MIND THE CURVATURE: CORRECTING STRUCTURAL BIAS IN BURES–WASSERSTEIN REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Tangent-space linearization makes scalar-on-distribution regression tractable for multivariate Gaussian inputs, but its predictions depend on the reference distribution. In Bures–Wasserstein space, nonzero curvature can turn reference mismatch into a nonlinear projection error that persists even when the input distributions are known exactly. We characterize the part of this error that remains after refitting the linear predictor and use it to motivate learning the reference from prediction residuals. A local risk expansion at general positive definite references quantifies the joint effect of reference displacement and data spread, with a signal-dependent commutator coefficient at the identity. The result distinguishes positive geometric bias from zero leading coefficients and exact predictive equivalence. An alternating procedure fits the regression and updates the reference covariance, while an empirical-risk bound links reference selection to the population target. Exact-covariance experiments verify the predicted coefficients and scaling quantitatively. Repeated prediction experiments assess validation-based selection and separate geometric from estimation error. Wearable, molecular, and sensing applications show task-dependent prediction gains with the same linear readout.

1 INTRODUCTION

Many modern learning problems associate each observational unit with an empirical distribution rather than a single vector. Examples include wearable measurements, glucose profiles, and subject-level imaging data, where distributional covariates retain variation and dependence beyond marginal summaries (Petersen & Müller, 2016; Ghosal et al., 2023; Ghosal & Matabuena, 2024; Ghosal et al., 2022; Matabuena et al., 2025; Niyogi et al., 2026). Scalar-on-distribution regression seeks to predict an outcome from such covariates. We study the geometric error that can arise when distributions in a space with nonzero curvature are projected into a common tangent space for linear prediction.

In one dimension, quantile functions give an isometric representation of Wasserstein distributions in $L^2(0,1)$ (Kloeckner, 2010; Panaretos & Zemel, 2020; Chen et al., 2023). For multivariate Gaussians, the covariance component instead has curved Bures–Wasserstein (BW) geometry (Takatsu, 2011). Figure 1 illustrates the consequence of this curvature. Projection at a task-aligned reference followed by parallel transport need not agree with direct projection at another reference, even after the constant reference shift is removed. The resulting nonlinear discrepancy can change the functions available to a linear predictor. Its component outside the fitted affine feature span creates approximation error even with exact distributions and arbitrarily many training observations.

Existing distribution representations include kernel mean embeddings (Szabó et al., 2016; Muandet et al., 2017), sliced Wasserstein features (Kolouri et al., 2016; Deshpande et al., 2019), and linearized optimal transport maps from a common reference (Wang et al., 2013; Moosmüller & Cloninger, 2023; Cloninger et al., 2025). Learnable transport references and supervised representations already incorporate task information (Mialon et al., 2021; Khurana et al., 2023). Our question concerns the approximation error induced by the reference within an explicit Gaussian tangent-feature family and linear readout. Related functional approaches and regression with geometric responses are discussed in Appendix I, with the direction of regression distinguished from the scalar response considered here.

Common references such as a standard Gaussian or a Wasserstein barycenter are selected independently of the prediction task. On a curved manifold, changing the reference need not be an affine

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

reparameterization that a refitted linear model can absorb. The Gaussian setting makes this distinction tractable through closed-form transport maps (Pitié & Kokaram, 2007; Chen et al., 2018; Delon & Desolneux, 2020; Flamary et al., 2019). We profile over the intercept and regression coefficients to identify the residual geometric component, then learn the reference covariance from this prediction criterion.

Our contributions are:

- We quantify the geometric bias that remains after affine refitting through a local risk expansion at general positive definite references, with a signal-dependent commutator coefficient at the identity. The analysis distinguishes positive bias, zero leading coefficients, and exact commuting equivalence.
- We establish an exact profiled-risk identity and an empirical-selection bound that connect geometric bias to residual minimization, and develop alternating updates of the reference covariance and regression coefficients. The fitted-risk decomposition separates geometric and estimation errors, motivating validation-based reference selection.
- We verify the predicted coefficients and joint scaling with exact-covariance experiments (Table 1), and assess prediction under controlled reference mismatch (Figure 2). Additional simulations examine validation selection and estimated-input robustness, while wearable, molecular, and sensing applications evaluate task-adaptive references.

2 Problem Setup

Let $\mathcal{G}_d$ be the nondegenerate Gaussian family, with $\mu_i = \mathcal{N}(m_i, \Sigma_i)$, $m_i \in \mathbb{R}^d$, $\Sigma_i \in \mathrm{Sym}_{++}(d)$, and scalar response $y_i \in \mathbb{R}$. Here $\mathrm{Sym}(d)$ and $\mathrm{Sym}_{++}(d)$ denote symmetric and symmetric positive definite $d \times d$ matrices, respectively. For measures in $\mathcal{P}_2(\mathbb{R}^d)$, the squared 2-Wasserstein distance is

$$W_2^2(\mu, \lambda) = \inf_{\pi \in \Pi(\mu,\lambda)} \int \|x - y\|^2 \, d\pi(x, y), \tag{1}$$

where $\Pi(\mu, \lambda)$ is the set of couplings (Chewi et al., 2025). For two nondegenerate Gaussians, Equation 1 has the closed form

$$W_2^2(\mu_i, \mu_j) = \|m_i - m_j\|^2 + \mathrm{tr}\Big\{\Sigma_i + \Sigma_j - 2(\Sigma_i^{1/2}\Sigma_j\Sigma_i^{1/2})^{1/2}\Big\}$$

(Chen et al., 2018; Delon & Desolneux, 2020). When the reference ρ is absolutely continuous, its optimal transport map defines the Wasserstein logarithmic map $\mathrm{Log}_\rho(\mu)(x) = T_{\rho\to\mu}(x) - x$, $x \sim \rho$, viewed in the $L^2(\rho)$ tangent space (Ambrosio et al., 2005; Wang et al., 2013; Moosmüller & Cloninger, 2023). At a Gaussian reference $\rho = \mathcal{N}(m_\rho, \Sigma_\rho)$, the optimal map to μ_i is affine, $T_{\rho\to\mu_i}(x) = m_i + A_{\rho,i}(x - m_\rho)$, where

$$A_{\rho,i} = \Sigma_\rho^{-1/2}(\Sigma_\rho^{1/2}\Sigma_i\Sigma_\rho^{1/2})^{1/2}\Sigma_\rho^{-1/2}.$$

The corresponding finite-dimensional tangent feature is

$$z_\rho(\mu_i) = \begin{bmatrix} m_i - A_{\rho,i}m_\rho \\ \mathrm{vech}(A_{\rho,i} - I_d) \end{bmatrix} \in \mathbb{R}^{d+d(d+1)/2}. \tag{2}$$

Thus the finite feature dimension is $p = d + d(d+1)/2$; half-vectorization avoids duplicating symmetric covariance coordinates. We fit a linear predictor by minimizing

$$(\widehat{\alpha}, \widehat{\boldsymbol{w}}) = \arg\min_{\alpha\in\mathbb{R}, \boldsymbol{w}\in\mathbb{R}^p} \sum_{i=1}^n \{y_i - \alpha - \boldsymbol{w}^\top z_\rho(\mu_i)\}^2.$$

The analysis focuses on how the choice of ρ changes the features presented to the learner. Throughout, μ denotes a predictor distribution, ρ a candidate reference, and $\rho_\star$ the task-aligned reference.

3 Methodology

3.1 Reference-dependent feature geometry

At the standard Gaussian reference, $A_{\rho,i} = \Sigma_i^{1/2}$. Thus $z_{\mathcal{N}(0,I_d)}(\mu_i)$ and the Euclidean feature $[m_i^\top, \mathrm{vech}(\Sigma_i^{1/2})^\top]^\top$ differ by a fixed translation.

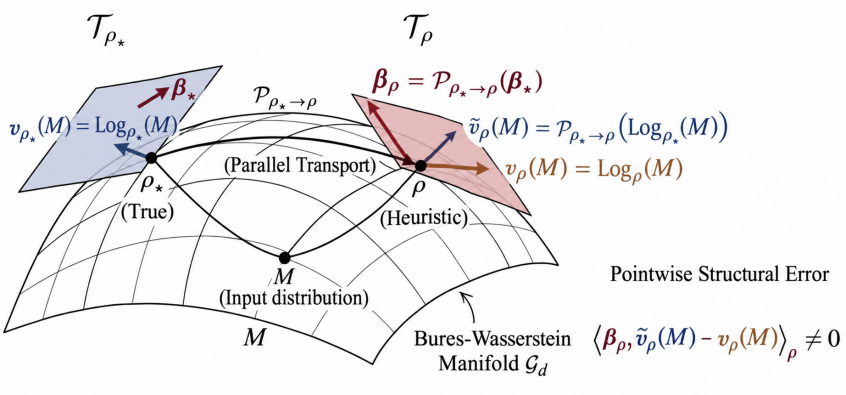

Figure 1: Schematic illustration of geometric structural bias induced by reference mismatch on the non-flat Bures–Wasserstein manifold. Projecting at $\rho_\star$ and then parallel transporting to ρ does not generally agree with direct projection at ρ, producing the displayed pointwise structural error.

Remark 1 (Standard-reference relationship)**.** With an intercept and unregularized coefficients, standard-reference tangent regression and this Euclidean embedding have the same augmented feature span and fitted values. This equivalence does not identify a generally optimal reference.

The invertible column transformation is given in Appendix C. The oracle comparator in Figure 2 uses this standard reference.

3.2 Geometric structural bias

Assume the conditional mean is linear at an unknown task-aligned reference $\rho_\star$,

$$Y = \alpha_\star + \langle \beta_\star, \mathrm{Log}_{\rho_\star}(M) \rangle_{\rho_\star} + \epsilon, \qquad \mathbb{E}(\epsilon \mid M) = 0, \quad \mathbb{E}(\epsilon^2) = \sigma^2, \tag{3}$$

where M is a random predictor distribution and $\beta_\star \in T_{\rho_\star}\mathcal{G}_d$. The tangent inner product is $\langle v, w \rangle_\rho = \int v(x)^\top w(x)\, d\rho(x)$, and $\mathcal{P}_{\rho_\star \to \rho}$ denotes parallel transport along the Gaussian Wasserstein geodesic. For a candidate reference ρ, let $v_\rho(M) = \mathrm{Log}_\rho(M)$, $\Delta_\rho = \mathrm{Log}_\rho(\rho_\star)$, and $\beta_\rho = \mathcal{P}_{\rho_\star \to \rho}\beta_\star$. The transported oracle feature $\tilde{v}_\rho(M) = \mathcal{P}_{\rho_\star \to \rho}\mathrm{Log}_{\rho_\star}(M)$ preserves the oracle signal, whereas $v_\rho(M)$ is observable once ρ is chosen. Define the corrected vector discrepancy $d_\rho(M) = \tilde{v}_\rho(M) - v_\rho(M) + \Delta_\rho$ and its signal component $b_\rho(M) = \langle \beta_\rho, d_\rho(M) \rangle_\rho$.

Figure 1 illustrates the unshifted path difference $\tilde{v}_\rho - v_\rho$. Adding Δ_ρ removes the constant reference shift to give d_ρ, whose local relation to curvature is derived in Appendix B. The intercept absorbs the constant shift, and coefficient refitting can absorb additional linear components. The risk characterization below identifies what remains, before Section 3.4 gives a BW-specific coefficient and remainder.

3.3 Why residual minimization targets structural bias

Pointwise discrepancy alone does not justify an optimization criterion because coefficient refitting can absorb a linear component of b_ρ. Let $f_\star(M) = \mathbb{E}(Y \mid M)$ and let $\mathcal{H}_\rho$ be the finite-dimensional subspace of $L^2(P_M)$ spanned by the constant function and the coordinates of $v_\rho(M)$, and let Π_ρ denote orthogonal projection onto this subspace. Define the profiled population risk

$$\mathcal{R}(\rho) = \inf_{\alpha \in \mathbb{R},\, w \in T_\rho \mathcal{G}_d} \mathbb{E}\{Y - \alpha - \langle w, v_\rho(M) \rangle_\rho\}^2. \tag{4}$$

Theorem 1 (Profiled prediction risk)**.** *Assume Equation 3 holds,* $\mathbb{E}\|v_\rho(M)\|_\rho^2 < \infty$*, and* $b_\rho(M) \in L^2(P_M)$*. Then*

$$\mathcal{R}(\rho) = \sigma^2 + \underbrace{\|(I - \Pi_\rho) b_\rho\|^2_{L^2(P_M)}}_{\mathcal{B}(\rho)}. \tag{5}$$

Hence $\mathcal{R}(\rho_\star) = \sigma^2$, and minimizing $\mathcal{R}$ is equivalent to minimizing $\mathcal{B}$. This characterization does not require a unique reference.

To see the identity, write $f_\star = \alpha_\rho + \langle \beta_\rho, v_\rho \rangle_\rho + b_\rho$. The first two terms belong to $\mathcal{H}_\rho$, and conditional mean-zero noise gives $\mathbb{E}(Y-h)^2 = \sigma^2 + \|f_\star - h\|_{L^2}^2$. Projecting onto $\mathcal{H}_\rho$ leaves exactly $(I - \Pi_\rho) b_\rho$ (Appendix C.3). The quantity $\mathcal{B}(\rho)$ is the geometric approximation error after refitting, rather than the entire pointwise discrepancy.

If the conditions hold on a class $\mathcal{K}$ containing $\rho_\star$ and an initial reference ρ_0, the recoverable risk is

$$\mathcal{R}(\rho_0) - \inf_{\rho \in \mathcal{K}} \mathcal{R}(\rho) = \|(I - \Pi_{\rho_0}) b_{\rho_0}\|_{L^2(P_M)}^2.$$

A mismatched reference can nevertheless attain the noise floor when $b_\rho \in \mathcal{H}_\rho$, because coefficient refitting absorbs its entire discrepancy. Residual minimization therefore identifies a predictively equivalent reference, not necessarily a unique geometric center.

3.4 When reference mismatch creates irreducible risk

The profiled-risk identity identifies the irreducible component but does not give its dependence on reference displacement. We quantify that dependence through a joint local expansion. Fix $\mathbf{S} \in \mathrm{Sym}_{++}(d)$ and $\mathbf{B}, \mathbf{H} \in \mathrm{Sym}(d)$, representing the generating covariance, response direction, and reference displacement. Let $\mathbf{U} \in \mathrm{Sym}(d)$ be random with $\|\mathbf{U}\|_{\mathrm{op}} \le L$ almost surely and $\mathrm{Cov}\{\mathrm{svec}(\mathbf{U})\} \succ 0$. Here svec is an orthonormal half-vectorization, equivalent to vech for unregularized regression. For sufficiently small $r > 0$ and real t, set

$$\begin{aligned} \mathbf{T}_r &= \mathbf{I} + r\mathbf{U}, \qquad & \mathbf{D}_t &= \mathbf{I} + t\mathbf{H}, \\ \mathbf{\Sigma}_r &= \mathbf{T}_r \mathbf{S} \mathbf{T}_r, \qquad & \mathbf{S}_t &= \mathbf{D}_t \mathbf{S} \mathbf{D}_t, \\ Y_r &= \alpha_\star + r\mathrm{tr}(\mathbf{B}\mathbf{U}) + \epsilon, \end{aligned} \tag{6}$$

where $\mathbb{E}(\epsilon \mid \mathbf{U}) = 0$, $\mathbb{E}\epsilon^2 = \sigma^2$, and $\mathbf{T}_r, \mathbf{D}_t \succ 0$. All means are zero, so only the complete covariance feature block and intercept are retained. The model is exactly tangent-linear at $\mathbf{S}$, since its transport to $\mathbf{\Sigma}_r$ is $\mathbf{T}_r$. At the candidate $\mathbf{S}_t$, the transport matrix $\mathbf{A}_{t,r}$ satisfies $\mathbf{A}_{t,r}\mathbf{S}_t\mathbf{A}_{t,r} = \mathbf{\Sigma}_r$. Write $R_r(t) = \inf_{a, \mathbf{W} = \mathbf{W}^\top} \mathbb{E}\{Y_r - a - \mathrm{tr}[\mathbf{W}(\mathbf{A}_{t,r} - \mathbf{I})]\}^2$.

Define $\mathcal{L}_{\mathbf{S}}(\mathbf{X}) = \mathbf{S}\mathbf{X} + \mathbf{X}\mathbf{S}$ and the linear and quadratic functions

$$\begin{aligned} \mathbf{K} &= \mathcal{L}_{\mathbf{S}}^{-1}(\mathbf{U}\mathbf{H}\mathbf{S} + \mathbf{S}\mathbf{H}\mathbf{U}), \\ \mathbf{F} &= \mathcal{L}_{\mathbf{S}}^{-1}\{\mathbf{K}\mathbf{S}\mathbf{U} + \mathbf{U}\mathbf{S}\mathbf{K} - \mathbf{U}(\mathbf{H}\mathbf{S} + \mathbf{S}\mathbf{H})\mathbf{U}\}. \end{aligned} \tag{7}$$

Let Π_U project onto the span of the constant and the coordinates of $\mathbf{U}$ in $L^2(P_U)$.

Theorem 2 (Local prediction risk at a general reference)**.** *Under the preceding conditions, as $(t, r) \to (0, 0)$,*

$$R_r(t) - \sigma^2 = t^2 r^4 \mathcal{C}_{\mathbf{S}} + O\{t^2 r^4 (|t| + r)\}, \qquad \mathcal{C}_{\mathbf{S}} = \|(I - \Pi_U)\mathrm{tr}(\mathbf{B}\mathbf{F})\|_{L^2}^2. \tag{8}$$

The remainder is uniform on a sufficiently small rectangle in (t, r). If $\mathcal{C}_{\mathbf{S}} > 0$, the excess risk is bounded between $\mathcal{C}_{\mathbf{S}} t^2 r^4/2$ and $3\mathcal{C}_{\mathbf{S}} t^2 r^4/2$ on a smaller rectangle. If $\mathcal{C}_{\mathbf{S}} = 0$, it is $O\{t^2 r^4 (|t| + r)^2\}$.

The coefficient depends on the response and data law as well as the geometry. The displacement satisfies $d_{\mathrm{BW}}^2(\mathbf{S}, \mathbf{S}_t) = t^2 \mathrm{tr}(\mathbf{H}\mathbf{S}\mathbf{H})$, so the theorem gives a quadratic reference-distance effect in directions with positive coefficient. It concerns exact covariances and population refitting, without assuming that an optimizer recovers a unique reference. The proof differentiates the OT equation directly, avoiding an affine-invariance assumption for the BW metric. It also normalizes the features before taking $r \to 0$, where the unnormalized Gram matrix degenerates (Appendix F).

Corollary 1 (Task dependence of the commutator term)**.** *At $\mathbf{S} = \mathbf{I}$, let $[\mathbf{X}, \mathbf{Y}] = \mathbf{X}\mathbf{Y} - \mathbf{Y}\mathbf{X}$ and $Q(\mathbf{U}) = \langle [\mathbf{H}, \mathbf{U}], [\mathbf{B}, \mathbf{U}] \rangle_F$. Then*

$$\mathcal{C}_{\mathbf{I}} = \frac{1}{16} \|(I - \Pi_U) Q\|_{L^2}^2. \tag{9}$$

If $\mathbf{U} \overset{d}{=} -\mathbf{U}$, this coefficient is $\mathrm{Var}(Q)/16$. If its law has positive density on a nonempty open subset of $\mathrm{Sym}(d)$, then $\mathcal{C}_{\mathbf{I}} > 0$ exactly when Q is a nonzero polynomial. In particular, $\mathbf{B} = \mathbf{H}$ with nonscalar $\mathbf{H}$ gives positive risk under this support condition.

Noncommuting directions affect prediction through their interaction with the signal and the part left after linear refitting. For $\mathbf{B} = \mathbf{I}$, the displayed coefficient is zero even if $[\mathbf{H}, \mathbf{U}] \neq 0$; only the higher-order bound follows. In contrast, if $[\mathbf{H}, \mathbf{U}] = 0$ almost surely, then $\mathbf{A}_{t,r} = \mathbf{T}_r \mathbf{D}_t^{-1}$ and $R_r(t) = \sigma^2$ exactly for every admissible t, r and any $\mathbf{S} \succ 0$. This separate commuting comparison permits a lower-dimensional support.

For a concrete positive example, take $d = 2$, $\mathbf{S} = \mathbf{I}$, $\mathbf{B} = \mathbf{H} = \mathrm{diag}(1, -1)$, and $\mathbf{U} = \left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$ with independent $a, b, c \sim \mathrm{Unif}[-1, 1]$. Then $Q = 8b^2$ and $\mathcal{C}_{\mathbf{I}} = 16/45$. The residual contains $b^2 - 1/3$, which a linear refit cannot absorb. This gives an explicit prediction penalty that remains with exact distributional inputs and arbitrarily many training distributions.

3.5 Alternating optimization for reference correction

An unsupervised BW barycenter minimizes covariate dispersion rather than prediction error. We use it to initialize real-data reference learning, with labels entering through the profiled prediction criterion.

We fix the reference mean at the training-set barycenter and learn its covariance. With the full covariance feature block and an intercept, a reference-mean change is an invertible affine reparameterization, leaving the profiled predictor unchanged (Appendix C.2). The curvature relevant here lies in the covariance component. The empirical counterpart of Equation 4 is

$$\widehat{\mathcal{R}}_n(\Sigma_\rho) = \min_{\alpha, \boldsymbol{w}} \mathcal{J}_n(\Sigma_\rho, \alpha, \boldsymbol{w}), \qquad \mathcal{J}_n(\Sigma_\rho, \alpha, \boldsymbol{w}) = n^{-1} \sum_{i=1}^{n} \left\{ y_i - \alpha - \boldsymbol{w}^\top z_{\Sigma_\rho}(\mu_i) \right\}^2, \quad (10)$$

where z_{Σ_ρ} denotes the coordinates in Equation 2 with the reference mean held fixed. Ordinary subject-level covariates, when present, are appended to z_{Σ_ρ} and profiled jointly; they do not depend on the reference. For a nonsingular augmented Gram matrix, the fitted coefficients are unique and the envelope theorem equates the profiled gradient with the reference gradient evaluated at those coefficients. Thus Equation 10 is the empirical counterpart of the population target in Equation 5.

On a compact reference class with continuous square-integrable features and uniformly nonsingular population Gram matrices, the empirical profiled risk converges uniformly to its population counterpart. Proposition 1 in Appendix C states the precise conditions and the resulting consistency for approximate global minimizers.

More explicitly, put $\delta_n = \sup_{\rho \in \mathcal{K}} |\widehat{\mathcal{R}}_n(\rho) - \mathcal{R}(\rho)|$. If $\widehat{\rho}$ is within η_n of the empirical global minimum, then

$$\mathcal{B}(\widehat{\rho}) - \inf_{\rho \in \mathcal{K}} \mathcal{B}(\rho) \leq 2\delta_n + \eta_n.$$

This follows by adding and subtracting the empirical risks at $\widehat{\rho}$ and a population minimizer. A small optimization gradient does not establish a small η_n. For a fitted predictor $\widehat{f} \in \mathcal{H}_{\widehat{\rho}}$ and independent test data, let $\mathcal{D}$ contain all data used for fitting and reference selection. Orthogonality then gives

$$\mathbb{E}\{(Y - \widehat{f}(M))^2 \mid \mathcal{D}\} - \sigma^2 = \mathcal{B}(\widehat{\rho}) + \|\widehat{f} - \Pi_{\widehat{\rho}} f_\star\|_{L^2}^2. \quad (11)$$

The additional estimation error explains why decreasing training RSS need not improve test prediction. We therefore assess checkpoint selection on independent validation data in the additional simulations of Appendix H.

Algorithm 1 uses $F_{\mathcal{I}}(\Sigma) = \min_{\alpha, \boldsymbol{w}: \boldsymbol{w}_{\mathcal{I}^c} = 0} \mathcal{J}_n(\Sigma, \alpha, \boldsymbol{w})$ for a fixed retained coordinate set $\mathcal{I}$. The full set recovers Equation 10. The line search uses an initial step $\bar{\eta} > 0$, contraction $\beta \in (0, 1)$, Armijo constant $c \in (0, 1)$, and a finite backtracking cap.

Here E_t is the symmetric Euclidean gradient, and the curve follows half the negative BW gradient. The positive transport-factor check keeps the simulation updates in the BW exponential-map domain. An endpoint-feasible variant used in the real-data implementation is described separately in Appendix E. Both variants admit the small feasible steps used by the following descent argument.

Armijo acceptance yields, for some $c \in (0, 1)$,

$$F_{\mathcal{I}}(\Sigma_\rho^{(t+1)}) \leq F_{\mathcal{I}}(\Sigma_\rho^{(t)}) - c\eta_t D_t, \qquad D_t = 2\,\mathrm{tr}\{\Sigma_\rho^{(t)} E_t^2\}. \quad (12)$$

Algorithm 1 Task-adaptive reference correction

Require: Training Gaussians $\{\mu_i\}_{i=1}^n$, responses $\{y_i\}_{i=1}^n$, maximum iterations $T_{\max}$, stopping tolerance τ_{stop}, and SPD floor ϵ_{spd}

Choose $\Sigma_\rho^{(0)}$ (training BW barycenter by default); set $t \leftarrow 0$
while $t < T_{\max}$ **do**
  $z_i^{(t)} \leftarrow z_{\Sigma_\rho^{(t)}}(\mu_i)$ for all i
  Profile OLS and compute $E_t \leftarrow \nabla F_{\mathcal{I}}(\Sigma_\rho^{(t)})$
  Stop if $\|\operatorname{grad} F_{\mathcal{I}}\|_{\text{BW}} < \tau_{\text{stop}}$
  Backtrack $\eta_t \in \{\bar{\eta}, \bar{\eta}\beta, \ldots\}$ until $I - \eta_t E_t \succ 0$ and $C_t = (I - \eta_t E_t)\Sigma_\rho^{(t)}(I - \eta_t E_t)$ has $\lambda_{\min}(C_t) \geq \epsilon_{\text{spd}}$ and satisfies Equation 12, refitting OLS at each trial
  If no trial is accepted within the backtracking cap, record failure and stop
  Set $\Sigma_\rho^{(t+1)} \leftarrow C_t$ and $t \leftarrow t + 1$
end while
return $\Sigma_\rho^{(t)}$ and least-squares coefficients refitted there

Under compact interior iterates, uniformly nonsingular retained Gram matrices, and local smoothness, the ideal uncapped algorithm has non-increasing loss and stationary accumulation points. Theorem 4 in Appendix D gives the precise statement and proof. Finite iteration caps and boundary termination do not certify stationarity or global optimality.

Computation and scope. With p retained features, spectral matrix roots and analytic Sylvester derivatives cost $O(nd^3)$, and an SVD least-squares fit costs $O(np^2)$ when $n \geq p$. Each accepted update may require several OLS refits during backtracking. The simulation implementation fixes the mean, includes an unpenalized intercept, and evaluates the full active covariance span at every reference. Rank failures, backtracking limits, and iteration caps are reported separately from gradient-based termination. The main risk theory assumes exact Gaussian parameters. Estimation of a linear transport map from sampled moments is a separate statistical problem (Flamary et al., 2019), examined here through an empirical robustness experiment.

4 Experiments

The simulation comparisons separate the geometric mechanism from finite-sample reference learning. We first test the local risk formula with exact covariance inputs, then assess prediction under reference mismatch and use repeated validation and risk-decomposition diagnostics to interpret reference learning. Real-data applications assess task-dependent reference adaptation with shared Gaussian inputs and fixed test splits.

4.1 Exact covariance mechanisms

For the two-dimensional model in Section 3.4, we compute the population OLS projection by tensor Gauss–Legendre quadrature. The inputs are exact covariances and the target is the known conditional mean, so the measured residual contains no response noise or distribution-estimation error. Table 1 compares $(R_r(t) - \sigma^2)/(t^2 r^4)$ along $t = r$ and $t = r^2$ with the theoretical coefficient $16/45$, reporting relative errors explicitly. A nonidentity reference checks the general Sylvester coefficient independently. Changing the off-diagonal entry from b to λb gives coefficient $16\lambda^4/45$ for each fixed $\lambda > 0$. The commuting endpoint is treated separately, since its active dimension drops. Changing the signal direction gives a further zero-leading-coefficient control (Appendix H).

At $t = r = 0.025$, normalized risk is 0.35566676 against $16/45 \approx 0.35555556$, and 0.00824125 against 0.00827645 at the general reference. The second joint path approaches the same coefficients. The commuting control has numerical-zero risk, whereas changing only the signal direction can remove the leading coefficient without imposing commutativity.

Table 1: Theoretical coefficient and numerical normalized excess risk $E(t,r) = \{R_r(t) - \sigma^2\}/(t^2 r^4)$. Relative errors are $100|E - \mathcal{C}_{\mathbf{S}}|/\mathcal{C}_{\mathbf{S}}$. All coefficients are computed after affine refitting.

$\mathbf{S}$	r	$\mathcal{C}_{\mathbf{S}}$	$E(r,r)$	Error (%)	$E(r^2,r)$	Error (%)
Identity	0.1	0.35555556	0.35735798	0.5069	0.35734714	0.5039
Identity	0.05	0.35555556	0.35600155	0.1254	0.35600085	0.1252
Identity	0.025	0.35555556	0.35566676	0.0313	0.35566672	0.0313
General	0.1	0.00827645	0.00816036	1.4026	0.00829052	0.1700
General	0.05	0.00827645	0.00821008	0.8019	0.00827991	0.0417
General	0.025	0.00827645	0.00824125	0.4253	0.00827731	0.0104

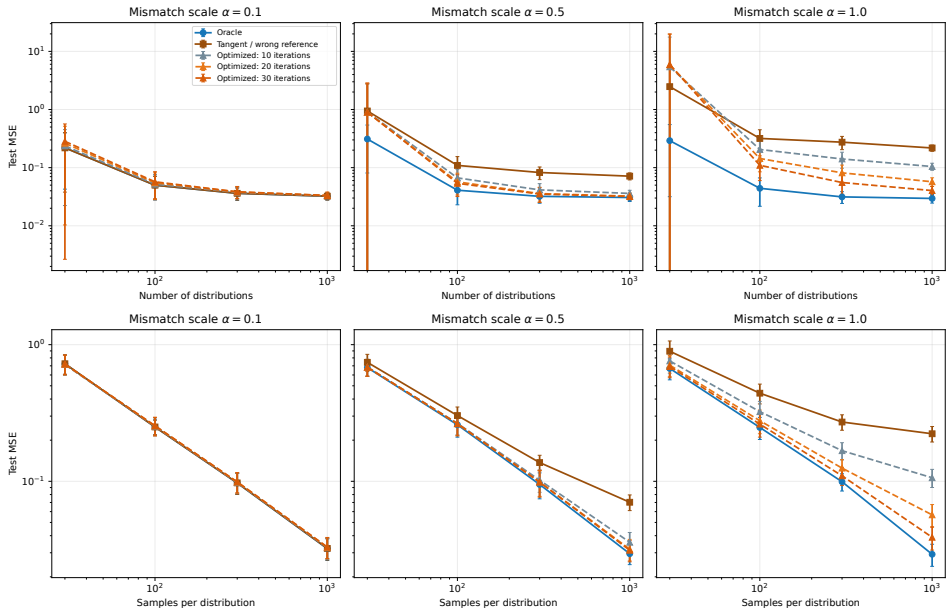

Figure 2: Mean test MSE $\pm$ one standard deviation over 20 replicates at $d = 5$. The top row varies N at $m = 1000$; the bottom row varies m at $N = 1000$. Columns correspond to mismatch scales $\alpha = 0.1, 0.5, 1$. Oracle uses the generating reference $\rho_\star = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

4.2 Controlled reference misspecification

The simulation investigates the prediction degradation caused by reference mismatch and the ability of reference updates to improve prediction. Responses are generated relative to the task-aligned reference $\rho_\star = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, whereas the fitted model starts from the misspecified reference $\rho = \mathcal{N}(\mathbf{0}, \mathbf{I}_d + \mathbf{e}\mathbf{e}^\top)$, with $\mathbf{e} = \alpha\mathbf{u}$ and one standard-normal direction $\mathbf{u}$ held fixed across conditions. We evaluate prediction MSE on held-out distributions. The oracle baseline uses the generating reference $\rho_\star$.

The $d = 5$ design crosses mismatch scales $\alpha \in \{0.1, 0.5, 1\}$ with two information regimes. The top row fixes $m = 1000$ observations per distribution and varies $N \in \{30, 100, 300, 1000\}$; the bottom row reverses these roles. This separates additional distributions from more accurate within-distribution estimation. In each replicate, 80% of the independently sampled distributions form the training set and 20% form the test set; all reference candidates use the same estimated Gaussian covariates and split. The initial mismatch direction is independent of the responses, while subsequent updates use training responses only.

Figure 2 shows little separation at $\alpha = 0.1$, where additional updates can slightly increase test error. At $\alpha = 0.5$ and 1, correction reduces the gap to the oracle baseline when $N \geq 100$. For $N = 30$, only 24 distributions train the 20-feature regression, and large-mismatch updates increase test error.

These prediction curves combine geometric, estimation, and numerical effects; Table 1 isolates the geometric expansion. Implementation details are in Appendix E.

Increasing m improves within-distribution estimation, whereas increasing N supplies independent units for fitting coefficients and selecting a reference. Large m therefore cannot resolve a nearly saturated regression when N is small.

Additional quantitative checks. With exact covariance inputs, twelve settings vary dimension, noise level, and training size, each with 50 independent replicates. Validation selects the update budget using data separate from the test set. Learning from a mismatched reference reduces mean test MSE by 0.62–19.56% across these settings, with paired pointwise intervals reported in Table 10. Independent noiseless calibration fits estimate the geometric component separately from the deployed predictor's estimation error (Table 11). These diagnostics test the statistical link in Equation 11. The full design, zero-case controls, stopping diagnostics, and estimated-input robustness are in Appendix H.

4.3 Real-data applications

We evaluate whether reference learning improves prediction beyond a fixed geometric representation and simple linear alternatives. Fixed and optimized tangent OLS share Gaussian inputs, preprocessing, ordinary covariates, and intercepts; only the reference covariance changes from the training BW barycenter. Mean + SD OLS uses marginal summaries, Euclidean OLS uses $(m_i, \mathrm{vech}(\Sigma_i^{1/2}))$, and Mean + Covariance RidgeCV uses $(m_i, \mathrm{vech}(\Sigma_i))$ with training-only penalty selection. Thus the tangent comparison isolates reference selection, whereas ridge tests a regularized alternative with the same Gaussian information.

Tasks. NHANES predicts released age in years from six-dimensional activity distributions, with sleep fraction and sex shared across all models (National Center for Health Statistics, 2020; John et al., 2019). ESOL predicts aqueous solubility, while OPERA BioHL and KOA predict biodegradation half-life and octanol–air partitioning (Delaney, 2004; Mansouri et al., 2018). Each molecular bag contains four atom descriptors in a single conformer, consisting of atomic number and distances to the centroid, nearest atom, and other atoms on average. Gas estimates carbon monoxide concentration from sensor-response distributions (Burgus, 2018). Training-fitted PCA reduces 14 log-resistance channels to dimension four; each bag is a fixed exposure window, so the task is concurrent calibration.

Evaluation protocol. Molecular splits group identical canonical structures; Gas uses chronological batches. Preprocessing and reference fitting use training data. NHANES uses a prespecified schedule; BioHL uses training-only grouped CV, and the remaining tasks use validation-selected schedules. Ridge selects among 15 penalties by training-only analytic leave-one-out error, using the same outer splits as OLS. Table 2 compares the five methods in terms of MSE, MAE, and R^2 on held-out test sets. Sample sizes and full protocols are in Appendix G.

Prediction performance. Optimized OLS has the lowest MSE among the five compared methods on the five applications (Table 2). On NHANES, reference learning reduces test MSE by 9.91% against fixed BW and lowers MAE from 11.554 to 10.822 years. On ESOL, Mean + Covariance RidgeCV improves on fixed BW, but optimized OLS has lower MSE, with reductions of 20.77% against BW and 12.09% against ridge. BioHL reduces MSE by 46.39% against fixed BW and 35.18% against Euclidean OLS, its lowest-error listed baseline. KOA reduces MSE by 10.49% against fixed BW. On these four tasks, optimized OLS also has the lowest MAE; the higher R^2 follows directly from its lower MSE on the same test outcomes.

Chronological test. On Gas, optimized OLS reduces MSE by 18.24% against fixed BW and by 14.68% against Mean + Covariance RidgeCV. It has the lowest MSE and the highest R^2, whereas Mean + SD yields lower MAE (0.756 versus 0.788 ppm). Additional results are in Appendix G.

Table 2: Test performance on applications from Gaussian summaries. Bold indicates the best unrounded value within each task and metric. MAE units: years (NHANES), ppm (Gas), and source-log units otherwise; MSE squares them and R^2 is unitless. Sample sizes and additional results are in Appendix G.

Dataset	Metric	Mean + SD	Euclidean	Mean+Cov. RidgeCV	Fixed BW	Optimized
NHANES	MSE ↓	240.995	212.022	220.359	201.390	**181.433**
	MAE ↓	12.771	11.914	12.115	11.554	**10.822**
	R^2 ↑	0.559	0.612	0.597	0.632	**0.668**
ESOL	MSE ↓	2.009	1.886	1.679	1.863	**1.476**
	MAE ↓	1.100	1.077	1.016	1.072	**0.924**
	R^2 ↑	0.520	0.550	0.599	0.555	**0.648**
BioHL	MSE ↓	0.139	0.115	0.122	0.139	**0.074**
	MAE ↓	0.295	0.260	0.280	0.275	**0.201**
	R^2 ↑	0.675	0.733	0.715	0.677	**0.827**
KOA	MSE ↓	1.181	0.990	1.279	0.848	**0.759**
	MAE ↓	0.808	0.812	0.794	0.754	**0.712**
	R^2 ↑	0.832	0.859	0.818	0.879	**0.892**
Gas	MSE ↓	1.162	1.368	1.091	1.138	**0.931**
	MAE ↓	**0.756**	0.955	0.793	0.885	0.788
	R^2 ↑	0.971	0.966	0.973	0.972	**0.977**

4.4 Optimization and sensitivity

Table 6 compares BioHL optimization budgets. Training CV selects the 30-update schedule with initial step 0.5. Increasing the cap to 100 lowers training MSE but raises inner-CV MSE, supporting validation-based stopping. The test comparison uses the CV-selected schedule.

The NHANES MSE reduction remains 9.08–9.91% as the SPD floor varies from 10^{-8} to 10^{-3}. Across four BioHL conformer conditions, validation MSE reductions relative to condition-matched fixed BW range from 13.03% to 56.92%. Complete sensitivity and validation-to-test comparisons are in Appendix G.4, Table 8. These diagnostics assess numerical and representation sensitivity within the same splits.

5 Conclusion and limitations

Reference selection changes the prediction space even when the regression remains linear in its features. In curved BW geometry, the resulting projection discrepancy produces geometric bias when its signal-dependent component survives affine refitting. Our local expansion quantifies this component, and the profiled-risk identity makes residual minimization a corresponding statistical target. The objective is predictive alignment rather than recovery of a unique geometric center.

The analysis concerns Gaussian distributions and linear tangent prediction, with a local curvature expansion. Validation addresses the estimation error introduced by reference learning, and estimated-input experiments assess empirical robustness. Extending the geometric-risk characterization beyond Gaussians and quantifying reference-selection uncertainty remain open directions.

REPRODUCIBILITY STATEMENT

The complete proofs and geometric derivations are provided in the appendix. The appendix records data construction, split and selection protocols, optimizer settings, numerical checks, and an SPD-boundary sensitivity analysis. The simulation package includes implementation scripts, environment metadata, seeded numerical checks, and machine-readable results. The original prediction figure is retained as a stored vector asset, with its provenance distinguished from the new exact-feature implementation. Real-data protocols and numerical settings are documented separately.

ETHICS STATEMENT

The empirical studies reuse public NHANES, ESOL, OPERA, and gas-sensor data; no new data were collected. Users of these datasets should follow the access, privacy, and reporting conditions specified by their respective data providers.

AI USE STATEMENT

We used generative AI tools to assist with writing and language polishing and with developing and checking mathematical proofs. We take responsibility for the mathematical arguments, code, experimental results, and final manuscript.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Birkhäuser Basel, 2005.

Satarupa Bhattacharjee and Hans-Georg Müller. Single index Fréchet regression. *The Annals of Statistics*, 51(4):1770–1798, 2023.

Javier Burgus. Gas sensor array temperature modulation. UCI Machine Learning Repository, 2018.

Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. Wasserstein regression. *Journal of the American Statistical Association*, 118(542):869–882, 2023.

Yongxin Chen, Tryphon T. Georgiou, and Allen Tannenbaum. Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278, 2018.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical optimal transport*. Springer, 2025.

Alexander Cloninger, Keaton Hamm, Varun Khurana, and Caroline Moosmüller. Linearized Wasserstein dimensionality reduction with approximation guarantees. *Applied and Computational Harmonic Analysis*, 74:101718, 2025.

Evangelina Dardati, Thibault Laurent, Paula Margaretic, Ean Paredes, and Christine Thomas-Agnan. Accounting for the full distribution of temperature to predict international migration. TSE Working Paper 26–1728, Toulouse School of Economics, 2026.

John S. Delaney. ESOL: Estimating aqueous solubility directly from molecular structure. *Journal of Chemical Information and Computer Sciences*, 44(3):1000–1005, 2004.

Julie Delon and Agnes Desolneux. A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G. Schwing. Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019.

Rémi Flamary, Karim Lounici, and André Ferrari. Concentration bounds for linear Monge mapping estimation and optimal transport domain adaptation. *arXiv preprint arXiv:1905.10155*, 2019. Revised 2020.

P. Thomas Fletcher. Geodesic regression and the theory of least squares on Riemannian manifolds. *International Journal of Computer Vision*, 105(2):171–185, 2013.

Rahul Ghosal and Marcos Matabuena. Multivariate scalar on multidimensional distribution regression with application to modeling the association between physical activity and cognitive functions. *Biometrical Journal*, 66(7):e202400042, 2024.

Rahul Ghosal, Vijay R. Varma, Dmitri Volfson, Jacek Urbanek, Jeffrey M. Hausdorff, Amber Watts, and Vadim Zipunnikov. Scalar on time-by-distribution regression and its application for modelling associations between daily-living physical activity and cognitive functions in Alzheimer's disease. *Scientific Reports*, 12(1):11558, 2022.

Rahul Ghosal, Vijay R. Varma, Dmitri Volfson, Inbar Hillel, Jacek Urbanek, Jeffrey M. Hausdorff, Amber Watts, and Vadim Zipunnikov. Distributional data analysis via quantile functions and its application to modeling digital biomarkers of gait in alzheimer's disease. *Biostatistics*, 24(3): 539–561, 2023.

Rahul Ghosal, Sujit K. Ghosh, Jennifer A. Schrack, and Vadim Zipunnikov. Distributional outcome regression via quantile functions and its application to modelling continuously monitored heart rate and physical activity. *Journal of the American Statistical Association*, 120(551):1347–1359, 2025.

Sonja Greven and Fabian Scheipl. A general framework for functional regression modelling. *Statistical Modelling*, 17(1–2):1–35, 2017.

Nicolas Guigui and Xavier Pennec. Numerical accuracy of ladder schemes for parallel transport on manifolds. *Foundations of Computational Mathematics*, 22:757–790, 2022.

Paulína Jašková, Javier Palarea-Albaladejo, Aleš Gába, Dorothea Dumuid, Željko Pedišić, Jana Pelclová, and Karel Hron. Compositional functional regression and isotemporal substitution analysis: Methods and application in time-use epidemiology. *Statistical Methods in Medical Research*, 32 (10):2064–2080, 2023.

Dinesh John, Qu Tang, Fahd Albinali, and Stephen Intille. An open-source monitor-independent movement summary for accelerometer data processing. *Journal for the Measurement of Physical Behaviour*, 2(4):268–281, 2019.

Varun Khurana, Harish Kannan, Alexander Cloninger, and Caroline Moosmüller. Supervised learning of sheared distributions using linearized optimal transport. *Sampling Theory, Signal Processing, and Data Analysis*, 21(1):1, 2023.

Benoît Kloeckner. A geometric study of Wasserstein spaces: Euclidean spaces. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 9(2):297–323, 2010.

Soheil Kolouri, Yang Zou, and Gustavo K. Rohde. Sliced Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5258–5267, 2016.

Andrew Leroux, Erjia Cui, Ekaterina Smirnova, John Muschelli, Jennifer A. Schrack, and Ciprian M. Crainiceanu. NHANES 2011–2014: Objective physical activity is the strongest predictor of all-cause mortality. *Medicine and Science in Sports and Exercise*, 56(10):1926–1934, 2024.

Zhenhua Lin and Hans-Georg Müller. Total variation regularized Fréchet regression for metric-space valued data. *The Annals of Statistics*, 49(6):3510–3533, 2021.

Kamel Mansouri, Chris M. Grulke, Richard S. Judson, and Antony J. Williams. OPERA models for predicting physicochemical properties and environmental fate endpoints. *Journal of Cheminformatics*, 10:10, 2018.

Marcos Matabuena, Paulo Félix, Ziad Akram Ali Hammouri, Jorge Mota, and Borja del Pozo Cruz. Physical activity phenotypes and mortality in older adults: a novel distributional data analysis of accelerometry in the NHANES. *Aging Clinical and Experimental Research*, 34(12):3107–3114, 2022.

Marcos Matabuena, Rahul Ghosal, Javier Enrique Aguilar, Ayya Keshet, Robert Wagner, Carmen Fernández Merino, Juan Sánchez Castro, Vadim Zipunnikov, Jukka-Pekka Onnela, and Francisco Gude. Glucodensity functional profiles outperform traditional continuous glucose monitoring metrics. *Scientific Reports*, 15(1):33662, 2025.

Grégoire Mialon, Dexiong Chen, Alexandre d'Aspremont, and Julien Mairal. A trainable optimal transport embedding for feature aggregation and its relationship to attention. In *International Conference on Learning Representations*, 2021.

Caroline Moosmüller and Alexander Cloninger. Linear optimal transport embedding: provable Wasserstein classification for certain rigid transformations and perturbations. *Information and Inference: A Journal of the IMA*, 12(1):363–389, 2023.

Jeffrey S. Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2(1): 321–359, 2015.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1–2):1–141, 2017.

National Center for Health Statistics. National Health and Nutrition Examination Survey 2013–2014 data documentation, codebook, and frequencies: Physical activity monitor – minute (PAXMIN_H). Centers for Disease Control and Prevention, 2020. First published November 2020; last revised October 2022.

Pratim Guha Niyogi, Muraleetharan Sanjayan, Kathryn C. Fitzgerald, Ellen M. Mowry, and Vadim Zipunnikov. Scalar-on-distribution regression via generalized odds with applications to accelerometry-assessed disability in multiple sclerosis. *arXiv preprint arXiv:2601.09126*, 2026.

Victor M. Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. Springer, 2020.

Alexander Petersen and Hans-Georg Müller. Functional data analysis for density functions by transformation to a Hilbert space. *The Annals of Statistics*, 44(1):183–218, 2016.

Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics*, 47(2):691–719, 2019.

François Pitié and Anil Kokaram. The linear Monge-Kantorovitch linear colour mapping for example-based colour transfer. In *IET 4th European Conference on Visual Media Production*, pp. 1–9. IET, 2007.

Zoltán Szabó, Bharath K. Sriperumbudur, Barnabás Póczos, and Arthur Gretton. Learning theory for distribution regression. *Journal of Machine Learning Research*, 17(152):1–40, 2016.

Asuka Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48 (4):1005–1026, 2011.

Renáta Talská, Karel Hron, and Tomáš Matys Grygar. Compositional scalar-on-function regression with application to sediment particle size distributions. *Mathematical Geosciences*, 53(7):1667–1695, 2021.

Yann Thanwerdas and Xavier Pennec. O(n)-invariant Riemannian metrics on SPD matrices. *Linear Algebra and Its Applications*, 661:163–201, 2023.

Huong Thi Trinh, Christine Thomas-Agnan, and Michel Simioni. Scalar-on-distribution regression for assessing the impact of climate change on rice yield in Vietnam. *Journal of Agricultural, Biological and Environmental Statistics*, 2026.

Wei Wang, Dejan Slepčev, Saurav Basu, John A. Ozolek, and Gustavo K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International Journal of Computer Vision*, 101(2):254–269, 2013.

Hojin Yang, Veerabhadran Baladandayuthapani, Arvind UK Rao, and Jeffrey S. Morris. Quantile function on scalar regression analysis for distributional data. *Journal of the American Statistical Association*, 115(529):90–106, 2020.

Ho Yun and Yoav Zemel. Gaussian optimal transport beyond Brenier's theorem. *arXiv preprint arXiv:2512.21464*, 2025.

A ADDITIONAL NHANES DETAILS

Accelerometry distributions and activity summaries have been used to study mortality in NHANES (Matabuena et al., 2022; Leroux et al., 2024). Our application uses released age as a prediction target and evaluates the reference-dependent representation.

Figure 3 gives a descriptive visual summary of the retained location–covariance information; the exact construction follows.

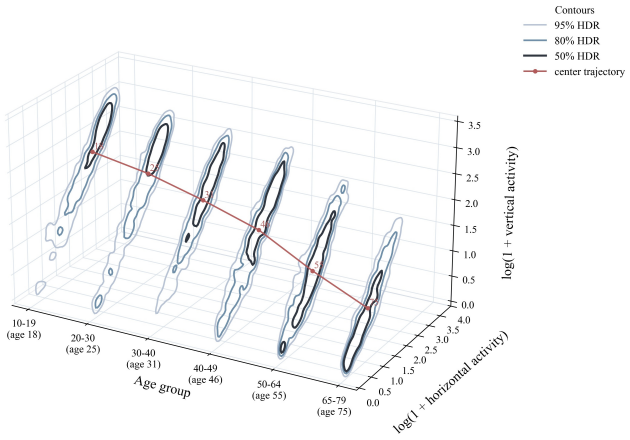

Figure 3: Illustrative subject-level density contours of log horizontal and vertical activity for selected participants from six age groups. Curves show the 50%, 80%, and 95% high-density regions.

We retain NHANES subjects with at least 600 valid wake minutes and cap each record at 3000 minutes. Each retained minute is encoded by the three axis-specific MIMS components, log triaxial MIMS activity, and sine/cosine recording-relative daily-phase features. The public age label is top-coded at 80 years. The subject-specific empirical measure is

$$P_i = n_i^{-1} \sum_{t=1}^{n_i} \delta_{X_{it}}, \qquad X_{it} \in \mathbb{R}^6.$$

We approximate P_i by a Gaussian and append sleep fraction and sex after the distributional embedding.

The 7597 subjects are split once with seed 42 into 6077 training and 1520 test subjects. Minute-level features, sleep fraction, and sex are standardized using training subjects only. The fixed reference is the equal-weight Gaussian Bures–Wasserstein barycenter of the 6077 training distributions; its covariance fixed-point iteration converges in seven steps at tolerance 10^{-9}. The optimized model fixes this barycenter mean, jointly profiles the two ordinary covariates with the tangent coordinates, and changes only the reference covariance. No test outcome enters preprocessing, reference construction, or optimization.

The five-method test comparison is in Table 2; supplementary RMSE values and the training-selected ridge penalty are in Tables 4 and 3, respectively. All five models include sleep fraction and sex.

B GEOMETRIC ORIGIN AND SCOPE OF THE LOCAL EXPANSION

Theorem 3 (Local curvature discrepancy)**.** *Suppose ρ, $\rho_\star$, and M lie in a common normal neighborhood. Using the curvature-sign convention of Guigui & Pennec (2022), and writing Rm_ρ for the Riemann curvature tensor, the neighboring-log expansion gives*

$$\widetilde{v}_\rho(M) = v_\rho(M) - \Delta_\rho + \frac{1}{6}\mathrm{Rm}_\rho(\Delta_\rho, v_\rho(M))\{2v_\rho(M) - \Delta_\rho\} + O_4, \tag{13}$$

where O_4 collects terms of joint degree at least four in $(\Delta_\rho, v_\rho(M))$. After the constant $-\Delta_\rho$ is absorbed by the intercept, this exact scalar discrepancy has the local expansion

$$b_\rho(M) = \frac{1}{6}\langle \beta_\rho, \mathrm{Rm}_\rho(\Delta_\rho, v_\rho(M))\{2v_\rho(M) - \Delta_\rho\}\rangle_\rho + \langle \beta_\rho, O_4\rangle_\rho. \tag{14}$$

The expansion in Theorem 3 is the neighboring-log expansion, not a claim that every pointwise discrepancy is irreducible. For a Riemannian manifold, take $p \in \mathcal{M}$ and sufficiently small $a, z \in T_p\mathcal{M}$. Write $p_a = \exp_p(a)$ and $p_z = \exp_p(z)$. The neighboring logarithm is

$$\ell_p(a, z) = \mathcal{P}_{p_a \to p}\mathrm{Log}_{p_a}(p_z).$$

Equation (2) of Guigui & Pennec (2022), based on the covariant Baker–Campbell–Hausdorff expansion, gives

$$\ell_p(a, z) = z - a + \frac{1}{6}\mathrm{Rm}_p(a, z)(2z - a) + O_4(a, z),$$

where the coefficients in O_4 are locally bounded by curvature and its covariant derivatives. Substituting $p = \rho$, $p_a = \rho_\star$, and $p_z = M$ proves Equation 13. Parallel transport is an isometry, so the signal in Equation 3 can be evaluated at ρ using β_ρ. The $-\Delta_\rho$ term is constant in M and is absorbed by the intercept. Taking the inner product of the remaining terms with β_ρ gives Equation 14.

For Gaussian covariance matrices, Bures–Wasserstein geometry has nonnegative sectional curvature and is generally non-flat (Takatsu, 2011). The important exception is a simultaneously diagonalizable family. In a common eigenbasis, the Bures–Wasserstein distance reduces to Euclidean distance between coordinatewise covariance square roots. Along such commuting directions the relevant curvature contraction vanishes, and the leading structural discrepancy is zero. Outside this regime, the expansion identifies a curvature-controlled nonlinear term, but irreducibility is determined only after projection onto the fitted feature span, as in Theorem 1.

C Proofs of theoretical results

This section proves the standard-reference equivalence, the profiled-risk bridge and empirical reference consistency. The BW local-risk proofs are in Appendix F.

C.1 Proof of the standard-reference relationship

Proof. Let $\mu = \mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$ be an arbitrary distribution in the manifold $\mathcal{G}_d$. The Euclidean embedding feature map is defined directly by the parameters of the distribution,

$$z_{EE}(\mu) = \begin{bmatrix} \boldsymbol{m} \\ \mathrm{vech}(\boldsymbol{\Sigma}^{1/2}) \end{bmatrix}.$$

At a reference $\rho = \mathcal{N}(\boldsymbol{m}_\rho, \boldsymbol{\Sigma}_\rho)$, the tangent-space feature uses the affine optimal transport map $T(x) = \boldsymbol{A}x + \boldsymbol{b}$. Its vectorized form is

$$z_\rho(\mu) = \begin{bmatrix} \boldsymbol{m} - \boldsymbol{A}\boldsymbol{m}_\rho \\ \mathrm{vech}(\boldsymbol{A} - \boldsymbol{I}_d) \end{bmatrix},$$

where $\boldsymbol{A} = \boldsymbol{\Sigma}_\rho^{-1/2}(\boldsymbol{\Sigma}_\rho^{1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}_\rho^{1/2})^{1/2}\boldsymbol{\Sigma}_\rho^{-1/2}$.

Substituting the standard normal reference $\rho = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ gives $\boldsymbol{m}_\rho = \boldsymbol{0}$ and $\boldsymbol{\Sigma}_\rho = \boldsymbol{I}_d$. The transport matrix $\boldsymbol{A}$ simplifies to

$$\boldsymbol{A} = \boldsymbol{I}_d^{-1/2}(\boldsymbol{I}_d^{1/2}\boldsymbol{\Sigma}\boldsymbol{I}_d^{1/2})^{1/2}\boldsymbol{I}_d^{-1/2} = \boldsymbol{\Sigma}^{1/2}.$$

Therefore, the tangent-space feature at this reference is

$$z_{\mathcal{N}(0,I_d)}(\mu) = \begin{bmatrix} \boldsymbol{m} - \boldsymbol{\Sigma}^{1/2}\boldsymbol{0} \\ \mathrm{vech}(\boldsymbol{\Sigma}^{1/2} - \boldsymbol{I}_d) \end{bmatrix} = \begin{bmatrix} \boldsymbol{m} \\ \mathrm{vech}(\boldsymbol{\Sigma}^{1/2}) \end{bmatrix} - \begin{bmatrix} \boldsymbol{0} \\ \mathrm{vech}(\boldsymbol{I}_d) \end{bmatrix}.$$

Let $\boldsymbol{c} = [\boldsymbol{0}^\top, \mathrm{vech}(\boldsymbol{I}_d)^\top]^\top$. Then

$$z_{\mathcal{N}(0,I_d)}(\mu) = z_{EE}(\mu) - \boldsymbol{c}.$$

This is an affine translation $\mathcal{T}(x) = x - \boldsymbol{c}$, which is invertible.

For linear regression with an intercept, the feature matrix is augmented by a column of ones $\mathbf{1}$. Subtracting $\boldsymbol{c}$ from every feature vector is an elementary column operation involving the intercept, so $\mathrm{Col}(\mathbf{Z}_{TSP}) = \mathrm{Col}(\mathbf{Z}_{EE})$. Consequently, the corresponding projection matrices satisfy $\mathbf{H}_{TSP} = \mathbf{H}_{EE}$, and the fitted values $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ are identical. □

C.2 Reference-mean invariance

For a fixed reference covariance, the transport matrices $A_{\rho,i}$ do not depend on the reference mean. Writing $h = m'_\rho - m_\rho \in \mathbb{R}^d$, the changed location block satisfies

$$m_i - A_{\rho,i} m'_\rho = (m_i - A_{\rho,i} m_\rho) - (A_{\rho,i} - I_d)h - h.$$

Each coordinate of $(A_{\rho,i} - I_d)h$ is a linear combination of the full covariance feature block, and the intercept absorbs the constant $-h$. Replacing h by $-h$ gives the inverse transformation. Thus the augmented feature span and unregularized profiled RSS are unchanged. This argument requires the full block and does not apply automatically to a restricted coordinate set.

C.3 Proof of Theorem 1

Proof. Parallel-transport isometry and the definition of b_ρ rewrite the conditional mean in Equation 3 as

$$f_\star(M) = \alpha_\rho + \langle \beta_\rho, v_\rho(M) \rangle_\rho + b_\rho(M), \qquad \alpha_\rho = \alpha_\star - \langle \beta_\rho, \Delta_\rho \rangle_\rho.$$

The first two terms belong to $\mathcal{H}_\rho$. For any $h \in \mathcal{H}_\rho$, conditional mean-zero noise gives the orthogonal decomposition

$$\mathbb{E}\{Y - h(M)\}^2 = \sigma^2 + \|f_\star - h\|^2_{L^2(P_M)}.$$

Profiling over h is therefore equivalent to projecting b_ρ onto $\mathcal{H}_\rho$, which yields $\mathcal{R}(\rho) = \sigma^2 + \|(I - \Pi_\rho) b_\rho\|^2$. At $\rho = \rho_\star$, $\Delta_\rho = 0$ and the transported and direct logarithms coincide, so $b_{\rho_\star} = 0$. □

For Equation 11, condition on all fitting and selection data so that $\widehat{\rho}$ and $\widehat{f}$ are fixed relative to the independent test draw. Decompose $f_\star - \widehat{f} = (I - \Pi_{\widehat{\rho}}) f_\star + (\Pi_{\widehat{\rho}} f_\star - \widehat{f})$. The two summands are orthogonal because the second belongs to $\mathcal{H}_{\widehat{\rho}}$. Adding conditional mean-zero response noise proves the identity, and Theorem 1 identifies the first squared norm with $\mathcal{B}(\widehat{\rho})$.

C.4 Uniform consistency of empirical reference selection

Proposition 1 (Consistency of empirical reference selection)**.** *Let (M_i, Y_i) be i.i.d., let $\mathcal{K}$ be compact, omit structurally constant coordinates, and write $x_\rho = [1, z_\rho(M)^\top]^\top$ and $G_\rho = \mathbb{E}(x_\rho x_\rho^\top)$. Assume x_ρ is measurable and almost surely continuous in ρ, $\mathbb{E}\{\sup_{\rho \in \mathcal{K}} \|x_\rho\|^2 + Y^2\} < \infty$, and $\inf_{\rho \in \mathcal{K}} \lambda_{\min}(G_\rho) > 0$. Then $\sup_{\rho \in \mathcal{K}} |\widehat{\mathcal{R}}_n(\rho) - \mathcal{R}(\rho)| \xrightarrow{p} 0$. Consequently, if $\widehat{\mathcal{R}}_n(\widehat{\rho}_n) \le \inf_{\rho \in \mathcal{K}} \widehat{\mathcal{R}}_n(\rho) + o_p(1)$, then $\mathrm{dist}\{\widehat{\rho}_n, \arg\min_{\rho \in \mathcal{K}} \mathcal{R}(\rho)\} \xrightarrow{p} 0$. The conclusion also holds for plug-in distributions whose empirical Gram and response-cross moments are uniformly $o_p(1)$-close to the direct versions.*

The compact class and moment envelope control the empirical objective uniformly, while nonsingularity keeps the profiled coefficients stable. The result requires approximate global minimization and does not follow from local stationarity alone. For estimated input distributions, the uniform moment approximation is an additional assumption, not a fixed-within-distribution-sample guarantee.

Proof. Let $x_\rho(M) = [1, z_\rho(M)^\top]^\top$, and define $G_\rho = \mathbb{E}[x_\rho x_\rho^\top]$, $c_\rho = \mathbb{E}[x_\rho Y]$, and $s = \mathbb{E}(Y^2)$. Uniform nonsingularity gives

$$\mathcal{R}(\rho) = s - c_\rho^\top G_\rho^{-1} c_\rho.$$

The empirical profiled risk has the same expression with empirical moments. Almost-sure continuity on the compact set and the integrable squared envelope give a uniform law of large numbers for the Gram matrices. The corresponding result for c_ρ follows from Cauchy–Schwarz and $\mathbb{E}(Y^2) <$

∞, while the ordinary law of large numbers applies to s. Uniform nonsingularity makes matrix inversion uniformly continuous over the resulting Gram matrices, hence $\sup_{\rho\in\mathcal{K}} |\widehat{\mathcal{R}}_n(\rho) - \mathcal{R}(\rho)| \xrightarrow{p} 0$. Compactness and continuity make the population argmin set nonempty and ensure separation away from it. The stated $o_p(1)$ empirical optimality condition and the standard argmin theorem then give

$$\operatorname{dist}\{\widehat{\rho}_n, \arg\min_{\rho\in\mathcal{K}} \mathcal{R}(\rho)\} \xrightarrow{p} 0.$$

If distributions are estimated, the assumed uniform $o_p(1)$ perturbations of the empirical Gram and response-cross-moment functions preserve the same uniform risk convergence and therefore the same argmin conclusion. □

D Proof of convergence for alternating optimization

Theorem 4 (Convergence of alternating optimization)**.** *Consider the exact-arithmetic algorithm without finite stopping or backtracking caps, with fixed $\bar{\eta} > 0$ and $\beta, c \in (0,1)$. Assume its covariance iterates lie in a compact subset of $\{\Sigma : \lambda_{\min}(\Sigma) > \epsilon_{\mathrm{spd}}\}$, with uniformly positive-definite augmented Gram matrices on a fixed retained feature set. Assume $F_{\mathcal{I}}$ has a locally Lipschitz gradient on an open neighborhood of this set. Profiled Armijo backtracking makes the objective nonincreasing, and every accumulation point, paired with its least-squares coefficients, is stationary for Equation 10 on these coordinates.*

Write $F = F_{\mathcal{I}}$ and let $\mathcal{J}_{n,\mathcal{I}}$ be the restriction of Equation 10 to $\boldsymbol{w}_{\mathcal{I}^c} = 0$. The vector $\theta(\Sigma)$ combines the fitted intercept and retained coefficients; $\mathcal{I}$ is fixed throughout optimization. The result concerns exact features and exact least squares; finite numerical approximations are distinguished below.

Proof. Let $X(\Sigma)$ be the augmented design matrix. Gaussian features are smooth on the SPD cone. Uniform positive definiteness of $n^{-1}X(\Sigma)^\top X(\Sigma)$ on the compact iterate set K therefore extends to a neighborhood of K. The formula $\theta(\Sigma) = (X^\top X)^{-1}X^\top y$ makes the coefficient map smooth there and bounded on K. The normal equations and the chain rule give

$$\nabla_\theta \mathcal{J}_{n,\mathcal{I}}(\Sigma, \theta(\Sigma)) = 0, \qquad E(\Sigma) := \nabla F(\Sigma) = \nabla_\Sigma \mathcal{J}_{n,\mathcal{I}}(\Sigma, \theta(\Sigma)). \tag{15}$$

All matrix gradients use the Frobenius inner product on symmetric matrices. For $C_\Sigma(\eta) = (I - \eta E)\Sigma(I - \eta E)$, direct differentiation gives

$$\frac{d}{d\eta} F(C_\Sigma(\eta))\Big|_{\eta=0} = -\langle E, E\Sigma + \Sigma E\rangle_F = -2\operatorname{tr}(\Sigma E^2) = -D(\Sigma).$$

In the BW metric, $\operatorname{grad} F = 2(E\Sigma + \Sigma E)$. To verify this, for $U = V\Sigma + \Sigma V$ with symmetric V, the metric pairing equals $2\operatorname{tr}(E\Sigma V) = \langle E, U\rangle_F$. Thus the initial curve velocity is half the negative BW gradient, although the additional positive-factor check in Algorithm 1 keeps the simulation steps in its exponential-map domain. The endpoint-only numerical variant also has this initial velocity.

We next verify a uniform sufficient step size. Compactness gives constants $m > \epsilon_{\mathrm{spd}}$, $M < \infty$ and $H < \infty$ such that $mI \preceq \Sigma \preceq MI$ and $\|E(\Sigma)\|_F \le H$ on K. If $H = 0$, all iterates are already stationary. Otherwise, choose a uniform radius $r > 0$ and Lipschitz bound L for ∇F on neighborhoods of K. For $0 \le \eta \le \bar{\eta}$, put $A = 2M + \bar{\eta}MH$. Then

$$\Delta = C_\Sigma(\eta) - \Sigma = -\eta(E\Sigma + \Sigma E) + \eta^2 E\Sigma E, \qquad \|\Delta\|_F \le A\eta\|E\|_F.$$

For all sufficiently small η, uniformly over K, this displacement is less than both r and $(m - \epsilon_{\mathrm{spd}})/2$. Hence the segment from Σ to $C_\Sigma(\eta)$ stays in the smooth neighborhood, and the candidate passes the eigenvalue check. Requiring also $\eta H < 1$ makes $I - \eta E \succ 0$ uniformly, so the extra simulation feasibility check preserves the lower step-size bound. Taylor's inequality gives

$$F(C_\Sigma(\eta)) \le F(\Sigma) - \eta D(\Sigma) + \eta^2 B\|E\|_F^2, \qquad B = MH + LA^2/2,$$

where $|\operatorname{tr}(\Sigma E^3)| \le MH\|E\|_F^2$ bounds the quadratic term in $\langle E, \Delta\rangle_F$. Since $D(\Sigma) \ge 2m\|E\|_F^2$, there is a uniform $\eta_0 > 0$ such that every $\eta \le \eta_0$ is feasible and satisfies Armijo; it suffices also to require $\eta B \le 2m(1-c)$. The geometric search consequently accepts $\eta_t \ge \min\{\bar{\eta}, \beta\eta_0\} > 0$.

Summing Equation 12, using $F \geq 0$, now gives $c\sum_t \eta_t D_t \leq F(\Sigma_\rho^{(0)})$. Thus

$$D_t \longrightarrow 0, \qquad \|E_t\|_F \longrightarrow 0, \qquad \|\Sigma_\rho^{(t+1)} - \Sigma_\rho^{(t)}\|_F \longrightarrow 0.$$

For any subsequence $\Sigma_\rho^{(t_j)} \to \Sigma^*$, continuity gives $\theta(\Sigma_\rho^{(t_j)}) \to \theta(\Sigma^*)$ and $E(\Sigma^*) = 0$. Equation 15 then makes both blocks of the restricted joint gradient zero at $(\Sigma^*, \theta(\Sigma^*))$. The same limit holds for coefficients carried from the preceding iterate, by the vanishing displacement and uniform continuity of θ on K. This proves the claimed stationary-accumulation result, not convergence to a global minimizer. □

The mean-invariance argument and Proposition 1 concern the full feature model. They do not automatically extend to an arbitrary sample-selected mask. A column constant at initialization can become active at another reference. The stationary result above is conditional on a fixed $\mathcal{I}$; it is a full-model statement when all feature coordinates are retained.

E Implementation and numerical checks

E.1 Exact-feature simulation implementation

The additional simulations in Appendix H use double precision, spectral SPD roots, and SVD least squares with an unpenalized intercept. Orthonormal half-vectorization is an invertible rescaling of the coordinates in Equation 2 and leaves the unregularized affine span unchanged. Columns constant in the current design are handled by centering and SVD; no initial feature mask is frozen across changing references. These additional prediction experiments retain the full covariance block, whose population covariance is positive definite, and omit structurally zero means. The reference mean stays fixed.

For a covariance S, OT matrix A_i, and fitted effective symmetric coefficient W, the transport derivative obeys

$$(\mathrm{d}A_i)SA_i + A_iS(\mathrm{d}A_i) = -A_i(\mathrm{d}S)A_i.$$

Solve the adjoint equation $SA_iZ_i + Z_iA_iS = W$. If e_i denotes the fitted residual, the profiled Euclidean gradient is

$$E = 2n^{-1}\sum_i e_iA_iZ_iA_i.$$

Writing $C_i = S^{1/2}A_iS^{1/2}$ reduces this to an SPD Sylvester equation, which remains well-defined at repeated eigenvalues. For nonzero fixed reference mean, W includes the contribution from the affine shift coordinates. This avoids eigenvector differentiation and fixed-iteration square-root approximations. The gradient is recomputed after each accepted OLS refit.

The initial step is 1, Armijo constant 10^{-4}, contraction $1/2$, maximum backtracks 40, covariance floor 10^{-8}, and BW gradient tolerance 10^{-8}. Trials require both $I - \eta E \succ 0$ and an endpoint above the floor. The recorded BW gradient norm is $\sqrt{4\mathrm{tr}(SE^2)}$. The default update cap is 60. Numerically invalid current designs raise an error; invalid trials are rejected. Every termination reason is saved.

Seeded checks compare analytic gradients with central finite differences at identity and non-identity references, both with and without mean features. They also check the transport equation, mean-reference invariance, the commuting barycenter, and Armijo descent. The maximum directional-derivative error is below 10^{-10} and the maximum checked transport-equation residual is below 10^{-12}. Exact values and software versions are recorded in the supplementary files `verification.json` and `environment.json`.

The scripts `run_geometric_simulations.py` and `render_simulation_results.py` generate the additional numerical outputs, diagnostic figure, and LaTeX tables. The renderer restores Figure 2 from its stored PDF asset.

E.2 Oracle-reference prediction results

Figure 2 reports the results at the oracle reference $\rho_\star = \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ for the baseline labeled *Oracle*. Under this simulation setting, oracle-reference regression and Euclidean embedding are exactly

equivalent with an intercept, yielding identical fitted values and predictions (Remark 1). The protocol uses estimated Gaussian inputs, 20 replicates, and fixed update budgets. The plotted values and error bars are retained from the original simulation results, and the vector figure is included as `reference_correction_oracle.pdf`.

E.3 Numerical implementation for the real-data experiments

The fixed-split real-data results retain their recorded numerical implementation and settings. They use the same affine-map feature family and profiled OLS target, with automatic differentiation, fixed retained coordinate sets, and endpoint-feasible covariance updates. Unlike the simulation implementation, their search does not require $I - \eta E \succ 0$ and is not asserted to follow an admissible BW exponential for every accepted endpoint. The proof above also covers the endpoint-only descent variant under its compactness, smoothness, and fixed-coordinate Gram assumptions.

NHANES uses 30 Newton–Schulz square-root iterations, Armijo constant 10^{-4}, contraction $1/2$, and at most 20 backtracks. Its reference mean is fixed at the training mean, with sleep fraction and sex appended and profiled jointly. The retained feature mask stays fixed during optimization; prediction uses spectral features with a nonconstant OLS refit. The initial step is 0.02, cap 30, step tolerance 10^{-6}, and reference floor 10^{-6}. The run reaches its cap with minimum reference eigenvalue 1.15×10^{-6}. The conservative 10^{-3} floor retains a 9.08% MSE reduction against the training barycenter (Table 8). Molecular and sensing settings are given in Appendix G.

F Proofs of the local BW risk results

F.1 A local profiling lemma

Let $\boldsymbol{x}_t \in L^2(P;\mathbb{R}^{p+1})$ include an intercept, with $\boldsymbol{x}_t = \boldsymbol{x}_0 + t\dot{\boldsymbol{x}}_0 + o_{L^2}(t)$, $\mathbf{G}_0 = \mathbb{E}(\boldsymbol{x}_0\boldsymbol{x}_0^\top) \succ 0$, and $f_\star = \boldsymbol{w}_0^\top \boldsymbol{x}_0$. Set $q = \boldsymbol{w}_0^\top \dot{\boldsymbol{x}}_0$ and let Π_0 project onto the coordinates of $\boldsymbol{x}_0$. Differentiating the normal equations for the population OLS coefficient gives

$$\dot{\boldsymbol{w}}_0 = -\mathbf{G}_0^{-1}\mathbb{E}(\boldsymbol{x}_0 q), \qquad f_\star - \boldsymbol{w}_t^\top \boldsymbol{x}_t = -t(I - \Pi_0)q + o_{L^2}(t).$$

Products of the L^2 expansions can be integrated by Cauchy–Schwarz. Consequently the excess risk is $t^2\|(I - \Pi_0)q\|_{L^2}^2 + o(t^2)$. This identity describes the effect of refitting and will be applied to normalized BW coordinates below.

F.2 Differentiating the transport equation

Use the notation and boundedness assumptions in Equation 6. The symmetric positive definite solution of $\mathbf{A}_{t,r}\mathbf{S}_t\mathbf{A}_{t,r} = \mathbf{T}_r\mathbf{S}\mathbf{T}_r$ is smooth on this domain, with derivatives uniform over the bounded support of $\mathbf{U}$ in a sufficiently small neighborhood. At $t = 0$ it equals $\mathbf{T}_r$. Writing $\mathbf{V}_r = \partial_t \mathbf{A}_{t,r}|_{t=0}$ gives

$$\mathbf{V}_r\mathbf{S}\mathbf{T}_r + \mathbf{T}_r\mathbf{S}\mathbf{V}_r = -\mathbf{T}_r(\mathbf{H}\mathbf{S} + \mathbf{S}\mathbf{H})\mathbf{T}_r.$$

Matching powers in $\mathbf{V}_r = \mathbf{V}_0 + r\mathbf{V}_1 + r^2\mathbf{V}_2 + O(r^3)$ yields

$$\begin{aligned}
\mathcal{L}_{\mathbf{S}}(\mathbf{V}_0) &= -(\mathbf{H}\mathbf{S} + \mathbf{S}\mathbf{H}),\\
\mathcal{L}_{\mathbf{S}}(\mathbf{V}_1) &= -(\mathbf{U}\mathbf{H}\mathbf{S} + \mathbf{S}\mathbf{H}\mathbf{U}),\\
\mathcal{L}_{\mathbf{S}}(\mathbf{V}_2) &= -\mathbf{U}(\mathbf{H}\mathbf{S} + \mathbf{S}\mathbf{H})\mathbf{U} - \mathbf{V}_1\mathbf{S}\mathbf{U} - \mathbf{U}\mathbf{S}\mathbf{V}_1.
\end{aligned}$$

Since $\mathbf{S} \succ 0$, the Sylvester operator is invertible. The solutions are $-\mathbf{H}$, $-\mathbf{K}$, and $\mathbf{F}$, respectively, proving

$$\partial_t \mathbf{A}_{t,r}|_{t=0} = -\mathbf{H} - r\mathbf{K} + r^2\mathbf{F} + O(r^3). \tag{16}$$

At $\mathbf{S} = \mathbf{I}$, this becomes

$$\mathbf{K} = (\mathbf{U}\mathbf{H} + \mathbf{H}\mathbf{U})/2, \qquad \mathbf{F} = [\mathbf{U}, [\mathbf{U}, \mathbf{H}]]/4.$$

The constant and linear terms can be absorbed by refitting. The next proof establishes the risk remaining from the quadratic term, including a uniform joint remainder.

F.3 Proof of Theorem 2

For $r \neq 0$, use the equivalent normalized feature

$$\mathbf{\Phi}_{t,r} = (\mathbf{A}_{t,r} - \mathbf{D}_t^{-1})/r.$$

At $r = 0$, the transport equation gives $\mathbf{A}_{t,0} = \mathbf{D}_t^{-1}$, even when $\mathbf{S}$ and $\mathbf{D}_t$ do not commute. Thus $\mathbf{\Phi}_{t,r} = \int_0^1 \partial_r \mathbf{A}_{t,sr}\, \mathrm{d}s$ extends smoothly through $r = 0$. At $t = 0$ it is exactly $\mathbf{U}$ for all small r. At $r = 0$, its value $\mathbf{J} = \mathbf{\Phi}_{t,0}$ solves

$$\mathbf{J}\mathbf{D}_t\mathbf{S} + \mathbf{S}\mathbf{D}_t\mathbf{J} = \mathbf{U}\mathbf{S} + \mathbf{S}\mathbf{U}.$$

Both operators are invertible at $t = 0$ and remain invertible nearby. Thus the normalized feature span on either coordinate axis is exactly the span of the constant and the entries of $\mathbf{U}$.

Let $g(\mathbf{U}) = \mathrm{tr}(\mathbf{B}\mathbf{U})$, and let $e(t, r)$ be its residual after projection onto $(1, \mathrm{svec}(\mathbf{\Phi}_{t,r})^\top)^\top$. The covariance assumption makes the augmented Gram matrix at $(0, 0)$ positive definite. Continuity preserves a uniform lower bound nearby. Uniform smoothness over the bounded support implies that the Gram matrix, coefficients, and L^2-valued residual have bounded derivatives of all required orders. In particular, e is three times continuously differentiable and $e(0, r) = e(t, 0) = 0$. The fundamental theorem of calculus in both variables gives

$$e(t, r) = t\, r \int_0^1 \int_0^1 \partial_t \partial_r e(at, br)\, \mathrm{d}a\, \mathrm{d}b.$$

By Equation 16, $\partial_t \mathbf{\Phi}_{t,r}|_{t=0} = -\mathbf{K} + r\mathbf{F} + O(r^2)$. The profiling calculation above therefore gives

$$\partial_t e(0, r) = -(I - \Pi_U)\mathrm{tr}\{\mathbf{B}(-\mathbf{K} + r\mathbf{F})\} + O_{L^2}(r^2) = -r\widetilde{q} + O_{L^2}(r^2),$$

where $\widetilde{q} = (I - \Pi_U)\mathrm{tr}(\mathbf{B}\mathbf{F})$, since $\mathbf{K}$ is linear in $\mathbf{U}$. Hence $\partial_t \partial_r e(0, 0) = -\widetilde{q}$, and the bounded third derivatives imply

$$e(t, r) = -t\, r\widetilde{q} + \eta(t, r), \qquad \|\eta(t, r)\|_{L^2} \leq C|t|r(|t| + r).$$

The actual conditional mean is $\alpha_\star + rg$, so $R_r(t) - \sigma^2 = r^2\|e(t, r)\|_{L^2}^2$. Squaring proves Equation 8 with a uniform remainder. If $\widetilde{q} = 0$, only the squared remainder remains. If $\|\widetilde{q}\|_{L^2}^2 > 0$, shrink the rectangle until the relative remainder is at most half this coefficient to obtain the stated two-sided bounds. At $r = 0$ both sides of the risk expansion vanish.

The operator representation is valid at arbitrary positive definite $\mathbf{S}$. Whitening is not a BW isometry, as already seen from $d_{\mathrm{BW}}(a/s, b/s) = d_{\mathrm{BW}}(a, b)/\sqrt{s}$ in dimension one. Our calculation therefore does not transfer the identity formula by whitening. General SPD metric conventions and Sylvester representations are discussed by Thanwerdas & Pennec (2023).

F.4 Proof of Corollary 1

At the identity, cyclicity of trace and skew symmetry of commutators give

$$4\mathrm{tr}(\mathbf{B}\mathbf{F}) = \mathrm{tr}\{\mathbf{B}[\mathbf{U}, [\mathbf{U}, \mathbf{H}]]\} = \langle [\mathbf{H}, \mathbf{U}], [\mathbf{B}, \mathbf{U}] \rangle_F = Q.$$

Equation 9 follows from Theorem 2. For centrally symmetric $\mathbf{U}$, Q is even and its linear coordinates are odd, so its projection onto their affine span is $\mathbb{E}Q$. This proves the variance formula. If the density is positive on a nonempty open set and $\mathcal{C}_\mathbf{I} = 0$, then the quadratic polynomial Q agrees there almost everywhere with an affine polynomial. Their difference must be identically zero. Homogeneity then implies $Q \equiv 0$. Conversely, $Q \equiv 0$ gives zero coefficient. For $\mathbf{B} = \mathbf{H}$, $Q = \|[\mathbf{H}, \mathbf{U}]\|_F^2$ is nonzero whenever $\mathbf{H}$ has two distinct eigenvalues and $\mathbf{U}$ has an off-diagonal entry between the corresponding eigenspaces. This supplies the stated strictly positive case.

Positive coordinate covariance alone does not imply this nonabsorbability condition. For example, on $p + 1$ affinely independent support points, an affine model with p coordinates can represent any scalar function on the support, including a quadratic one. The support condition excludes such cancellations.

For the two-dimensional example, $Q = 8b^2$, $\mathbb{E}b^2 = 1/3$, and $\mathbb{E}b^4 = 1/5$. Thus $\mathcal{C}_\mathbf{I} = 4(1/5 - 1/9) = 16/45$. Replacing the off-diagonal entry by λb gives coefficient $16\lambda^4/45$ for each fixed $\lambda > 0$. Replacing $\mathbf{B}$ by $\cos(\theta)\mathbf{H} + \sin(\theta)\mathbf{I}$ gives coefficient $16\cos^2(\theta)/45$. These formulas underlie the mechanism experiments.

F.5 Exact equality in compatible commuting directions

If $[\mathbf{H}, \mathbf{U}] = 0$ almost surely, the product $\mathbf{T}_r \mathbf{D}_t^{-1}$ is symmetric positive definite and solves the transport equation. Consequently $r\mathbf{U} = \mathbf{D}_t \mathbf{A}_{t,r} - \mathbf{I}$. For $\mathbf{W}_t = (\mathbf{B}\mathbf{D}_t + \mathbf{D}_t\mathbf{B})/2$ and $a_t = \alpha_\star + \mathrm{tr}\{\mathbf{B}(\mathbf{D}_t - \mathbf{I})\}$,

$$a_t + \mathrm{tr}\{\mathbf{W}_t(\mathbf{A}_{t,r} - \mathbf{I})\} = \alpha_\star + \mathrm{tr}\{\mathbf{B}(\mathbf{D}_t\mathbf{A}_{t,r} - \mathbf{I})\} = \alpha_\star + r\mathrm{tr}(\mathbf{B}\mathbf{U}).$$

The conditional mean is represented exactly, and the profiled excess risk is zero. This argument needs neither full-dimensional support nor pairwise commutativity of every raw covariance matrix at a general $\mathbf{S}$.

G Real-data protocols and additional results

G.1 Evaluation protocol

All comparisons use the five methods defined in Section 4.3. Mean + SD OLS uses marginal means and standard deviations; Euclidean OLS uses means and upper-triangular covariance-square-root entries. Mean + Covariance RidgeCV uses $(m_i, \mathrm{vech}(\Sigma_i))$, with sleep fraction and sex included for NHANES as in the other models. Its penalty is selected among 15 values from 10^{-2} to 10^5 by analytic leave-one-out error after scaling fitted once on the training set. Fixed and optimized tangent OLS use Equation 2, identical preprocessing, and an intercept; only the reference covariance changes.

NHANES uses a prespecified schedule, BioHL uses training-only grouped CV, and ESOL, KOA, and Gas use validation-selected schedules.

MSE is the primary metric. RMSE is its square root; RMSE and MAE retain the response units, and R^2 is dimensionless. Relative MSE reduction is $100(1 - \mathrm{MSE}_{\mathrm{opt}}/\mathrm{MSE}_{\mathrm{baseline}})$. The strongest listed baseline is the comparator with the lowest reported MSE. Sample sizes, Gaussian dimensions, and ridge penalties are given in Table 3. Table 4 supplements the main test comparison with RMSE; Table 5 reports the corresponding validation metrics.

Table 3: Sample sizes, Gaussian dimension d, and training-selected ridge penalty λ. NHANES uses a prespecified optimization schedule without a validation partition.

Dataset	d	Train	Validation	Test	Ridge λ
NHANES	6	6077	—	1520	1.0000
ESOL	4	675	226	227	0.3162
BioHL	4	90	30	30	31.6228
KOA	4	162	54	54	0.3162
Gas	4	700	300	300	0.0316

Table 4: Test RMSE for the five methods in Table 2. Units are years (NHANES), ppm (Gas), and source-log units for the molecular tasks. Bold marks the lowest unrounded RMSE in each row. Sample sizes are in Table 3.

Dataset	Mean + SD	Euclidean	Mean+Cov. RidgeCV	Fixed BW	Optimized
NHANES	15.5240	14.5610	14.8445	14.1912	**13.4697**
ESOL	1.4176	1.3734	1.2958	1.3649	**1.2149**
BioHL	0.3732	0.3387	0.3499	0.3724	**0.2727**
KOA	1.0867	0.9948	1.1308	0.9209	**0.8713**
Gas	1.0779	1.1695	1.0445	1.0670	**0.9648**

G.2 Molecular construction and selection

ESOL contains 1128 records on the released logarithmic aqueous-solubility scale (Delaney, 2004). The curated OPERA endpoints are BioHL, $\log_{10}$ biodegradation half-life in days, and KOA, the

Table 5: Validation performance with the same five methods as Table 2. Validation sizes are ESOL 226, BioHL 30, KOA 54, and Gas 300. RMSE and MAE use source-log units for molecular tasks and ppm for Gas; MSE squares these units. Bold marks the best unrounded value within each task and metric.

Method	MSE ↓	RMSE ↓	MAE ↓	R^2 ↑
ESOL				
Mean + SD OLS	2.6105	1.6157	1.2564	0.4392
Euclidean OLS	2.4155	1.5542	1.1744	0.4811
Mean + Cov. RidgeCV	2.1357	1.4614	1.1142	0.5412
Fixed BW tangent OLS	2.3756	1.5413	1.1848	0.4896
Optimized tangent OLS	**1.8659**	**1.3660**	**1.0076**	**0.5991**
BioHL				
Mean + SD OLS	0.2866	0.5353	0.4047	0.4316
Euclidean OLS	0.3479	0.5898	0.4071	0.3100
Mean + Cov. RidgeCV	0.2889	0.5375	0.3475	0.4270
Fixed BW tangent OLS	0.4266	0.6531	0.4555	0.1539
Optimized tangent OLS	**0.1838**	**0.4287**	**0.3012**	**0.6355**
KOA				
Mean + SD OLS	1.2280	1.1082	0.8399	0.8498
Euclidean OLS	0.8461	0.9198	0.6958	0.8965
Mean + Cov. RidgeCV	1.1848	1.0885	0.8116	0.8551
Fixed BW tangent OLS	0.7786	0.8824	**0.6610**	0.9048
Optimized tangent OLS	**0.7507**	**0.8664**	0.6652	**0.9082**
Gas				
Mean + SD OLS	0.5377	0.7333	0.5496	0.9868
Euclidean OLS	0.3165	0.5626	0.4315	0.9922
Mean + Cov. RidgeCV	0.4331	0.6581	0.5074	0.9894
Fixed BW tangent OLS	0.2853	0.5341	0.4200	0.9930
Optimized tangent OLS	**0.2682**	**0.5179**	**0.3981**	**0.9934**

$\log_{10}$ octanol–air partition coefficient (Mansouri et al., 2018). Each molecule has one hydrogen-added ETKDGv3 conformer, with an initial UFF budget of 200 iterations. Its atoms form a bag of four descriptors, consisting of atomic number, distance to the geometric centroid, distance to the nearest other atom, and mean distance to other atoms. This representation is invariant to atom permutation and rigid transformations; it is a distribution of atoms within one conformer, not a distribution over conformers. The common transformation is fitted on training bags.

Canonical SMILES groups are randomly assigned with seed 42 to approximately 60/20/20 training/validation/test partitions. Identical canonical structures stay together.

ESOL and KOA use validation-selected 30-update schedules with initial step 5. BioHL uses three-fold GroupKFold within the training set, with shuffled groups at seed 20260915. Every fold refits scaling, the barycenter, and the regressions. Pooled out-of-fold MSE selects among zero updates and budgets $\{30, 100\}$ crossed with initial steps $\{0.05, 0.5, 5\}$. The selected schedule is $(30, 0.5)$, as shown in Table 6. The primary conformer condition uses ETKDG seed 23 plus the source-row index; alternative conformers assess sensitivity.

Table 6: BioHL MSE across optimization budgets ($d = 4$). Bold identifies the training-CV-selected schedule; nonzero budgets use initial step 0.5. Errors are in squared source-log units.

Update cap	Train	Inner CV	Validation
0	0.1957	0.7636	0.4266
30	**0.1523**	**0.3748**	**0.1838**
100	0.1443	0.3950	0.1837

G.3 GAS-SENSOR CONSTRUCTION AND SELECTION

The gas-sensor temperature-modulation data (UCI 487) (Burgus, 2018) comprise 13 dated files with 100 exposures per file. The first seven files supply 700 training bags, the next three supply 300 validation bags, and the final three supply 300 test bags. For exposure $k = 1, \ldots, 100$, the fixed interval $[900k + 300, 900k + 600)$ seconds supplies the bag, subject to at least 800 valid points; all 300 test exposures satisfy this criterion. The 14 log-resistance channels are standardized and projected by training-fitted PCA to dimension four. The response is mean CO concentration in ppm over the same window, without using concentration to choose the window or predictors. The task is concurrent sensor calibration.

Gas selects on validation among the six nonzero schedules in Table 7 and a zero-update candidate. The selected schedule is $(100, 5)$; optimization stops after 84 updates by its reference-step tolerance. The terminal minimum eigenvalue is approximately 1.00×10^{-6}, at the imposed SPD floor.

Table 7: Gas schedule comparison on the same 300 validation exposures. Entries are RMSE in ppm; zero-update performance is in Table 5.

Maximum updates / initial step	Validation RMSE
30 / 0.05	0.534048
30 / 0.5	0.533232
30 / 5	0.525051
100 / 0.05	0.533841
100 / 0.5	0.530974
100 / 5	0.517864

G.4 REFERENCE AND REPRESENTATION SENSITIVITY

Table 8(A) compares validation and test performance of the same fitted predictors. BioHL's MSE reduction changes from 56.92% to 46.39%, whereas Gas's changes from 6.00% to 18.24%. References, coefficients, preprocessing, and schedules are identical between the two evaluations; only the evaluation sets differ. Panel (B) records the NHANES SPD-floor sensitivity described in Appendix E.

The four BioHL conditions in Table 8(C) share the outer split. They use ETKDG seeds 23, 10023, and 20023 plus the source-row index, and a refinement of the seed-23 condition with up to 800 additional UFF iterations for non-converged structures. Among the 120 training and validation molecules, UFF non-convergence counts are 16, 15, 14, and 0, respectively. For seed 10023, optimized RMSE is 0.5005 versus 0.4888 for Mean + SD OLS.

Table 8: Additional reference-learning diagnostics. (A) Validation and test MSE reductions for the same fitted models. (B) NHANES sensitivity to the SPD floor; Δ is the test MSE reduction. (C) BioHL validation RMSE across conformer conditions, with training CV within each condition. All reductions are relative to matched fixed BW; BioHL errors use source-log units.

(A) Validation-to-test transfer

Dataset	Cap / initial step	Validation reduction (%)	Test reduction (%)
BioHL	30 / 0.5	56.92	46.39
KOA	30 / 5	3.58	10.49
Gas	100 / 5	6.00	18.24

(B) Sensitivity to the SPD floor

Floor / model	Updates	MSE (yr^2) $\downarrow$	MAE (yr) $\downarrow$	R^2 $\uparrow$	Δ (%)
Fixed BW	—	201.3899	11.5544	0.6317	—
10^{-8}	13	181.4293	10.8229	0.6682	9.91
10^{-6}	30	181.4332	10.8224	0.6682	9.91
10^{-5}	13	181.6387	10.8315	0.6678	9.81
10^{-4}	11	182.1127	10.8512	0.6669	9.57
10^{-3}	19	183.0974	10.8872	0.6651	9.08

(C) Conformer sensitivity: BioHL

Condition	Fixed BW RMSE $\downarrow$	Optimized RMSE $\downarrow$	MSE reduction (%)
Seed 23	0.6531	0.4287	56.92
Seed 10023	0.5366	0.5005	13.03
Seed 20023	0.4324	0.3999	14.46
UFF refined	0.6937	0.4574	56.52

G.5 Numerical implementation and data sources

Molecular and sensing experiments use double precision, training-fitted common scaling, population covariance normalization, and $10^{-6}I$ regularization of transformed bag covariances. Reference optimization uses a 10^{-6} SPD floor and 60 iterative matrix-square-root steps, with exact eigendecomposition-based tangent features and an OLS refit for prediction. The covariance search profiles training squared error with Armijo acceptance. Simulation and NHANES settings are specified in Appendix E.

ESOL uses the released processed solubility file, and the molecular endpoints use the curated OPERA repository. Gas is UCI 487 (DOI 10.24432/C5S302); NHANES processing is specified in Appendix A. The supplementary records provide split identifiers, preprocessing and model settings, paired predictions, and source-file hashes for reproducibility.

H Simulation protocols and additional results

H.1 Population mechanism experiment

For the identity construction in Section 3.4, let the three distinct entries of $\mathbf{U}$ be independent $\mathrm{Unif}[-1,1]$ variables. We approximate expectations with the tensor product of 12-point Gauss–Legendre rules, giving 1728 nodes with normalized product weights. Both the signal projection and its risk evaluation use this deterministic integration rule; this experiment estimates a population quantity rather than test prediction from a finite training set. Features are normalized as $(\mathbf{A}_{t,r} - \mathbf{D}_t^{-1})/r$ to avoid numerical Gram degeneration. The radius is $r \in \{0.1, 0.05, 0.025, 0.0125\}$, and the joint paths are $t = r$ and $t = r^2$. The general-reference construction uses the same law of $\mathbf{U}$ with

$$\mathbf{S} = \begin{pmatrix} 1.7 & 0.35 \\ 0.35 & 0.9 \end{pmatrix}, \quad \mathbf{H} = \begin{pmatrix} 0.4 & 0.6 \\ 0.6 & -0.2 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0.8 & -0.3 \\ -0.3 & 0.1 \end{pmatrix}.$$

Table 9: Noncommuting-direction control. The coefficient is computed from the Sylvester expression and compared with its closed form. The final column instead uses the complete OT map at $t = r = 0.025$. The endpoint $\lambda = 0$ is a separate commuting case.

λ	$16\lambda^4/45$	Computed coefficient	$E(0.025, 0.025)$
0	0.00000000	0.00000000	0.00000000
0.25	0.00138889	0.00138889	0.00138932
0.5	0.02222222	0.02222222	0.02222918
0.75	0.11250000	0.11250000	0.11253521
1	0.35555556	0.35555556	0.35566676

Its coefficient is computed directly from the Sylvester expression in Equation 7, followed by weighted affine projection. Table 1 reports both paths for $r \in \{0.1, 0.05, 0.025\}$; saved results include all four radii. Increasing the quadrature order from 12 to 16 changes the identity-reference risk at $(t, r) = (0.05, 0.1)$ by approximately 1.93×10^{-14} in relative terms.

Noncommutativity and signal controls. At $\mathbf{S} = \mathbf{I}$ and $\mathbf{H} = \mathbf{B} = \mathrm{diag}(1, -1)$, replacing the off-diagonal entry by λb yields the coefficient $16\lambda^4/45$ for each fixed $\lambda > 0$. The grid is $\lambda \in \{0, 1/4, 1/2, 3/4, 1\}$; the zero endpoint is a separate commuting case and does not satisfy the full coordinate-covariance assumption. For fixed full-dimensional $\mathbf{U}$, replacing the signal by $\mathbf{B}_\theta = \cos(\theta)\mathbf{H} + \sin(\theta)\mathbf{I}$ gives $16\cos^2(\theta)/45$. The computed coefficients at $\theta = 0, \pi/6, \pi/3, \pi/2$ are 0.35555556, 0.26666667, 0.08888889, and numerical zero, respectively. All checked coefficient errors are below 10^{-12}. Table 9 reports the noncommuting-direction control numerically. The last signal direction has zero leading coefficient; this does not assert exact zero risk.

For an exact equivalence check at a general reference, use the displayed $\mathbf{S}, \mathbf{B}$, set $\mathbf{H} = \mathrm{diag}(1, -1)$ and $\mathbf{U} = \mathrm{diag}(a, c)$, and take $(t, r) = (0.3, 0.2)$. The profiled risk is 4.07×10^{-32} numerically, agreeing with the exact commuting-transport identity. The construction concerns commutation of transport directions relative to the generating reference; raw covariance matrices need not commute at general $\mathbf{S}$.

H.2 Additional exact-feature prediction diagnostics

These diagnostics use the exact-feature implementation in Appendix E and a separate generator from the original experiment in Figure 2. Let $p = d(d+1)/2$ and let $\mathbf{V}$ be symmetric with independent upper-triangular entries distributed as $\mathrm{Unif}[-1, 1]$. Define

$$\mathbf{S} = \mathrm{diag}\{\mathrm{linspace}(0.7, 1.7, d)\} + 0.15\mathbf{1}\mathbf{1}^\top/d, \qquad \mathbf{H} = \mathrm{diag}\{\mathrm{linspace}(-1, 1, d)\},$$
$$\mathbf{T} = \mathbf{I} + 0.65\mathbf{V}/d, \qquad \mathbf{\Sigma} = \mathbf{TST}, \qquad f_\star(\mathbf{\Sigma}) = \sqrt{3}\mathrm{tr}(\mathbf{HV})/\|\mathbf{H}\|_F,$$
$$Y = f_\star(\mathbf{\Sigma}) + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 1/\mathrm{SNR}),$$

with independent noise and all means zero. Here $\mathrm{linspace}(a, b, d)$ means d equally spaced values including a, b. The bound $\|\mathbf{V}\|_{\mathrm{op}} \le d$ gives $\mathbf{T} \succeq 0.35\mathbf{I}$, and $\mathrm{Var}(f_\star) = 1$ by construction. At the generating reference the conditional mean is exactly tangent-linear.

We use $d \in \{2, 5\}$, $n = k(p+1)$ with $k \in \{2, 5, 20\}$, and $\mathrm{SNR} \in \{5, 20\}$. Each setting has 50 replicates and independent training, validation, calibration, and test distributions. Validation size is $\max(128, n)$, and both calibration and test sizes are 4096. The oracle-reference comparator fixes $\mathbf{S}$ but estimates its intercept and coefficients from training data. The mismatched initialization is $(\mathbf{I} + 0.6\mathbf{H})\mathbf{S}(\mathbf{I}+0.6\mathbf{H})$. It is optimized for at most 60 updates with the settings in Appendix E. Validation MSE selects among updates 0, 10, 30, and 60. If optimization stops early, unavailable checkpoints are replaced by the last available iterate, duplicates are removed, and the earliest checkpoint breaks an exact tie. The selected predictor retains its training-only OLS coefficients; validation observations are not pooled into its coefficient fit. The last feasible mismatch-initialized iterate is also recorded as an unselected comparator. All methods within a replicate share the same observations.

For replicate $j = 0, \ldots, 49$, the seed is $20260923 + 100000d + 10000\mathrm{SNR} + 100k + j$. NumPy `SeedSequence.spawn(5)` provides independent streams for training, validation, test, calibra-

Table 10: Finite-training-sample prediction over 50 replicates. The final column is the paired MSE decrease from the misspecified reference to its validation-selected learned version, with a 95% t interval.

d	n	SNR	Oracle	Misspecified	Learned	Paired decrease
2	8	5	0.4810	0.4600	0.4358	0.0242 ± 0.0198
2	8	20	0.1020	0.1362	0.1222	0.0140 ± 0.0049
2	20	5	0.2472	0.2638	0.2551	0.0087 ± 0.0024
2	20	20	0.0625	0.0827	0.0695	0.0133 ± 0.0017
2	80	5	0.2091	0.2249	0.2130	0.0119 ± 0.0014
2	80	20	0.0527	0.0678	0.0545	0.0133 ± 0.0005
5	32	5	0.3785	0.3908	0.3862	0.0046 ± 0.0028
5	32	20	0.1031	0.1107	0.1086	0.0021 ± 0.0010
5	80	5	0.2518	0.2571	0.2555	0.0016 ± 0.0008
5	80	20	0.0630	0.0680	0.0657	0.0023 ± 0.0003
5	320	5	0.2113	0.2160	0.2141	0.0019 ± 0.0003
5	320	20	0.0527	0.0573	0.0549	0.0024 ± 0.0001

tion, and within-distribution observations, in that order. The full seed and environment metadata, replicate-level metrics, and optimization histories accompany the simulation code.

Across the twelve settings, learning from the misspecified reference reduces mean test MSE by 0.62–19.56%. The pointwise paired 95% intervals for its MSE decrease exclude zero in 12 of twelve settings (Table 10). Validation selection also improves over the last iterate on average in the smallest training sets (Appendix H).

Uncertainty and validation selection. The per-setting intervals use the sample standard deviation of 50 paired MSE differences and the Student t_{49} critical value. They are pointwise Monte Carlo intervals, without simultaneous-coverage adjustment across the twelve settings. These intervals concern the additional exact-feature experiments. The original Figure 2 displays standard deviations across 20 replicates under its separate protocol. At $n/(p+1) = 2$, the equally weighted paired improvement from validation selection over the last mismatch-initialized iterate is 0.0168 ± 0.0072 (approximate 95% interval). The finite-sample oracle-reference fit can also have higher test error in an individual setting, since it still estimates coefficients from a small training set. The pooled validation comparison averages the four setting-specific means at $k = 2$ equally. Its approximate normal interval has half-width $1.96\sqrt{\sum_{s=1}^{4} \widehat{v}_s/50/4}$, where $\widehat{v}_s$ is the within-setting sample variance of the paired difference.

Separating geometric and estimation errors. After a reference is selected, we fit a separate OLS predictor $\widehat{h}_{\rho,\mathrm{cal}}$ to the noiseless signals of the independent calibration sample. On the exact-covariance test sample, define

$$\widehat{Q}_\rho = n_{\mathrm{test}}^{-1} \sum_i \{f_\star(\mathbf{\Sigma}_i) - \widehat{h}_{\rho,\mathrm{cal}}(\mathbf{\Sigma}_i)\}^2,$$

$$\widehat{R}_{\mathrm{sig},\rho} = n_{\mathrm{test}}^{-1} \sum_i \{f_\star(\mathbf{\Sigma}_i) - \widehat{f}_{\rho,\mathrm{train}}(\mathbf{\Sigma}_i)\}^2.$$

The first estimates $\mathcal{B}(\rho)$ with finite calibration and evaluation error, while the second includes estimation error from the smaller noisy training set. Neither calibration nor test signals enter reference selection. Table 11 reports these quantities at the largest training size; all sizes are in the saved results.

Optimization diagnostics. The 600 reference-learning runs terminate at the iteration cap in 574 cases and the backtracking limit in 26 cases. The terminal reference lies within 1% of the imposed 10^{-8} covariance floor in 29 runs, compared with 5 runs for the validation-selected reference used for prediction. Validation retains the initial reference in 111 runs. The median optimizer time is 0.049 seconds per run in the recorded single-thread environment; times exclude validation and test evaluation. All saved accepted steps satisfy the recorded Armijo decrease and positive transport-factor checks.

Table 11: Separating approximation and fitting error at $n/(p+1) = 20$. Each pair gives the independent-calibration projection residual $\widehat{Q}$ and the training-fitted signal MSE $\widehat{R}_{\mathrm{sig}}$, averaged over 50 replicates. The calibration estimate is not an exact population projection.

d	SNR	Reference	$\widehat{Q}$	$\widehat{R}_{\mathrm{sig}}$
2	5	Oracle reference	0.000000	0.008871
2	5	Misspecified	0.014397	0.024416
2	5	Learned from mismatch	0.003544	0.012645
2	20	Oracle reference	0.000000	0.002380
2	20	Misspecified	0.014353	0.017499
2	20	Learned from mismatch	0.001761	0.004180
5	5	Oracle reference	0.000000	0.011061
5	5	Misspecified	0.004288	0.015502
5	5	Learned from mismatch	0.002532	0.013699
5	20	Oracle reference	0.000000	0.002659
5	20	Misspecified	0.004294	0.007311
5	20	Learned from mismatch	0.001989	0.004867

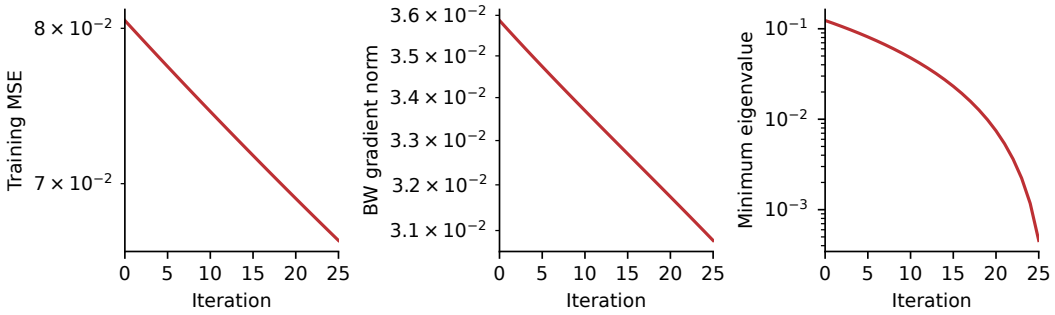

Figure 4: Optimization diagnostics at $d = 2$, $n = 8$, SNR 5, replicate 3, initialized at the misspecified reference. The panels show training MSE, BW gradient norm, and the minimum reference eigenvalue.

Figure 4 illustrates the evolution of training loss, gradient norm, and the minimum reference eigenvalue during optimization.

H.3 Robustness to estimated distributional inputs

At $d = 2$, $n = 80$, and SNR 5, estimate covariances from $m \in \{20, 100, 500\}$ Gaussian observations, using sample centering and denominator $m - 1$. Means remain structurally zero. For each m, 30 replicates reuse the latent distributions and responses of the first 30 exact-input replicates. All methods receive the same estimated training, validation, and test inputs, including the oracle-reference comparator. Optimization and validation selection are rerun for each m. Table 12 includes all sampling levels and the paired exact-input comparison. The saved projection diagnostic still uses exact covariances and does not decompose noisy-input risk.

Table 12: Distribution-estimation robustness at $d = 2$, $n = 80$, and SNR 5 over 30 replicates. Entries are mean test MSE $\pm$ 95% confidence half-width. The generating reference is known in the oracle-reference comparator, but its input covariances are also estimated when m is finite.

m	Oracle reference	Misspecified	Learned mismatch
Exact	0.2089 ± 0.0024	0.2250 ± 0.0028	0.2131 ± 0.0026
20	0.6455 ± 0.0070	0.6698 ± 0.0075	0.6552 ± 0.0087
100	0.3365 ± 0.0051	0.3555 ± 0.0061	0.3424 ± 0.0060
500	0.2390 ± 0.0029	0.2557 ± 0.0033	0.2431 ± 0.0030

I FURTHER CONNECTIONS TO DISTRIBUTION AND MANIFOLD REGRESSION

Functional and compositional representations. Transforming a univariate density or quantile function into linear coordinates connects distributional analysis to functional regression (Petersen & Müller, 2016; Morris, 2015; Greven & Scheipl, 2017). The direction of regression matters. Yang et al. (2020) regress quantile functions on scalar covariates, while Ghosal et al. (2025) model distributional responses using scalar and distributional predictors. These models complement scalar-on-distribution regression but do not study the same response space. Compositional scalar-on-function regression uses the relative information in densities, with applications to particle-size distributions and movement behavior (Talská et al., 2021; Jašková et al., 2023). Recent temperature-distribution applications address crop yield and migration (Trinh et al., 2026; Dardati et al., 2026). Their Bayes-space representations differ from the transport geometry of multivariate Gaussian covariates considered here. In particular, a Hilbert representation under one density geometry does not imply a flat representation under the BW metric.

Regression with geometric responses. Geodesic regression fits manifold-valued responses along geodesics (Fletcher, 2013), whereas Fréchet regression defines conditional means through expected squared distances (Petersen & Müller, 2019; Lin & Müller, 2021; Bhattacharjee & Müller, 2023). The present response is scalar, and the reference determines the linear functions available on the predictor space. This distinction locates the approximation error in the feature representation rather than in a geodesic model for the response. Wasserstein regression using tangent bundles provides a closer connection for univariate distributional predictors (Chen et al., 2023). Our local calculation concerns the additional curvature of multivariate Gaussian covariance geometry.

Gaussian transport and statistical error. Classical affine transport formulas make the nondegenerate Gaussian family explicit (Pitié & Kokaram, 2007; Chen et al., 2018; Delon & Desolneux, 2020). Concentration bounds for linear Monge-map estimation address uncertainty in sampled first and second moments (Flamary et al., 2019). Our population risk expansion isolates reference-induced approximation error with those moments known exactly. The SPD assumption keeps the Sylvester operators invertible and the local feature maps smooth. Transport between degenerate Gaussian measures requires a broader formulation (Yun & Zemel, 2025), which is outside the local SPD analysis.