_{a,k}(x; V)\| \leq M.$$

Moreover,

$$\max_k \sum_{a=0}^{1} \left\{ \|\hat{\pi}_k(a \mid \cdot) - \pi_0(a \mid \cdot)\|_{L_2(P_X)} + \|\hat{m}_{a,k}(\cdot; V) - m_a(\cdot; V)\|_{L_2(P_X;\mathbb{R}^J)} \right\} = o_p(1),$$

and

$$\max_k \sum_{a=0}^{1} \|\hat{\pi}_k(a \mid \cdot) - \pi_0(a \mid \cdot)\|_{L_2(P_X)} \, \|\hat{m}_{a,k}(\cdot; V) - m_a(\cdot; V)\|_{L_2(P_X;\mathbb{R}^J)} = o_p(n^{-1/2}).$$

Assumption C.3 (Nondegenerate fixed-location covariance)**.** For the fixed location set V, the efficient covariance $\Sigma_V = \mathrm{Var}_{P_0}\{\psi_V(Z)\}$ is positive definite.

Assumption C.3 is only needed for inverse-covariance whitening and chi-square limits. If Σ_V is singular, one may work in the nonzero eigenspace, but we do not pursue that extension.

C.3 Regular local paths

The local testing results are formulated along regular quadratic-mean differentiable paths. This is the primitive local smoothness condition; the LAN expansion used in the proofs follows from it.

Assumption C.4 (Regular QMD local path)**.** For each fixed $g \in T_{P_0}$, there exists a path $t \mapsto P_{t,g} \subset \mathcal{P}$ through P_0 with densities $p_{t,g}$ relative to a dominating measure ν, such that

$$\int \left(\sqrt{p_{t,g}} - \sqrt{p_0} - \frac{t}{2} g \sqrt{p_0} \right)^2 d\nu = o(t^2).$$

Under Assumption C.4, the standard LAN expansion holds for $P_{n,h,g} = P_{h/\sqrt{n},g}^{\otimes n}$:

$$\log \frac{dP_{n,h,g}}{dP_0^{\otimes n}} = h\,\mathbb{G}_n g - \frac{1}{2} h^2 \|g\|_{L_2(P_0)}^2 + o_{P_0}(1), \qquad \mathbb{G}_n g := \frac{1}{\sqrt{n}} \sum_{i=1}^n \{g(Z_i) - \mathbb{E}_{P_0} g(Z)\}.$$

It also implies contiguity of $P_{n,h,g}$ with respect to $P_0^{\otimes n}$. Thus $o_{P_0}(1)$ remainders in the fixed-location first-order expansion transfer to $o_{P_{n,h,g}}(1)$ remainders along the local path.

For Proposition 3.2, we additionally assume that $P \mapsto \Sigma_V(P)$ is continuous along $t \mapsto P_{t,g}$ at $t = 0$, and that $\Sigma_V(P_{t,g})$ remains nonsingular for t in a neighborhood of zero.

C.4 Euclidean location-learning conditions

For Theorem 4.1, Assumption A3 requires $\mathcal{Y} \subset \mathbb{R}^d$ and a compact search class $\mathcal{V} \subset [-R, R]^{Jd}$. Let

$$\Pi(\mathcal{V}) := \{v \in \mathbb{R}^d : \ v = v_j \text{ for some } V = (v_1, \ldots, v_J) \in \mathcal{V}\}.$$

Assumption C.5 (Euclidean location regularity)**.** There exist constants $B_k, L_k, B_m, L_m < \infty$ such that the following hold.

(i) For all $v, v' \in \Pi(\mathcal{V})$ and $y \in \mathcal{Y}$,

$$|k_Y(v, y)| \le B_k, \qquad |k_Y(v, y) - k_Y(v', y)| \le L_k \|v - v'\|.$$

(ii) With probability tending to one, for all $a \in \{0, 1\}$, $x \in \mathcal{X}$, and $v, v' \in \Pi(\mathcal{V})$,

$$|\hat{m}_a(x; v)| \le B_m, \qquad |\hat{m}_a(x; v) - \hat{m}_a(x; v')| \le L_m \|v - v'\|.$$

(iii) The estimated propensity is uniformly bounded away from zero with probability tending to one:

$$\inf_{a,x} \hat{\pi}(a \mid x) \ge \underline{\varepsilon}$$

for some $\underline{\varepsilon} > 0$.

The true regressions $m_a(\cdot; v)$ inherit the boundedness and Lipschitz properties from the kernel by Jensen's inequality, so no separate smoothness assumption on m_a is needed for the learning theorem. The nuisance error entering Theorem 4.1 is

$$\rho_{n_\eta} := \sum_{a=0}^{1} \left[\|\hat{\pi}(a \mid \cdot) - \pi_0(a \mid \cdot)\|_{L_2(P_X)} + \sup_{v \in \Pi(\mathcal{V})} \|\hat{m}_a(\cdot; v) - m_a(\cdot; v)\|_{L_2(P_X)} \right].$$

The learning theorem only requires $\rho_{n_\eta} = o_p(1)$, because it controls uniform consistency of the learning criterion rather than a root-n expansion of the final test statistic.

D Background on local asymptotic testing for fixed locations

This appendix reviews the local asymptotic tools used in Sections 2 and 3. Once the location set V is fixed, the target is the finite-dimensional parameter $\mu_V(P) \in \mathbb{R}^J$. Proposition 2.1 identifies its observed-data canonical gradient ψ_V, and Theorem 3.1 studies the local testing problem generated by this gradient.

The relevant objects are the local drift $\eta_V(g)$, the efficient covariance Σ_V, and the whitened local signal

$$\lambda_V(g) := \eta_V(g)^\top \Sigma_V^{-1} \eta_V(g).$$

The role of this appendix is to explain why these three quantities govern both regular estimation of μ_V and first-order local power of tests at the selected locations.

D.1 QMD paths and the local Gaussian approximation

The local theory is pathwise. We do not assume a finite-dimensional parametric model for the whole data-generating law. Instead, for each score direction $g \in T_{P_0}$, Assumption C.4 postulates a regular quadratic-mean differentiable path $t \mapsto P_{t,g}$ through P_0. Quadratic mean differentiability means that, for densities $p_{t,g}$ relative to a dominating measure ν,

$$\int \left(\sqrt{p_{t,g}} - \sqrt{p_0} - \frac{t}{2} g \sqrt{p_0} \right)^2 d\nu = o(t^2).$$

This condition is the standard local smoothness assumption behind Le Cam's asymptotic theory [2, 20, 39].

For the contiguous alternatives $P_{n,h,g} := P_{h/\sqrt{n},g}^{\otimes n}$, QMD implies the LAN expansion [39, Theorem 7.2]

$$\log \frac{dP_{n,h,g}}{dP_0^{\otimes n}} = h\,\mathbb{G}_n g - \frac{1}{2} h^2 \|g\|_{L_2(P_0)}^2 + o_{P_0}(1), \qquad \mathbb{G}_n g := \frac{1}{\sqrt{n}} \sum_{i=1}^n g(Z_i),$$

where $g \in L_2^0(P_0)$. Thus, along each regular path, the original experiment is locally approximated by a Gaussian shift experiment. This is the only LAN input used in the paper. No least favorable submodel needs to be constructed for the main results; least favorable paths are useful for interpretation, but the proofs use arbitrary regular QMD paths and tangent-space projection.

D.2 Le Cam's third lemma and the local drift

Le Cam's third lemma describes how an asymptotically linear statistic shifts under a contiguous alternative. If

$$T_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi(Z_i) + o_{P_0}(1), \qquad \mathbb{E}_{P_0} \phi(Z) = 0,$$

then, under $P_{n,h,g}$,

$$T_n \xrightarrow{d} N(h\,\mathbb{E}_{P_0}\{\phi(Z)g(Z)\}, \mathrm{Var}_{P_0}\{\phi(Z)\}).$$

The covariance is unchanged to first order; the mean shifts in the direction correlated with the score g.

For the fixed-location statistic, $\phi = \psi_V$. Since ψ_V is the canonical gradient of μ_V, the pathwise derivative of μ_V along g is

$$\eta_V(g) := \left. \frac{d}{dt} \right|_{t=0} \mu_V(P_{t,g}) = \mathbb{E}_{P_0}\{\psi_V(Z) g(Z)\}.$$

Thus $\eta_V(g)$ is not an auxiliary definition. It is exactly the mean shift of the canonical-gradient statistic. If $\eta_V(g) = 0$, the selected locations are locally blind to direction g; if $\eta_V(g) \neq 0$, the local alternative produces a first-order shift in the witness coordinates.

D.3 Regularity, convolution, and the efficient finite-signal experiment

In the nonparametric observed-data model, the canonical gradient of the finite signal μ_V is unique. Thus the efficiency statement is not a nontrivial comparison among different regular influence functions for the same target. Rather, it is a statement about the efficient limit experiment for regular estimation of μ_V.

By the Hájek–Le Cam convolution theorem [39, Theorem 25.20], if T_n is any regular estimator of μ_V and

$$\sqrt{n}\{T_n - \mu_V(P_0)\} \xrightarrow{d} L \quad \text{under } P_0^{\otimes n},$$

then

$$L = N(0, \Sigma_V) * M$$

for some probability law M. The canonical-gradient estimator $\bar{z}^{\mathrm{dr}}_{n,V}$ attains the no-extra-noise case $M = \delta_0$, since Theorem 2.2 gives

$$\sqrt{n}(\bar{z}^{\mathrm{dr}}_{n,V} - \mu_V) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_V(Z_i) + o_p(1).$$

Hence $N(0, \Sigma_V)$ is the efficient Gaussian limit for the finite witness. The covariance Σ_V is therefore not one possible normalization among many; it is the covariance of the efficient finite-signal experiment.

D.4 From efficient estimation to local testing geometry

Under Theorem 2.2 and the local path assumptions,

$$\sqrt{n}\,\bar{z}^{\mathrm{dr}}_{n,V} \xrightarrow{d} N(h\eta_V(g), \Sigma_V) \qquad \text{under } P_{n,h,g}.$$

After whitening, the efficient finite-signal experiment is

$$\Sigma_V^{-1/2}\sqrt{n}\,\bar{z}^{\mathrm{dr}}_{n,V} \xrightarrow{d} N(h\Sigma_V^{-1/2}\eta_V(g), I_J).$$

For any scalar contrast $a^\top \mu_V$, the efficient squared local signal-to-noise ratio along g is

$$\frac{(a^\top \eta_V(g))^2}{a^\top \Sigma_V a}.$$

If the local direction g were known, optimizing this quantity over $a \neq 0$ gives the Rayleigh quotient

$$\max_{a \neq 0} \frac{(a^\top \eta_V(g))^2}{a^\top \Sigma_V a} = \eta_V(g)^\top \Sigma_V^{-1} \eta_V(g) = \lambda_V(g),$$

with optimizer proportional to $\Sigma_V^{-1}\eta_V(g)$. This is the directional Neyman–Pearson benchmark in the finite-signal Gaussian experiment.

When the direction is unknown, no uniformly most powerful test exists over all drift directions in the multivariate Gaussian shift experiment. The natural omnibus reduction is the whitened problem $G \sim N(m, I_J)$, with $H_0 : m = 0$ and $H_1 : m \neq 0$. In this limit experiment, the likelihood-ratio, Wald, and score statistics all reduce to $\|G\|^2$. Transported back to the finite witness, this is the Hotelling statistic based on Σ_V^{-1}, with noncentrality $h^2\lambda_V(g)$. This is why the population location-learning criterion is

$$\Lambda_V(P) := \mu_V(P)^\top \Sigma_V(P)^{-1} \mu_V(P),$$

and why the practical criterion uses $\mathcal{P}_\tau(V) = \mu_V^\top(\Sigma_V + \tau I_J)^{-1}\mu_V$.

D.5 Marginal and conditional distributional targets

The finite witness μ_V compares the marginal interventional laws, integrating over the covariate distribution P_X. This differs from a conditional distributional effect at $X = x$. Park et al. [28] study conditional distributional effects using kernel conditional mean embeddings, conditional

witness functions, and U-statistic regression. Näf et al. [25] estimate conditional kernel treatment effects using distributional random forests and provide asymptotic inference for that target. These methods address how the outcome law changes for a given covariate profile. DR-ME instead asks whether treatment changes a finite set of expected kernel similarities in the target population. Averaging a conditional estimator over X changes the inferential target and requires a corresponding analysis of its estimation error.

For function-valued risk minimizers, Hudson et al. [14] test risk derivatives over restricted direction classes. Global MMD tests for unknown functions use higher-order pathwise differentiability [22], while Hilbert-valued one-step theory gives efficient inference for RKHS-valued parameters [21]. DR-ME uses the first-order canonical gradient of a finite projection of the interventional embedding difference. The resulting covariance governs both inference on these coordinates and the criterion used to select them.

D.6 Relation to existing semiparametric kernel inference

This paper uses kernel embeddings, but the local problem is different from global MMD testing with unknown nuisance functions. In the unknown-function MMD setting of Luedtke et al. [22], the squared MMD has a degenerate first-order derivative under the null, and valid testing requires a second-order U-statistic analysis. Here, for fixed V, the target $\mu_V \in \mathbb{R}^J$ is first-order pathwise differentiable. The null limit is therefore a finite-dimensional chi-square law rather than a degenerate infinite weighted chi-square limit.

The paper is also related to Hilbert-valued semiparametric one-step estimation. General Hilbert-valued theory shows that, when an efficient influence function exists, one-step estimators can achieve root-n Hilbert-norm inference; interventional kernel mean embeddings are a key example [21]. Our construction can be viewed as applying bounded linear evaluation maps to the interventional RKHS discrepancy, yielding a finite-dimensional parameter with canonical gradient ψ_V. The contribution here is not Hilbert-valued estimation per se, but the testing geometry induced by this finite projection and its use for interpretable location learning.

D.7 Why the orthogonal DR geometry is the relevant geometry

The regression contrast $m_1(X;V) - m_0(X;V)$ has mean μ_V. If m_0 and m_1 were known, its covariance could be smaller than Σ_V. That comparison, however, corresponds to an oracle problem in which nuisance regressions are given. It is not the observed-data semiparametric problem faced by a regular procedure using flexible nuisance estimates.

A feasible plug-in regression contrast is first-order sensitive to regression error. Without stronger nuisance conditions, it does not provide the stable local expansion needed for regular local testing. By contrast, the orthogonal pseudo-feature $z_V^{\mathrm{dr}}(Z;\eta)$ is built from the canonical gradient. Theorem 2.2 gives

$$\sqrt{n}(\bar{z}_{n,V}^{\mathrm{dr}} - \mu_V) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_V(Z_i) + o_p(1), \qquad S_{n,V}^{\mathrm{dr}} \to_p \Sigma_V.$$

Thus the fixed-location test enters the efficient finite-signal Gaussian experiment with drift $\eta_V(g)$ and covariance Σ_V. This is the covariance geometry used both for local testing and for location learning.

E Proofs for Section 2

Throughout this appendix, fix $V = (v_1, \ldots, v_J) \in \mathcal{Y}^J$ and write

$$k_V(y) := k_Y(V, y) = \big(k_Y(v_1, y), \ldots, k_Y(v_J, y)\big)^\top \in \mathbb{R}^J.$$

We use $\|\cdot\|$ for the Euclidean norm on $\mathbb{R}^J$. Since k_Y is bounded and J is fixed,

$$B_V := \sup_{y \in \mathcal{Y}} \|k_V(y)\| < \infty.$$

Consequently, $\|m_a(x;V)\| \le B_V$ for $a \in \{0,1\}$ and all x. We prove the cross-fitted case. The independent split case is the same argument with a single validation split independent of the nuisance-fitting sample.

E.1 Identification

Proof of the identification claim. Fix $a \in \{0,1\}$. By iterated expectation,

$$\mu_{a,V} = \mathbb{E}\big[k_V(Y(a))\big] = \mathbb{E}[\mathbb{E}\{k_V(Y(a)) \mid X\}]\,.$$

Conditional exchangeability gives

$$\mathbb{E}\{k_V(Y(a)) \mid X\} = \mathbb{E}\{k_V(Y(a)) \mid A = a, X\},$$

and consistency gives $Y = Y(a)$ on $\{A = a\}$. Hence

$$\mathbb{E}\{k_V(Y(a)) \mid A = a, X\} = \mathbb{E}\{k_V(Y) \mid A = a, X\} = m_a(X;V).$$

Therefore $\mu_{a,V} = \mathbb{E}[m_a(X;V)]$, and subtracting the two arm-specific identities yields

$$\mu_V = \mu_{1,V} - \mu_{0,V} = \mathbb{E}\{m_1(X;V) - m_0(X;V)\}.$$

□

E.2 A basic doubly robust identity

Lemma E.1 (Doubly robust algebra)**.** *Fix $a \in \{0,1\}$, let $\eta = (\pi, r_0, r_1)$, and define*

$$D_a(Z;\eta) := \phi_V^a(Z;\eta) - \phi_V^a(Z;\eta_0), \qquad \eta_0 = (\pi_0, m_0, m_1).$$

Then

$$\begin{aligned} D_a(Z;\eta) = &\left(1 - \frac{\mathbf{1}\{A=a\}}{\pi_0(a \mid X)}\right)\{r_a(X;V) - m_a(X;V)\} \\ &+ \mathbf{1}\{A=a\}\left\{\frac{1}{\pi(a \mid X)} - \frac{1}{\pi_0(a \mid X)}\right\}\{k_V(Y) - r_a(X;V)\}, \end{aligned} \tag{25}$$

and

$$P_0 D_a(\cdot;\eta) = P_0\left[\frac{\pi_0(a \mid X) - \pi(a \mid X)}{\pi(a \mid X)}\{m_a(X;V) - r_a(X;V)\}\right]. \tag{26}$$

Moreover, on any event on which $\pi(a \mid X) \ge \underline{\pi} > 0$ almost surely and $\sup_x \|r_a(x;V)\| \le M$,

$$P_0\|D_a(\cdot;\eta)\|^2 \le C\left\{\|r_a(\cdot;V) - m_a(\cdot;V)\|^2_{L_2(P_X;\mathbb{R}^J)} + \|\pi(a \mid \cdot) - \pi_0(a \mid \cdot)\|^2_{L_2(P_X)}\right\}, \tag{27}$$

where $C < \infty$ depends only on the true positivity constant, $\underline{\pi}$, B_V, and M.

Proof. Write $I_a := \mathbf{1}\{A = a\}$, $\pi_a(X) := \pi(a \mid X)$, and $\pi_{0,a}(X) := \pi_0(a \mid X)$. Then

$$\begin{aligned} D_a(Z;\eta) &= \frac{I_a}{\pi_a(X)}\{k_V(Y) - r_a(X;V)\} + r_a(X;V) \\ &\quad - \frac{I_a}{\pi_{0,a}(X)}\{k_V(Y) - m_a(X;V)\} - m_a(X;V) \\ &= I_a\left\{\frac{1}{\pi_a(X)} - \frac{1}{\pi_{0,a}(X)}\right\}\{k_V(Y) - r_a(X;V)\} \\ &\quad + \left(1 - \frac{I_a}{\pi_{0,a}(X)}\right)\{r_a(X;V) - m_a(X;V)\}, \end{aligned}$$

which proves (25).

Taking conditional expectation given X, the second term in (25) has mean zero. Also,

$$\mathbb{E}[I_a\{k_V(Y) - r_a(X;V)\} \mid X] = \pi_{0,a}(X)\{m_a(X;V) - r_a(X;V)\}.$$

Thus

$$\mathbb{E}\{D_a(Z;\eta) \mid X\} = \frac{\pi_{0,a}(X) - \pi_a(X)}{\pi_a(X)}\{m_a(X;V) - r_a(X;V)\},$$

and (26) follows.

For the L_2 bound, use $(u+v)^2 \le 2u^2 + 2v^2$ in (25). True positivity implies

$$\mathbb{E}\left[\left|1 - \frac{I_a}{\pi_{0,a}(X)}\right|^2 \middle| X\right] \le C,$$

so the first squared term is bounded by

$$C\,\|r_a(\cdot;V) - m_a(\cdot;V)\|^2_{L_2(P_X;\mathbb{R}^J)}.$$

On the stated event,

$$\left|\frac{1}{\pi_a(X)} - \frac{1}{\pi_{0,a}(X)}\right| \le C|\pi_a(X) - \pi_{0,a}(X)|, \qquad \|k_V(Y) - r_a(X;V)\| \le B_V + M.$$

Hence the second squared term is bounded by

$$C\,\|\pi(a \mid \cdot) - \pi_0(a \mid \cdot)\|^2_{L_2(P_X)}.$$

Combining the two bounds proves (27). □

E.3 Proof of Proposition 2.1

Proof of Proposition 2.1. The bias identity is exactly (26). Since

$$\mathbb{E}\{\phi^a_V(Z;\eta)\} - \mu_{a,V} = P_0 D_a(\cdot;\eta),$$

we obtain

$$\mathbb{E}\{\phi^a_V(Z;\eta)\} - \mu_{a,V} = \mathbb{E}\left[\frac{\pi_0(a \mid X) - \pi(a \mid X)}{\pi(a \mid X)}\{m_a(X;V) - r_a(X;V)\}\right].$$

Therefore $\mathbb{E}\{\phi^a_V(Z;\eta)\} = \mu_{a,V}$ if either $\pi(a \mid \cdot) = \pi_0(a \mid \cdot)$ or $r_a(\cdot;V) = m_a(\cdot;V)$. Subtracting the two arm-specific identities gives the corresponding double-robustness statement for z^{dr}_V.

It remains to identify the canonical gradient. We prove the arm-specific statement; the contrast follows by linearity. Consider a regular path $t \mapsto P_t$ through P_0 with score $s \in T_{P_0}$. Write

$$m_{a,t}(x;V) := \mathbb{E}_t[k_V(Y) \mid A = a, X = x], \qquad \mu_{a,V}(P_t) := \mathbb{E}_t[m_{a,t}(X;V)].$$

Let $s_X(X) := \mathbb{E}[s(Z) \mid X]$. Standard conditional-score calculus gives

$$\left.\frac{d}{dt}\right|_{t=0} \mu_{a,V}(P_t) = \mathbb{E}[m_a(X;V)s_X(X)] + \mathbb{E}[\mathbb{E}\{(k_V(Y) - m_a(X;V))s(Z) \mid A = a, X\}].$$

The first term can be written as

$$\mathbb{E}[m_a(X;V)s_X(X)] = \mathbb{E}[\{m_a(X;V) - \mu_{a,V}\}s(Z)],$$

because $\mathbb{E}[s(Z)] = 0$. The second term can be represented in observed-data form as

$$\mathbb{E}\left[\frac{\mathbf{1}\{A = a\}}{\pi_0(a \mid X)}\{k_V(Y) - m_a(X;V)\}s(Z)\right].$$

Hence

$$\left.\frac{d}{dt}\right|_{t=0} \mu_{a,V}(P_t) = \mathbb{E}\left[\left\{\frac{\mathbf{1}\{A = a\}}{\pi_0(a \mid X)}\{k_V(Y) - m_a(X;V)\} + m_a(X;V) - \mu_{a,V}\right\}s(Z)\right].$$

Thus the arm-specific canonical gradient is

$$\phi^a_V(Z;\eta_0) - \mu_{a,V}.$$

Since $\mu_V = \mu_{1,V} - \mu_{0,V}$, the canonical gradient of μ_V is

$$\{\phi^1_V(Z;\eta_0) - \mu_{1,V}\} - \{\phi^0_V(Z;\eta_0) - \mu_{0,V}\} = z^{\mathrm{dr}}_V(Z;\eta_0) - \mu_V = \psi_V(Z).$$

□

E.4 Proof of Theorem 2.2

Proof of Theorem 2.2. Let $\mathcal{I}_1, \dots, \mathcal{I}_K$ be the validation folds, let $n_k := |\mathcal{I}_k|$, and let $P_{n,k}$ denote the empirical measure on fold k. For $i \in \mathcal{I}_k$, write

$$\hat{z}_i := z_V^{\mathrm{dr}}(Z_i; \hat{\eta}_k), \qquad z_0(Z_i) := z_V^{\mathrm{dr}}(Z_i; \eta_0), \qquad \psi_V(Z_i) = z_0(Z_i) - \mu_V.$$

Because K is fixed, all $o_p(\cdot)$ statements below may be read uniformly over k.

Let $\mathcal{E}_n$ be the event on which the fitted propensities are uniformly bounded away from zero and the fitted regressions are uniformly bounded, as required by Assumption C.2. Then $P_0(\mathcal{E}_n) \to 1$. On this event, define

$$d_{a,k}(Z) := \phi_V^a(Z; \hat{\eta}_k) - \phi_V^a(Z; \eta_0), \qquad d_k(Z) := d_{1,k}(Z) - d_{0,k}(Z),$$

and

$$\alpha_{a,k} := \|\hat{\pi}_k(a \mid \cdot) - \pi_0(a \mid \cdot)\|_{L_2(P_X)}, \qquad \beta_{a,k} := \|\hat{m}_{a,k}(\cdot; V) - m_a(\cdot; V)\|_{L_2(P_X; \mathbb{R}^J)}.$$

By Lemma E.1 and Cauchy–Schwarz,

$$\|P_0 d_k\| \le C \sum_{a=0}^{1} \alpha_{a,k} \beta_{a,k}, \tag{28}$$

$$P_0 \|d_k\|^2 \le C \sum_{a=0}^{1} (\alpha_{a,k}^2 + \beta_{a,k}^2). \tag{29}$$

The assumed nuisance consistency and product-rate conditions therefore imply, uniformly over k,

$$\|P_0 d_k\| = o_p(n^{-1/2}), \qquad P_0 \|d_k\|^2 = o_p(1).$$

First-order expansion. Since

$$\bar{z}_{n,V}^{\mathrm{dr}} = \sum_{k=1}^{K} \frac{n_k}{n} P_{n,k}(z_0 + d_k),$$

we have

$$\sqrt{n}(\bar{z}_{n,V}^{\mathrm{dr}} - \mu_V) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_V(Z_i) + R_{1n} + R_{2n},$$

where

$$R_{1n} := \sqrt{n} \sum_{k=1}^{K} \frac{n_k}{n} (P_{n,k} - P_0) d_k, \qquad R_{2n} := \sqrt{n} \sum_{k=1}^{K} \frac{n_k}{n} P_0 d_k.$$

For R_{2n}, (28) and the product-rate condition give

$$\|R_{2n}\| \le C\sqrt{n} \sum_{k=1}^{K} \frac{n_k}{n} \sum_{a=0}^{1} \alpha_{a,k} \beta_{a,k} = o_p(1).$$

For R_{1n}, condition on the training data used to construct $\hat{\eta}_k$. Then d_k is fixed and independent of the validation observations in fold k, so

$$\mathbb{E}_0 \left[\left\| \sqrt{n} \, \frac{n_k}{n} (P_{n,k} - P_0) d_k \right\|^2 \middle| \hat{\eta}_k \right] \le \frac{n_k}{n} P_0 \|d_k\|^2 \le P_0 \|d_k\|^2 = o_p(1).$$

A conditional Markov inequality gives each fold contribution as $o_p(1)$. Since K is fixed, $R_{1n} = o_p(1)$. Hence

$$\sqrt{n}(\bar{z}_{n,V}^{\mathrm{dr}} - \mu_V) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_V(Z_i) + o_p(1).$$

Covariance consistency. Let

$$M_n := \frac{1}{n}\sum_{i=1}^{n} \hat{z}_i \hat{z}_i^\top, \qquad M_0 := P_0[z_0 z_0^\top].$$

The ordinary law of large numbers gives

$$\frac{1}{n}\sum_{i=1}^{n} z_0(Z_i) z_0(Z_i)^\top \to_p M_0,$$

because z_0 is bounded under the bounded-kernel and positivity assumptions. It remains to show that replacing $z_0(Z_i)$ by $\hat{z}_i$ is negligible. Define

$$A_n := \frac{1}{n}\sum_{i=1}^{n} \|\hat{z}_i - z_0(Z_i)\|^2 = \sum_{k=1}^{K} \frac{n_k}{n} P_{n,k}\|d_k\|^2.$$

Conditionally on the nuisance fits,

$$\mathbb{E}_0[A_n \mid \hat{\eta}_1, \ldots, \hat{\eta}_K] = \sum_{k=1}^{K} \frac{n_k}{n} P_0\|d_k\|^2 = o_p(1),$$

so $A_n = o_p(1)$ by conditional Markov. Therefore,

$$\begin{aligned}\left\| M_n - \frac{1}{n}\sum_{i=1}^{n} z_0(Z_i) z_0(Z_i)^\top \right\|_F &\le A_n + 2A_n^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\|z_0(Z_i)\|^2\right)^{1/2} \\ &= o_p(1).\end{aligned}$$

Thus $M_n \to_p M_0$. Since the first-order expansion implies $\bar{z}^{\mathrm{dr}}_{n,V} \to_p \mu_V$,

$$S^{\mathrm{dr}}_{n,V} = \frac{n}{n-1}\left\{M_n - \bar{z}^{\mathrm{dr}}_{n,V}\bar{z}^{\mathrm{dr}\top}_{n,V}\right\} \to_p M_0 - \mu_V\mu_V^\top = \mathrm{Var}_{P_0}\{z_V^{\mathrm{dr}}(Z;\eta_0)\} = \Sigma_V.$$

If Σ_V is positive definite and $\gamma_n \downarrow 0$, then

$$S^{\mathrm{dr}}_{n,V} + \gamma_n I_J \to_p \Sigma_V,$$

and continuity of matrix inversion at positive definite matrices gives

$$(S^{\mathrm{dr}}_{n,V} + \gamma_n I_J)^{-1} \to_p \Sigma_V^{-1}.$$

□

F Proofs for Section 3

Throughout this appendix, $V \in \mathcal{Y}^J$ is fixed and $P_0 \in \mathcal{P}$ satisfies

$$\mu_V(P_0) = 0.$$

Let $t \mapsto P_{t,g}$ be a quadratic-mean differentiable regular path through P_0 with score $g \in T_{P_0}$, and let

$$P_{n,h,g} := P^{\otimes n}_{h/\sqrt{n},g}.$$

By the standard QMD-to-LAN implication for i.i.d. experiments [39, Theorem 7.2],

$$\ell_{n,h,g} := \log \frac{dP_{n,h,g}}{dP_0^{\otimes n}} = h\,\mathbb{G}_n g - \frac{1}{2}h^2\|g\|^2_{L_2(P_0)} + o_{P_0}(1), \tag{30}$$

where

$$\mathbb{G}_n g := \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(Z_i),$$

and $P_{n,h,g}$ is contiguous with respect to $P_0^{\otimes n}$. Since $g \in T_{P_0} \subset L_2^0(P_0)$, $\mathbb{E}_{P_0}\{g(Z)\} = 0$.

F.1 A contiguity transfer lemma

Lemma F.1 (Contiguity transfer)**.** *Let Q_n be contiguous with respect to $P_0^{\otimes n}$. If $R_n = o_{P_0}(1)$, then $R_n = o_{Q_n}(1)$. If $A_n \to_{P_0} A$, then $A_n \to_{Q_n} A$. In particular, if A is positive definite and $\gamma_n \downarrow 0$, then*

$$(A_n + \gamma_n I)^{-1} \to_{Q_n} A^{-1}.$$

Proof. For any $\epsilon > 0$,

$$P_0^{\otimes n}(\|R_n\| > \epsilon) \to 0.$$

Contiguity implies

$$Q_n(\|R_n\| > \epsilon) \to 0,$$

so $R_n = o_{Q_n}(1)$. The convergence $A_n \to A$ transfers by applying the same argument to the events

$$\{\|A_n - A\| > \epsilon\}.$$

Since $\gamma_n \downarrow 0$, $A_n + \gamma_n I \to_{Q_n} A$. The inverse statement follows from continuity of matrix inversion at positive definite matrices. □

F.2 Proof of Theorem 3.1

Proof of Theorem 3.1. Define

$$W_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_V(Z_i).$$

Since $\mu_V(P_0) = 0$, the first-order representation in Theorem 2.2 gives, under $P_0^{\otimes n}$,

$$\sqrt{n}\, \bar{z}_{n,V}^{\mathrm{dr}} = W_n + r_n, \qquad r_n = o_{P_0}(1), \tag{31}$$

and

$$S_{n,V}^{\mathrm{dr}} \to_{P_0} \Sigma_V.$$

By QMD, $P_{n,h,g}$ is contiguous with respect to $P_0^{\otimes n}$. Lemma F.1 therefore implies

$$r_n = o_{P_{n,h,g}}(1), \qquad S_{n,V}^{\mathrm{dr}} \to_{P_{n,h,g}} \Sigma_V, \tag{32}$$

and, since $\Sigma_V \succ 0$ and $\gamma_n \downarrow 0$,

$$(S_{n,V}^{\mathrm{dr}} + \gamma_n I_J)^{-1} \to_{P_{n,h,g}} \Sigma_V^{-1}.$$

We next determine the local law of W_n. Under $P_0^{\otimes n}$, the joint vector

$$(W_n, \mathbb{G}_n g) = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_V(Z_i), \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(Z_i) \right)$$

converges by the multivariate central limit theorem to a centered Gaussian vector (W, G) with covariance

$$\begin{pmatrix} \Sigma_V & \eta_V(g) \\ \eta_V(g)^\top & \|g\|^2_{L_2(P_0)} \end{pmatrix}.$$

Indeed,

$$\mathrm{Cov}_{P_0}\{\psi_V(Z), g(Z)\} = \mathbb{E}_{P_0}\{\psi_V(Z) g(Z)\} = \eta_V(g).$$

Combining this joint central limit theorem with the LAN expansion (30), Le Cam's third lemma gives, under $P_{n,h,g}$,

$$W_n \xrightarrow{d} N(h\eta_V(g), \Sigma_V).$$

Together with (32), this yields

$$\sqrt{n}\, \bar{z}_{n,V}^{\mathrm{dr}} \xrightarrow{d} N(h\eta_V(g), \Sigma_V) \qquad \text{under } P_{n,h,g}.$$

For the quadratic statistic, write

$$\hat{\lambda}_{n,V}^{\mathrm{dr}} = (\sqrt{n}\, \bar{z}_{n,V}^{\mathrm{dr}})^\top (S_{n,V}^{\mathrm{dr}} + \gamma_n I_J)^{-1} (\sqrt{n}\, \bar{z}_{n,V}^{\mathrm{dr}}).$$

By Slutsky's theorem,

$$\hat{\lambda}^{\mathrm{dr}}_{n,V} \xrightarrow{d} Y^\top \Sigma_V^{-1} Y, \qquad Y \sim N(h\eta_V(g), \Sigma_V).$$

Equivalently, if $Z \sim N(0, I_J)$, then

$$Y = \Sigma_V^{1/2} Z + h\eta_V(g),$$

and therefore

$$Y^\top \Sigma_V^{-1} Y = \left\| Z + h\Sigma_V^{-1/2}\eta_V(g) \right\|^2.$$

Thus

$$\hat{\lambda}^{\mathrm{dr}}_{n,V} \xrightarrow{d} \chi^2_J\big(h^2 \eta_V(g)^\top \Sigma_V^{-1} \eta_V(g)\big) = \chi^2_J(h^2 \lambda_V(g)).$$

The whitened representation follows from the same argument:

$$\Sigma_V^{-1/2}\sqrt{n}\,\bar{z}^{\mathrm{dr}}_{n,V} \xrightarrow{d} N(h\Sigma_V^{-1/2}\eta_V(g), I_J),$$

and

$$\hat{\lambda}^{\mathrm{dr}}_{n,V} = \left\| \Sigma_V^{-1/2}\sqrt{n}\,\bar{z}^{\mathrm{dr}}_{n,V} \right\|^2 + o_{P_{n,h,g}}(1).$$

□

F.3 Proof of Proposition 3.2

Proof of Proposition 3.2. Let $t_n := h/\sqrt{n}$. Since μ_V is pathwise differentiable at P_0 along $t \mapsto P_{t,g}$, with derivative $\eta_V(g)$, and since $\mu_V(P_0) = 0$,

$$\mu_V(P_{t,g}) = t\,\eta_V(g) + o(t) \qquad \text{as } t \to 0.$$

Substituting $t = t_n$ gives

$$\mu_V(P_{h/\sqrt{n},g}) = \frac{h}{\sqrt{n}}\eta_V(g) + o(n^{-1/2}).$$

Let

$$\Sigma_n := \Sigma_V(P_{t_n,g}).$$

By continuity along the path and nonsingularity in a neighborhood of P_0,

$$\Sigma_n \to \Sigma_V(P_0) = \Sigma_V, \qquad \Sigma_n^{-1} \to \Sigma_V^{-1}.$$

Set

$$u_n := \sqrt{n}\,\mu_V(P_{t_n,g}).$$

Then

$$u_n = h\eta_V(g) + o(1).$$

Therefore

$$n\,\Lambda_V(P_{t_n,g}) = u_n^\top \Sigma_n^{-1} u_n \to h^2\,\eta_V(g)^\top \Sigma_V^{-1} \eta_V(g) = h^2 \lambda_V(g).$$

□

F.4 Proof of Corollary 3.3

Proof of Corollary 3.3. Under the global null $H_0 : \Delta = 0$, every finite-location signal is zero:

$$\mu_V(P_0) = 0 \qquad \text{for all } V \in \mathcal{Y}^J.$$

In particular, conditional on the learning split,

$$\mu_{\hat{V}}(P_0) = 0.$$

Conditional on the learning split, the selected $\hat{V}$ is fixed relative to the final testing data. The final statistic is therefore a fixed-location statistic at $V = \hat{V}$, computed with $\gamma_{n_{\mathrm{te}}} \downarrow 0$. By the assumed conditional fixed-location first-order representation and covariance consistency, together with the multivariate central limit theorem and Slutsky's theorem,

$$\hat{\lambda}^{\mathrm{dr}}_{n_{\mathrm{te}},\hat{V}} \xrightarrow{d} \chi^2_J \qquad \text{conditionally on the learning split.}$$

Thus the conditional rejection probability converges to α at the chi-square critical value. Since the rejection indicator is bounded and the limiting rejection probability is nonrandom, the unconditional rejection probability also converges to α. □

G Additional local-theory refinements

Throughout this appendix, fix $V \in \mathcal{Y}^J$ and let $P_0 \in \mathcal{P}$ satisfy

$$\mu_V(P_0) = 0.$$

Write T_{P_0} for the tangent space at P_0. For $g \in T_{P_0}$, let

$$P_{n,h,g} := P_{h/\sqrt{n},g}^{\otimes n}$$

denote the corresponding contiguous local alternatives. Recall that

$$\psi_V(Z) = z_V^{\mathrm{dr}}(Z;\eta_0) - \mu_V$$

is the canonical gradient of the fixed-location signal μ_V, and define

$$\Sigma_V := \mathrm{Var}_{P_0}\{\psi_V(Z)\}, \qquad \eta_V(g) := \mathbb{E}_{P_0}\{\psi_V(Z)g(Z)\}, \qquad \lambda_V(g) := \eta_V(g)^\top \Sigma_V^{-1}\eta_V(g).$$

The main text identifies the efficient finite-signal Gaussian experiment and the omnibus Hotelling statistic. This appendix records the corresponding known-direction benchmark, which explains the role of $\lambda_V(g)$ as a directional local signal-to-noise ratio.

Remark G.1 (Directional and omnibus local benchmarks)**.** Within the finite-signal Gaussian experiment, if the local direction g were known, the Neyman–Pearson linear test projects along $\Sigma_V^{-1}\eta_V(g)$ and attains the directional benchmark above [26, 39]. When g is unknown, no uniformly most powerful test exists over all drift directions in the multivariate Gaussian shift experiment. The Hotelling statistic is instead the canonical omnibus quadratic test in the whitened efficient Gaussian experiment. We therefore claim efficiency of the finite-signal experiment and its covariance geometry, not uniformly most powerful testing over all distributional alternatives.

G.1 Known-direction score benchmark

Theorem 3.1 treats the omnibus alternative $\mu_V \neq 0$. If the local direction g were known, the finite-signal Gaussian experiment also gives a one-dimensional Neyman–Pearson benchmark.

Corollary G.2 (Directional efficient score test)**.** *Assume the conditions of Theorem 3.1, and suppose that $\eta_V(g) \neq 0$. Define*

$$L_{n,V,g} := \frac{\eta_V(g)^\top \big(S_{n,V}^{\mathrm{dr}} + \gamma_n I_J\big)^{-1} \sqrt{n}\, \bar{z}_{n,V}^{\mathrm{dr}}}{\Big[\eta_V(g)^\top \big(S_{n,V}^{\mathrm{dr}} + \gamma_n I_J\big)^{-1} \eta_V(g)\Big]^{1/2}}.$$

Then, under $P_0^{\otimes n}$,

$$L_{n,V,g} \xrightarrow{d} N(0,1),$$

and, under $P_{n,h,g}$,

$$L_{n,V,g} \xrightarrow{d} N\big(h\sqrt{\lambda_V(g)}, 1\big).$$

Consequently, the one-sided level-α rule that rejects for

$$L_{n,V,g} > z_{1-\alpha}$$

has asymptotic power

$$1 - \Phi\big(z_{1-\alpha} - h\sqrt{\lambda_V(g)}\big).$$

Moreover, in the finite-signal Gaussian experiment

$$Y \sim N(h\eta_V(g), \Sigma_V), \qquad h \in \mathbb{R},$$

this directional rule is Neyman–Pearson optimal for testing $h = 0$ against any fixed simple alternative $h = h_1 > 0$.

Proof. Let

$$A_n := \big(S^{\mathrm{dr}}_{n,V} + \gamma_n I_J\big)^{-1}, \qquad W_n := \sqrt{n}\,\bar{z}^{\mathrm{dr}}_{n,V}.$$

By Theorem 3.1,

$$W_n \xrightarrow{d} N\big(h\eta_V(g), \Sigma_V\big) \qquad \text{under } P_{n,h,g},$$

and

$$A_n \to_p \Sigma_V^{-1}.$$

Therefore, by Slutsky's theorem,

$$\eta_V(g)^\top A_n W_n = \eta_V(g)^\top \Sigma_V^{-1} W_n + o_{P_{n,h,g}}(1).$$

Hence

$$\eta_V(g)^\top A_n W_n \xrightarrow{d} N\big(h\,\eta_V(g)^\top \Sigma_V^{-1}\eta_V(g),\, \eta_V(g)^\top \Sigma_V^{-1}\Sigma_V\Sigma_V^{-1}\eta_V(g)\big)\,,$$

or equivalently,

$$\eta_V(g)^\top A_n W_n \xrightarrow{d} N\big(h\lambda_V(g), \lambda_V(g)\big).$$

Also,

$$\eta_V(g)^\top A_n \eta_V(g) \to_p \eta_V(g)^\top \Sigma_V^{-1}\eta_V(g) = \lambda_V(g).$$

Dividing numerator and denominator yields

$$L_{n,V,g} \xrightarrow{d} N\big(h\sqrt{\lambda_V(g)}, 1\big).$$

The null statement is obtained by setting $h = 0$, and the displayed power formula follows immediately.

For the Neyman–Pearson statement, in the finite-signal Gaussian experiment

$$Y \sim N(h\eta_V(g), \Sigma_V),$$

the log-likelihood ratio between $h = h_1$ and $h = 0$ is, up to an additive constant,

$$h_1\,\eta_V(g)^\top \Sigma_V^{-1} Y - \frac{1}{2}h_1^2\,\eta_V(g)^\top \Sigma_V^{-1}\eta_V(g).$$

Thus the most powerful level-α test rejects for large values of

$$\eta_V(g)^\top \Sigma_V^{-1} Y,$$

or equivalently for large values of its standardized version. This is the limiting test generated by $L_{n,V,g}$. □

G.2 Interpretation

For any scalar contrast $a^\top \mu_V$, the squared local signal-to-noise ratio along g is

$$\frac{(a^\top \eta_V(g))^2}{a^\top \Sigma_V a}.$$

The known-direction benchmark optimizes this quantity over $a \neq 0$, yielding the Rayleigh quotient

$$\lambda_V(g) = \eta_V(g)^\top \Sigma_V^{-1}\eta_V(g).$$

When g is unknown, no uniformly most powerful test exists over all drift directions in the multivariate Gaussian shift experiment. The fixed-location DR-ME statistic instead uses the omnibus quadratic statistic in the whitened efficient Gaussian experiment, whose local power is governed by the same noncentrality $h^2\lambda_V(g)$. This is why the population criterion

$$\Lambda_V(P) = \mu_V(P)^\top \Sigma_V(P)^{-1}\mu_V(P)$$

is the relevant target for learning informative locations.

H Proofs for Section 4

Throughout this appendix, write $n := n_{\mathrm{tr}}$ and

$$D_\eta := \sigma\big((Z_i)_{i\in I_\eta}, \hat\pi, \hat m_0, \hat m_1\big).$$

For $V \in \mathcal{V}$ and $z = (x, a, y) \in \mathcal{Z}$, define the nuisance-fitted pseudo-feature

$$\hat z_V(z) := z_V^{\mathrm{dr}}(z; \hat\eta), \qquad \hat\eta = (\hat\pi, \hat m_0, \hat m_1).$$

Thus, for $i \in I_{\mathrm{tr}}$, $\hat z_{i,V}^{\mathrm{dr}} = \hat z_V(Z_i)$. Conditional on D_η, the variables $\{\hat z_V(Z_i) : i \in I_{\mathrm{tr}}\}$ are i.i.d. for each fixed V.

Define the conditional population quantities

$$\mu_V^\eta := \mathbb{E}[\hat z_V(Z) \mid D_\eta], \qquad M_V^\eta := \mathbb{E}[\hat z_V(Z)\hat z_V(Z)^\top \mid D_\eta], \qquad \Sigma_V^\eta := M_V^\eta - \mu_V^\eta \mu_V^{\eta\top},$$

and the conditional criterion

$$\mathcal{P}_\tau^\eta(V) := \mu_V^{\eta\top}(\Sigma_V^\eta + \tau I_J)^{-1}\mu_V^\eta.$$

The proof of Theorem 4.1 uses

$$\sup_{V\in\mathcal{V}} \big|\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V) - \mathcal{P}_\tau(V)\big| \le \sup_{V\in\mathcal{V}} \big|\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V) - \mathcal{P}_\tau^\eta(V)\big| + \sup_{V\in\mathcal{V}} \big|\mathcal{P}_\tau^\eta(V) - \mathcal{P}_\tau(V)\big|. \tag{33}$$

The first term is the training-split empirical fluctuation. The second term is the nuisance-induced discrepancy between the conditional and oracle population criteria.

H.1 Uniform bounds for pseudo-features

Let

$$\bar\varepsilon := \min\{\varepsilon, \underline{\varepsilon}\},$$

where ε is the true positivity constant and $\underline{\varepsilon}$ is the fitted-propensity lower bound in Assumption C.5.

Lemma H.1 (Uniform boundedness and Lipschitzness)**.** *Under Assumption C.5, there exist finite constants*

$$B_z := 2\sqrt{J}\Big[\bar\varepsilon^{-1}(B_k + B_m) + B_m\Big], \qquad L_z := 2\Big[\bar\varepsilon^{-1}(L_k + L_m) + L_m\Big],$$

such that, with probability tending to one, for all $z \in \mathcal{Z}$ and all $V, V' \in \mathcal{V}$,

$$\|\hat z_V(z)\| \le B_z, \qquad \|\hat z_V(z) - \hat z_{V'}(z)\| \le L_z\|V - V'\|.$$

The same bounds hold for the oracle feature $z_V^0(z) := z_V^{\mathrm{dr}}(z; \eta_0)$.

Proof. For $V = (v_1, \dots, v_J)$,

$$\|k_V(y)\|^2 = \sum_{j=1}^J |k_Y(v_j, y)|^2 \le JB_k^2, \qquad \|\hat m_a(x; V)\|^2 \le JB_m^2.$$

Hence the arm-specific fitted feature is bounded by

$$\left\| \frac{\mathbf{1}\{A = a\}}{\hat\pi(a \mid X)}\{k_V(Y) - \hat m_a(X; V)\} + \hat m_a(X; V) \right\| \le \sqrt{J}\Big[\bar\varepsilon^{-1}(B_k + B_m) + B_m\Big].$$

Subtracting the two arm-specific features gives the bound for $\hat z_V$.

For Lipschitzness, Assumption C.5 gives

$$\|k_V(y) - k_{V'}(y)\| \le L_k\|V - V'\|, \qquad \|\hat m_a(x; V) - \hat m_a(x; V')\| \le L_m\|V - V'\|.$$

Therefore each arm-specific fitted feature is Lipschitz with constant

$$\bar\varepsilon^{-1}L_k + (\bar\varepsilon^{-1} + 1)L_m \le \bar\varepsilon^{-1}(L_k + L_m) + L_m.$$

Subtracting the two arms gives the displayed L_z. The oracle feature satisfies the same bounds because the true regressions inherit boundedness and Lipschitzness from k_Y, and the true propensities are bounded below by ε. □

H.2 Uniform empirical control

Lemma H.2 (Uniform concentration for bounded Lipschitz Euclidean classes)**.** *Let $\mathcal{V} \subset [-R, R]^m$, and let*

$$\mathcal{F} := \{f_V : V \in \mathcal{V}\}$$

be a class of measurable real-valued functions such that, conditionally on D_η,

$$|f_V(z)| \le B, \qquad |f_V(z) - f_{V'}(z)| \le L\|V - V'\|$$

for all z and all V, V'. Then

$$\sup_{V \in \mathcal{V}} \left|(P_n^{\mathrm{tr}} - P_0) f_V\right| = O_p\left(\sqrt{\frac{m \log n}{n}}\right),$$

where $P_n^{\mathrm{tr}} := n^{-1} \sum_{i \in I_{\mathrm{tr}}} \delta_{Z_i}$, and the hidden constant depends only on R, B, L, m.

Proof. Let $\mathcal{N}_\delta$ be a δ-net of $\mathcal{V}$. Since $\mathcal{V} \subset [-R, R]^m$,

$$|\mathcal{N}_\delta| \le \left(\frac{C_0 R}{\delta}\right)^m$$

for a universal constant C_0. For each V, choose $\Pi_\delta(V) \in \mathcal{N}_\delta$ with $\|V - \Pi_\delta(V)\| \le \delta$. Then

$$\sup_{V \in \mathcal{V}} |(P_n^{\mathrm{tr}} - P_0) f_V| \le \max_{U \in \mathcal{N}_\delta} |(P_n^{\mathrm{tr}} - P_0) f_U| + 2L\delta.$$

Conditionally on D_η, Hoeffding's inequality and a union bound give

$$\Pr\left(\max_{U \in \mathcal{N}_\delta} |(P_n^{\mathrm{tr}} - P_0) f_U| > t \,\middle|\, D_\eta\right) \le 2|\mathcal{N}_\delta| \exp\left(-\frac{n t^2}{2B^2}\right).$$

Taking $\delta = n^{-1/2}$ and $t = M\sqrt{m \log n / n}$, with M large enough, proves the result. □

Lemma H.3 (Uniform mean and covariance concentration)**.** *Under Assumption C.5,*

$$\sup_{V \in \mathcal{V}} \|\bar{z}_{\mathrm{tr},V}^{\mathrm{dr}} - \mu_V^\eta\| = O_p\left(\sqrt{\frac{Jd \log n}{n}}\right),$$

and, with

$$\widehat{M}_{\mathrm{tr},V} := \frac{1}{n} \sum_{i \in I_{\mathrm{tr}}} \hat{z}_{i,V}^{\mathrm{dr}} \hat{z}_{i,V}^{\mathrm{dr}\top},$$

$$\sup_{V \in \mathcal{V}} \|\widehat{M}_{\mathrm{tr},V} - M_V^\eta\|_F = O_p\left(\sqrt{\frac{Jd \log n}{n}}\right).$$

Consequently,

$$\sup_{V \in \mathcal{V}} \|S_{\mathrm{tr},V}^{\mathrm{dr}} - \Sigma_V^\eta\|_F = O_p\left(\sqrt{\frac{Jd \log n}{n}}\right).$$

Proof. Set $m = Jd$. By Lemma H.1, each coordinate class

$$\{z \mapsto \hat{z}_{V,j}(z) : V \in \mathcal{V}\}$$

is conditionally bounded and Lipschitz. Applying Lemma H.2 coordinatewise and using fixed J yields the mean bound.

For second moments, apply the same lemma to

$$\{z \mapsto \hat{z}_{V,j}(z)\hat{z}_{V,\ell}(z) : V \in \mathcal{V}\}.$$

This class is conditionally bounded by B_z^2 and Lipschitz with constant $2B_z L_z$. Applying the scalar concentration bound entrywise and using fixed J gives the second-moment bound.

Finally,

$$S_{\mathrm{tr},V}^{\mathrm{dr}} = \frac{n}{n-1}\left(\widehat{M}_{\mathrm{tr},V} - \bar{z}_{\mathrm{tr},V}^{\mathrm{dr}} \bar{z}_{\mathrm{tr},V}^{\mathrm{dr}\top}\right), \qquad \Sigma_V^\eta = M_V^\eta - \mu_V^\eta \mu_V^{\eta\top}.$$

The uniform boundedness of $\bar{z}_{\mathrm{tr},V}^{\mathrm{dr}}$ and μ_V^η, together with

$$\|uu^\top - vv^\top\|_F \le (\|u\| + \|v\|)\|u - v\|,$$

gives the covariance bound. □

H.3 Nuisance transfer

Lemma H.4 (Uniform nuisance transfer)**.** *Under Assumption C.5,*

$$\sup_{V\in\mathcal{V}} \|\mu_V^\eta - \mu_V\| = O_p(\rho_{n_\eta}), \qquad \sup_{V\in\mathcal{V}} \|\Sigma_V^\eta - \Sigma_V\|_F = O_p(\rho_{n_\eta}).$$

Proof. For $V \in \mathcal{V}$, write

$$z_V^0(Z) := z_V^{\mathrm{dr}}(Z;\eta_0), \qquad d_V(Z) := \hat{z}_V(Z) - z_V^0(Z).$$

Then $\mu_V^\eta - \mu_V = P_0 d_V$. For each arm a, let

$$D_{a,V}(Z) := \phi_V^a(Z;\hat{\eta}) - \phi_V^a(Z;\eta_0).$$

Lemma E.1 gives, uniformly in V,

$$P_0\|D_{a,V}\|^2 \le C\left\{\|\hat{\pi}(a\mid\cdot) - \pi_0(a\mid\cdot)\|_{L_2(P_X)}^2 + \|\hat{m}_a(\cdot;V) - m_a(\cdot;V)\|_{L_2(P_X;\mathbb{R}^J)}^2\right\}.$$

Since

$$\|\hat{m}_a(\cdot;V) - m_a(\cdot;V)\|_{L_2(P_X;\mathbb{R}^J)}^2 \le J \sup_{v\in\Pi(\mathcal{V})} \|\hat{m}_a(\cdot;v) - m_a(\cdot;v)\|_{L_2(P_X)}^2,$$

and $d_V = D_{1,V} - D_{0,V}$, we obtain

$$\sup_{V\in\mathcal{V}} P_0\|d_V\|^2 \le C\rho_{n_\eta}^2.$$

Thus

$$\sup_{V\in\mathcal{V}} \|\mu_V^\eta - \mu_V\| \le \sup_{V\in\mathcal{V}} (P_0\|d_V\|^2)^{1/2} = O_p(\rho_{n_\eta}).$$

For covariance, let $M_V^0 := P_0[z_V^0(Z)z_V^0(Z)^\top]$. Since

$$\hat{z}_V\hat{z}_V^\top - z_V^0 z_V^{0\top} = d_V d_V^\top + d_V z_V^{0\top} + z_V^0 d_V^\top,$$

and $\sup_{V,z}\|z_V^0(z)\| \le B_z$,

$$\sup_{V\in\mathcal{V}} \|M_V^\eta - M_V^0\|_F = O_p(\rho_{n_\eta}).$$

Finally,

$$\Sigma_V^\eta - \Sigma_V = (M_V^\eta - M_V^0) - (\mu_V^\eta \mu_V^{\eta\top} - \mu_V\mu_V^\top),$$

and the outer-product term is $O_p(\rho_{n_\eta})$ uniformly because the means are uniformly bounded. This proves the result. □

H.4 Perturbation of the ridge criterion

Lemma H.5 (Perturbation bound)**.** *Fix $\tau > 0$ and define*

$$Q_\tau(m, A) := m^\top (A + \tau I_J)^{-1} m$$

for $m \in \mathbb{R}^J$ and symmetric positive semidefinite $A \in \mathbb{R}^{J\times J}$. If

$$\|m\| \vee \|\tilde{m}\| \le B,$$

then

$$|Q_\tau(m, A) - Q_\tau(\tilde{m}, \tilde{A})| \le C(B,\tau)\big(\|m - \tilde{m}\| + \|A - \tilde{A}\|_F\big),$$

where one may take $C(B,\tau) = 2B/\tau + B^2/\tau^2$.

Proof. Let $B_A := A + \tau I_J$ and $B_{\tilde{A}} := \tilde{A} + \tau I_J$. Since $A, \tilde{A} \succeq 0$,

$$\|B_A^{-1}\|_{\mathrm{op}} \le \tau^{-1}, \qquad \|B_{\tilde{A}}^{-1}\|_{\mathrm{op}} \le \tau^{-1}.$$

Decompose

$$Q_\tau(m, A) - Q_\tau(\tilde{m}, \tilde{A}) = m^\top B_A^{-1} m - \tilde{m}^\top B_A^{-1}\tilde{m} + \tilde{m}^\top (B_A^{-1} - B_{\tilde{A}}^{-1})\tilde{m}.$$

The first term is bounded by $2B\tau^{-1}\|m - \tilde{m}\|$. For the second, use the resolvent identity

$$B_A^{-1} - B_{\tilde{A}}^{-1} = B_A^{-1}(\tilde{A} - A)B_{\tilde{A}}^{-1},$$

which gives

$$|\tilde{m}^\top (B_A^{-1} - B_{\tilde{A}}^{-1})\tilde{m}| \le B^2\tau^{-2}\|A - \tilde{A}\|_F.$$

Combining the bounds proves the lemma. □

H.5 Proof of Theorem 4.1

Proof of Theorem 4.1. By Lemma H.5,

$$\sup_{V\in\mathcal{V}}\left|\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V)-\mathcal{P}_\tau^\eta(V)\right|\le C\left[\sup_{V\in\mathcal{V}}\|\bar{z}_{\mathrm{tr},V}^{\mathrm{dr}}-\mu_V^\eta\|+\sup_{V\in\mathcal{V}}\|S_{\mathrm{tr},V}^{\mathrm{dr}}-\Sigma_V^\eta\|_F\right].$$

Lemma H.3 gives

$$\sup_{V\in\mathcal{V}}\left|\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V)-\mathcal{P}_\tau^\eta(V)\right|=O_p\left(\sqrt{\frac{Jd\log n_{\mathrm{tr}}}{n_{\mathrm{tr}}}}\right).$$

Similarly, Lemmas H.5 and H.4 imply

$$\sup_{V\in\mathcal{V}}\left|\mathcal{P}_\tau^\eta(V)-\mathcal{P}_\tau(V)\right|=O_p(\rho_{n_\eta}).$$

Combining these two bounds with (33) yields

$$\Delta_{\mathrm{tr}}:=\sup_{V\in\mathcal{V}}\left|\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V)-\mathcal{P}_\tau(V)\right|=O_p\left(\sqrt{\frac{Jd\log n_{\mathrm{tr}}}{n_{\mathrm{tr}}}}+\rho_{n_\eta}\right).$$

It remains to prove the deterministic optimization inequality. Let

$$V_\tau^\star\in\arg\max_{V\in\mathcal{V}}\mathcal{P}_\tau(V).$$

By definition of the empirical optimization gap,

$$\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V_\tau^\star)-\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(\hat{V})\le\varepsilon_{\mathrm{tr}}(\hat{V}).$$

Therefore

$$\begin{aligned}\mathcal{P}_\tau(V_\tau^\star)-\mathcal{P}_\tau(\hat{V})&=\{\mathcal{P}_\tau(V_\tau^\star)-\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V_\tau^\star)\}\\&\quad+\{\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V_\tau^\star)-\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(\hat{V})\}\\&\quad+\{\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(\hat{V})-\mathcal{P}_\tau(\hat{V})\}\\&\le 2\Delta_{\mathrm{tr}}+\varepsilon_{\mathrm{tr}}(\hat{V}).\end{aligned}$$

If $\varepsilon_{\mathrm{tr}}(\hat{V})=O_p(e_{\mathrm{tr}})$, the stochastic near-optimality statement follows immediately. □

I Finite-dictionary location learning

For structured outcomes, Euclidean optimization may be inappropriate. A finite dictionary provides a direct alternative. Let

$$\mathcal{C}=\{c_1,\dots,c_M\}\subset\mathcal{Y},\qquad\mathcal{V}_J(\mathcal{C})\subset\mathcal{C}^J.$$

For any output $\hat{V}\in\mathcal{V}_J(\mathcal{C})$, define its empirical dictionary optimization gap

$$\varepsilon_{\mathrm{tr}}^{\mathcal{C}}(\hat{V}):=\sup_{V\in\mathcal{V}_J(\mathcal{C})}\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V)-\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(\hat{V}).$$

For exhaustive search over the dictionary, this gap is zero.

Assume there exist constants $B_k,B_m,\varepsilon,\tau>0$ such that:

(i) for all $c\in\mathcal{C}$ and $y\in\mathcal{Y}$, $|k_Y(c,y)|\le B_k$;
(ii) for all $a\in\{0,1\}$ and $x\in\mathcal{X}$, both $\pi_0(a\mid x)$ and $\hat{\pi}(a\mid x)$ are at least ε;
(iii) for all a,x,c, $|m_a(x;c)|\vee|\hat{m}_a(x;c)|\le B_m$.

Define

$$\rho_{n_\eta}^{\mathcal{C}}:=\sum_{a=0}^{1}\left[\|\hat{\pi}(a\mid\cdot)-\pi_0(a\mid\cdot)\|_{L_2(P_X)}+\max_{c\in\mathcal{C}}\|\hat{m}_a(\cdot;c)-m_a(\cdot;c)\|_{L_2(P_X)}\right].$$

Theorem I.1 (Uniform consistency over a finite dictionary)**.** *Fix J, suppose $n_\eta, n_{\mathrm{tr}} \to \infty$, and assume $\rho^{\mathcal{C}}_{n_\eta} = o_p(1)$. Then*

$$\Delta^{\mathcal{C}}_{\mathrm{tr}} := \sup_{V \in \mathcal{V}_J(\mathcal{C})} \left| \hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V) - \mathcal{P}_\tau(V) \right| = O_p\left(\sqrt{\frac{\log |\mathcal{V}_J(\mathcal{C})|}{n_{\mathrm{tr}}}} + \rho^{\mathcal{C}}_{n_\eta} \right).$$

In particular, since $|\mathcal{V}_J(\mathcal{C})| \le M^J$,

$$\Delta^{\mathcal{C}}_{\mathrm{tr}} = O_p\left(\sqrt{\frac{J \log M}{n_{\mathrm{tr}}}} + \rho^{\mathcal{C}}_{n_\eta} \right).$$

Moreover, for any output $\hat{V}$ and any

$$V^\star_{\tau,\mathcal{C}} \in \arg\max_{V \in \mathcal{V}_J(\mathcal{C})} \mathcal{P}_\tau(V),$$

$$\mathcal{P}_\tau(V^\star_{\tau,\mathcal{C}}) - \mathcal{P}_\tau(\hat{V}) \le 2\Delta^{\mathcal{C}}_{\mathrm{tr}} + \varepsilon^{\mathcal{C}}_{\mathrm{tr}}(\hat{V}).$$

Thus, if $\varepsilon^{\mathcal{C}}_{\mathrm{tr}}(\hat{V}) = O_p(e^{\mathcal{C}}_{\mathrm{tr}})$,

$$\mathcal{P}_\tau(V^\star_{\tau,\mathcal{C}}) - \mathcal{P}_\tau(\hat{V}) = O_p\left(\sqrt{\frac{J \log M}{n_{\mathrm{tr}}}} + \rho^{\mathcal{C}}_{n_\eta} + e^{\mathcal{C}}_{\mathrm{tr}} \right).$$

Proof. Let $N_{\mathcal{C}} := |\mathcal{V}_J(\mathcal{C})|$. For $i \in I_{\mathrm{tr}}$, define

$$\hat{z}_{i,V} := z^{\mathrm{dr}}_V(Z_i; \hat{\eta}), \qquad z^0_{i,V} := z^{\mathrm{dr}}_V(Z_i; \eta_0).$$

Conditionally on I_η, the variables $\{\hat{z}_{i,V} : i \in I_{\mathrm{tr}}\}$ are i.i.d. for each fixed V.

The boundedness and positivity assumptions imply

$$\|\hat{z}_{i,V}\| \vee \|z^0_{i,V}\| \le \sqrt{J}\, B_z$$

for a constant $B_z < \infty$ depending only on B_k, B_m, ε. A union bound and Bernstein's inequality over the finite class give

$$\sup_{V \in \mathcal{V}_J(\mathcal{C})} \|\bar{z}^{\mathrm{dr}}_{\mathrm{tr},V} - \mu^{\hat{\eta}}_V\| = O_p\left(\sqrt{\frac{\log N_{\mathcal{C}}}{n_{\mathrm{tr}}}} \right),$$

and the same entrywise argument for second moments gives

$$\sup_{V \in \mathcal{V}_J(\mathcal{C})} \|S^{\mathrm{dr}}_{\mathrm{tr},V} - \Sigma^{\hat{\eta}}_V\|_F = O_p\left(\sqrt{\frac{\log N_{\mathcal{C}}}{n_{\mathrm{tr}}}} \right).$$

The nuisance transfer bound is the same as Lemma H.4, with $\sup_{v \in \Pi(\mathcal{V})}$ replaced by $\max_{c \in \mathcal{C}}$. Hence

$$\sup_{V \in \mathcal{V}_J(\mathcal{C})} \|\mu^{\hat{\eta}}_V - \mu_V\| = O_p(\rho^{\mathcal{C}}_{n_\eta}), \qquad \sup_{V \in \mathcal{V}_J(\mathcal{C})} \|\Sigma^{\hat{\eta}}_V - \Sigma_V\|_F = O_p(\rho^{\mathcal{C}}_{n_\eta}).$$

The perturbation Lemma H.5 then yields

$$\Delta^{\mathcal{C}}_{\mathrm{tr}} = O_p\left(\sqrt{\frac{\log N_{\mathcal{C}}}{n_{\mathrm{tr}}}} + \rho^{\mathcal{C}}_{n_\eta} \right).$$

Since $N_{\mathcal{C}} \le M^J$, the displayed simplified rate follows.

Finally, by definition of the dictionary optimization gap,

$$\hat{\mathcal{P}}_{\tau,\mathrm{tr}}(V^\star_{\tau,\mathcal{C}}) - \hat{\mathcal{P}}_{\tau,\mathrm{tr}}(\hat{V}) \le \varepsilon^{\mathcal{C}}_{\mathrm{tr}}(\hat{V}).$$

Adding and subtracting the empirical criterion gives

$$\mathcal{P}_\tau(V^\star_{\tau,\mathcal{C}}) - \mathcal{P}_\tau(\hat{V}) \le 2\Delta^{\mathcal{C}}_{\mathrm{tr}} + \varepsilon^{\mathcal{C}}_{\mathrm{tr}}(\hat{V}).$$

The stochastic bound follows immediately. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the finite-location causal testing target, the orthogonal DR construction, the local testing-efficiency result, the location-learning criterion, and the empirical scope. The claims are matched by the theory in Sections 2–4 and the experiments in Section 5 and Appendix B.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The discussion states the main scope restrictions: point-treatment causal inference under observed confounding, fixed kernels, and the dependence of power on the learned finite locations. It also notes that structured outcomes may require constrained/regularized location-search classes for practical applications.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The main theoretical results state their assumptions in Sections 2–4. Full assumptions and proofs are provided in Appendices C, E, F, H, and I.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the data-generating processes, sample splits, nuisance estimators, kernels, candidate dictionaries, location-learning rules, baselines, and Monte Carlo repetitions in Section 5 and Appendix B. An implementation is included in the supplementary material.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: An implementation is included in the supplementary material. The experiments use synthetic data generated by the provided scripts and OCTMNIST [42], which is publicly available and cited in the paper.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main text gives the core experimental setup, including the split protocol, number of locations, candidate dictionary size, nominal level, and number of Monte Carlo replications. Further details on the data-generating processes, nuisance models, bandwidth choices, OCTMNIST construction, and implementation are given in Appendices A and B.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experiments report rejection rates over repeated Monte Carlo runs, and the text states the Monte Carlo standard error at nominal level 0.05. Several figures include variability bands or error bars over repetitions, and the appendix reports additional diagnostic tables and plots.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix B reports the local hardware for the runtime study. The original hardware and package versions of the additional archived simulation runs were not recorded; their scripts, seeds, and raw outputs are available.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is theoretical and experimental research on causal distributional testing and uses synthetic data and a public medical-image benchmark. We are not aware of any deviation from the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The method may have positive impact by improving interpretable causal distributional analysis, including settings where mean effects are insufficient. Possible health applications with medical imaging or with DNA, protein sequences, etc. Potential negative impacts arise if the method is used in high-stakes domains without complying to ethics.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models, pre-trained generative systems, scraped datasets, or other assets with substantial misuse risk. The released material consists of code for reproducing statistical experiments.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use the OCTMNIST subset of MedMNIST v2. The MedMNIST dataset is licensed under Creative Commons Attribution 4.0 International (CC BY 4.0), except DermaMNIST, which is licensed under CC BY-NC 4.0; our use of OCTMNIST therefore falls under CC BY 4.0. The MedMNIST code is licensed under Apache-2.0. We cite MedMNIST v2 and the corresponding OCTMNIST source-data paper as requested by the MedMNIST project.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new asset is the implementation included in the supplementary material. The paper and appendices document the algorithms, data-generating processes, experimental settings, and matrix implementation needed to use the code.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing experiments or new human-subject data collection. OCTMNIST is used only as an existing public benchmark dataset.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject experiments requiring IRB or equivalent approval. OCTMNIST is used as an existing public benchmark.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used as part of the core methodology, theoretical results, experiments, data generation, or evaluation pipeline. They were used only for general brainstorming, editorial feedback on drafts, and language/formatting assistance; all scientific claims and conclusions were developed and verified by the authors.