# MARGINAL SCORES UNDER STRUCTURED MISSINGNESS: COVERAGE AND JOINT RECOVERY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Incomplete records constrain marginal scores, while joint predictions depend on relationships that may never be observed together. We characterize when marginal Fisher error controls joint-score error, and at what cost. For Gaussians with a fixed number of observed coordinates per record, we determine the worst recovery constant up to universal factors throughout the budget range. The conditioning cost is cubic at fixed missing fractions and decreases toward one as observations become sufficiently complete. For density-ratio ANOVA models, we prove a global comparison governed by interaction coverage and a matching quadratic gradient cost. A shared background can increase this cost without changing the observations informative about the contrast. Statistical rates weight each interaction's inverse sample count by its gradient energy, and retain the quadratic cost when the background amplitude is unknown. Mixture experiments examine recovery without a prescribed interaction basis. Full-waveform ECG experiments show that identical acquired marginals can support different joint scores and conditional predictions, with acquisition gains depending on the target and cohort.

## 1 INTRODUCTION

Incomplete records reveal different coordinate subsets of a common joint distribution. Marginal score matching learns from such records by applying the Hyvärinen score to each observed marginal (Givens et al., 2025). Its population excess loss is a weighted marginal Fisher divergence. The intended output is a joint model that represents dependencies across all coordinates. Its observable loss is useful for guaranteeing joint-score accuracy only if the marginal constraints control the joint score. Givens et al. (2025, Remark 4.7) leave this connection open.

The answer depends on what the model is allowed to hide. Every covariance entry of a Gaussian becomes visible as soon as its coordinate pair is observed. A non-Gaussian third-order interaction can leave every pairwise marginal unchanged. Even when all relevant interactions are visible, their effect on observed scores can be much smaller than their effect on the joint score. We therefore ask which masks identify the allowed interactions and how accurately the observed marginal scores determine the joint score.

Our two main results quantify this inverse problem in different model classes. For *Gaussians* with condition numbers at most $\kappa > 1$, complete pair coverage is necessary and sufficient. Among designs observing exactly $k$ coordinates per record, uniform acquisition is minimax. We prove its matching recovery law for arbitrary covariance orientations and means: a fixed missing fraction costs a cubic conditioning factor, whereas observing all but one coordinate costs $\Theta\{(1 + \kappa/d)^3\}$.

For *density-ratio ANOVA models*, we prove a global Fisher comparison with a quadratic gradient cost that is necessary even in a model whose singleton interactions are all covered. Its coverage factor comes from an exact local identity: at a product reference, each canonical interaction contributes its gradient energy multiplied by its observation probability. A common background can amplify joint-score error without changing the observations informative about the contrast. An uncovered interaction prevents control even within smooth, strongly log-concave models. Thus the two model classes answer the same recovery question through different representations of dependence.

For a known finite basis, the minimax rate weights inverse effective sample counts by each interaction's gradient energy. Mixture experiments then examine recovery when the interaction features



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.

are not supplied to the learner, including acquisition guided by a pilot whose observations count toward the budget. ECG experiments assess the consequences of unobserved dependence for conditional prediction, whose average Fisher error is controlled by joint Fisher error for balanced queries (equation (3)). The appendix guide (Table 1) maps the models to the results.

## 2 Related work

**Learning from incomplete records.** Marginal score matching (Givens et al., 2025) extends the Hyvärinen objective (Hyvärinen, 2005) to observed coordinate subsets and motivates the reverse comparison studied here. Imputation offers another route for unnormalized models (Uehara et al., 2020). In incomplete covariance estimation, jointly observed pair counts already determine statistical rates (Cai & Zhang, 2016; Park et al., 2021). We study the corresponding joint spatial-score loss, its conditioning cost, and the higher-order observation requirements of non-Gaussian models.

**Interaction structure and effective samples.** Functional ANOVA gives classical orthogonal coordinate decompositions (Kuo et al., 2010). Centered product perturbations appear in fixed-marginal constructions (Sarmanov, 1966). Order-$r$ log-density ANOVA models are identified by $r$-dimensional marginals (Jeon, 2012). For classification with missing covariates, Sell et al. (2024) obtain minimax rates governed by ANOVA-specific available-case counts. Our density-ratio model yields a gradient-energy identity, global Fisher comparison, and local joint-score coefficient $\sum_j \lambda_j / w_{B_j}$.

**Joint-score geometry and estimation.** Gradient information quantifies dependence through joint and marginal scores (Ding et al., 2019). General score-estimation lower bounds also expose the need for regularity (Lewis et al., 2025). Our matching construction keeps the informative marginal for a parameter fixed while varying the joint score's sensitivity to that parameter. The efficiency analysis uses composite proper scoring rules and Godambe information (Dawid et al., 2016; Varin et al., 2011). Appendix F compares nearby losses and model classes.

## 3 Setup

Let $P, Q$ have positive differentiable densities on $\mathbb{R}^d$, with $d \geq 2$. In the Gaussian specialization, write $P = N(\mu, S)$ and $Q = N(\nu, T)$ with positive-definite covariances. A mask $A \subseteq [d]$ is drawn independently of $X$ with probability $\pi_A$ (missing completely at random). We write $P_A, Q_A$ for the corresponding coordinate marginals and define

$$D_{\mathrm{F}}(P, Q) = \mathbb{E}_P \|\nabla \log p(X) - \nabla \log q(X)\|_2^2, \tag{1}$$

$$D_{\mathrm{M}}^{\pi}(P, Q) = \sum_A \pi_A D_{\mathrm{F}}(P_A, Q_A). \tag{2}$$

We call these the joint and marginal Fisher divergences, respectively. When $Q = Q_{\widehat{\theta}}$ is fitted from data, their values conditional on the training sample are population risks. Any additional expectation over that sample is written explicitly. For a deployment context $C$ drawn independently of $X$, let $U = [d] \setminus C$ be the unobserved coordinates. If $\mathbb{P}(i \in U) = v$ for every $i$, the conditionals of the same joint density $Q$ satisfy

$$\mathbb{E}_C \mathbb{E}_{P_C} D_{\mathrm{F}}(P_{U|C}, Q_{U|C}) = v D_{\mathrm{F}}(P, Q), \tag{3}$$

because differentiating a conditional log density in its unobserved coordinates removes the conditioning normalizer. The identity applies to conditionals derived from $Q$; separately trained conditional predictors need not correspond to a common joint density. Under the usual integration-by-parts conditions, the excess expected Hyvärinen score of the revealed marginals equals $D_{\mathrm{M}}^{\pi}(P, Q)$. For coordinates and covariance entries, define

$$w_i = \sum_{A \ni i} \pi_A, \qquad w_{ij} = \sum_{A \supseteq \{i,j\}} \pi_A, \tag{4}$$

$$w_{\mathrm{vertex}} = \min_i w_i, \qquad w_{\mathrm{cov}} = \min_{i<j} w_{ij}. \tag{5}$$

Thus $w_{\mathrm{cov}}$ is the minimum pair-observation probability. Since $d \geq 2$ and $w_{ij} \leq w_i$, it always satisfies $w_{\mathrm{cov}} \leq w_{\mathrm{vertex}}$.

# 4 GAUSSIAN RECONSTRUCTION: COVERAGE AND CONDITIONING

We begin with the population question: when does the observable marginal divergence control its joint counterpart uniformly? Pair coverage determines whether such control is possible, while covariance conditioning determines its cost.

**Theorem 1** (Global comparison and coverage boundary). *For $P = N(\mu, S)$ and $Q = N(\nu, T)$ with $w_{\mathrm{cov}} > 0$,*

$$D_{\mathrm{F}}(P, Q) \le \frac{\mathrm{cond}(S)\,\mathrm{cond}(T)^2}{w_{\mathrm{cov}}} D^{\pi}_{\mathrm{M}}(P, Q). \tag{6}$$

*Conversely, if a covariance diagonal or pair has zero coverage, no finite comparison holds uniformly over any sufficiently small unrestricted Gaussian neighborhood of the identity. Thus complete pair coverage is necessary and sufficient on classes containing such a neighborhood and having uniformly bounded condition numbers. Necessity need not hold on restricted classes, such as diagonal covariances. The quantitative coverage factors are also exact near the identity: for $w_i > 0$, the equal-covariance mean perturbation $P = N(0, I)$, $Q = N(\epsilon e_i, I)$ satisfies $D^{\pi}_{\mathrm{M}}(P, Q) = w_i D_{\mathrm{F}}(P, Q)$. For $w_{ij} > 0$, $i < j$, the covariance perturbation*

$$P = N(0, I), \qquad Q = N\{0, I + \epsilon(e_i e_j^{\top} + e_j e_i^{\top})\}, \qquad 0 < |\epsilon| < 1, \tag{7}$$

*satisfies $D^{\pi}_{\mathrm{M}}(P, Q) = w_{ij} D_{\mathrm{F}}(P, Q)$.*

Under the usual score-matching conditions, the same bound controls Fisher excess over the best fixed-center Gaussian working model even for non-Gaussian data. This follows from classical score matching (Hyvärinen, 2005) and the moment-projection identity in Appendix E.6. With only finite second moments, the corresponding quadratic-risk statement remains valid.

Appendix A proves the comparison by Gaussian score algebra and weighted entry counting. Fisher divergence is directed: $S$ governs the expectation and $T^{-1}$ the candidate score, accounting for the asymmetric condition factors.

## 4.1 HOW OBSERVATION BUDGET MODERATES CONDITIONING

Here every record reveals exactly $k$ coordinates, as in a fixed-size sensor panel. The following result gives the matching order across these per-record budgets, including near-complete observation.

**Theorem 2** (Fixed-budget minimax recovery). *Let $d \ge 3$, $2 \le k < d$, and $\kappa > 1$. Over all Gaussian laws with arbitrary means and covariance condition numbers at most $\kappa$, let $C^*_{d,k,\kappa}$ be the supremum of $D_{\mathrm{F}}(P, Q)/D^{\pi}_{\mathrm{M}}(P, Q)$ over distinct pairs in this class, with uniform $k$-masks. Put $r = k/d$, $p_2 = k(k-1)/\{d(d-1)\}$, and*

$$\Psi_{d,k,\kappa} = \max\left\{\frac{1}{p_2}, \frac{[r + (1-r)\kappa]^3}{r^2}\right\}. \tag{8}$$

*For a universal numerical constant $C_0$,*

$$\Psi_{d,k,\kappa} \le C^*_{d,k,\kappa} \le C_0 \Psi_{d,k,\kappa}. \tag{9}$$

*Uniform masks minimize the actual worst-case comparison constant among all cardinality-$k$ designs. Moreover, if $\kappa(1-r) < 1/6$, then $C^*_{d,k,\kappa} \le \{1 - 6\kappa(1-r)\}^{-1}$.*

An isolated covariance pair gives the $1/p_2$ lower bound. For the conditioning term, let $u = d^{-1/2}\mathbf{1}$ and take $1 < a \le \kappa$:

$$S = \kappa(I - uu^{\top}) + uu^{\top}, \qquad T = \kappa(I - uu^{\top}) + auu^{\top}. \tag{10}$$

Every $k$-mask sees the same informative all-ones direction. Its exact joint/marginal ratio approaches $[r + (1-r)\kappa]^3/r^2$ as $a \downarrow 1$ (Theorem 6). The upper bound applies to arbitrary covariances: it controls small missingness, then uses an isotropic ridge to reduce conditioning. Classical matrix Bregman and relative-curvature tools (Dhillon & Tropp, 2008; Lu et al., 2018) transfer this local argument to finite Gaussian pairs. Appendix A.3 gives the proof and an explicit certificate. The result yields $\Theta\{(1 + \kappa/d)^3\}$ for $k = d - 1$, and the constant tends to one when $\kappa(1-r) \to 0$. At the restricted endpoint $\kappa = 1$, only isotropic covariances are admissible and the exact constant is $1/r$;

for fixed $d, k$, the limit as $\kappa \downarrow 1$ through $\kappa > 1$ is instead $1/p_2$, since arbitrarily small anisotropic perturbations remain admissible. This minimax result concerns the population comparison over mask designs. Section 6 treats finite-sample estimation risk.

The target loss also matters. For the same Gaussian class and masks, the optimal joint-to-marginal KL comparison replaces the cubic term in $\Psi_{d,k,\kappa}$ by $[r+(1-r)\kappa]^2/r^2$ (Proposition 7). Equivalently, this is the cost for Fisher error integrated along a common Ornstein–Uhlenbeck flow. Thus the observation budget affects score accuracy at the original data distribution differently from distributional accuracy.

**Fixed cardinality versus an average budget.** If only $\mathbb{E}|A| \le k$ is imposed and empty masks are allowed, observing all coordinates with probability $r = k/d$ and none otherwise gives $D_{\mathrm{M}}^{\pi} = rD_{\mathrm{F}}$, with empty-mask divergence zero. Its constant $1/r$ is optimal for the same Gaussian class: $\sum_i w_i \le k$ implies some $w_i \le r$, and a mean shift along $i$ with covariance $I$ has ratio $1/w_i \ge 1/r$. This different acquisition experiment explains why Theorem 2 fixes each record's cardinality. Within that design class, uniform masks also maximize minimum pair coverage (Chee et al., 2013). Their minimax optimality follows from the symmetry of the full Gaussian class and need not extend to a restricted model.

## 5 Beyond pairs: coverage of model interactions

**The local reconstruction problem.** For a bounded, mean-zero perturbation $h$, let $dQ_t = (1 + th)dP$. Marginalization replaces $h$ by $h_A = \mathbb{E}_P[h \mid X_A]$. Under the differentiability and finite-energy conditions of Proposition 9, the second-order joint and marginal Fisher errors are respectively

$$\|\nabla h\|^2_{L_2(P)}, \qquad \|\mathcal{T}_\pi h\|^2 = \sum_A \pi_A \|\nabla_A \mathbb{E}_P[h \mid X_A]\|^2_{L_2(P_A)}. \tag{11}$$

Thus stable reconstruction requires the second energy to bound the first from below on the allowed tangents. Injectivity alone need not give a positive uniform bound in an infinite-dimensional model.

**Regularity alone does not suffice.** For $d \ge 3$, take a standard Gaussian product density $p$ and $q_\epsilon = p\{1 + \epsilon \prod_i \tanh(x_i)\}$. Integrating out any coordinate removes the perturbation. For small nonzero $\epsilon$, both laws are smooth and strongly log-concave with covariance $I$ and Lipschitz scores, yet every proper marginal agrees while their joint scores differ (Proposition 8). The missing ingredient is a restriction on interactions, even when the distributions are well conditioned in these senses.

**An exact interaction-coverage law.** At a product reference, canonical ANOVA components have zero integral in each of their coordinates. An observed marginal retains precisely the components contained in its mask. This makes the local operator diagonal in gradient energy and also supports a global comparison.

**Theorem 3** (Interaction coverage for density-ratio ANOVA)**.** *Let $\mu = \bigotimes_{i=1}^d \mu_i$ have a positive $C^1$ density, and suppose each factor satisfies a Poincaré inequality with constant $C_{\mathrm{P}}$. Let $\mathcal{B}$ be a collection of nonempty subsets of $[d]$. For $u \in \{f, g\}$, write*

$$u(x) = \sum_{B \in \mathcal{B}} u_B(x_B), \qquad \int u_B(x_B)\, d\mu_i(x_i) = 0 \quad (i \in B), \tag{12}$$

*where the centering holds for every fixed value of the other coordinates in $B$, and the components are bounded and $C^1$. Let $P_u$ have density $1 + u$ relative to $\mu$; its coordinates may be dependent even though $\mu$ is a product measure. Suppose $0 < m < 1 < M < \infty$ and $0 < L < \infty$ satisfy*

$$m \le 1 + f, 1 + g \le M, \qquad \|\nabla f\|_\infty \vee \|\nabla g\|_\infty \le L. \tag{13}$$

*For each interaction define*

$$w_B = \mathbb{P}(B \subseteq A) = \sum_{A \supseteq B} \pi_A, \qquad w_{\mathcal{B}} = \min_{B \in \mathcal{B}} w_B. \tag{14}$$

*If $w_{\mathcal{B}} > 0$, then*

$$D_{\mathrm{F}}(P_f, P_g) \le \frac{C_{\mathrm{A}}}{w_{\mathcal{B}}} D_{\mathrm{M}}^{\pi}(P_f, P_g). \tag{15}$$

*Writing $R = M/m$ and $\gamma = L\sqrt{C_{\mathrm{P}}}/m$, one may take $C_{\mathrm{A}} = R\{1+\gamma(\sqrt{R}+R)\}^2$; equation (74) also retains a complementary bound for small $\gamma$. The quadratic dependence on $L$ is necessary even in a two-dimensional, fully covered first-order density-ratio model (Proposition 10). At $f = g = 0$, the best local comparison constant over this interaction class is exactly $1/w_{\mathcal{B}}$. More precisely, for a canonical tangent $h = \sum_{B\in\mathcal{B}} h_B$,*

$$\lim_{t\to 0} \frac{D_{\mathrm{F}}(P_0, P_{th})}{t^2} = \sum_{B\in\mathcal{B}} \|\nabla h_B\|^2_{L_2(\mu_B)}, \tag{16}$$

$$\lim_{t\to 0} \frac{D^{\pi}_{\mathrm{M}}(P_0, P_{th})}{t^2} = \sum_{B\in\mathcal{B}} w_B \|\nabla h_B\|^2_{L_2(\mu_B)}. \tag{17}$$

*The inverse-coverage dependence is also attained at finite perturbations: if $h_B$ is a nonzero canonical component and $|\epsilon|$ is small enough, then*

$$D^{\pi}_{\mathrm{M}}(P_0, P_{\epsilon h_B}) = w_B D_{\mathrm{F}}(P_0, P_{\epsilon h_B}). \tag{18}$$

*Consequently, an allowed component with $w_B = 0$ rules out even local control.*

A Poincaré argument transfers the gradient identity to Fisher error. Appendix B.3 gives the proof and an extension to approximate interaction structure. It retains the quadratic gradient cost and adds a term for the energy of poorly observed interactions.

**Why coverage alone does not determine the cost.** The matching construction uses only two singleton interactions. Set $a(x,y) = 1 + \beta\sin(Ny) + \epsilon\sin x$ and $b(y) = 1 + \beta\sin(Ny)$, with $\epsilon = \beta = 1/4$, and compare $P = a\,\varphi\otimes\varphi$ with $Q = b\,\varphi\otimes\varphi$. Observe $x$ with probability $w \le 1/2$ and $y$ otherwise. The $y$ marginals agree, and the $x$ marginal discrepancy is independent of $N$. Nevertheless the joint score difference includes

$$\partial_y \log(p/q) = -\frac{\epsilon\beta N \sin x\cos(Ny)}{a(x,y)b(y)}. \tag{19}$$

Thus $D^{\pi}_{\mathrm{M}}(P,Q)$ is of order $w\epsilon^2$, while $D_{\mathrm{F}}(P,Q)$ is of order $\epsilon^2(1+N^2)$, with fixed density-ratio bounds. The first-order density ratio in this example need not have an additive log density or independent coordinates. The common background changes joint-score sensitivity while the observed $x$ distribution and the contrast $a-b$ stay fixed. This yields the quadratic regularity cost in Proposition 10 and its learning cost in Section 6.

The gradient cost describes recovery of the joint score. In the same density-ratio class, the joint-to-marginal KL comparison has factor at most $M/(mw_{\mathcal{B}})$, independent of $L$ (Appendix B.5). Thus interaction coverage identifies both targets, while their stability costs differ.

For the complete order-$r$ density-ratio hierarchy, every $r$-tuple must be covered. Uniform cardinality-$k$ masks give $w_{\mathcal{B}} = \binom{k}{r}/\binom{d}{r}$ for $k \ge r$, and zero otherwise.

## 6 Statistical Cost of Observing Interactions

The same interaction coordinates connect reconstruction to learning from samples. Consider a finite canonical basis $\psi_1,\dots,\psi_s$ under the product reference $\mu$, with supports $B_j$, bounded values and first derivatives, and

$$\mathbb{E}_\mu \psi_j\psi_k = \mathbf{1}_{j=k}, \qquad \mathbb{E}_\mu\langle\nabla\psi_j, \nabla\psi_k\rangle = \lambda_j \mathbf{1}_{j=k}, \quad \lambda_j > 0. \tag{20}$$

Such a basis is obtained by orthogonalizing within each interaction support and diagonalizing its gradient Gram matrix. Let $b_j = \|\psi_j\|_\infty$, fix $0 < \rho < 1$, and define

$$p_\theta = \mu\Big(1 + \sum_j \theta_j\psi_j\Big), \qquad \Theta_\rho = \Big\{\theta : \sum_j b_j|\theta_j| \le \rho\Big\}, \quad D = \mathrm{diag}(\lambda_j). \tag{21}$$

All these densities are positive and normalized. For a fixed mask schedule, let $N_j$ count the records containing $B_j$. Estimate $\theta_j$ by the mean of $\psi_j$ over those records, using zero if $N_j = 0$, and project the resulting vector onto $\Theta_\rho$ in the $D$ norm. Denote this observable-moment estimator by $\widehat{\theta}_{\mathrm{mom}}$.

**Theorem 4** (Interaction sample complexity). *Suppose $\mu$ has Poincaré constant $C_{\mathrm{P}}$ and the independent records have law $P_\theta$ from (21). There are constants $c, C > 0$, depending only on the fixed model region and reference, such that*

$$\mathbb{E}_\theta D_{\mathrm{F}}(P_\theta, P_{\widehat{\theta}_{\mathrm{mom}}}) \le C \left\{ \sum_{j:N_j>0} \frac{\lambda_j}{N_j} + \sum_{j:N_j=0} \lambda_j \theta_j^2 \right\}. \tag{22}$$

*For independent masks with $w_{B_j} > 0$, the minimax joint-score risk over $\Theta_\rho$, allowing arbitrary score estimators, satisfies*

$$\frac{c}{n} \sum_j \frac{\lambda_j}{w_{B_j}} \;\le\; \inf_{\widehat{s}} \sup_{\theta \in \Theta_\rho} \mathbb{E}_\theta \|\widehat{s} - \nabla \log p_\theta\|^2_{L^2(P_\theta)} \;\le\; \frac{C}{n} \sum_j \frac{\lambda_j}{w_{B_j}} \tag{23}$$

*for all sufficiently large $n$ for the fixed model and coverage law. An uncovered allowed interaction gives a nonvanishing minimax lower bound.*

Records contribute to an interaction whenever they contain its support, regardless of which larger mask they use. Appendix C gives finite-$n$ bounds and the constants for the fixed model and coverage law.

The background regularity also has a sample cost. In the same family with two sine terms $p_{\theta,\beta} = \varphi \otimes \varphi\{1 + \theta \sin x + \beta \sin(Ny)\}$, let each record reveal $x$ with probability $w$ and $y$ otherwise. With both amplitudes unknown in $[-1/4, 1/4]$ and the basis known, minimax joint-score risk is of order

$$\frac{1 + N^2}{1 + nw} + \frac{N^2}{1 + n(1 - w)} \tag{24}$$

uniformly in $n \ge 1$, $N \ge 1$, and $0 \le w \le 1$ (Corollary 12). Clipped acquired moments attain this rate. The observed $x$ distribution depends on neither $N$ nor $\beta$. The second term accounts for learning the background amplitude from $y$. For $w \le 1/2$, the order remains $(1 + N^2)/(1 + nw)$, as when the background is known. Geometry therefore raises the learning cost without making the observable $\theta$ experiment harder.

**Local information and estimator efficiency.** At the product reference, the observed parameter Fisher information is $\mathrm{diag}(w_{B_j})$, while the joint spatial-score metric is $D$. The local asymptotic minimax coefficient is exactly

$$\mathcal{R}_{\mathrm{eff}} = \sum_j \frac{\lambda_j}{w_{B_j}}, \tag{25}$$

attained by the projected moment estimator (Theorem 13). More generally, this coefficient is $\mathrm{tr}(\mathsf{G} I_\pi^{-1})$, where $\mathsf{G}$ measures joint-score sensitivity to parameter changes and $I_\pi$ is the observed parameter information. We evaluate it for the fitted mixture below. Marginal score matching can have an additional efficiency cost, weighted by the same inverse coverages (Proposition 14).

## 7 Experiments

Controlled studies test coverage and estimation cost. ECGs assess the consequences for joint models and conditional predictions. Table 4 records the training, pilot, development, selection and evaluation information used in each study.

### 7.1 Learning without a prescribed interaction basis

We fit eight-component Gaussian mixtures in 32 dimensions, learning all 791 mean, weight, and shared-covariance parameters from incomplete records. The generating weights contain a three-group parity contrast of strength $t \in \{0, 3/4\}$. Neither the feature nor its groups are supplied to the learner. Acquisition combines triple-hole and uniform masks, each observing four of eight groups: $\pi_\eta = (1-\eta)\pi_{\mathrm{h}} + \eta\pi_{\mathrm{u}}$, $\eta \in \{0, 1/4, 1/2, 1\}$. Every design has exactly the same scalar and pair counts. The signal triple is observed with probability $\eta/14$. At these truths, the joint law is unidentifiable at $\eta = 0$ and identifiable at every $\eta > 0$ within the fitted family (Appendix D).

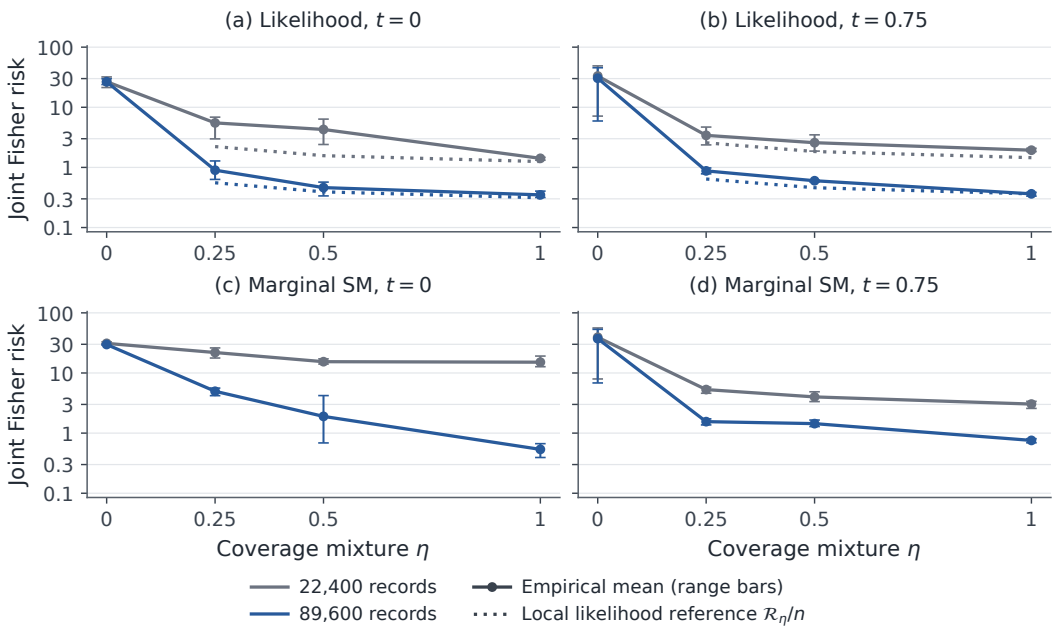


Figure 1: Joint-score learning as interaction coverage increases at fixed single-coordinate and pair counts. Top: acquired likelihood; bottom: likelihood-initialized marginal score matching at its final iterate. Colors indicate sample size. Circles and bars show means and ranges over three training repetitions. Solid lines connect empirical means. Dotted curves show the full-model local likelihood reference $\mathcal{R}_\eta/n$ for $\eta > 0$ (Proposition 15). The horizontal axis is linear and the vertical axis logarithmic. The twice-budget fully observed diagnostic remains in Table 2.

For each configuration and repetition, we select one of six starts by acquired validation likelihood, then optimize exact marginal score matching from that fit. Figure 1 compares both objectives across coverage levels. Increasing sample size reduces joint risk in all 36 paired comparisons at positive coverage. At $n = 89,600$, $t = 3/4$, increasing $\eta$ from $1/4$ through $1/2$ to 1 reduces mean likelihood joint risk from $0.875$ through $0.600$ to $0.363$, while acquired-marginal risks stay between $0.046$ and $0.054$. Hole acquisition has joint risk $30.46$ despite marginal risk $0.058$. Table 2 reports all fitted risks, which include estimation and optimization costs.

The full fitted family has 45 locally invisible directions under hole acquisition. Some mean directions require four groups and have coverage $\eta/70$, beyond the planted triple's $\eta/14$. Proposition 15 gives the full-model coefficient $\mathcal{R}_\eta = \operatorname{tr}(\mathsf{G} I_\eta^{-1}) = K_t/\eta + O(1)$ as coverage vanishes. Its large-sample likelihood reference $\mathcal{R}_\eta/n$ accounts for nuisance information from records missing the triple. At the larger sample size, mean likelihood risks are $1.00$–$1.62$ times this reference across both truths and all positive coverages.

Both objectives support the acquisition comparison: higher coverage reduces joint risk in 65 of 72 adjacent-design comparisons. Continuing score matching lowers training loss but increases population risk in these fits. Selecting fixed checkpoints by acquired validation score instead retains the likelihood initialization in 48 of 60 cases and substantially reduces this gap (Table 2). The coverage ordering persists. The population comparison therefore explains the cost of marginal error, while estimation and model selection determine how much error remains. Appendix D gives the selection and optimization controls. Appendix E.2 directly tests the known ANOVA class.

**Using a pilot to guide acquisition.** In the same fitted family, a pilot uses univariate fourth cumulants and the known number of signal groups to select three groups to observe together more often. Every scheme counts its pilot and validation records toward the budget and preserves all singleton and pair counts. With acquired-likelihood fitting, we compare uniform acquisition, 560- and 4,480-record pilots, and a 560-record pilot followed by a quarter-uniform mixture. Across twelve training repetitions in each of eight signal-strength and mean-perturbation conditions, the larger pilot reduces

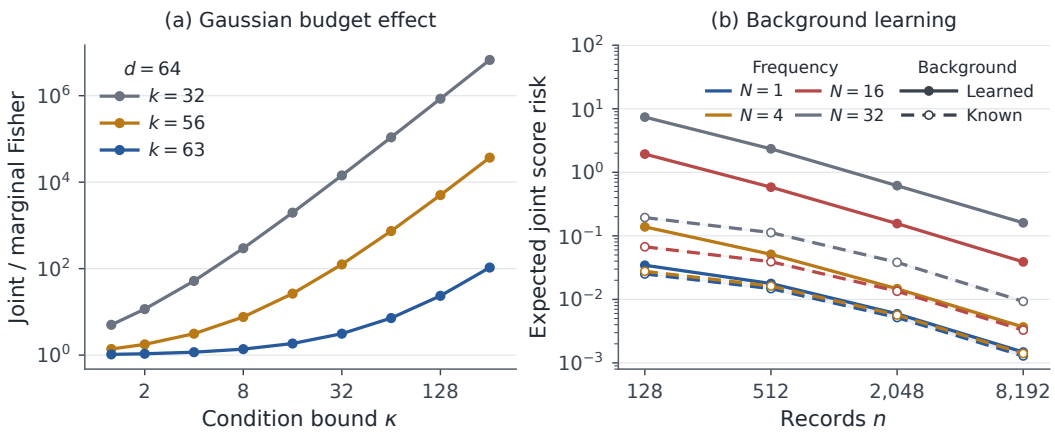


Figure 2: Quantitative costs of joint-score recovery. (a) Exact joint/marginal Fisher ratios for the Gaussian pair in (10), with $d = 64$, $a = 1.125$, and $k$ observed coordinates. (b) Moment estimation in (109), with $\theta = 0$, $\beta = w = 1/8$. Solid curves learn both amplitudes; dashed curves use the same $\widehat{\theta}$ and the true background amplitude. Means and $\pm 1.96$ Monte Carlo standard-error bands use 2,048 paired repetitions. All axes are logarithmic and points use their numerical coordinates.

mean paired joint-Fisher risk by $4.3\%$–$22.8\%$, with six of the eight pointwise 95% intervals excluding zero. The smaller pilot increases mean risk in two weak-signal conditions. Table 3 reports all conditions and ablations, and Appendix D.3 gives the allocation and numerical checks.

### 7.2 Observation budget and joint-score geometry

Figure 2a evaluates the Gaussian hard pair by direct matrix inversion. Observing 63 of 64 coordinates delays the large conditioning penalty compared with observing half the coordinates, as quantified by Theorem 2. Appendix E.3 separately assesses finite-sample estimation at a fixed covariance orientation, including all failed fits.

Figure 2b tests the background-learning mechanism. The two observed moments estimate both amplitudes from incomplete records. Increasing $N$ leaves the observations informative about $\theta$ unchanged but raises joint-score risk. Estimating the background adds a further cost. The full 192-configuration grid also varies the truth and acquisition probability. At $n = 8192$, interior-parameter risks are between $0.946$ and $1.040$ times their analytically predicted local coefficients divided by $n$; boundary cases use the appropriate projection correction. Appendix C.7 gives the complete grid, numerical integration checks, and finite-sample interpretation.

A separate 14-interaction experiment with a logistic reference tests (25) and the predicted score-matching inefficiency. All three estimators benefit from increased triple coverage. The precision check gives SM/MLE risk ratios $1.178$ and $1.166$, consistent with the predicted $1.180$ (Appendix C).

### 7.3 Joint ambiguity under fixed ECG acquisitions

PTB-XL (Wagner et al., 2020) supplies eight-lead, 1,000-point waveforms from 12,880 training and 1,904 test patients. Each training patient reveals four permanent leads, with exactly 6,440 patients per lead. Separated, cyclic, uniform and matched-pair triple-hole designs vary co-observation at this fixed budget. Testing reconstructs six leads from two. These budgets cover training acquisitions. Model development and selection of the main CNN checkpoint use additional development and validation information (Table 4, Appendix E.4).

Cyclic windows never reveal four lead pairs. Holding every observed covariance entry fixed, classical covariance completion (Dempster, 1972) produces distinct Gaussian joint scores with *zero* acquired-marginal Fisher divergence. Yet their test quadratic scores differ by $-188.33$ per preconditioned scalar (paired interval $[-206.12, -171.73]$; Figure 3b). Its population expectation equals the joint Fisher-risk difference under the integration-by-parts conditions in Appendix E.6. Conditional NLL

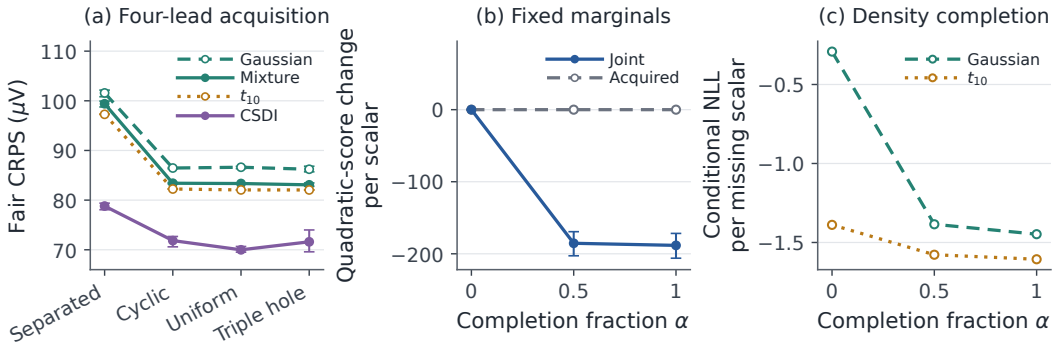


Figure 3: Full-waveform ECGs: acquisition value and joint ambiguity. (a) Patient-averaged fair CRPS from 32 draws, with three-fit range bars. (b,c) Post-hoc cyclic completion preserves every acquired marginal. (b) Paired test quadratic-score contrast from $\alpha = 0$, per preconditioned scalar; bars give paired patient-bootstrap 95% intervals conditional on three fits. (c) Conditional NLL per missing mV scalar. Fraction $\alpha$ interpolates toward a feasible entropy-directed completion. Axes are linear; points share coordinates.

changes from $-0.292$ to $-1.448$, largely closing the original Gaussian cycle–uniform gap. A Student-$t$ model exhibits the same ambiguity (Figure 3c). This exploratory check shows that the original gap partly reflects a choice of unseen dependence.

The acquisition effect extends to conditional diffusion. CSDI (Tashiro et al., 2021), trained for 200 epochs at the same budget, reduces fair CRPS by $11.2\%$ and fair energy by $15.4\%$ under uniform versus separated acquisition, with all three fits agreeing (Figure 3a). The pair-matched triple-hole contrast is smaller ($2.25\%/1.55\%$), with one of three fits reversing direction for both scores. These predictive metrics assess acquisition value separately from the joint-score target (Appendix E.5). Fixed-marginal completion also lowers conditional NLL on two external ECG cohorts without refitting. Subsequent exploratory quadratic-score analysis favors completion in both cohorts, but reverses the PTB-XL uniform–separated ranking: separated has lower quadratic scores on CS/N and Georgia in both coordinate systems (Appendix E.7). The value of an acquisition design therefore depends on the cohort, the fitted model and the prediction target.

## 8 Conclusion

Marginal-score accuracy controls the joint score when observations cover the dependencies allowed by the model, with a cost determined by the target's sensitivity to those dependencies. We quantify this cost for Gaussians across fixed per-record budgets and for density-ratio ANOVA models with bounded gradients, giving matching orders in both cases. Statistical bounds translate coverage and background sensitivity into sample requirements. Mixture experiments demonstrate the value of coverage beyond a prescribed interaction basis. ECG completion changes joint scores and conditional predictions while preserving every acquired marginal. A next step is to allocate observations using estimated interaction strength while accounting for uncertainty in the pilot.

### AI use statement

AI tools were used to assist with brainstorming, developing theoretical models and hypotheses, formulating mathematical claims, developing and writing proofs, and designing experiments. They also assisted with implementation, synthetic-data generation, data preprocessing, result interpretation, literature review, figures, and manuscript preparation. Checks included comparison with prior work and executable mathematical and numerical verification. The authors take responsibility for the final content, correctness, originality, and attribution.

### Reproducibility statement

Appendices A–C contain the proofs; Appendices D–E give experimental protocols. The anonymous supplement provides code, fixed configurations, result summaries, numerical checks, and commands to regenerate the figures and tables. Its README distinguishes verification of saved results from rerunning experiments and documents public-data requirements.

## References


Andrea Basteri, Carlo Ciliberto, and Alessandro Rudi. Distributionally faithful imputation via positive semi-definite kernel density estimation, 2026. arXiv preprint arXiv:2607.07767.

T. Tony Cai and Anru Zhang. Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *Journal of Multivariate Analysis*, 150:55–74, 2016. doi: 10.1016/j.jmva.2016.05.002.

Yichao Cai and Javen Qinfeng Shi. An identifiability theory of masked prediction: Mode blindness and mask schedules. arXiv:2608.01383v3, 2026.

Yeow Meng Chee, Charles J. Colbourn, Alan C. H. Ling, and Richard M. Wilson. Covering and packing for pairs. *Journal of Combinatorial Theory, Series A*, 120(7):1440–1449, 2013. doi: 10.1016/j.jcta.2013.04.005.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alex Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 288–313, 2023. doi: 10.52202/075280-0015.

A. Philip Dawid, Monica Musio, and Laura Ventura. Minimum scoring rule inference. *Scandinavian Journal of Statistics*, 43(1):123–138, 2016. doi: 10.1111/sjos.12168.

A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972. doi: 10.2307/2528966.

Inderjit S. Dhillon and Joel A. Tropp. Matrix nearness problems with Bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146, 2008. doi: 10.1137/060649021.

Jie Ding, Robert Calderbank, and Vahid Tarokh. Gradient information for representation and modeling. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

C. A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140:1917–1923, 2014. doi: 10.1002/qj.2270.

Peter G. M. Forbes and Steffen Lauritzen. Linear estimating equations for exponential families with application to Gaussian linear concentration models. *Linear Algebra and its Applications*, 473: 261–283, 2015. doi: 10.1016/j.laa.2014.08.015.

Christian Genest, Karel Hron, and Johanna G. Nešlehová. Orthogonal decomposition of multivariate densities in Bayes spaces and relation with their copula-based representation. *Journal of Multivariate Analysis*, 198:105228, 2023. doi: 10.1016/j.jmva.2023.105228.

Zoubin Ghahramani and Michael I. Jordan. Supervised learning from incomplete data via an EM approach. In *Advances in Neural Information Processing Systems*, volume 6, pp. 120–127, 1993.

Richard D. Gill and Boris Y. Levit. Applications of the van Trees inequality: a Bayesian Cramér–Rao bound. *Bernoulli*, 1(1–2):59–79, 1995.

Josh Givens, Song Liu, and Henry W. J. Reeve. Score matching with missing data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19523–19561, 2025.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Yongho Jeon. A characterization of the log-density smoothing spline ANOVA model. *Communications in Statistics—Theory and Methods*, 41(12):2081–2087, 2012. doi: 10.1080/03610926.2011.606485.

Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. In *International Conference on Learning Representations*, 2023.

Mladen Kolar and Eric P. Xing. Estimating sparse precision matrices from data with missing values. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 635–642, 2012.

Frances Y. Kuo, Ian H. Sloan, Grzegorz W. Wasilkowski, and Henryk Woźniakowski. On decompositions of multivariate functions. *Mathematics of Computation*, 79(270):953–966, 2010. doi: 10.1090/S0025-5718-09-02319-9.

Rebecca M. Lewis, Oliver Y. Feng, Henry W. J. Reeve, Min Xu, and Richard J. Samworth. Learning the score under shape constraints. arXiv:2512.14624, 2025.

Yuchen Li, Alexandre Kirchmeyer, Aashay Mehta, Yilong Qin, Boris Dadachev, Kishore Papineni, Sanjiv Kumar, and Andrej Risteski. Promises and pitfalls of generative masked language modeling: Theoretical framework and practical guidelines. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27969–28017, 2024.

Roderick J. A. Little. Robust estimation of the mean and covariance matrix from data with missing values. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 37(1):23–38, 1988. doi: 10.2307/2347491.

Chuanhai Liu and Donald B. Rubin. ML estimation of the t distribution using EM and its extensions, ECM and ECME. *Statistica Sinica*, 5(1):19–39, 1995.

Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018. doi: 10.1137/16M1099546.

Siwei Lyu. Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 359–366, 2009.

Mitsuki Nagai and Keisuke Yano. The geometry of Stein's method of moments: A canonical decomposition via score matching. arXiv:2603.12843, 2026.

Seongoh Park, Xinlei Wang, and Johan Lim. Estimating high-dimensional covariance and precision matrices under general missing dependence. *Electronic Journal of Statistics*, 15(2):4868–4915, 2021. doi: 10.1214/21-EJS1892.

Eduardo Pavez and Antonio Ortega. Covariance matrix estimation with non uniform and data dependent missing observations. *IEEE Transactions on Information Theory*, 67(2):1201–1215, 2021. doi: 10.1109/TIT.2020.3039118.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, 2026. doi: 10.1038/s44360-026-00096-z.

Matthew Reyna, Nadi Sadr, Annie Gu, Erick Andres Perez Alday, Chengyu Liu, Salman Seyedi, Amit Shah, and Gari Clifford. Will two do? varying dimensions in electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021. *PhysioNet*, 2022. doi: 10.13026/34va-7q14. Version 1.0.3.

Matthew A. Reyna, Nadi Sadr, Erick A. Perez Alday, Annie Gu, Amit J. Shah, Chad Robichaux, Ali Bahrami Rad, Andoni Elola, Salman Seyedi, Sardar Ansari, Hamid Ghanbari, Qiao Li, Ashish Sharma, and Gari D. Clifford. Will two do? varying dimensions in electrocardiography: The PhysioNet/Computing in Cardiology Challenge 2021. In *Computing in Cardiology*, volume 48, pp. 1–4, 2021. doi: 10.22489/CinC.2021.134.

O. V. Sarmanov. Generalized normal correlation and two-dimensional Fréchet classes. *Doklady Akademii Nauk SSSR*, 168(1):32–35, 1966.

Torben Sell, Thomas B. Berrett, and Timothy I. Cannings. Nonparametric classification with missing data. *The Annals of Statistics*, 52(3):1178–1200, 2024.

Devin Smedira, Abhijith Jayakumar, Sidhant Misra, Marc Vuffray, and Andrey Y. Lokhov. Finite sample bounds for learning with score matching. *arXiv preprint arXiv:2605.14168*, 2026.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59, 2017.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24804–24816. Curran Associates, Inc., 2021.

Masatoshi Uehara, Takeru Matsuda, and Jae Kwang Kim. Imputation estimators for unnormalized models with missing data. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 831–841, 2020.

Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7:154, 2020. doi: 10.1038/s41597-020-0495-6.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *PhysioNet*, 2022. doi: 10.13026/kfzx-aw45. Version 1.0.3.

Jianwei Zheng, Huimin Chu, Daniele Struppa, Jianming Zhang, Sir Magdi Yacoub, Hesham El-Askary, Anthony Chang, Louis Ehwerhemuepha, Islam Abudayyeh, Alexander Barrett, Guohua Fu, Hai Yao, Dongbo Li, Hangyuan Guo, and Cyril Rakovski. Optimal multi-stage arrhythmia classification approach. *Scientific Reports*, 10:2898, 2020. doi: 10.1038/s41598-020-59821-7.

Jianwei Zheng, Hangyuan Guo, and Huimin Chu. A large scale 12-lead electrocardiogram database for arrhythmia study. *PhysioNet*, 2022. doi: 10.13026/wgex-er52. Version 1.0.0.

Zihan Zhou, Chenguang Wang, Hongyi Ye, Yongtao Guan, and Tianshu Yu. Incomplete data, complete dynamics: A diffusion approach. In *The Fourteenth International Conference on Learning Representations*, 2026.

## GUIDE TO THE APPENDIX

Table 1 maps the models to the results they support. The appendices follow the main argument: Gaussian recovery, non-Gaussian recovery, learning costs, mixture learning, and ECG protocols.

Table 1: Models, observed constraints, and the role of each result.

| **Model** | **Observed constraints** | **Result and role** |
|---|---|---|
| Unrestricted Gaussian | Mean and covariance entries | Global Fisher comparison; matched fixed-cardinality budget law (Section 4). |
| Density-ratio ANOVA | Canonical components contained in each mask | Global Fisher comparison and matching gradient cost (Section 5). |
| Known finite basis | Acquired interaction moments | Statistical explanation: effective counts and target-weighted risk (Section 6). |
| Eight-component mixture | Marginals of the full fitted family | Learning study: identification at specified truths, local information and estimator diagnostics (Section 7). |
| ECG working models | Acquired leads | Application: marginally equivalent joint choices and conditional prediction (Section 7). |

## A GAUSSIAN RECOVERY PROOFS

### A.1 GLOBAL GAUSSIAN COMPARISON

Put $\delta = \mu - \nu$. Gaussian score algebra and centering at $\mu$ give the orthogonal decomposition

$$D_{\mathrm{F}}(P, Q) = \|S^{-1/2}(S - T)T^{-1}\|_F^2 + \|T^{-1}\delta\|_2^2. \tag{26}$$

Consequently, its covariance part is at most

$$\frac{\|S - T\|_F^2}{\lambda_{\min}(S)\lambda_{\min}(T)^2}. \tag{27}$$

For every supported mask, principal-submatrix interlacing and the same identity give the lower bound

$$D_{\mathrm{F}}(P_A, Q_A)_{\mathrm{cov}} \geq \frac{\|(S - T)_A\|_F^2}{\lambda_{\max}(S)\lambda_{\max}(T)^2}. \tag{28}$$

Weighted entry counting yields

$$\begin{aligned}\sum_A \pi_A \|(S - T)_A\|_F^2 &= \sum_i w_i (S - T)_{ii}^2 + 2\sum_{i<j} w_{ij}(S - T)_{ij}^2 \\ &\geq w_{\mathrm{cov}} \|S - T\|_F^2.\end{aligned} \tag{29}$$

Comparing this inequality with (27) proves the covariance coefficient in (6). Likewise,

$$\sum_A \pi_A \|T_A^{-1}\delta_A\|_2^2 \geq \frac{w_{\mathrm{vertex}}}{\lambda_{\max}(T)^2}\|\delta\|_2^2, \tag{30}$$

whereas the mean term in (26) is at most $\|\delta\|_2^2/\lambda_{\min}(T)^2$. Since $w_{\mathrm{cov}} \leq w_{\mathrm{vertex}}$ and $\mathrm{cond}(S) \geq 1$, the covariance coefficient also bounds the mean ratio, proving the forward assertion.

For the converse, if $w_i = 0$, take $S = I$ and perturb only the $i$th diagonal of $T$. No supported marginal changes. If $w_{ij} = 0$, take

$$S = I, \qquad T = I + \epsilon(e_i e_j^\top + e_j e_i^\top), \qquad 0 < |\epsilon| < 1. \tag{31}$$

No supported mask contains both $i$ and $j$, so every observed principal submatrix is unchanged although the joint Fisher divergence is positive. Sending $\epsilon$ to zero places the counterexample in every sufficiently small neighborhood of the identity and completes the proof.

For the quantitative statements, if only the mean differs by $\epsilon e_i$ and both covariances equal $I$, precisely the masks containing $i$ have nonzero Fisher divergence, and each such marginal divergence equals the joint divergence. Hence $D_{\mathrm{M}}^{\pi} = w_i D_{\mathrm{F}}$. Under (7), a mask not containing both $i, j$ sees the identity under both distributions. A mask containing both sees the same nontrivial $2 \times 2$ block as the joint pair, with identity coordinates added. Thus every nonzero marginal divergence equals the joint divergence and $D_{\mathrm{M}}^{\pi} = w_{ij} D_{\mathrm{F}}$.

**Corollary 5** (Lower bound for richer models)**.** *Let $\mathcal{P}$ be any distribution class that contains a sufficiently small centered Gaussian neighborhood of the identity and the pairs in (10). If some covariance pair has zero coverage, then there are distinct $P, Q \in \mathcal{P}$ for which every supported observed marginal agrees, so $D_{\mathrm{M}}^{\pi}(P, Q) = 0 < D_{\mathrm{F}}(P, Q)$. If instead every supported mask has cardinality $k$, any constant $C_\kappa$ satisfying*

$$D_{\mathrm{F}}(P, Q) \le C_\kappa D_{\mathrm{M}}^{\pi}(P, Q) \qquad \text{for all } P, Q \in \mathcal{P} \tag{32}$$

*must be at least the ratio in (33). Hence $C_\kappa = \Omega(\kappa^3)$ for fixed $a > 1$ and observation fraction bounded away from one.*

Corollary 5 concerns population divergences and therefore remains valid with unlimited data and exact optimization. The same lower bound also applies by approximation: if pairs from $\mathcal{P}$ converge to each Gaussian example in both $D_{\mathrm{F}}$ and $D_{\mathrm{M}}^{\pi}$, a uniform reverse inequality on $\mathcal{P}$ passes to the limit and inherits the same lower bound.

### A.2 Exact exchangeable hard pair

**Theorem 6** (Cubic Gaussian hard pair)**.** *For every distribution over masks of cardinality $k$, the pair in (10) satisfies*

$$\frac{D_{\mathrm{F}}\{N(0, S), N(0, T)\}}{D_{\mathrm{M}}^{\pi}\{N(0, S), N(0, T)\}} = \frac{[ka + (d-k)\kappa]^2 [k + (d-k)\kappa]}{dk^2 a^2}. \tag{33}$$

*In particular, for any fixed $a > 1$ independent of $\kappa$ (and $\kappa \ge a$), the ratio is $\Omega(\kappa^3)$ whenever the observation fraction $k/d$ stays bounded away from one.*

The full covariances in (10) differ only along $u$. Their joint Fisher divergence is therefore

$$D_{\mathrm{F}}\{N(0, S), N(0, T)\} = \frac{(a-1)^2}{a^2}. \tag{34}$$

For any cardinality-$k$ mask $A$, the two principal covariances agree on the orthogonal complement of $\mathbf{1}_A$ and have respective eigenvalues

$$s = \frac{k + (d-k)\kappa}{d}, \qquad t = \frac{ka + (d-k)\kappa}{d} \tag{35}$$

along $\mathbf{1}_A/\sqrt{k}$. Hence every such mask has the same marginal Fisher divergence,

$$D_{\mathrm{F}}(P_A, Q_A) = s(1/t - 1/s)^2 = \frac{k^2 (a-1)^2}{d^2 s t^2}. \tag{36}$$

Averaging over any distribution of cardinality-$k$ masks changes nothing. Dividing the joint expression by the marginal expression and substituting $s, t$ gives (33). Its cubic conclusion follows by keeping $a$ fixed independently of $\kappa$ and reading the leading power in the displayed polynomial ratio.

### A.3 Fixed-budget recovery, including near-complete observations

We prove Theorem 2. The proof starts with a local perturbation estimate, then uses a ridge argument to extend it beyond small missingness. The numerical constants are sufficient bounds and are not optimized.

**From curvature to finite pairs.** For $f(C) = \mathrm{tr}(C^{-1})$ and symmetric $H$, define

$$g_C(H) = \mathrm{tr}(C^{-1} H C^{-2} H) = \tfrac{1}{2} D^2 f(C)[H, H]. \tag{37}$$

The centered Gaussian divergence is exactly $f(S) - f(T) - Df(T)[S-T]$. Thus, with $H = S - T$ and $C_t = T + tH$,

$$D_{\mathrm{F}}\{N(0,S), N(0,T)\} = 2\int_0^1 (1-t) g_{C_t}(H)\, dt. \tag{38}$$

The same formula holds on every principal marginal. The class $\{C \succ 0 : \mathrm{cond}(C) \le \kappa\}$ is convex, since

$$\lambda_{\max}(C_t) \le \kappa\{(1-t)\lambda_{\min}(T) + t\lambda_{\min}(S)\} \le \kappa\lambda_{\min}(C_t). \tag{39}$$

Consequently a uniform curvature comparison on this class integrates to the identical finite-pair comparison. This is the classical relative-curvature reduction (Lu et al., 2018, Proposition 1.1), applied to a matrix Bregman divergence (Dhillon & Tropp, 2008).

**Mean shifts are controlled by covariance directions.** Fix $C \succ 0$ and suppose $g_C(H) \le c\mathbb{E}_A g_{C_A}(H_A)$ for every symmetric $H$. For $v \neq 0$, put $q = v^\top C^{-1} v$ and $q_A = v_A^\top C_A^{-1} v_A$. Writing $B = [d] \setminus A$, the Schur complement gives

$$q - q_A = (v_B - C_{BA}C_A^{-1}v_A)^\top (C_B - C_{BA}C_A^{-1}C_{AB})^{-1}(v_B - C_{BA}C_A^{-1}v_A) \ge 0. \tag{40}$$

Taking $H = vv^\top$ in the curvature bound yields

$$q\|C^{-1}v\|^2 \le c\mathbb{E}_A\{q_A\|C_A^{-1}v_A\|^2\} \le cq\mathbb{E}_A\|C_A^{-1}v_A\|^2. \tag{41}$$

Dividing by $q > 0$ proves the same comparison for the mean-score loss. The case $v = 0$ is immediate. Thus a uniform covariance-curvature bound already controls arbitrary means with the identical constant.

**Small missingness.** For this step masks can have any distribution. Put $\delta = \max_i \Pr(i \notin A)$, $K = C^{-1}$, $m = \lambda_{\min}(K)$, $M = \lambda_{\max}(K)$, and $X = C^{-1/2}HC^{-1/2}$, so $g_C(H) = \|X\|_K^2 := \mathrm{tr}(X^\top KX)$. If $B = A^c$ and $E_B$ is coordinate inclusion, define

$$L_B = E_B(K_B)^{-1}E_B^\top, \qquad R_B = K^{1/2}L_BK^{1/2}, \qquad P_B = I - R_B. \tag{42}$$

Here $R_B$ is an orthogonal projector and

$$g_{C_A}(H_A) = \|P_BXP_B\|_K^2. \tag{43}$$

Indeed, $U = C^{1/2}E_AC_A^{-1/2}$ satisfies $UU^\top = P_B$, $U^\top KU = C_A^{-1}$, and $U^\top XU = C_A^{-1/2}H_AC_A^{-1/2}$. Writing $L = \mathbb{E}L_B$ and $\overline{R} = \mathbb{E}R_B$, interlacing gives

$$L \preceq (\delta/m)I, \qquad \overline{R} \preceq (\delta/m)K, \qquad \|\overline{R}\|_{\mathrm{op}} \le \delta\kappa. \tag{44}$$

Use $\langle V, W\rangle_K = \mathrm{tr}(V^\top KW)$ and put $g = \|X\|_K^2$. We have

$$\mathbb{E}\langle X, R_BX\rangle_K \le \delta\kappa g, \qquad \mathbb{E}\langle X, XR_B\rangle_K \le \delta\kappa g, \tag{45}$$

$$\mathbb{E}\|XR_B\|_K^2 \le \delta\kappa g, \qquad \mathbb{E}\|R_BXR_B\|_K^2 \le \delta\kappa g. \tag{46}$$

For the first inequality in (45), the left side is $\mathrm{tr}(X^2K\overline{R})$ and

$$K^{-1/2}\,\mathrm{sym}(K\overline{R})K^{-1/2} = \mathrm{sym}(KL) \preceq \delta\kappa I, \tag{47}$$

where $\mathrm{sym}(V) = (V + V^\top)/2$ and $\|KL\|_{\mathrm{op}} \le M\delta/m$. The other linear inequality and the first squared inequality follow from $\mathrm{tr}(XKX\overline{R}) \le \delta\kappa g$. For the last one, right multiplication by $R_B$ contracts Frobenius norm, so

$$\mathbb{E}\|R_BXR_B\|_K^2 \le \mathbb{E}\|R_BX\|_K^2 \le M\,\mathrm{tr}(X^2\overline{R}) \le \delta\kappa g. \tag{48}$$

Since $\langle X, R_BXR_B\rangle_K = \langle XR_B, R_BXR_B\rangle_K$, Cauchy–Schwarz and (46) bound its expected absolute value by $\delta\kappa g$. Expanding the square with $Y_B = X - P_BXP_B = R_BX + XR_B - R_BXR_B$ now yields

$$\mathbb{E}g_{C_A}(H_A) = g - 2\mathbb{E}\langle X, Y_B\rangle_K + \mathbb{E}\|Y_B\|_K^2 \ge (1 - 6\delta\kappa)g. \tag{49}$$

Together with (38) and (41), this proves the small-missingness bound for arbitrary Gaussian means and covariances.

**Ridge transfer.** Let $C' = C + (a-1)\lambda_{\min}(C)I$, $a \geq 1$. Then $\mathrm{cond}(C') \leq 1 + (\kappa - 1)/a$. Diagonalizing $C$ and, separately, each $C_A$ shows

$$g_C(H) \leq a^3 g_{C'}(H), \qquad g_{C'_A}(H_A) \leq g_{C_A}(H_A). \tag{50}$$

In fact curvature is a sum of $H_{ij}^2/(\lambda_i \lambda_j^2)$, and the ridge preserves each relevant eigenbasis. Each principal block receives the same scalar shift $(a-1)\lambda_{\min}(C)$ as the original matrix. Applying (49) at $C'$ gives, whenever $\delta < 1/6$ and $(1-6\delta)a > 6\delta(\kappa - 1)$,

$$C^*_{d,k,\kappa} \leq \min\left\{\frac{\kappa^3}{p_2}, \frac{a^4}{(1-6\delta)a - 6\delta(\kappa-1)}\right\}, \qquad \delta = 1 - r. \tag{51}$$

The same factor holds for finite pairs and arbitrary means by (38) and (41). The scalar ridge may be chosen separately at every original path point $C_t$. The ridge term is minimized at $a_* = \max\{1, 8\delta(\kappa-1)/(1-6\delta)\}$. For $a_* > 1$ its value is $2048\delta^3(\kappa-1)^3/(1-6\delta)^4$. If $\delta < 1/18$, this is at most $10368\{\delta(\kappa-1)\}^3 \leq 10368\Psi_{d,k,\kappa}$; when $a_* = 1$ the value is at most $4/(1-6\delta) < 6$. If instead $\delta \geq 1/18$, use

$$\frac{\kappa^3}{p_2} \leq \frac{2}{\delta^3}\frac{[r + \delta\kappa]^3}{r^2} \leq 11664\Psi_{d,k,\kappa}, \tag{52}$$

where $r^2/p_2 \leq 2$ for $k \geq 2$. Thus $C_0 = 11664$ suffices.

**Lower bounds and minimax design.** The isolated pair in (7) is feasible for every $\kappa > 1$ at sufficiently small $\epsilon$, giving $C^* \geq 1/p_2$. Sending $a \downarrow 1$ in (33) gives the other term of $\Psi$. For an arbitrary fixed-$k$ design with comparison constant $C_\pi < \infty$, apply its inequality to every simultaneous coordinate permutation of $P, Q$. The joint divergence and condition numbers are unchanged. Averaging these inequalities makes the mask distribution uniform, proving $C^* \leq C_\pi$; the conclusion is immediate when $C_\pi$ is infinite. Finally, at $\kappa = 1$ the covariance matrices are scalar identities, and direct calculation makes both the mean and covariance marginal losses equal to $r$ times their joint losses.

**The comparison rate and the exact constant.** The lower witnesses establish the parameter order but need not maximize the ratio. For $d = 3, k = 2, \kappa = 2$, let $J = \mathbf{1}\mathbf{1}^\top$, $C = 2I - J/3$, and $H = J - I$. Then

$$g_C(H) = \frac{17}{4}, \qquad g_{C_A}(H_A) = \frac{35}{64} \quad (|A| = 2), \qquad \frac{g_C(H)}{\mathbb{E}_A g_{C_A}(H_A)} = \frac{272}{35} > \frac{16}{3} = \Psi_{3,2,2}. \tag{53}$$

These identities follow by simultaneous diagonalization: $C, H$ have eigenvalues $(1, 2, 2)$ and $(2, -1, -1)$, while each $C_A, H_A$ has eigenvalues $(4/3, 2)$ and $(1, -1)$. The pair $T = C, S = C + \epsilon H$ is feasible for small positive $\epsilon$, and its divergence ratio tends to $272/35$. Thus exact optimality of the uniform design is compatible with a strict gap between $C^*$ and the displayed lower envelope.

### A.4 A COMPANION COMPARISON FOR DISTRIBUTIONAL ACCURACY

The same compression argument distinguishes the conditioning costs of spatial-score and distributional recovery. Define $\mathcal{K}_\pi(P, Q) = \mathbb{E}_A \mathrm{KL}(P_A \| Q_A)$.

**Proposition 7** (Gaussian KL recovery under a fixed budget)**.** *Under the Gaussian class and uniform masks of Theorem 2, the smallest constant $C^{\mathrm{KL}}_{d,k,\kappa}$ satisfying $\mathrm{KL}(P\|Q) \leq C^{\mathrm{KL}}_{d,k,\kappa}\mathcal{K}_\pi(P, Q)$ obeys*

$$\Phi_{d,k,\kappa} \leq C^{\mathrm{KL}}_{d,k,\kappa} \leq 92\Phi_{d,k,\kappa}, \qquad \Phi_{d,k,\kappa} = \max\left\{p_2^{-1}, \frac{[r + (1-r)\kappa]^2}{r^2}\right\}. \tag{54}$$

*Uniform masks minimize this worst comparison constant among fixed-$k$ designs. If $2\kappa(1-r) < 1$, the constant is at most $\{1 - 2\kappa(1-r)\}^{-1}$. The endpoints $k = d$ and $\kappa = 1$ have constants $1$ and $1/r$, respectively. For fixed $d, k < d$, the right limit is instead $\lim_{\kappa\downarrow 1} C^{\mathrm{KL}}_{d,k,\kappa} = 1/p_2$.*

*Proof.* Use the log-determinant potential $f(C) = -\log\det C$, whose Hessian is $h_C(H) = \mathrm{tr}(C^{-1}HC^{-1}H)$. Centered Gaussian KL is half its Bregman divergence, so

$$\mathrm{KL}\{N(0, S)\|N(0, T)\} = \frac{1}{2}\int_0^1 (1-t)h_{T+t(S-T)}(S - T)\, dt. \tag{55}$$

The covariance condition-number class is convex as before. Means are again controlled by rank-one covariance directions: writing $q = v^\top C^{-1} v$ and $q_A = v_A^\top C_A^{-1} v_A \le q$, we have $h_C(vv^\top) = q^2$ and $h_{C_A}(v_A v_A^\top) = q_A^2 \le q q_A$. Consequently a curvature comparison also bounds the mean KL term.

Keep $X, P_B, R_B$ from Appendix A.3. Here the compressed curvature is unweighted:

$$h_C(H) = \|X\|_{\mathrm{F}}^2, \qquad h_{C_A}(H_A) = \|P_B X P_B\|_{\mathrm{F}}^2. \tag{56}$$

For any mask law, set $\delta = \max_i \mathbb{P}(i \notin A)$. Since $\mathbb{E} R_B \preceq \delta \kappa I$, the exact identity

$$\|P_B X P_B\|_{\mathrm{F}}^2 = \|X\|_{\mathrm{F}}^2 - 2\,\mathrm{tr}(X^2 R_B) + \|R_B X R_B\|_{\mathrm{F}}^2 \tag{57}$$

gives $\mathbb{E}_A h_{C_A}(H_A) \ge (1 - 2\delta\kappa) h_C(H)$. This proves the small-missingness comparison after integrating and adding the mean term. No covariance matrices need commute.

For uniform fixed-$k$ masks, entry counting gives the elementary bound $\kappa^2/p_2$. For $C' = C + (a-1)\lambda_{\min}(C) I$, $a \ge 1$, the same eigenbasis calculation as before gives

$$h_C(H) \le a^2 h_{C'}(H), \qquad h_{C'_A}(H_A) \le h_{C_A}(H_A), \qquad \mathrm{cond}(C') \le 1 + (\kappa - 1)/a. \tag{58}$$

Hence, for $\delta = 1 - r < 1/2$ and $(1 - 2\delta)a > 2\delta(\kappa - 1)$,

$$C^{\mathrm{KL}}_{d,k,\kappa} \le \min\left\{ \frac{\kappa^2}{p_2}, \frac{a^3}{(1-2\delta)a - 2\delta(\kappa - 1)} \right\}. \tag{59}$$

The ridge term is minimized at $a_* = \max\{1, 3\delta(\kappa-1)/(1-2\delta)\}$. When $a_* > 1$ its value is $27\delta^2(\kappa-1)^2/(1-2\delta)^3$. For $\delta < 1/6$, this is at most $(729/8)\delta^2(\kappa-1)^2 \le 92\Phi$; the case $a_* = 1$ gives at most $3/(1-2\delta) < 9/2$. For $\delta \ge 1/6$, use $r^2/p_2 \le 2$ to obtain $\kappa^2/p_2 \le (2/\delta^2)\Phi \le 72\Phi$.

An isolated pair perturbation at the identity gives the lower bound $1/p_2$. For the other term take $u = d^{-1/2}\mathbf{1}$, $C = \kappa I - (\kappa - 1) u u^\top$, and $H = u u^\top$. Then $h_C(H) = 1$ and $h_{C_A}(H_A) = \{r/[r + (1-r)\kappa]\}^2$ for every $k$-mask. The feasible finite pair $T = C, S = C + \epsilon H$ approaches this curvature ratio as $\epsilon \downarrow 0$. Permutation averaging proves optimality of the uniform design. The endpoints follow directly from isotropic Gaussian KL and full observation. The isolated-pair bound and $C^{\mathrm{KL}}_{d,k,\kappa} \le \kappa^2/p_2$ give the stated right limit. At $\kappa = 1$ all nonisotropic perturbation directions are excluded, explaining the discontinuity. □

**Noise-integrated scores.** Let $P_t, Q_t$ follow the same Ornstein–Uhlenbeck flow with generator $\Delta - x \cdot \nabla$. Gaussian relative entropy dissipation gives $\frac{d}{dt}\mathrm{KL}(P_t \| Q_t) = -D_{\mathrm{F}}(P_t, Q_t)$, and both laws converge to the same standard Gaussian. Coordinate marginalization commutes with this flow. Therefore

$$\mathrm{KL}(P\|Q) = \int_0^\infty D_{\mathrm{F}}(P_t, Q_t)\, dt, \qquad \mathcal{K}_\pi(P, Q) = \int_0^\infty D^\pi_{\mathrm{M}}(P_t, Q_t)\, dt. \tag{60}$$

The connection between score discrepancy and entropy dissipation is classical (Lyu, 2009); equation (54) supplies its acquisition cost here. In particular, for $k = d - 1$ this cost is $\Theta\{(1 + \kappa/d)^2\}$, compared with the cubic cost for scores at the original distribution. This comparison concerns the integrated target. With arbitrary covariance scales, a fixed finite noise time alone cannot improve the worst comparison constant: any interior Gaussian pair, after a sufficiently large common rescaling, is the OU image of another pair in the same condition-number class. Specifically, write $b_t = 1 - e^{-2t}$. A target with covariances $cS, cT$ and means $\sqrt{c}\mu, \sqrt{c}\nu$ has preimage covariances $e^{2t}(cS - b_t I), e^{2t}(cT - b_t I)$ and means $e^t\sqrt{c}\mu, e^t\sqrt{c}\nu$. For sufficiently large $c$, both preimage condition numbers are at most $\kappa$ whenever the target pair is strictly inside this class. Common rescaling preserves each joint-to-marginal divergence ratio. Boundary pairs are approximated by replacing $S, T$ with $S + \eta I, T + \eta I$ as $\eta \downarrow 0$; the divergences are continuous and the acquired divergence is positive for distinct Gaussians when $k \ge 2$. Since OU also preserves the condition-number class, the worst comparison constant is therefore exactly unchanged at each fixed finite time, for either KL or spatial Fisher divergence. Pairs approaching the supremum may depend on the noise time.

# B Non-Gaussian recovery: proofs and sharpness

## B.1 A non-Gaussian obstruction

Recovering the true Gaussian law uses the fact that means and covariances determine the joint law. Pair coverage is not sufficient on unrestricted non-Gaussian classes, even under strong regularity.

**Proposition 8** (Pair coverage is insufficient beyond Gaussians). *Let $d \geq 3$ and suppose the mask distribution is supported on proper subsets of $[d]$. There exist distinct smooth, strongly log-concave distributions $P, Q$ on $\mathbb{R}^d$, both with covariance $I$ and globally Lipschitz scores, such that $P_A = Q_A$ for every proper $A \subsetneq [d]$. Consequently, $D_{\mathrm{M}}^{\pi}(P, Q) = 0 < D_{\mathrm{F}}(P, Q)$ for any such mask distribution, even when $w_{\mathrm{cov}} > 0$.*

*Proof.* Let $p(x) = \prod_{i=1}^{d} \varphi(x_i)$ be the standard Gaussian density, put $h(t) = \tanh(t)$ and $H(x) = \prod_{i=1}^{d} h(x_i)$, and define

$$q_\epsilon(x) = p(x)\{1 + \epsilon H(x)\}, \qquad 0 < |\epsilon| < 1. \tag{61}$$

The function $h$ is bounded and has zero mean under the standard Gaussian, so $q_\epsilon$ is a positive probability density. Integrating out any missing coordinate kills the perturbation term. Hence every proper marginal of $Q_\epsilon$ equals the corresponding marginal of $P$, and both covariance matrices equal $I$. On the other hand,

$$D_{\mathrm{F}}(P, Q_\epsilon) = \epsilon^2 \mathbb{E}_P \frac{\|\nabla H(X)\|_2^2}{\{1 + \epsilon H(X)\}^2} > 0. \tag{62}$$

For $h = \tanh$, we have $|H| \leq 1$, $\|\nabla H\|_2^2 \leq d$, and $\|\nabla^2 H\|_{\mathrm{op}} \leq d + 1$. If $0 < |\epsilon| \leq 1/\{8(d+1)\}$, then

$$\|\nabla^2 \log(1 + \epsilon H)\|_{\mathrm{op}} \leq 2|\epsilon|(d+1) + 4\epsilon^2 d < \frac{1}{2}. \tag{63}$$

It follows that $-\nabla^2 \log q_\epsilon \succeq I/2$ and $\|\nabla^2 \log q_\epsilon\|_{\mathrm{op}} \leq 3/2$. Thus $Q_\epsilon$ is strongly log-concave and its score is globally Lipschitz with a constant that remains bounded as $\epsilon \to 0$. □

If $\pi_{[d]} > 0$, the inequality $D_{\mathrm{M}}^{\pi}(P, Q) \geq \pi_{[d]} D_{\mathrm{F}}(P, Q)$ is immediate. Without full observations, positive non-Gaussian results instead need model restrictions that rule out the interaction in Proposition 8 and provide a stable reconstruction of the joint score from observed marginals.

### B.2 The local score operator

We first state the local problem without imposing an interaction model. The result identifies the analytic condition that must replace Gaussian covariance algebra.

**Proposition 9** (Local marginal-score operator). *Let $P$ have a positive $C^1$ density $p$ on $\mathbb{R}^d$. Let $h$ be bounded and $C^1$, with $\mathbb{E}_P h = 0$, and assume $\nabla h \in L_2(P)$ and $\nabla_A h_A \in L_2(P_A)$. Suppose that $h_A(x_A) = \mathbb{E}_P[h(X) \mid X_A = x_A]$ is $C^1$ and may be differentiated under the conditional integral for every mask with $\pi_A > 0$. For sufficiently small $t$, let $Q_t$ have density $q_t = p(1 + th)$. Then*

$$\lim_{t \to 0} \frac{D_{\mathrm{F}}(P, Q_t)}{t^2} = \mathbb{E}_P \|\nabla h(X)\|_2^2, \tag{64}$$

$$\lim_{t \to 0} \frac{D_{\mathrm{M}}^{\pi}(P, Q_t)}{t^2} = \sum_A \pi_A \mathbb{E}_{P_A} \|\nabla_A h_A(X_A)\|_2^2. \tag{65}$$

*Thus the observable tangent operator is*

$$\mathcal{T}_\pi h = \left(\sqrt{\pi_A}\, \nabla_A \mathbb{E}_P[h \mid X_A]\right)_A. \tag{66}$$

*For any linear class $\mathcal{H}$ of admissible tangents, a finite comparison of the limiting quadratic forms, uniform over $h \in \mathcal{H}$, exists if and only if*

$$\alpha_{\mathcal{H}} := \inf_{\substack{h \in \mathcal{H} \\ \|\nabla h\|_{L_2(P)} > 0}} \frac{\sum_A \pi_A \|\nabla_A \mathbb{E}_P[h \mid X_A]\|_{L_2(P_A)}^2}{\|\nabla h\|_{L_2(P)}^2} > 0. \tag{67}$$

*The smallest constant for these quadratic forms is $1/\alpha_{\mathcal{H}}$. Uniform control over a finite neighborhood additionally requires a common bound on the density perturbations.*

*Proof.* Marginalizing $q_t$ gives $q_{t,A} = p_A(1 + th_A)$. Hence

$$\nabla \log p - \nabla \log q_t = -\frac{t \nabla h}{1 + th}, \qquad \nabla_A \log p_A - \nabla_A \log q_{t,A} = -\frac{t \nabla_A h_A}{1 + th_A}. \tag{68}$$

The boundedness assumptions and dominated convergence prove (64)–(65). Taking the infimum of the ratio of the two limiting quadratic forms proves the final claim. In particular, a nonconstant tangent in $\ker(\mathcal{T}_\pi)$ rules out local control. □

The operator condition need not reduce to a combinatorial rule at a dependent reference law. It does at a product law when the admissible density ratios have bounded interaction order. The classical functional ANOVA decomposition and its Sobolev orthogonality (Kuo et al., 2010) let us compute how marginal scores weight each component under an arbitrary mask law.

### B.3 Global comparison for density-ratio ANOVA

**Lemma 1** (A common score–energy comparison)**.** *Let $\mu$ satisfy a Poincaré inequality with constant $C_{\mathrm{P}}$. Let $f, g$ be centered $C^1$ functions satisfying (13), and set $\gamma = L\sqrt{C_{\mathrm{P}}}/m$. Then*

$$\frac{\|\nabla(f-g)\|^2_{L_2(\mu)}}{M(1+\gamma\sqrt{M/m})^2} \le D_{\mathrm{F}}(P_f, P_g) \le \frac{(1+\gamma)^2}{m}\|\nabla(f-g)\|^2_{L_2(\mu)}. \tag{69}$$

*Proof.* Write $a = 1+f, b = 1+g, \delta = f-g, s = \nabla\log(a/b)$, and $D = \int a\|s\|^2 d\mu$. All unqualified norms below are in $L_2(\mu)$. Since $\mathbb{E}_\mu\delta = 0$, the identity $s = (\nabla\delta - \delta\nabla g/b)/a$ gives

$$\sqrt{D} \le m^{-1/2}\{\|\nabla\delta\|_2 + (L/m)\|\delta\|_2\} \le m^{-1/2}(1+\gamma)\|\nabla\delta\|_2. \tag{70}$$

For the lower bound set $v = \delta/b$ and $Q = b\mu$. Then $\mathbb{E}_Q v = 0$, and bounded density perturbation gives

$$\mathrm{Var}_Q(v) \le C_{\mathrm{P}}(M/m)\int \|\nabla v\|^2 dQ. \tag{71}$$

Indeed, $\mathrm{Var}_Q(v) \le M\,\mathrm{Var}_\mu(v)$ and $b \ge m$. Because $\nabla v = (a/b)s$,

$$\int \|\nabla v\|^2 dQ \le (M/m)D, \qquad \|v\|_2^2 \le m^{-1}\,\mathrm{Var}_Q(v) \le C_{\mathrm{P}}M^2m^{-3}D. \tag{72}$$

Finally, $\nabla\delta = as + v\nabla g$ and $\|as\|_2^2 \le MD$ imply

$$\|\nabla\delta\|_2 \le \sqrt{M}(1+\gamma\sqrt{M/m})\sqrt{D}. \tag{73}$$

This proves the lower bound for every finite $L$. □

The projection identity and Lemma 1, together with a separate zero-order estimate, prove Theorem 3 with

$$C_{\mathrm{A}} = \min\left\{R(1+\gamma)^2(1+\gamma\sqrt{R})^2, R[1+\gamma(\sqrt{R}+R)]^2\right\}, \quad R = M/m, \quad \gamma = L\sqrt{C_{\mathrm{P}}}/m. \tag{74}$$

*Proof.* Canonical centering and independence give

$$u_A := \mathbb{E}_\mu[u \mid X_A] = \sum_{B\subseteq A} u_B. \tag{75}$$

If $B \ne C$, only derivatives in $B \cap C$ can contribute to $\langle\nabla h_B, \nabla h_C\rangle$. Integrating over a coordinate in $B \triangle C$ kills each cross term. Consequently

$$\|\nabla h\|^2_{L_2(\mu)} = \sum_B \|\nabla h_B\|^2_{L_2(\mu_B)}, \qquad \sum_A \pi_A\|\nabla_A h_A\|^2_{L_2(\mu_A)} = \sum_B w_B\|\nabla h_B\|^2_{L_2(\mu_B)}. \tag{76}$$

All these operations are justified by the bounded gradients of the sums: their canonical components are finite sums of conditional projections, whose derivatives are bounded. Proposition 9 proves the local identities.

For the global comparison put $\delta = f - g$. Conditional expectation preserves $m, M, L$ and the common product Poincaré constant. Applying the lower bound of Lemma 1 to each marginal and using (76) gives

$$D^\pi_{\mathrm{M}}(P_f, P_g) \ge \frac{w_{\mathcal{B}}}{M(1+\gamma\sqrt{M/m})^2}\|\nabla\delta\|_2^2. \tag{77}$$

The upper bound of the same lemma gives the first term of (74). For the second term, retain the zero-order discrepancy separately. The proof of (71) also gives

$$\|\delta\|_2^2 = \mathbb{E}_Q[bv^2] \le M \operatorname{Var}_Q(v) \le C_{\mathrm{P}} M^3 m^{-2} D_{\mathrm{F}}(P_f, P_g). \tag{78}$$

Apply this inequality to each marginal. Ordinary ANOVA orthogonality, as well as its gradient version, then gives

$$w_{\mathcal{B}} \|\delta\|_2^2 \le C_{\mathrm{P}} M^3 m^{-2} D_{\mathrm{M}}^{\pi}(P_f, P_g), \tag{79}$$

$$w_{\mathcal{B}} \|\nabla\delta\|_2^2 \le M(1+\gamma\sqrt{R})^2 D_{\mathrm{M}}^{\pi}(P_f, P_g). \tag{80}$$

Substitution into the first inequality of (70) proves

$$\sqrt{D_{\mathrm{F}}(P_f, P_g)} \le \sqrt{\frac{R}{w_{\mathcal{B}}}} \{1 + \gamma(\sqrt{R} + R)\} \sqrt{D_{\mathrm{M}}^{\pi}(P_f, P_g)}. \tag{81}$$

This proves the second term without replacing the zero-order discrepancy by another gradient bound.

Finally, for $f = 0$ and $g = \epsilon h_B$, a mask containing $B$ gives the full divergence; every other mask removes the perturbation. This proves (18), including the zero-coverage obstruction and sharp reciprocal-coverage factor. The order-$r$ hierarchy follows from $w_B \ge w_C$ whenever $B \subseteq C$. Nonzero bounded smooth centered products supply witnesses for every $B$ under the positive product density. □

**Approximate interaction structure.** The same argument quantifies departures from an exact order restriction. Allow all canonical components and let $E_{<\tau} = \sum_{B: w_B < \tau} \|\nabla(f_B - g_B)\|_2^2$ for $\tau > 0$. With the same $C_{\mathrm{A}}$ as in (74),

$$\sqrt{D_{\mathrm{F}}(P_f, P_g)} \le \sqrt{\frac{C_{\mathrm{A}}}{\tau} D_{\mathrm{M}}^{\pi}(P_f, P_g)} + \frac{1+\gamma}{\sqrt{m}} \sqrt{E_{<\tau}}. \tag{82}$$

In particular, for any $\eta > 0$, the divergence is at most $(1+\eta) C_{\mathrm{A}} D_{\mathrm{M}}^{\pi}/\tau + (1+\eta^{-1})(1+\gamma)^2 E_{<\tau}/m$. Both coefficients grow quadratically in $L$. The remainder measures energy in poorly observed or unobserved interactions. An entirely invisible component cannot be bounded from the acquired data.

To prove the bound, split $\delta = f - g = \delta_+ + \delta_-$ by $w_B \ge \tau$ and $w_B < \tau$. Orthogonality and the two marginal estimates above give

$$\|\nabla\delta_+\|_2^2 \le \frac{M(1+\gamma\sqrt{R})^2}{\tau} D_{\mathrm{M}}^{\pi}, \qquad \|\delta_+\|_2^2 \le \frac{C_{\mathrm{P}} M^3}{m^2 \tau} D_{\mathrm{M}}^{\pi}. \tag{83}$$

Poincaré gives $\|\delta_-\|_2 \le \sqrt{C_{\mathrm{P}} E_{<\tau}}$. Substituting these bounds and $\|\nabla\delta_-\|_2 = \sqrt{E_{<\tau}}$ into (70), before its last Poincaré step, proves (82) with $R[1+\gamma(\sqrt{R}+R)]^2$. Applying the gradient-energy comparison alone gives the same residual coefficient and $C_{\mathrm{E}} = R(1+\gamma)^2(1+\gamma\sqrt{R})^2$ for the observed term. Taking the smaller coefficient proves the stated $C_{\mathrm{A}}$ bound.

Using gradient orthogonality before taking square roots also gives the complementary bound $D_{\mathrm{F}} \le C_{\mathrm{E}} D_{\mathrm{M}}^{\pi}/\tau + (1+\gamma)^2 E_{<\tau}/m$. It has no cross term and can be smaller when $L$ is small. The square-root bound preserves the quadratic cost for large $L$.

The class contains smooth non-Gaussian examples with $h_B(x_B) = \prod_{i \in B} \tanh(x_i)$ at a Gaussian reference. Small perturbations are strongly log-concave by Proposition 8. The global theorem also permits larger perturbations satisfying its density and gradient bounds. The supplementary certificate checks the intermediate energy bounds, including cases with $\gamma > 1$.

### B.4 Sharpness of the gradient cost

**Proposition 10** (Covered interactions with unstable joint scores). *Fix $\mu = N(0, I_2)$, $m = 1/2$, $M = 3/2$, and the first-order density-ratio class $\mathcal{B} = \{\{1\}, \{2\}\}$. Observe $\{1\}$ and $\{2\}$ with probabilities $w$ and $1-w$, where $0 < w \le 1/2$. For every $L \ge 1$,*

$$\sup_{f,g} \frac{D_{\mathrm{F}}(P_f, P_g)}{D_{\mathrm{M}}^{\pi}(P_f, P_g)} \ge \frac{\tanh(1)}{144} \frac{L^2}{w}, \tag{84}$$

*where the supremum is over this class satisfying (13) with gradient budget $L$ and positive denominator. Together with Theorem 3, the worst-case large-$L$ order is $L^2/w$.*

*Proof.* Put $\epsilon = \beta = 1/4$ and, for an integer $N \geq 1$, let

$$b(x, y) = 1 + \beta \sin(Ny), \qquad a(x, y) = 1 + \beta \sin(Ny) + \epsilon \sin x. \tag{85}$$

Both perturbations are centered sums of singleton functions. Their ratios lie in $[1/2, 3/2]$ uniformly in $N$, and

$$\|\nabla b\|_\infty \leq N/4, \qquad \|\nabla a\|_\infty \leq \frac{\sqrt{1+N^2}}{4} \leq N/2. \tag{86}$$

The $y$ marginals agree, while the $x$ density ratios are $1 + \epsilon \sin x$ and 1. Therefore

$$D_{\mathrm{M}}^{\pi}(a\mu, b\mu) = w\epsilon^2 \mathbb{E}_\varphi \frac{\cos^2 X}{1 + \epsilon \sin X} \leq \frac{w\epsilon^2}{1-\epsilon} \mathbb{E}_\varphi \cos^2 X, \tag{87}$$

which is positive and independent of $N$. Direct differentiation gives

$$a\, \partial_x \log(a/b) = \epsilon \cos x, \tag{88}$$

$$a\, \partial_y \log(a/b) = -\epsilon\beta N \sin x \cos(Ny)/b. \tag{89}$$

The nonnegative $y$ contribution alone implies

$$D_{\mathrm{F}}(a\mu, b\mu) \geq \frac{\epsilon^2 \beta^2 N^2}{M^3} \mathbb{E}_\varphi \sin^2 X\, \mathbb{E}_\varphi \cos^2(NY). \tag{90}$$

The Gaussian characteristic function gives $\mathbb{E} \cos^2(NY) \geq 1/2$ and $\mathbb{E} \sin^2 X / \mathbb{E} \cos^2 X = \tanh(1)$. Dividing by (87) yields $D_{\mathrm{F}}/D_{\mathrm{M}}^{\pi} \geq \tanh(1) N^2/(144w)$. Finally choose $N = \lfloor 2L \rfloor$, for which $N/2 \leq L$ and $N \geq L$. □

The construction keeps $a - b = \epsilon \sin x$ and both observation probabilities fixed. Only the common background's frequency changes, establishing the quadratic gradient cost within the covered first-order density-ratio class.

For a product candidate $Q = Q_X \otimes Q_Y$, conditional centering of the joint score gives

$$D_{\mathrm{F}}(P, Q) = D_{\mathrm{F}}(P, P_X \otimes P_Y) + D_{\mathrm{F}}(P_X, Q_X) + D_{\mathrm{F}}(P_Y, Q_Y). \tag{91}$$

The first term is the gradient mutual information of Ding et al. (2019, Definition 1), up to their factor $1/2$. In (85), $Q$ is a product and the marginal discrepancy is fixed; the proposition quantifies the growing dependence-score term inside the specified density-ratio class.

### B.5 KL RECOVERY WITHOUT A GRADIENT COST

Retain the canonical density-ratio representation and bounds $m, M$ of Theorem 3, with $w_{\mathcal{B}} > 0$. Then

$$\mathrm{KL}(P_f \| P_g) \leq \frac{M}{m w_{\mathcal{B}}} \sum_A \pi_A \,\mathrm{KL}((P_f)_A \| (P_g)_A). \tag{92}$$

This comparison requires neither a gradient bound nor a Poincaré inequality. It follows from ordinary ANOVA orthogonality and the scalar entropy curvature, as we verify directly. For density ratios $a, b \in [m, M]$ with unit integrals, the Taylor remainder of $z \log z$ gives

$$\mathrm{KL}(a\mu \| b\mu) = \int_0^1 (1-t) \mathbb{E}_\mu \frac{(a-b)^2}{b + t(a-b)}\, dt, \qquad \frac{\|a-b\|_2^2}{2M} \leq \mathrm{KL}(a\mu \| b\mu) \leq \frac{\|a-b\|_2^2}{2m}. \tag{93}$$

Writing $\delta = f - g$, marginalization preserves these density bounds and

$$\sum_A \pi_A \|\mathbb{E}_\mu[\delta \mid X_A]\|_2^2 = \sum_{B \in \mathcal{B}} w_B \|\delta_B\|_2^2 \geq w_{\mathcal{B}} \|\delta\|_2^2. \tag{94}$$

Combining the two displays proves (92). The reciprocal coverage factor is necessary: a single canonical perturbation $g = \epsilon h_B$ with $f = 0$ gives the exact ratio $1/w_B$, because masks containing $B$ retain the full divergence and all other masks give zero. The argument does not determine the optimal density-bound prefactor.

Thus both KL and Fisher recovery require coverage of the allowed interactions, but the quadratic gradient cost in Proposition 10 is specific to score recovery. In its two-sine example with $w = 1/2$, the KL ratio stays at most six for every background frequency, whereas the Fisher ratio grows quadratically.

# C LEARNING FROM COVERED INTERACTIONS

We prove the statistical statements in Section 6. Throughout, the reference $\mu$ and finite canonical basis are known. The records are independent, and masks are fixed in advance or drawn independently of the data. Write

$$m = 1 - \rho, \quad M = 1 + \rho, \quad L = \rho \max_j \frac{\|\nabla \psi_j\|_\infty}{b_j}, \qquad \gamma = \frac{L\sqrt{C_{\mathrm{P}}}}{m},$$
$$c_+ = \frac{(1+\gamma)^2}{m}, \quad c_- = \frac{1}{M(1+\gamma\sqrt{M/m})^2}. \tag{95}$$

Every $f_\theta = \sum_j \theta_j \psi_j$ has mean zero, satisfies $m \le 1 + f_\theta \le M$, and has gradient bounded by $L$. For different canonical supports $B, C$, a common nonzero gradient coordinate $i$ belongs to $B \cap C$. Integration over a coordinate in $B \triangle C$ then makes the gradient inner product vanish. Thus a finite gradient Gram is block diagonal by support. An orthogonal rotation within each block preserves canonicity and $L^2(\mu)$ orthonormality and yields (20).

## C.1 FISHER LOSS AND DIRICHLET ERROR

**Lemma 2.** *For $\theta, \eta \in \Theta_\rho$,*

$$c_- \|\theta - \eta\|_D^2 \le D_{\mathrm{F}}(P_\theta, P_\eta) \le c_+ \|\theta - \eta\|_D^2. \tag{96}$$

*Proof.* Apply Lemma 1 to $f_\theta, f_\eta$. Equation (20) gives $\|\nabla(f_\theta - f_\eta)\|_{L_2(\mu)}^2 = \|\theta - \eta\|_D^2$, yielding the constants in (95). □

## C.2 OBSERVABLE MOMENTS AND THE FINITE-SAMPLE UPPER BOUND

For $N_j > 0$, let $z_j = N_j^{-1} \sum_{i: B_j \subseteq A_i} \psi_j(X_i)$, and otherwise let $z_j = 0$. Orthogonality gives $\mathbb{E}_\theta \psi_j = \theta_j$, while $\mathbb{E}_\theta \psi_j^2 \le M$. Conditional on a fixed mask schedule,

$$\mathbb{E}_\theta \|z - \theta\|_D^2 \le M \sum_{j: N_j > 0} \frac{\lambda_j}{N_j} + \sum_{j: N_j = 0} \lambda_j \theta_j^2. \tag{97}$$

Correlations between coefficient estimates do not enter this diagonal loss. Projection onto the closed convex set $\Theta_\rho$ in the $D$ norm cannot increase its squared distance from $\theta$. Lemma 2 therefore proves (22), with the two explicit coefficients $c_+ M$ and $c_+$, respectively.

For independent masks, $N_j \sim \mathrm{Binomial}(n, w_{B_j})$. The inequality $1/k \le 2/(k+1)$ for $k \ge 1$ and the binomial identity for $\mathbb{E}[1/(N_j + 1)]$ imply

$$\mathbb{E} \frac{\mathbf{1}_{N_j > 0}}{N_j} \le \frac{2}{(n+1) w_{B_j}} \qquad (w_{B_j} > 0). \tag{98}$$

Thus

$$\mathbb{E}_\theta D_{\mathrm{F}}(P_\theta, P_{\widehat{\theta}_{\mathrm{mom}}}) \le c_+ \left\{ \frac{2M}{n+1} \sum_{j: w_{B_j} > 0} \frac{\lambda_j}{w_{B_j}} + \sum_j \lambda_j \theta_j^2 (1 - w_{B_j})^n \right\}. \tag{99}$$

When every $w_{B_j} > 0$, the second term is uniformly $O(\sum_j \lambda_j / (n w_{B_j}))$: use $\theta_j^2 \le \rho^2 / b_j^2 \le \rho^2$ and $(1-w)^n \le e^{-nw} \le 1/(enw)$. This proves the minimax upper bound in Theorem 4.

## C.3 A LOWER BOUND FOR ARBITRARY SCORE ESTIMATORS

Choose amplitudes $a_j > 0$ satisfying

$$\sum_j b_j a_j \le \rho, \qquad n w_{B_j} a_j^2 \le m/16. \tag{100}$$

The second restriction is void if $w_{B_j} = 0$. All vertices $\theta_j \in \{-a_j, a_j\}$ belong to the model. Two neighbors in coordinate $j$ have the same observed law unless $B_j \subseteq A$. When this interaction is visible, their marginal density ratios differ by $2a_j\psi_j$. The inequality $\mathrm{KL} \le \chi^2$ bounds their one-record observed KL by $4w_{B_j}a_j^2/m$. The $n$-record KL is at most $1/4$, hence total variation is at most $1/\sqrt{8} < 1/2$.

Under uniform independent signs, every coordinate sign test has average error at least $1/4$; convexity of total variation gives the same conclusion after mixing the nuisance signs. A sign error entails squared coordinate error at least $a_j^2$. It follows that every parameter estimator has worst-case $D$-risk at least $\frac{1}{4}\sum_j \lambda_j a_j^2$.

For an arbitrary estimated score field $\widehat{s}$, choose an $L^2(\mu)$ nearest score $s_{\widehat{\eta}}$ among $s_\eta = \nabla \log p_\eta$, $\eta \in \Theta_\rho$. Compactness and continuity give existence. Measurable approximate minimizers yield the same bound by a limit. Subtracting the common reference score makes this definition valid whenever the risk is finite. The true score is an admissible competitor, so

$$\|\widehat{s} - s_\theta\|^2_{L^2(P_\theta)} \ge \frac{m}{4}\|s_{\widehat{\eta}} - s_\theta\|^2_{L^2(\mu)} \tag{101}$$

$$\ge \frac{mc_-}{4M}\|\widehat{\eta} - \theta\|_D^2. \tag{102}$$

Consequently, the minimax score risk is at least

$$\frac{mc_-}{16M}\sum_j \lambda_j a_j^2. \tag{103}$$

Taking $a_j = \min\{\rho/(sb_j), \sqrt{m/(16nw_{B_j})}\}$, with the square root interpreted as infinity at zero coverage, proves a nonvanishing lower bound for any uncovered interaction. Once the square-root choice is active for every $j$, (103) becomes $\frac{m^2c_-}{256Mn}\sum_j \lambda_j/w_{B_j}$, proving Theorem 4. This threshold specifies the finite-model regime of the rate statement. For a fixed schedule replace $nw_{B_j}$ by $N_j$ throughout.

### C.4 Background regularity has a finite-sample cost

The family used in Proposition 10 also separates observation counts from the geometry of score estimation. Let $\mu = \varphi \otimes \varphi$ and let $b$ be a known positive $C^1$ function with $\mathbb{E}_\varphi b = 1$, $3/4 \le b \le 5/4$, and $\mathbb{E}_\varphi(b')^2 < \infty$. Consider

$$p_\theta(x, y) = \mu(x, y)\{b(y) + \theta \sin x\}, \quad |\theta| \le 1/4. \tag{104}$$

Only $\theta$ is unknown. Independently for each of $n$ records, observe $x$ with probability $w$ and $y$ otherwise, where $0 \le w \le 1$; the mask is also observed. Put $c_s = (1 - e^{-2})/2$, $c_c = (1 + e^{-2})/2$, and

$$G_b = c_c\mathbb{E}_\varphi[b^{-1}] + c_s\mathbb{E}_\varphi[(b')^2/b^3]. \tag{105}$$

**Proposition 11** (Sample cost of the common background)**.** *There are universal positive constants, independent of $n, b, w$, such that*

$$\inf_{\widehat{s}} \sup_{|\theta|\le 1/4} \mathbb{E}_\theta\|\widehat{s} - \nabla \log p_\theta\|^2_{L^2(P_\theta)} \asymp \frac{G_b}{1 + nw}. \tag{106}$$

*For each known $b$, the infimum allows arbitrary estimated vector fields that may depend on $b$. A density estimate in (104) attains the upper bound. At $\theta = 0$, the observed parameter information per record is $wc_s$, whereas the joint spatial-score curvature is $G_b$. In particular, for $b(y) = 1 + \frac{1}{4}\sin(Ny)$, $N \ge 1$, the risk is $\asymp (1 + N^2)/(1 + nw)$ with universal constants.*

*Proof.* Write $a_\theta = b + \theta h$, where $h(x) = \sin x$. The score difference is $(\theta - \eta)(b\nabla h - h\nabla b)/(a_\theta a_\eta)$. Since $|\theta h/b| \le 1/3$, integration under $P_\theta$ gives

$$\frac{27}{64}G_b(\theta - \eta)^2 \le D_{\mathrm{F}}(P_\theta, P_\eta) \le \frac{27}{8}G_b(\theta - \eta)^2. \tag{107}$$

The two numerator components are orthogonal coordinate vectors, which yields (105). For the sinusoidal background, $3/4 \le b \le 5/4$ and $\mathbb{E}\cos^2(NY) \ge 1/2$ imply $G_b \asymp 1 + N^2$ uniformly in $N$.

Let $K \sim \mathrm{Binomial}(n, w)$ count the observed $x$ values. Their marginal is $\varphi(x)(1 + \theta \sin x)$, with $\mathbb{E}_\theta \sin X = \theta c_s$ and $\mathrm{Var}_\theta(\sin X) \le c_s$. For $K > 0$, clip their mean divided by $c_s$ to $[-1/4, 1/4]$; use zero when $K = 0$. This estimator satisfies

$$\mathbb{E}_\theta(\widehat{\theta} - \theta)^2 \le \frac{\mathbb{P}(K = 0)}{16} + \frac{1}{c_s}\mathbb{E}\left[\frac{\mathbf{1}_{K>0}}{K}\right] \lesssim \frac{1}{1 + nw}. \tag{108}$$

For the last bound use $1/K \le 2/(K+1)$ on $K > 0$, $\mathbb{E}[1/(K+1)] = \{1-(1-w)^{n+1}\}/\{(n+1)w\}$ for $w > 0$, and the trivial squared-error bound $1/4$. For $w = 0$, the estimator is identically zero and the same upper bound is immediate. Equation (107) proves the upper bound.

For the lower bound, take $\theta_\pm = \pm u$ with $u = 1/(8\sqrt{1 + nw})$. The $y$ observations have identical laws, and the one-record observed KL divergence is at most $4wu^2 c_s/(1 - u) \le 6wu^2$. Thus the $n$-record KL is at most $6/64$ and the total variation is less than $1/4$. Since $1/2 \le a_\theta \le 3/2$, the squared $L^2(\mu)$ separation of the two score fields is at least $(2/3)(27/64)G_b(2u)^2$. Any estimated field induces a test by selecting the nearer score in $L^2(\mu)$. An error in this test costs at least a quarter of that squared separation, and risk under either $P_\theta$ is at least half its $L^2(\mu)$ risk. The average testing error is at least $(1 - \mathrm{TV})/2$, proving the lower bound for arbitrary fields. Finally, differentiating the observed likelihood at zero gives the parameter score $\mathbf{1}_{A=\{1\}} \sin X$, with variance $wc_s$; differentiating the joint spatial score gives curvature $G_b$. □

**Learning the background amplitude.** The observed $x$ distribution for the unknown coefficient does not depend on $b$, whereas its joint-score metric $G_b$ does. We now also estimate the background amplitude, keeping the basis and frequency known.

**Corollary 12** (Unknown background amplitude)**.** *Let $N \ge 1$, $n \ge 1$, and $0 \le w \le 1$. For the family*

$$p_{\theta,\beta}(x, y) = \mu(x, y)\{1 + \theta \sin x + \beta \sin(Ny)\}, \qquad (\theta, \beta) \in [-1/4, 1/4]^2, \tag{109}$$

*observe $x$ with probability $w$ and $y$ otherwise, independently in each record. With universal constants independent of $n, N, w$,*

$$\inf_{\widehat{s}} \sup_{\theta,\beta} \mathbb{E}_{\theta,\beta} \|\widehat{s} - \nabla \log p_{\theta,\beta}\|^2_{L^2(P_{\theta,\beta})} \asymp \frac{1 + N^2}{1 + nw} + \frac{N^2}{1 + n(1 - w)}. \tag{110}$$

*The infimum permits arbitrary estimated vector fields. Clipped estimates of the two acquired sine moments attain the upper bound. For $w \le 1/2$, the rate simplifies to $(1 + N^2)/(1 + nw)$.*

*Proof.* Write $h = \sin x$, $k = \sin(Ny)$, $a = 1 + \theta h + \beta k$, and $\widetilde{a} = 1 + \eta h + \beta' k$. Put $\Delta_\theta = \theta - \eta$, $\Delta_\beta = \beta - \beta'$, and $D = \beta\Delta_\theta - \theta\Delta_\beta$. The exact Fisher divergence is

$$D_{\mathrm{F}}(P_{\theta,\beta}, P_{\eta,\beta'}) = \mathbb{E}_\mu \frac{\cos^2 x(\Delta_\theta + Dk)^2 + N^2 \cos^2(Ny)(\Delta_\beta - Dh)^2}{a\widetilde{a}^2}. \tag{111}$$

Let $c_N = (1 + e^{-2N^2})/2$, $s_N = 1 - c_N$. Since $1/2 \le a, \widetilde{a} \le 3/2$, reflection symmetry gives

$$\mathcal{Q} = c_c\Delta_\theta^2 + N^2 c_N \Delta_\beta^2 + (c_c s_N + N^2 c_N c_s)D^2, \tag{112}$$

$$\frac{8}{27}\mathcal{Q} \le D_{\mathrm{F}}(P_{\theta,\beta}, P_{\eta,\beta'}) \le 8\mathcal{Q} \le 10\{(1 + N^2)\Delta_\theta^2 + N^2\Delta_\beta^2\}. \tag{113}$$

The last step uses $D^2 \le (\Delta_\theta^2 + \Delta_\beta^2)/8$ and $N \ge 1$. The acquired moments have expectations $\theta c_s$ and $\beta s_N$, and second moments $c_s$ and $s_N$, respectively. Estimate each amplitude from its own acquired records, use zero for an empty sample, and clip to $[-1/4, 1/4]$. Since $c_s \le s_N \le 1/2$, the preceding binomial-count argument bounds their mean squared errors by $C/(1 + nw)$ and $C/(1 + n(1 - w))$. Equation (113) proves the upper bound, without requiring independence of the two counts.

For the first lower-bound term, restrict to $\beta = 1/4$ and apply Proposition 11. Revealing this fixed background to the estimator can only reduce risk. For the second, set $\theta = 0$ and $\beta_\pm = \pm v$, where $v = 1/(8\sqrt{1 + n(1 - w)})$. Only acquired $y$ values distinguish them; the $n$-record observed KL is at most $6/64$. Their squared $L^2(\mu)$ score separation is bounded below by a universal multiple of $N^2v^2$, by (113) and $c_N \ge 1/2$. The same nearest-score testing argument used above therefore gives $cN^2/(1 + n(1 - w))$ for arbitrary fields. The maximum of the two lower bounds is at least half their sum after choosing the smaller constant. □

### C.5 SHARP LOCAL COEFFICIENT: A DIRECT INFORMATION PROOF

**Theorem 13.** *Under the conditions of Theorem 4, with every $w_{B_j} > 0$, the local asymptotic minimax joint-score coefficient at $\theta = 0$ is $\mathcal{R}_{\text{eff}}$ in (25). More precisely, allowing arbitrary score estimators,*

$$\lim_{c\to\infty} \liminf_{n\to\infty} \inf_{\widehat{s}_n} \sup_{\|h\|\le c} n\mathbb{E}_{h/\sqrt{n}} \|\widehat{s}_n - s_{h/\sqrt{n}}\|^2_{L^2(P_{h/\sqrt{n}})} \ge \mathcal{R}_{\text{eff}}. \tag{114}$$

*The projected moment estimator attains this coefficient uniformly on every fixed bounded set of local parameters $h$.*

*Proof.* We give the information argument explicitly, following the classical van Trees principle (Gill & Levit, 1995). Let $W = \operatorname{diag}(w_{B_j})$. At zero the observed parameter score is $\{\mathbf{1}_{B_j \subseteq A}\psi_j(X)\}_j$, with information $W$. Bounded features and positivity imply that the information $I(\theta)$ is continuous at zero. In the local parameter $h$, the $n$-record information is $I(h/\sqrt{n})$.

Let $a = c/\sqrt{s}$ and give each $h_j$ independent density $a^{-1}\cos^2(\pi h_j/(2a))\mathbf{1}_{|h_j|<a}$. The prior is supported on $\|h\| \le c$, vanishes at its boundary, and has information $J_c = \pi^2 s c^{-2} I_s$. For the joint data–prior density, let $U$ be its $h$ score. If $g(h)$ is a differentiable vector target and $\widehat{g}$ any estimator, integration by parts and matrix Cauchy–Schwarz give

$$\mathbb{E}[(\widehat{g} - g)U^\top] = \mathbb{E}\nabla g =: B, \qquad \mathbb{E}[(\widehat{g} - g)(\widehat{g} - g)^\top] \succeq B\{\mathbb{E}I(h/\sqrt{n}) + J_c\}^{-1}B^\top. \tag{115}$$

These identities require no unbiasedness of the estimator.

Define $t_j(\theta) = \lambda_j^{-1}\langle s_\theta - s_0, \nabla\psi_j\rangle_{L^2(\mu)}$ and $g_n(h) = \sqrt{n}\, t(h/\sqrt{n})$. The score formula gives $\nabla t(0) = I_s$, so $\nabla g_n \to I_s$ uniformly for bounded $h$. Projecting $\widehat{s} - s_0$ onto the orthogonal span of $\nabla\psi_j$ induces an estimator $\widehat{g}_n$. Bessel's inequality gives

$$n\|\widehat{s} - s_{h/\sqrt{n}}\|^2_{L^2(\mu)} \ge \|\widehat{g}_n - g_n(h)\|^2_D. \tag{116}$$

Moreover $p_{h/\sqrt{n}}/\mu = 1 + O_c(n^{-1/2})$ uniformly. Applying (115) to the integrated risk, then taking $n \to \infty$, gives the lower bound $\operatorname{tr}\{D(W + J_c)^{-1}\}$. Finally $c \to \infty$ yields $\operatorname{tr}(DW^{-1})$.

For attainment, conditional on the masks each visible moment mean is unbiased with variance $\operatorname{Var}_\theta(\psi_j)/N_j$. Uniformly for $\theta = h/\sqrt{n}$ with bounded $h$, $\operatorname{Var}_\theta(\psi_j) = 1 + O_c(n^{-1/2})$ and $n\mathbb{E}[\mathbf{1}_{N_j>0}/N_j] \to 1/w_{B_j}$. Binomial concentration controls small counts, and $\mathbb{P}(N_j = 0)$ decays exponentially. Thus $n\mathbb{E}\|z - \theta\|^2_D \to \operatorname{tr}(DW^{-1})$ uniformly on each bounded local set. Bounded-variable concentration makes the projection inactive with exponentially high probability locally and gives $\mathbb{E}\|z - \theta\|^4 = O_c(n^{-2})$. Finally the score formula implies

$$D_{\text{F}}(P_\theta, P_\eta) = \|\eta - \theta\|^2_D + O\{(\|\eta\| + \|\theta\|)\|\eta - \theta\|^2\} \tag{117}$$

in a fixed neighborhood of zero. Its expected remainder is $o(n^{-1})$. Outside that neighborhood, exponential concentration and bounded loss on $\Theta_\rho$ control the remainder. This proves attainment. A deterministic schedule with $N_j/n \to w_{B_j} > 0$ has the same proof. □

### C.6 THE SCORE-MATCHING COEFFICIENT

Group an orthonormal basis by support $B$. Let $G_B$ be its gradient Gram and $T_B$ the Gram of its image under the reference generator $\mathcal{L}_\mu = \Delta + \nabla \log\mu \cdot \nabla$. The score-matching coefficient becomes an estimator statement under standard local M-estimation conditions. Specifically, the true parameter is an interior, locally unique minimizer of each population criterion. The reported local minimizers are consistent and satisfy their estimating equations. The criteria are twice continuously differentiable on a fixed neighborhood, with integrable envelopes that yield a uniform law for their Hessians. The estimating equations satisfy a central limit theorem, and the limiting Hessians and observed information are nonsingular. We call an estimator that satisfies these conditions *regular*. Uniform integrability is needed only to convert distributional limits into limits of expected risk.

**Proposition 14.** *Use a finite $L^2(\mu)$-orthonormal canonical basis, grouped by support $B$, with $w_B > 0$ on each fitted support. Suppose its functions are $C^2$, integration by parts is valid for $\mathcal{L}_\mu$, and the generator images are square integrable and canonical on the same supports. Let $G_B$ and $T_B$*

*be the Grams defined above. For consistent regular roots of the marginal-score estimating equations at $\theta = 0$, with uniform integrability for expected-risk limits,*

$$\mathcal{R}_{\mathrm{SM}} = \mathcal{R}_{\mathrm{eff}} + \sum_B \frac{\mathrm{tr}\{G_B^{-1}(T_B - G_B^2)\}}{w_B}, \qquad T_B - G_B^2 \succeq 0. \tag{118}$$

*Equality with $\mathcal{R}_{\mathrm{eff}}$ holds exactly when each fitted interaction span is invariant under $\mathcal{L}_\mu$.*

*Proof.* For a feature visible under $A$, the Hyvärinen-loss derivative at zero is $2\mathcal{L}_\mu\psi_j$; otherwise it vanishes. Its sensitivity block is $H_B = 2w_B G_B$, and its variance block is $V_B = 4w_B T_B$. Different supports have zero cross blocks by canonicity. The sandwich covariance and joint-score Gram therefore give coefficient $w_B^{-1}\,\mathrm{tr}(G_B^{-1}T_B)$ for block $B$. Integration by parts gives $\mathbb{E}_\mu \psi_B(\mathcal{L}_\mu\psi_B)^\top = -G_B$. The $L^2(\mu)$ projection residual of $\mathcal{L}_\mu\psi_B$ onto the span of $\psi_B$ has Gram $T_B - G_B^2 \succeq 0$. Summing the blocks proves the decomposition. Since every $G_B$ is positive definite, its extra trace vanishes exactly when this residual does. □

### C.7 Background-learning experiment

We evaluate the two clipped observable-moment estimators from Corollary 12. The fixed grid is $\theta \in \{0, 1/8\}$, $\beta \in \{1/8, 1/4\}$, $N \in \{1, 4, 16, 32\}$, $w \in \{1/8, 1/2, 7/8\}$, and $n \in \{128, 512, 2048, 8192\}$, with 2,048 repetitions per configuration. All 192 configurations are retained. The paired oracle comparator uses the identical estimate of $\theta$ and the true $\beta$. Every training record supplies only its acquired coordinate. Exact normal-proposal rejection sampling generates each coordinate marginal. Independent masks are shared across configurations; estimates of $\theta$ are shared across $\beta, N$, estimates of $\beta$ across $\theta$, and sample sizes use nested prefixes. These shared draws couple configurations without increasing the number of independent repetitions.

Population risk is evaluated by (111). Periodic quadrature under wrapped Gaussian measures and a degree-24 Chebyshev representation of the candidate-dependent metric accelerate evaluation. Direct positive integration at interior and boundary pairs, 128/256-point quadrature, degree-18/24 interpolation, and an independent Gaussian-line quadrature agree to within $3.1 \times 10^{-15}$ relatively on the checked configurations. Every simulated pair satisfies (113). These checks assess numerical resolution. The risk comparison itself is proved analytically. Bands show the mean $\pm 1.96$ Monte Carlo standard errors.

For interpreting constants, let $J$ be the joint spatial-score parameter metric at the truth. In the interior the moment estimator has asymptotic risk coefficient

$$J_{11}\frac{1/c_s - \theta^2}{w} + J_{22}\frac{1/s_N - \beta^2}{1 - w}. \tag{119}$$

The parameter covariance is asymptotically diagonal because the acquired coordinate contributions are disjoint and centered. At $\beta = 1/4$, with $\theta$ interior, projection halves the second term. At $n = 8192$ the interior cases have risk/prediction ratios between $0.946$ and $1.040$. The complete result record includes the boundary and smaller-sample outcomes. These pointwise risks test the estimator's local approximation. The uniform minimax order in (110) follows from its separate proof and does not require equal constants for the learned and known backgrounds.

### C.8 Logistic example and experimental protocol

Let $\mu$ be the four-dimensional standard logistic product law and put $u_i = \tanh(x_i/2)$. Under $\mu$, the $u_i$ are independent uniforms on $[-1, 1]$. The normalized feature $\phi_i = \sqrt{3}u_i$ satisfies

$$\mathbb{E}\phi_i^2 = 1, \quad \mathbb{E}(\phi_i')^2 = 2/5, \quad \mathcal{L}_\mu\phi_i = -\phi_i(1 - u_i^2), \quad \mathbb{E}(\mathcal{L}_\mu\phi_i)^2 = 8/35. \tag{120}$$

For $\psi_B = \prod_{i\in B}\phi_i$, $r = |B|$, direct product integration yields

$$\lambda_B = 2r/5, \qquad \tau_B = 8r/35 + 4r(r-1)/25, \qquad \frac{\tau_B}{\lambda_B^2} = 1 + \frac{3}{7r}. \tag{121}$$

The logistic reference has a finite Poincaré constant. For example, its one-dimensional Cheeger constant is $1/2$, so the Cheeger–Poincaré bound and tensorization give the sufficient bound $C_{\mathrm{P}} \le 16$.

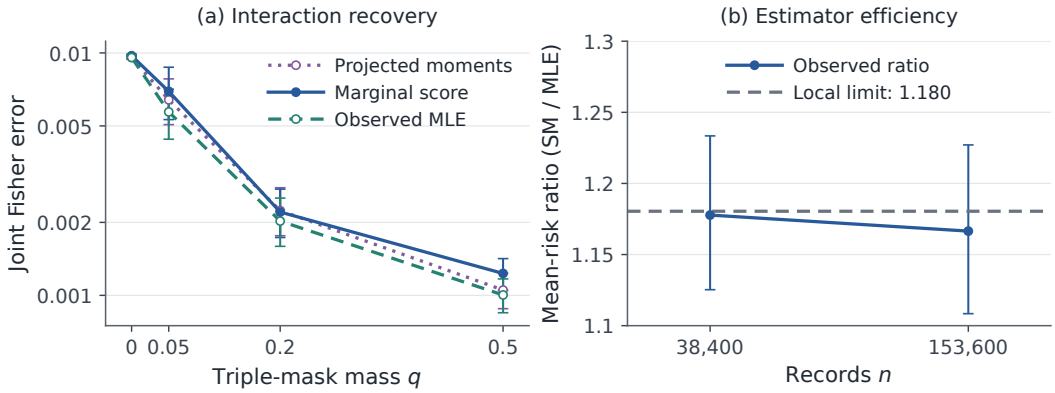


Figure 4: Statistical cost of interaction coverage under logistic reference. (a) Mean joint Fisher error for the nonzero-interaction truth at $n = 38400$. The x-axis is the total probability of a size-three mask; the measurement budget stays fixed. (b) Ratio of mean SM and MLE risks at the product truth with $q = .5$, using 100 fresh paired repetitions per sample size. The dashed line is the theoretical local ratio $71.7714/60.8$. Bars in both panels are 95% bootstrap Monte Carlo intervals. Axes are logarithmic only for risk in (a) and sample size in (b); points use their actual numerical coordinates.

For computation we use unnormalized features $\varphi_B = \prod_{i \in B} u_i$ for all 14 supports of orders one through three, with raw coefficients $\sum_B |\alpha_B| \le 0.85$. Thus $\theta_B = \alpha_B 3^{-r/2}$, $v_B = \mathbb{E}\varphi_B^2 = 3^{-r}$, and $g_B = \mathbb{E}\|\nabla\varphi_B\|^2 = (2r/5)3^{-r}$. The two truths are the product law and the interaction model with raw coefficients $\alpha_{\{1\}} = .04$, $\alpha_{\{2,4\}} = .08$, $\alpha_{\{1,2,3\}} = .42$, and $\alpha_{\{1,3,4\}} = -.12$. The Gaussian-reference study in Appendix E.2 uses the same coefficient pattern. The logistic reference here permits explicit efficiency constants.

Masks have size-one, size-two and size-three masses $q, 1-2q, q$, uniform within each size, for $q \in \{0, .05, .2, .5\}$. Each 2400-record block realizes these counts exactly and observes each coordinate exactly half the time. For $n \in \{2400, 9600, 38400\}$ and 30 paired repetitions, all three estimators share the same simulated records and masks. Sample sizes are nested within a repetition. The moment estimate is projected in the diagonal $g_B$ norm. Marginal SM starts at zero and at the moment estimate, with training loss selecting the fit. Observed MLE minimizes its convex marginal negative log-likelihood. Both use the same coefficient region and projected-gradient tolerance $2 \times 10^{-7}$. Unobserved coefficients use the common zero tie-break. The main grid contains 2160 estimates: 1440 optimized fits and 720 direct moment estimates. All pass the specified numerical checks.

After the main grid, a separate precision check fixes the product truth and $q = .5$, with $n \in \{38400, 153600\}$ and 100 fresh repetitions. It adds 600 estimates and leaves the main grid unchanged. This design has $\mathcal{R}_{\text{eff}} = 60.8$ and $\mathcal{R}_{\text{SM}} = 71.7714$. The corresponding observed mean $nD_F$ values are $62.34, 57.45$ for MLE and $73.42, 67.02$ for SM. No estimate in this check reaches the coefficient boundary. The two stages use separate random streams. Exact seeds are listed in the supplement.

Fisher risks use tensor Gauss–Legendre integration in $u$. Comparing orders 16 and 24 on the first and last repetition of every cell changes risk by at most $6.46 \times 10^{-10}$ relatively. Independent finite differences, marginal information integrals, exact rational efficiency identities and direct observed-KL checks audit the formulas. An additional exact diagnostic uses $\mathbb{E}_\alpha \varphi_B^2 = v_B$: the raw moment estimator's expected Dirichlet risk is

$$\sum_{B:N_B>0} g_B \left\{ \frac{1}{N_B v_B} - \frac{\alpha_B^2}{N_B} \right\} + \sum_{B:N_B=0} g_B \alpha_B^2. \tag{122}$$

The supplement preserves all outcomes, boundary indicators, counts and convergence records. Reported intervals use 20000 paired bootstrap resamples of entire simulation repetitions and are unadjusted Monte Carlo intervals.

Figure 4 separates coverage gains from estimator efficiency. At $n = 38400$, increasing triple-mask mass from 0 to .5 reduces mean SM joint error from .00974 to .00123, an 87.4% reduction. Moment

estimation and MLE recover the same improvement from visibility. Both SM/MLE intervals contain the predicted ratio 1.180. These test estimator efficiency separately from panel (a)'s coverage effect.

# D LEARNING A MIXTURE WITHOUT A PRESCRIBED INTERACTION BASIS

## D.1 MODEL, ACQUISITION AND IDENTIFIABILITY

**Population and acquired records.** There are eight groups of four coordinates. Index the eight mixture components by $z \in \{-1, 1\}^3$ and set

$$p_t(x) = \sum_z \frac{1 + t z_1 z_2 z_3}{8} \varphi_{32}(x; m_z, S), \qquad t \in \{0, 3/4\}. \tag{123}$$

For $j = 1, 2, 3$, group $j$ of $m_z$ is $1.4 z_j v$, where $v = (1, .8, -.6, .4)/\|(1, .8, -.6, .4)\|_2$; other group means are zero. The shared covariance is $S_{ij} = 0.65^2 \cdot 0.35^{|i-j|}$. The fitted family has eight components, arbitrary means, positive weights, and an arbitrary shared positive-definite covariance: 791 free parameters. The learner knows the component count, shared-covariance assumption and acquisition groups. It estimates the means, weights and covariance without knowing the signal groups, their directions or the parity contrast. The family provides exact joint densities and marginal scores without restricting the fit to the generating interaction feature.

Uniform acquisition uses each of the $\binom{8}{4} = 70$ four-group masks equally often. Write $T = \{1, 2, 3\}$. For the triple-hole design, a mask containing $s$ groups in $T$ occurs with multiplicity $4, 1, 3, 0$ for $s = 0, 1, 2, 3$, respectively, in a 140-row library. Both designs observe 16 scalar coordinates per record, each group in exactly half the records, and each distinct group pair in $3/14$ of the records. Consequently all scalar and scalar-pair counts also match. Uniform acquisition covers $T$ with probability $1/14$; the triple-hole design never covers it. The fully observed diagnostic uses the same number of records and twice the measurement budget. Mask libraries are independently permuted across rows. For intermediate coverage we use $\pi_\eta = (1 - \eta)\pi_{\mathrm{h}} + \eta\pi_{\mathrm{u}}$ at $\eta = 1/4, 1/2$. Its multiplicities by $s = 0, 1, 2, 3$ are $(14, 5, 11, 2)$ in a 560-row library and $(6, 3, 5, 2)$ in a 280-row library, respectively. The scalar and pair quotas remain exactly unchanged, while the signal triple has coverage $\eta/14$ and any particular four-group mask containing it has probability $\eta/70$.

**Identification and ambiguity.** If a mask omits any group in $T$, summing over that latent sign cancels $t$ from its marginal density. This remains exact with the dense $S$. The two truths therefore induce identical entire acquisition experiments under the triple-hole design, for every sample size. Every positive $\eta$ identifies the true joint law within the fitted family. Each such design includes every uniform mask with positive probability. To see this, for an acquired coordinate set $A$ and any $a \in \mathbb{R}^{|A|}$, the marginal moment generating function satisfies

$$\log M_A(ra)/r^2 \;\longrightarrow\; a^\top S_A a/2 \quad (r \to \infty),$$

since the finite log-sum-exp of component means grows only linearly. Thus each acquired marginal identifies $S_A$, and pair coverage identifies $S$. Dividing each marginal characteristic function by its nonvanishing Gaussian factor identifies the projected atomic measure of means and weights. The eight true means projected onto $T$ are distinct. The observed masks $T \cup \{J\}$ align these eight atoms across every remaining group $J$, even for equal weights, and hence determine all means and weights. This establishes uniqueness at the specified truth; optimization is assessed separately below.

For an estimator-independent lower bound, fix $t = 3/4$ and write $p_t = p_0(1 + th)$, where $h$ is the posterior mean of $z_1 z_2 z_3$ under $p_0$. Under an equal prior on $p_0, p_t$, the triple-hole data leave the prior unchanged. The best estimated vector field is the score of $p_{t/2}$, with Bayes risk

$$\frac{1}{2} D_{\mathrm{F}}(P_0, P_{t/2}) + \frac{1}{2} D_{\mathrm{F}}(P_t, P_{t/2}) = \frac{t^2}{2} \mathbb{E}_{P_0} \frac{\|\nabla h\|^2}{(2 + th)(1 + th)} > 0. \tag{124}$$

This lower bound applies to the worst expected risk over the two truths, rather than to every fit at either truth. Direct evaluation at $t = 3/4$ gives 0.150 (Monte Carlo standard error 0.00148, 131,072 draws).

**Mean perturbations requiring four groups.** The generating weight contrast involves three groups, but arbitrary component means permit four-group perturbations. Fix $t$, put $e_z = z_1 z_2 z_3$, and perturb coordinate $j$ in a nonsignal group $J$ by $m_z(c) = m_z + c(e_z - t)\mathbf{e}_j$, with $\mathbf{e}_j$ the coordinate unit vector. Write $P_{t,c}$ for this perturbed mixture. Because $w_z(e_z - t) = (1 - t^2)e_z/8$, the marginal density derivative at $c = 0$ vanishes unless the mask contains $T \cup \{J\}$. Under uniform acquisition this is exactly one of 70 masks. For coordinate $j = 13$ (one-based), the tridiagonal AR(1) precision implies that the parameter tangent $u = \partial_c \log p_{t,c}|_{c=0}$ depends only on the first 14 coordinates. It is therefore measurable in the first four groups and its gradient is supported there. Along this path,

$$\lim_{c \to 0} \frac{D_{\mathrm{F}}(P_t, P_{t,c})}{D_{\mathrm{M}}^{\pi_\eta}(P_t, P_{t,c})} = 70/\eta, \qquad \eta > 0. \tag{125}$$

Indeed the joint gradient energy and that of the sole visible marginal agree, while marginal averaging multiplies the latter by $\eta/70$. Differentiation of this finite Gaussian mixture is justified by Gaussian tail domination. This local calculation applies to the fitted mixture family. Its dependent reference differs from that of the global ANOVA bound.

**The full space of hidden directions.** The acquisition kernel can be computed for the entire fitted family. Let $I_{\mathrm{h}}, I_{\mathrm{u}}$ denote its observed parameter Fisher information under hole and uniform masks at either truth in (123). Let $\mathsf{G}$ measure the local joint-score change: for a parameter increment $b$, $D_{\mathrm{F}}(P_\theta, P_{\theta+b}) = b^\top \mathsf{G} b + o(\|b\|^2)$.

**Proposition 15** (The fitted mixture's rare-coverage cost)**.** *At each specified truth,* $\dim\ker I_{\mathrm{h}} = 45$*, while* $I_{\mathrm{u}}$ *and* $\mathsf{G}$ *are positive definite on the 791-dimensional local parameter space. The hole marginal spatial-score operator has the same 45-dimensional kernel. For* $\pi_\eta = (1-\eta)\pi_{\mathrm{h}} + \eta\pi_{\mathrm{u}}$*, let* $I_\eta = (1-\eta)I_{\mathrm{h}} + \eta I_{\mathrm{u}}$ *and* $\mathcal{R}_\eta = \mathrm{tr}(\mathsf{G} I_\eta^{-1})$*. Then*

$$\mathcal{R}_\eta = \frac{K_t}{\eta} + O(1), \qquad K_t = \mathrm{tr}\{\mathsf{G}_{VV}(I_{\mathrm{u}})^{-1}_{VV}\} > 0, \quad V = \ker I_{\mathrm{h}}, \quad \eta \downarrow 0. \tag{126}$$

*Restrictions are expressed in any fixed basis of* $V$ *and a complement.*

*Proof.* Write a parameter tangent as weight increments $a_z$, mean increments $v_z$, and a shared-covariance increment $H$, where $\sum_z a_z = 0$. Put $g_{zj} = w_z v_{zj}$. For an acquired coordinate set $A$, let $T_A \subseteq T$ be its acquired signal groups. After removing the common Gaussian factor, the derivative of its moment generating function is a sum of distinct exponentials indexed by $z_{T_A}$, with polynomial coefficients

$$\sum_{z:\, z_{T_A} \text{ fixed}} \left( a_z + y^\top g_{z,A} + \frac{w_z}{2} y^\top H_A y \right).$$

Distinct exponentials with polynomial coefficients are independent: restrict to a generic ray, separate its distinct exponential rates, then vary the ray to identify the coefficients. A zero acquired derivative thus forces $H_A = 0$, since each aggregated weight is positive. Pair coverage gives $H = 0$.

Expand $a_z$ and each $g_{zj}$ in Walsh characters $\chi_B(z) = \prod_{i \in B} z_i$, $B \subseteq T$. A projected character survives exactly when $B \subseteq T_A$. For weights, hole masks see all proper $B$; normalization excludes the constant, leaving one hidden direction $B = T$. If $j$ belongs to signal group $g$, an acquired mask must contain $g$. Its only invisible characters are $T \setminus \{g\}$ and $T$, giving $12 \times 2$ hidden mean directions. For each of the other 20 coordinates, only $T$ is invisible. The nullity is therefore $1 + 24 + 20 = 45$. Uniform masks reveal every character, so their information is positive definite. A zero marginal spatial-score derivative makes its parameter score constant, and centering makes that constant zero; thus the parameter and spatial-score kernels agree. A tangent with zero joint-score energy has a constant parameter score, which vanishes by centering. Local identification of the full mixture then makes the tangent zero. Thus $\mathsf{G}$ is positive definite too.

In coordinates $V \oplus W$, write

$$I_\eta = \begin{pmatrix} \eta A & \eta B \\ \eta B^\top & C + \eta(D - C) \end{pmatrix}, \qquad A \succ 0,\ C \succ 0.$$

Block inversion gives $\eta I_\eta^{-1} \to \mathrm{diag}(A^{-1}, 0)$, with an $O(1)$ remainder after subtracting the leading inverse. Taking the trace against $\mathsf{G}$ proves (126). □

Here $\mathcal{R}_\eta$ is the efficient joint-score coefficient in the local Gaussian estimation experiment. The finite-mixture derivatives have locally polynomial envelopes under Gaussian tails, so the information and target metric are finite and continuous. The reference $\mathcal{R}_\eta/n$ first takes the local large-sample limit at fixed $\eta > 0$ and then decreases $\eta$. It approximates local estimation risk, without asserting equality for a finite-sample fit. The supplement independently checks the full 791-column acquired Jacobian (ranks 746 and 791), all types of hidden directions, and the 45-dimensional restricted metrics.

### D.2 Fitting and numerical evaluation

**Fitting, selection, and evaluation.** Training sizes are 22,400 and 89,600, with three independent training repetitions at each truth and design. Each of the 60 cells uses six initializations computed only from acquired entries. The acquired-data EM update includes conditional second moments of missing coordinates (Ghahramani & Jordan, 1993), pooled over components for the shared covariance. It runs for at most 180 iterations, followed by at most 400 full-batch L-BFGS likelihood iterations. We select the start with highest acquired validation likelihood on 2,800 independent records of the same design. Records for all 360 starts are retained. No complete validation or test quantity is used for selection.

Each selected likelihood estimate initializes a further 400-iteration optimization of the exact acquired-marginal Hyvärinen score. If $r_z$ are posterior responsibilities, $K = S_A^{-1}$ and $v_z = Km_{z,A}$, its analytic score and divergence are

$$s_A = \mathbb{E}_r v_z - Kx_A, \qquad \nabla \cdot s_A = -\operatorname{tr} K + \mathbb{E}_r\|v_z\|^2 - \|\mathbb{E}_r v_z\|^2.$$

We minimize the empirical mean of $\|s_A\|^2 + 2\nabla \cdot s_A$, divided by the constant acquired dimension. We retain the final iterate without selecting intermediate checkpoints. The two objectives use the same family but different initialization procedures and optimization budgets. Both optimizers use FP64, strong-Wolfe searches, gradient tolerance $10^{-7}$, objective-change tolerance $10^{-11}$, covariance floor $.02^2$, and weight floor $10^{-5}$. Iteration-cap hits are recorded.

Joint Fisher risk and joint log-density excess use the exact true scores on 16,384 independent complete test rows, shared across designs and objectives within each truth, sample size, and training repetition. Acquired marginal Fisher risk uses a separate 28,000-row sample with exact mask quotas. Training values and validation values are also shared across designs, but the fitter receives NaNs in every unacquired entry. The original 36-cell protocol and the subsequent 24-cell positive-coverage extension were separately fixed after development runs; development is not counted as additional independent confirmation. The extension retains the original sample sizes, seeds, fitting budgets, and selection rule. Figure intervals show the three training-repetition ranges. The supplement additionally reports paired test-sample intervals conditional on each fitted pair.

**Results and numerical checks.** The 72 adjacent-design comparisons in Section 7 span two sample sizes, two truths, three repetitions and two fitting objectives. Uniform acquisition has lower joint risk than hole acquisition in all 12 likelihood and all 12 score-matching comparisons. The likelihood iteration cap is reached by 17 of 36 original selected starts and 15 of 24 extension selections; the score cap is reached by 19 of 36 original and 15 of 24 extension endpoints. All remain included.

Table 2 reports both risks for likelihood, the final score-matching iterate, and validation-selected score matching. These fits measure finite-sample performance. The logistic coefficient isolates local estimator efficiency within its own model.

**Acquired validation for score matching.** A selection control uses the same 60 likelihood initializations, training records, and 2,800 acquired validation records. We fixed a 400-iteration budget and a ten-iteration checkpoint interval before running this control. The candidate set contains the initialization, every tenth accepted L-BFGS iterate, and the final iterate. Minimum acquired validation Hyvärinen score selects one candidate, with earlier iterates breaking exact ties. Checkpoint evaluation leaves the optimizer state unchanged, and all final parameters reproduce the original fits exactly. Neither complete records nor population risks enter selection. This control was specified after the original results had been examined.

Validation selects the likelihood initialization in 48 of 60 cases. For the other twelve, selected joint risk is 0.997–2.051 times initial risk. At $n = 89,600$, $t = 3/4$, the selected mean risks

Table 2: Mixture learning: estimated population joint and acquired-marginal Fisher risks, averaged over three training repetitions. ML denotes acquired likelihood. SM final uses the optimizer endpoint; SM val selects among fixed checkpoints by acquired validation score, including the ML initialization. All use the same model and acquired records. Full observation uses twice the measurement budget.

| | | | Joint risk | | | Marginal risk | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | $t$ | Acquisition | ML | SM final | SM val | ML | SM final | SM val |
| 22,400 | 0 | Hole ($\eta = 0$) | 26.7 | 31.1 | 26.7 | 0.216 | 0.265 | 0.216 |
| | | $\eta = 1/4$ | 5.50 | 21.9 | 5.50 | 0.214 | 0.328 | 0.214 |
| | | $\eta = 1/2$ | 4.28 | 15.4 | 4.28 | 0.209 | 0.296 | 0.209 |
| | | Uniform ($\eta = 1$) | 1.42 | 15.1 | 1.42 | 0.187 | 0.298 | 0.187 |
| | | Full ($2\times$) | 0.185 | 0.211 | 0.185 | 0.185 | 0.211 | 0.185 |
| 22,400 | 0.75 | Hole ($\eta = 0$) | 33.1 | 39.3 | 33.1 | 0.209 | 0.248 | 0.209 |
| | | $\eta = 1/4$ | 3.41 | 5.30 | 3.48 | 0.250 | 0.310 | 0.255 |
| | | $\eta = 1/2$ | 2.58 | 4.00 | 2.58 | 0.224 | 0.290 | 0.224 |
| | | Uniform ($\eta = 1$) | 1.93 | 3.04 | 1.93 | 0.214 | 0.291 | 0.214 |
| | | Full ($2\times$) | 0.197 | 0.220 | 0.197 | 0.197 | 0.220 | 0.197 |
| 89,600 | 0 | Hole ($\eta = 0$) | 26.5 | 29.8 | 26.7 | 0.0579 | 0.0658 | 0.0628 |
| | | $\eta = 1/4$ | 0.899 | 4.96 | 0.899 | 0.0477 | 0.0687 | 0.0477 |
| | | $\eta = 1/2$ | 0.461 | 1.90 | 0.461 | 0.0464 | 0.0604 | 0.0464 |
| | | Uniform ($\eta = 1$) | 0.349 | 0.536 | 0.349 | 0.0480 | 0.0553 | 0.0480 |
| | | Full ($2\times$) | 0.0472 | 0.0516 | 0.0485 | 0.0472 | 0.0517 | 0.0485 |
| 89,600 | 0.75 | Hole ($\eta = 0$) | 30.5 | 37.2 | 30.9 | 0.0582 | 0.0685 | 0.0644 |
| | | $\eta = 1/4$ | 0.875 | 1.55 | 1.16 | 0.0536 | 0.0699 | 0.0631 |
| | | $\eta = 1/2$ | 0.600 | 1.44 | 0.831 | 0.0504 | 0.0734 | 0.0599 |
| | | Uniform ($\eta = 1$) | 0.363 | 0.756 | 0.363 | 0.0462 | 0.0603 | 0.0462 |
| | | Full ($2\times$) | 0.0446 | 0.0494 | 0.0446 | 0.0447 | 0.0496 | 0.0447 |

are $1.159, 0.831, 0.363$ for $\eta = 1/4, 1/2, 1$, compared with $1.551, 1.438, 0.757$ at the final iterate. Thus selection substantially reduces the fitting gap while preserving the acquisition comparison. The selected fits retain lower mean joint risk with greater positive coverage in all eight adjacent comparisons across the two sample sizes and truths. Initialization identifiers, checkpoint losses, selected iterations, and both population risks are provided for every condition.

The numerical audit regenerated the original test samples and reconstructed every selected risk array from fitted parameters. It verified exact mask counts, missing-value boundaries, EM conditional second moments, marginal scores, spatial divergences, parameter derivatives, objective values, and selection. An independent population check compares the change in Hyvärinen score with the change in marginal Fisher risk on the separate 28,000-row sample; the largest absolute discrepancy is 3.64 stratified Monte Carlo standard errors across the 60 comparisons. The supplement provides objective and convergence summaries, the check results, and code to rerun the fits.

**Full-model information reference.** We integrate the analytic parameter scores and their spatial derivatives over the full 791-dimensional parameter family and all 70 masks. At $\eta = (1/4, 1/2, 1)$, the resulting $\mathcal{R}_\eta$ values are $(49{,}741, 35{,}116, 28{,}036)$ for $t = 0$ and $(57{,}462, 41{,}157, 32{,}521)$ for $t = 3/4$. The displayed reference divides these coefficients by $n$; it has no fitted scale factor. With 16,384 population integration draws per truth, halving the integration sample changes these coefficients by at most $0.79\%$ (a numerical resolution diagnostic). Mean likelihood risks are $1.00$–$1.62$ times this reference at the larger sample size and $1.13$–$2.73$ times it at the smaller size.

Records missing the triple still carry information about other parameters. In distribution, the $(n, \eta) = (89{,}600, 1/4)$ acquisition experiment factors into the $(22{,}400, 1)$ experiment plus 67,200 independent hole records: the exact per-mask quotas add accordingly. These extra records inform nuisance directions, although they are ancillary in the submodel with only $t$ unknown. Indeed, at fixed $m = n\eta$ the total information is $nI_\eta = mI_{\mathrm{u}} + (n - m)I_{\mathrm{h}}$. Its joint-score local risk decreases as hole records are added. For the full fitted model, the corresponding local-risk ratios are $0.444$ and $0.442$ for the

two truths; the observed likelihood mean-risk ratios are $0.635$ and $0.453$. The factorization holds in distribution. The saved training arrays are not nested subsets.

**Sensitivity to the optimization budget.** We refine every selected endpoint that hits its original iteration cap for up to 1,600 additional L-BFGS iterations on its own acquired objective: 36 original endpoints and 30 extension endpoints. Eligibility uses only the recorded cap, and all endpoints are retained. The other fits keep their original endpoint. All 24 uniform-versus-hole comparisons preserve their direction; 67 of 72 adjacent-coverage comparisons favor higher coverage after refinement. Individual uniform-design risks change by at most $5.1\%$. The largest relative change across the extended grid is $45.5\%$, in a small-sample $\eta = 1/4$ likelihood fit whose joint risk increases as acquired likelihood improves. Four refined endpoints reach the additional cap, all from the original hole design; none of the new refinements does. The figure and table retain the original optimization budgets. The supplement gives risks and convergence summaries for both versions.

### D.3 Pilot-guided acquisition at matched budgets

**Population and pilot observations.** We retain the full fitted family in (123), combining signal-mean multipliers $a \in \{1, 1/2\}$ with nonsignal-mean perturbation RMS $\epsilon \in \{0, .05\}$, at both $t$ values. For each $t$, an independent $8 \times 20$ Gaussian perturbation is centered using the component weights, normalized to unit entrywise RMS, and then scaled by $\epsilon$. This direction is fixed across training repetitions and signal strengths. All eight truths remain in the fitted family.

For this acquisition experiment, the pilot additionally knows that there are three signal groups, but not their identities or generating parameters. It observes $n_0 \in \{560, 4480\}$ uniform records, with $n_0/2$ observations per coordinate. From each coordinate's unbiased sample fourth cumulant $\widehat{\kappa}_{4,i}$, it forms $s_g = \max_{i \in g}(-\widehat{\kappa}_{4,i})$ for each group $g$. The three groups with largest $s_g$ form $\widehat{T}$. This uses univariate acquired values only. A signal coordinate with mean $\pm b$ has fourth cumulant $-2b^4$ under the shared Gaussian noise, which motivates the rule in this construction.

**Allocation and exact budgets.** After the pilot, each four-group mask $A$ receives probability

$$q_{\widehat{T}}(A) = \frac{c_{|A \cap \widehat{T}|}}{140}, \qquad (c_0, c_1, c_2, c_3) = (0, 3, 1, 4). \tag{127}$$

These probabilities retain group frequency $1/2$ and distinct-pair frequency $3/14$, while doubling coverage of $\widehat{T}$ from $1/14$ to $1/7$. The latter is the maximum under those quotas: with $H = |A \cap \widehat{T}|$, the fixed first two factorial moments imply $\mathbb{P}(H = 0) + \mathbb{P}(H = 3) = 1 - \mathbb{E}H + \mathbb{E}\binom{H}{2} = 1/7$. The comparison arms are uniform throughout, each pilot followed by $q_{\widehat{T}}$, and the 560-record pilot followed by $\frac{1}{4}\pi_{\mathrm{u}} + \frac{3}{4}q_{\widehat{T}}$. All use 89,600 training records, including the pilot, and 2,800 independent uniform records for validation: 92,400 records and 1,478,400 measured scalars. All per-mask counts are integers and all lower-order quotas match exactly, including validation. Pilots are nested and consist of complete 70-mask cycles. Later masks depend on the pilot but are assigned independently of their own records' values, so the acquired likelihood remains valid conditionally on the acquisition history.

For a true triple $B$ with $j = |B \cap \widehat{T}|$, its coverage under $q_{\widehat{T}}$ is $(9, 11, 8, 20)_j/140$, indexed by $j = 0, 1, 2, 3$. Thus a wrong selection still observes every triple with probability at least $2/35$. The five four-group masks disjoint from $\widehat{T}$ receive no later observations, but remain present in the reused uniform pilot. The quarter-uniform arm also observes these masks after the pilot.

**Independent learning comparison.** After development, we fixed twelve new training repetitions for every condition and arm: 384 final models and 2,304 initializations using only acquired entries. Each uses up to 180 batched FP64 EM iterations and 400 likelihood iterations, with the same family, constraints and tolerances as above. Acquired validation likelihood selects one of six starts; a selected start reaching the 400-step cap receives up to 1,600 further likelihood iterations. No complete observation is passed to the fitter or selector. Joint Fisher risk and KL are evaluated on 65,536 fresh complete draws. Generating values and evaluation draws are shared across arms within each repetition; the eight conditions are coupled and are not pooled as independent repetitions.

Table 3: Pilot-guided acquisition at identical total, singleton and pair budgets. Entries are mean paired joint-Fisher risk reductions (percent) relative to uniform, with pointwise 95% intervals across twelve training repetitions. $a$ scales the signal means and $\epsilon$ is the RMS of added nonsignal means. Every condition and all three candidate arms are shown; negative entries indicate larger risk.

| $a$ | $\epsilon$ | $t$ | 560 pilot | 4,480 pilot | 560 pilot + 1/4 uniform |
|---|---|---|---|---|---|
| | | | Mean [95% CI] | Mean [95% CI] | Mean [95% CI] |
| 1 | 0 | 0 | 10.63 [ 4.49, 16.77] | 12.43 [ 5.02, 19.83] | 9.41 [ 2.19, 16.63] |
| 1 | 0 | 0.75 | 20.90 [ 12.61, 29.20] | 22.73 [ 16.45, 29.01] | 15.94 [ 5.85, 26.02] |
| 1 | 0.05 | 0 | 3.42 [ −2.71, 9.55] | 7.25 [ 2.11, 12.40] | 1.90 [ −4.14, 7.94] |
| 1 | 0.05 | 0.75 | 22.66 [ 8.77, 36.55] | 22.80 [ 9.76, 35.84] | 16.93 [ 0.54, 33.32] |
| 0.5 | 0 | 0 | −33.75 [−64.72, −2.77] | 14.97 [−9.16, 39.11] | −24.53 [−56.40, 7.34] |
| 0.5 | 0 | 0.75 | 6.33 [ 0.98, 11.67] | 8.51 [ 3.26, 13.76] | 4.39 [ −1.24, 10.01] |
| 0.5 | 0.05 | 0 | −6.78 [−13.19, −0.37] | 4.33 [−2.43, 11.09] | −6.09 [−13.50, 1.31] |
| 0.5 | 0.05 | 0.75 | 6.88 [ −1.35, 15.12] | 12.31 [ 7.53, 17.09] | 3.49 [ −4.21, 11.19] |

**Results and numerical sensitivity.** Table 3 reports mean per-repeat percentage risk reductions, with pointwise Student-$t$ intervals across the twelve training repetitions. Both pilots select correctly in all strong-signal repetitions; in each weak-signal condition, the smaller pilot succeeds in 4 or 7 of 12, and the larger in 11 or 12. The larger pilot improves mean risk in all eight conditions, with intervals crossing zero in two; its mean KL changes also favor acquisition. The table also includes all risk increases for the smaller pilot and the quarter-uniform ablation. In particular, one wrong larger-pilot selection nearly doubles risk in the weak, unperturbed $t = 0$ condition, despite improvement in the other eleven repetitions.

The supplement provides selection records for every start, risk summaries, and all five declared contrasts in each condition. Source and quota checks, pilot selection, acquired validation and joint risks are independently recomputed. Of the selected starts, 114 receive the prescribed additional optimization; all four remaining capped endpoints are separately extended by up to 6,400 iterations. None then reaches the cap, and their risks change by at most $0.079\%$. The table retains the original endpoints; the supplement provides risk summaries for both versions. These extensions assess numerical sensitivity and do not establish global optimality of the mixture fits.

# E Controlled learning and ECG protocols

## E.1 Protocol and scope

Table 4 collects the information used by the mixture and ECG studies. The density-ratio and Gaussian diagnostics below specify their own known-model fitting rules and uncertainty measures.

## E.2 Non-Gaussian interaction learning

We fit the four-dimensional density-ratio model $p_\theta = \varphi_4\{1 + \sum_{1\le|B|\le3} \theta_B \prod_{i\in B} \tanh x_i\}$, with $\|\theta\|_1 \le 0.85$. The order-three learner estimates all 14 coefficients; the order-two comparator estimates ten. Masks are uniform within sizes one, two, and three, with total probabilities $q, 1-2q, q$. Every design observes two coordinates on average and every mask is proper. Pair coverage is $(1+q)/6$, and triple coverage is $q/4$.

Figure 5 compares the two model orders and the joint and marginal targets. At $n = 16000$, the order-three model has median marginal error $0.00123$ with pair-only masks, but joint error $0.0489$. Adding triple observations reduces its median joint error to $0.0167$ at $q = 0.1$ and $0.00832$ at $q = 0.4$. The order-two model remains near $0.048$ under all designs. All 540 fits are included.

The reference is $N(0, I_4)$. In one-based coordinate notation the nonzero truth coefficients are $\theta_{\{1\}} = 0.04$, $\theta_{\{2,4\}} = 0.08$, $\theta_{\{1,2,3\}} = 0.42$, and $\theta_{\{1,3,4\}} = -0.12$. Their absolute sum is $0.66$, so the truth is interior to the candidate $\ell_1$ ball of radius $0.85$. We use exact rejection sampling from

Table 4: Information used by the learning and ECG studies. Mixture confirmation budgets include acquired training, pilot and selection records. ECG budgets match formal training acquisitions, with additional development and validation information as listed. Test outcomes do not select the reported fits; later mechanism analyses are explicitly exploratory.

| Stage | Mixture studies | ECG studies |
|---|---|---|
| Formal fitting | 16 of 32 scalars per record; exact singleton/pair quotas. The fully observed diagnostic uses twice this budget. | 12,880 PTB-XL patients, four of eight leads each; unacquired values excluded from fitting. |
| Pilot | Acquisition study: 560 or 4,480 uniform records included in the 89,600 training total; known signal-group count, acquired univariate cumulants. | No acquisition pilot. |
| Development | Separate development runs fix the model and protocol before confirmation; outside the reported confirmation budget. | Development fits on folds 1–2; fold-8 acquired powers set common scales, complete reconstruction targets inform model choice. Outside the formal training budget. |
| Start/checkpoint selection | 2,800 additional acquired validation records: same design for the coverage study, uniform for the pilot study. Six starts selected by likelihood; SM control checkpoints selected by acquired score. No complete validation. | Main CNN uses complete fold-9 validation MAE; control CNN and CSDI use fixed epochs 40 and 200. Validation information is additional to training acquisition. |
| Sampling calibration | Exact densities/scores for evaluation; no predictive ensemble-size choice. | 64 fold-9 patients with evaluation targets calibrate 32 CSDI draws before test-score inspection; additional validation use. |
| Final evaluation | Complete simulated test draws; the coverage study also uses a separate acquired-risk sample. Truths are not used for fit selection. | Complete fold 10 and fixed-model external cohorts; patient and record bootstrap units, respectively. |
| Later analyses | Positive-coverage and pilot confirmations follow separately fixed protocols; SM checkpoint control follows the original fits. | Completion/tail and CNN selection controls reuse examined PTB-XL folds; external quadratic-score analysis follows external prediction outcomes. |

the Gaussian reference, with envelope $1 + \|\theta_0\|_1$. Within each of 30 repetitions, designs share the same complete sample and sample sizes $1000, 4000, 16000$ are nested. Masks are independent of the sample values, with $q \in \{0, 0.1, 0.4\}$.

For a mask, let $u$ be the sum of the retained interaction terms. Up to a parameter-independent term, its exact Hyvärinen loss is

$$2\frac{\Delta u - x_A^\top \nabla u}{1+u} - \frac{\|\nabla u\|^2}{(1+u)^2}. \tag{128}$$

Both model orders use the same projected-gradient method with Armijo backtracking, initialized at zero. Projection onto the $\ell_1$ ball ensures positive densities everywhere. A fit passes when $\|\theta - \mathrm{proj}(\theta - \nabla\widehat{J})\|_2 \le 2 \cdot 10^{-7}$; all final iterates are retained. This is a stationarity criterion for the nonlinear score objective. Invisible coefficients are fixed at zero to select among equivalent fits.

Population risks are integrated directly under the known non-Gaussian truth using product Gauss–Hermite quadrature of order 24 per coordinate. The code checks the analytic objective gradient by central differences and re-evaluates a deterministic selection of fitted candidates at order 32. The maximum objective-gradient discrepancy is $7.4 \cdot 10^{-12}$. Increasing quadrature order from 24 to 32 changes the audited joint risks by at most $0.034\%$. All 540 fits pass the stationarity threshold.

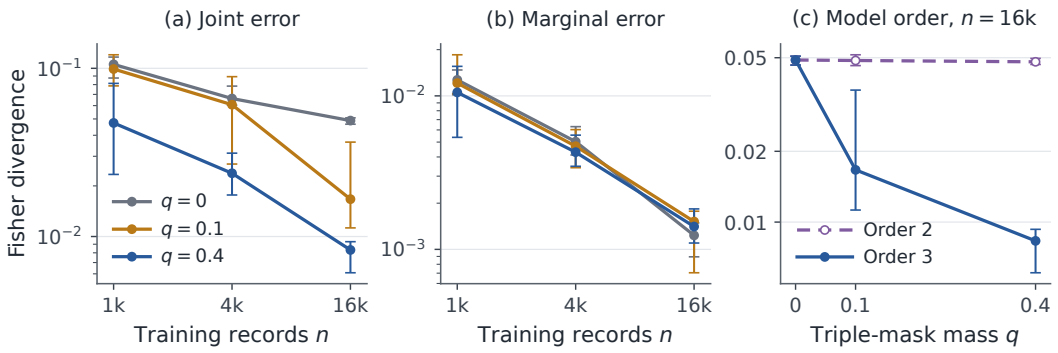


Figure 5: Learning non-Gaussian interactions with two observed coordinates per record on average. Medians and interquartile ranges over 30 paired repetitions; risks use the known non-Gaussian population. Left: joint error for the order-three model. Center: its observable marginal error. Right: order-two and order-three joint errors at the largest sample size. All vertical axes are logarithmic; horizontal axes are logarithmic in (a,b) and linear in (c). The masks are always proper.

### E.3 Conditioning and coverage: convergence audit

For Gaussian estimation diagnostics we use a zero-mean model with positive-definite precision $\Omega(\theta)$. Its marginal precision and exact marginal Hyvärinen loss are

$$K_A(\theta) = \{(\Omega(\theta)^{-1})_{AA}\}^{-1}, \qquad \ell_\theta(A, x_A) = x_A^\top K_A(\theta)^2 x_A - 2\,\mathrm{tr}\, K_A(\theta). \tag{129}$$

This specializes Givens et al. (2025, Definition 4.1 and Proposition 4.3). Each marginal precision is recomputed; substituting the full precision submatrix $\Omega_{AA}$ generally changes the objective.

Marginal score matching and observed-data MLE use the same pairwise initialization projected onto the positive semidefinite cone, parameter region, sample and convergence test. Figure 6 reports medians and interquartile ranges of repeated-sample risks. For the estimator comparison, we first compute the risk ratio within each jointly successful repetition, then take the median. Its bootstrap interval resamples the paired log ratios.

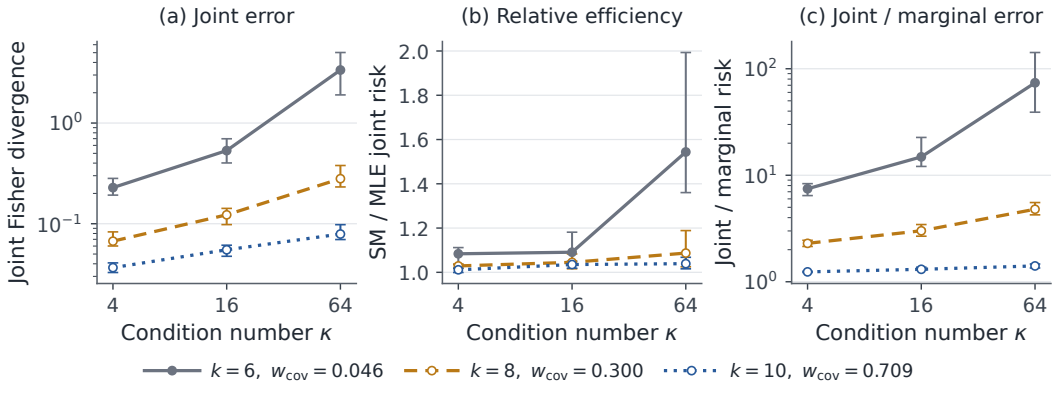


Figure 6: Conditioning and structured-mask coverage for Gaussian estimation. Panels (a) and (c) show medians and interquartile ranges; panel (b) gives 95% paired-bootstrap intervals. Paired comparisons use jointly converged fits. Horizontal axes and vertical axes in (a,c) are logarithmic; the vertical axis in (b) is linear. Including the sole failed fit changes the hardest-cell median joint risk from 3.363 to 3.414 and the paired SM/MLE ratio from 1.544 to 1.616; its joint risk is 68.41.

The shared setup uses a fixed $d = 11$ covariance orientation, $n = 4800$, $\kappa \in \{4, 16, 64\}$, and nonuniform cyclic panels with minimum pair coverage $\{0.046, 0.300, 0.709\}$.

At the highest coverage, the median marginal-score/MLE joint-risk ratios are $1.011$, $1.035$, and $1.039$ as $\kappa$ increases, while the joint-to-marginal ratio stays between $1.24$ and $1.41$. At the lowest coverage, these ranges become $1.084$–$1.544$ and $7.47$–$73.80$.

Figure 6 contains 40 repetitions for each of nine configurations. All 360 likelihood fits and 359 of 360 marginal-score fits passed the prespecified convergence test. The sole failure occurred at the hardest setting $(\kappa, k) = (64, 6)$; the caption reports its effect on the all-fit summaries. The primary paired ratio uses only jointly successful repetitions, so the MLE fit paired with the failed marginal-score run is also excluded from panel (b).

### E.4 Full-waveform ECG acquisition and reconstruction

**Data and patient partitions.** We use PTB-XL version 1.0.3, distributed through PhysioNet (Wagner et al., 2020; 2022; Pollard et al., 2026). Each 100 Hz recording contains ten seconds. We retain I, II, and V1–V6, excluding the four deterministically derived limb leads. The retained leads may still be statistically dependent. Header gains and baselines convert the original integer signals to mV. No beat alignment, diagnostic filter, interpolation, or frequency truncation is applied. The official folds separate patients. Folds 1–7 contain 13,142 candidate training patients; a fixed random subset of 12,880 permits exact acquisition quotas. All their 14,945 records are retained. Fold 8 supplies 2,173 development records from 1,881 patients, fold 9 supplies 2,183 validation records from 1,942 patients, and fold 10 supplies 2,198 test records from 1,904 patients. Repeated recordings share the patient's acquisition mask. Losses weight each record inversely by that patient's record count; evaluation first averages within patients and then across patients.

**Exactly matched acquisition budgets.** Order the leads as $(\mathrm{I}, \mathrm{II}, \mathrm{V1}, \ldots, \mathrm{V6})$. Every training patient reveals four entire waveforms, and every lead is revealed for exactly 6,440 patients. Separated panels alternate the first four and last four leads. Cyclic windows rotate four consecutive leads on the eight-cycle; opposite pairs $(1, 5), (2, 6), (3, 7), (4, 8)$ are never observed together. Uniform acquisition balances all 70 four-subsets. A fourth, triple-hole design never contains $T = \{\mathrm{I}, \mathrm{II}, \mathrm{V1}\}$. It matches *every* pair's 2,760 observed patients under uniform acquisition, whereas $T$ occurs for 920 patients under uniform acquisition and zero under the triple-hole design. Saved integer quotas and masks specify the latter design completely. Each design has three patient assignments. Estimators share the acquired values, and CNN initializations are paired across designs within each repetition.

**Spectral joint models.** An orthonormal real Fourier transform preserves all 8,000 waveform coordinates. Eleven bands share $8 \times 8$ covariance matrices. The nonnegative frequency-index boundaries are $0, 1, 6, 16, 31, 51, 81, 121, 181, 251, 351, 501$; the DC and Nyquist coefficients count once and interior real/imaginary coefficients count twice. Let $c_b$ be the resulting real-coefficient count in band $b$, so that $\sum_b c_b = 1000$. One scalar scale per band, estimated from acquired univariate development powers, preconditions the coefficients. Its use of fold 8 is common to all formal spectral fits. Coordinates within a component are conditionally independent across real Fourier coefficients, with zero means and band covariances $\Sigma_{kb} \succeq .01 I_8$ in these preconditioned units. We use either one Gaussian component or a mixture of two components with one component label for the entire record. Taking principal covariance submatrices gives exact normalized acquired marginals. The two-component choice was made using development reconstruction results.

Both observed likelihood and marginal score matching fit each family and design, for 48 spectral fits. Likelihood uses two-start EM with a 1,000-step budget. Score matching starts from that solution and uses up to four 1,000-iteration refinement rounds. The supplement records stationarity residuals and all starts. Two score fits that reached the initial budget were continued on their unchanged training objectives; the largest MAE change over evaluated queries was $0.0064\,\mu$V. The original fits remain the primary estimates. Whitening, submatrix inverses, and the orthonormal transform are accounted for when reporting densities in mV coordinates. These working families approximate the ECG distribution.

**Nonlinear conditional comparator.** The CNN receives zero-filled context waveforms and an explicit context mask. A width-40 input convolution is followed by six residual blocks with dilations $1, 2, 4, 8, 16, 32$, group normalization and GELU activations, then an eight-channel output. During training, two acquired leads serve as context and the other two provide targets for the squared-error

Table 5: PTB-XL point prediction on the same 1,904 test patients and queries, in $\mu$V. MAE averages three fits; RMSE takes the square root after averaging MSE across patients and fits. CNN validation uses complete validation for selection; CNN 40 epochs uses fixed stopping. CSDI predicts with the mean of 32 samples. All designs acquire four leads per training patient. Lower is better.

| Model | Separated | Cyclic | Uniform | Triple hole |
|---|---|---|---|---|
| | | **MAE** | | |
| Gaussian NLL | 116.45 | 100.26 | 99.77 | 99.52 |
| Gaussian SM | 116.45 | 100.42 | 99.92 | 100.05 |
| Mixture NLL | 115.39 | 98.66 | 97.98 | 97.81 |
| Mixture SM | 115.43 | 98.68 | 98.07 | 97.96 |
| $t_{10}$ NLL | 115.09 | 99.00 | 98.35 | 98.32 |
| CNN (validation) | 116.27 | 91.08 | 85.03 | 84.96 |
| CNN (40 epochs) | 119.03 | 91.10 | 85.09 | 85.02 |
| CSDI (32 draws) | 102.96 | 94.16 | 93.27 | 94.97 |
| | | **RMSE** | | |
| Gaussian NLL | 231.06 | 200.58 | 199.69 | 199.41 |
| Gaussian SM | 231.06 | 200.75 | 199.86 | 201.06 |
| Mixture NLL | 230.31 | 198.55 | 197.49 | 197.14 |
| Mixture SM | 230.29 | 198.84 | 197.76 | 197.65 |
| $t_{10}$ NLL | 232.52 | 204.13 | 201.92 | 201.89 |
| CNN (validation) | 225.15 | 180.27 | 172.80 | 172.68 |
| CNN (40 epochs) | 226.47 | 180.08 | 172.86 | 172.77 |
| CSDI (32 draws) | 208.10 | 189.22 | 186.44 | 188.48 |

loss. Unacquired labels never enter training. AdamW uses learning rate $10^{-3}$, weight decay $10^{-4}$, batch size 64, gradient clipping at 5, and cosine decay over at most 40 epochs. Main CNN checkpoints minimize the complete fold-9 validation MAE, with patience 10 after epoch 20. This comparator therefore uses complete validation beyond its acquired training entries. A control uses the fixed 40th epoch for all four designs and three repetitions, without reading validation records during training. Training histories agree exactly with the corresponding main runs over their common epochs. The control retains prior development choices and was completed after the initial test results had been examined. The CNN predicts conditionals without specifying a joint density.

**Evaluation and uncertainty.** The primary query supplies two context leads and reconstructs the other six. Every test patient receives one fixed query, shared by all methods; the 28 lead pairs each have 68 test patients. Prespecified secondary queries provide I+II and also evaluate V1 alone. MAE, RMSE, and marginal CRPS use $\mu$V. Conditional NLL subtracts the context log density from the joint log density, negates the result and divides by 6,000 missing scalars. Densities use mV coordinates. The reported 90% coverage and width refer to marginal intervals at individual time points, not simultaneous bands for an entire ECG. MAE averages the three fits; RMSE is the square root of MSE averaged across patients and fits. Bootstrap contrasts first average each patient's statistics over the fits, then resample patients: 4,000 resamples for the initial waveform study and 2,000 for the density follow-up. These intervals quantify patient sampling uncertainty conditional on the trained models and their acquisition assignments. No test patient is removed for a large error or a large NLL.

Table 5 reports both point metrics on the common query. Fixing the CNN at epoch 40 preserves the design ordering and changes uniform MAE from $85.03$ to $85.09\,\mu$V. The largest change is for separated panels, from $116.27$ to $119.03\,\mu$V.

**The matched-pair comparison.** Uniform and triple-hole CNNs differ little on the overall query average (Table 5). For the specific I+II→V1 task, the hole increases MAE by $0.788\,\mu$V ($0.98\%$), with patient interval $[0.580, 0.984]$; the three-fit increases are $0.890$, $0.211$, and $1.261\,\mu$V. Fixed-epoch selection gives $0.798\,\mu$V with interval $[0.595, 0.991]$, and one of its three repetitions is essentially zero. The effect is small and specific to this query, with variability across training repetitions despite the narrower patient interval.

Table 6: Cyclic acquisition: conditional NLL per missing scalar under two tail families and two completions. The fixed-shape $t_{10}$ transformation preserves each Gaussian's conditional means. All patients are included.

| | Gaussian | | Fixed-shape $t_{10}$ | |
|---|---|---|---|---|
| Partition | Original | Completed | Original | Completed |
| Test | −0.292 | −1.448 | −1.172 | −1.607 |
| Validation | 12.799 | 3.770 | −1.170 | −1.608 |

**A coherent heavy-tailed density check.** A shared-precision Student-$t$ model separates covariance completion from tail sensitivity. For one record, draw $\lambda \sim \mathrm{Gamma}(\nu/2, \nu/2)$ with rate parameterization, and conditionally draw all real Fourier coefficients with covariance $\Sigma_b/\lambda$. One latent precision is shared by the whole record. The resulting 8,000-dimensional joint law and every acquired marginal are multivariate $t$ laws with the same $\nu$. For $D$ observed scalars, Mahalanobis energy $\delta$, and block covariance $S$, its log density is

$$\log p(x) = \log\Gamma\left(\frac{\nu+D}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) - \frac{D}{2}\log(\nu\pi) - \frac{1}{2}\log\det S - \frac{\nu+D}{2}\log(1+\delta/\nu).$$

Incomplete-data $t$ estimation is classical (Little, 1988; Liu & Rubin, 1995). We fit $\nu \in \{3, 10, 30\}$ by observed likelihood, with posterior precision $(\nu+D)/(\nu+\delta)$, conditional residual covariances in the EM update, and a common-scale likelihood maximization. All 36 likelihood fits pass the final training gradient check. The $\nu = 10$ row is the prespecified reference degree of freedom; uniform-design test NLL is $-1.6244, -1.6237, -1.6234$ for $\nu = 3, 10, 30$. The supplement also records all 12 $t$ score fits, including four that fail the original gradient threshold after the documented extensions. These fits do not support an efficiency ranking.

**Completion at exactly fixed acquired marginals.** For the cyclic design, each band has four unobserved covariance entries. For each fitted Gaussian covariance $\Sigma_b^{(0)}$, optimize $\log\det S_b$ over these entries while fixing all others, as in classical covariance selection (Dempster, 1972). Damped Newton iterations check positive definiteness. If this unconstrained maximum-determinant endpoint violates the original eigenvalue floor .01, halve its displacement from $\Sigma_b^{(0)}$ until feasible. Three band-level endpoints across the Gaussian and $t$ fits require such backtracking. Denote the feasible endpoint by $\widetilde{\Sigma}_b$ and define

$$\Sigma_b^{(\alpha)} = (1-\alpha)\Sigma_b^{(0)} + \alpha\widetilde{\Sigma}_b, \qquad \alpha \in \{0, 1/2, 1\}.$$

The feasible completion moves toward higher entropy but may stop short of the maximum because of the eigenvalue floor. The same operation on $t$ scale matrices preserves its acquired marginals at fixed $\nu$.

For every acquired mask $A$, the principal matrices $\Sigma_{b,A}^{(\alpha)}$ are exactly independent of $\alpha$. Hence the entire acquired marginal densities agree, as do their population Fisher divergences and both empirical training objectives. For the Gaussian fits, define $\Delta_b = (\Sigma_b^{(0)})^{-1} - (\Sigma_b^{(\alpha)})^{-1}$. The joint divergence between the two fitted models in preconditioned coordinates is

$$\frac{1}{8000} D_{\mathrm{F}}(P^{(0)}, P^{(\alpha)}) = \frac{1}{8000}\sum_b c_b \,\mathrm{tr}(\Delta_b \Sigma_b^{(0)} \Delta_b). \tag{130}$$

This divergence uses the *original fitted Gaussian* as its reference. At $\alpha = 1$, the three values are 18.34, 17.51 and 18.42, while acquired-marginal divergence remains zero. Uniform and triple-hole designs leave no free covariance entries and are unchanged by this operation. The separated fit is already block diagonal and also remains unchanged. All 30 completion models pass these equality and positive-definiteness checks.

**Tail assumptions and covariance completion.** Replacing a fitted Gaussian by $t_{10}$ with scale .8 times its covariance preserves its unconditional covariance and every conditional mean. This fixed-shape comparator isolates density changes from point prediction. Table 6 applies it before and after

Table 7: PTB-XL common probabilistic scores ($\mu$V), averaged within patients and over three fits. Every entry uses 32 draws on the same 1,904 patients. Lower is better. Density models use acquired likelihood; CSDI uses acquired-target denoising. These are working-model comparisons.

| Model | Score | Separated | Cyclic | Uniform | Triple hole |
|---|---|---|---|---|---|
| Gaussian | CRPS | 101.58 | 86.46 | 86.62 | 86.22 |
| | Energy | 152.60 | 128.60 | 128.25 | 127.91 |
| Mixture | CRPS | 99.40 | 83.40 | 83.34 | 83.08 |
| | Energy | 151.53 | 126.38 | 125.80 | 125.52 |
| $t_{10}$ | CRPS | 97.28 | 82.24 | 82.06 | 82.05 |
| | Energy | 150.97 | 126.81 | 125.82 | 125.82 |
| CSDI | CRPS | 78.80 | 71.85 | 69.99 | 71.60 |
| | Energy | 137.59 | 120.06 | 116.37 | 118.20 |

completion. Completion changes Gaussian test MAE from $100.26$ to $99.54\,\mu$V; the larger change is in joint conditional density. The completed cyclic model nearly matches uniform Gaussian NLL ($-1.443$), showing that the original gap partly reflected its covariance completion. A heavy-tailed family additionally reduces sensitivity to large amplitudes: on fold 9, one patient's contribution to the mean cyclic Gaussian NLL is $13.004$, whereas the full mean is $12.799$. All patients remain in every reported metric.

**Study sequence and scope.** The main waveform protocol, patient acquisitions, queries and model configurations were fixed before test evaluation. The Student-$t$ and completion analyses are exploratory follow-ups on already examined folds. Their parameters were fixed before their own evaluations, and completion uses only fitted covariances and mask support. Thus the evaluations do not select the completion parameters, but they reuse the same test set. The supplement contains all result grids, optimizer statuses and scripts. The study assesses acquisition and reconstruction of waveforms, without evaluating clinical diagnostic fidelity or claiming state-of-the-art imputation. The external comparison below tests whether the completion benefits persist on other cohorts.

### E.5 Conditional diffusion at the same acquisition budget

We additionally fit CSDI (Tashiro et al., 2021) under the same fixed patient acquisitions: four layers, 64 channels, eight attention heads, and 50 diffusion steps with the original quadratic noise schedule. Each training example supplies two acquired leads as context and the other two as targets. Unacquired values are removed before spectral preconditioning and never enter the loss. All twelve models use the fixed 200th epoch, without validation-based checkpoint selection. Adam uses learning rate $10^{-3}$, reduced tenfold after epochs 150 and 180, batch 16 and gradient clipping at 5. Initialization, training order and random draws are paired across designs within each repetition. Neural evaluation uses BF16 autocast with FP32 diffusion states, and scores are accumulated in FP64.

We compare predictive distributions using 32 independent ancestral draws per record and their nested eight-draw prefix. For missing-coordinate truth $y \in \mathbb{R}^{6000}$ and draws $z_1, \dots, z_S$, both primary scores have the unbiased ensemble form (Ferro, 2014)

$$\widehat{\mathcal{S}}_\rho = \frac{1}{S}\sum_s \rho(z_s, y) - \frac{1}{2S(S-1)}\sum_{s\neq t}\rho(z_s, z_t),$$
$$\rho_1(u,v) = \frac{\|u-v\|_1}{6000}, \quad \rho_2(u,v) = \frac{\|u-v\|_2}{\sqrt{6000}}.$$

These give coordinate-averaged fair CRPS and normalized fair energy, respectively. All models use the same patients, queries, physical units and ensemble size in Table 7; density rows reuse the 32-draw prefixes of their exact conditional samplers. The original 100-draw CSDI budget was reduced before inspecting its test scores, following nested 8/16/32-draw calibration on 64 validation patients. The calibration assesses the precision of the proper scores, without testing tail-interval calibration.

Table 5 also evaluates the CSDI sample mean as a point prediction. Increasing from eight to 32 draws reduces MAE by 2.28–2.90 $\mu$V across designs without changing their ordering. Uniform acquisition gives MAE 93.27 $\mu$V and RMSE 186.44 $\mu$V with 32 draws. These point metrics retain finite-sampling error, whereas the fair scores above are unbiased in ensemble size. The fixed-epoch CNN has lower point error under cyclic, uniform and triple-hole acquisition; CSDI has lower error under separated acquisition. Sample-median MAE and the eight-draw results are also retained in the supplement.

Uniform–separated CSDI differences are $-8.81\,\mu$V for fair CRPS (patient-bootstrap interval $[-9.28, -8.35]$) and $-21.22\,\mu$V for fair energy ($[-22.42, -20.06]$). Their three-fit relative improvements range from $10.50\%$ to $12.43\%$ and $14.35\%$ to $16.25\%$, respectively. The eight-draw prefix gives $11.19\%$ and $15.42\%$ average improvements, close to the 32-draw values. Relative to cyclic acquisition, the 32-draw improvements are $2.59\%$ and $3.08\%$, positive in every fit. The matched-pair triple-hole contrast is smaller: $2.25\%$ and $1.55\%$ on average, with fit ranges $[-0.29, 4.49]\%$ and $[-0.65, 3.32]\%$. The first fit favors triple-hole acquisition on both scores. Intervals use the 2.5th and 97.5th percentiles of 10,000 patient resamples after averaging the three fits and condition on those fits. The supplement retains all designs, repetitions, fair variogram scores and ensemble-size diagnostics. This extension reuses the previously studied PTB-XL cohort. Its predictive scores assess conditional accuracy, separately from the joint Fisher target of the recovery theorems.

### E.6 Working Gaussian scores under a non-Gaussian data law

Let $R$ be the data law, fix a center $m$, and suppose $S = \mathbb{E}_R[(X-m)(X-m)^\top] \succ 0$ is finite. Set $G = N(m, S)$ and $Q_T = N(m, T)$ for $T \succ 0$. The expected Gaussian quadratic score on coordinates $A$ is $J_{R_A}(T_A) = \operatorname{tr}(T_A^{-1} S_A T_A^{-1}) - 2\operatorname{tr}(T_A^{-1})$. Completing the square gives its excess over the minimum at $S_A$:

$$\mathcal{E}_A(T) = \operatorname{tr}\{(T_A^{-1} - S_A^{-1}) S_A (T_A^{-1} - S_A^{-1})\} = D_{\mathrm{F}}\{N(0, S_A), N(0, T_A)\}. \tag{131}$$

This identity requires only second moments, including for empirical data. It also has a population Fisher interpretation. If $R$ and its relevant marginals have finite Fisher information and satisfy the usual score-matching integration-by-parts conditions, then Hyvärinen (2005, Theorem 1 and Section 3.1) gives

$$\mathcal{E}_A(T) = D_{\mathrm{F}}(R_A, (Q_T)_A) - D_{\mathrm{F}}(R_A, G_A) = D_{\mathrm{F}}(G_A, (Q_T)_A). \tag{132}$$

Indeed, integration by parts gives $D_{\mathrm{F}}(R_A, (Q_T)_A) = \mathbb{E}_{R_A}\|\nabla \log r_A\|^2 + J_{R_A}(T_A)$; the first term cancels when subtracting the value at $S_A$. Thus $G_A$ is the best Gaussian with the fixed center for each marginal, and all these optima are marginals of the same joint $G$. Theorem 1 yields

$$\mathcal{E}_{[d]}(T) \le \frac{\operatorname{cond}(S)\operatorname{cond}(T)^2}{w_{\mathrm{cov}}} \sum_A \pi_A \mathcal{E}_A(T). \tag{133}$$

Equation (3) transfers this excess to balanced conditional-score queries with factor $v$. The approximation error $D_{\mathrm{F}}(R, G)$ remains separate from the Gaussian-oracle excess.

For the spectral working family, let $x_{bf} \in \mathbb{R}^8$ index the $c_b$ real coefficient vectors in band $b$ and set $S_b = c_b^{-1} \sum_f \mathbb{E}_R[x_{bf} x_{bf}^\top]$. The Gaussian oracle repeats $S_b$ across that band; its excess is $\sum_b c_b D_{\mathrm{F}}\{N(0, S_b), N(0, T_b)\}$. The quadratic risk involves only these pooled within-vector moments, so this statement allows dependence across coefficients under $R$. Pooling commutes with taking a lead submatrix. Apply the eight-dimensional comparison to each band and sum. The masks select lead groups, so a uniform mask law over all scalar coordinates would describe a different acquisition scheme. The oracle permits all positive-definite band matrices without the fitting floor, and comparisons use fixed coordinate scales. It differs from the fitted-model reference in (130). Substituting held-out moments gives an empirical excess whose benchmark may be biased for the population optimum. Paired quadratic-score differences between fixed candidates cancel that benchmark. Under the stated integration-by-parts conditions, they estimate the population Fisher-risk difference.

**Observable paired risk along the completion path.** Write $K_b^{(\alpha)} = (\Sigma_b^{(\alpha)})^{-1}$ and let $Q_{ib} = \sum_f x_{ibf} x_{ibf}^\top$ be the band scatter, averaged within patient $i$ on PTB-XL. Figure 3b averages

$$\Delta_i(\alpha) = \frac{1}{8000} \sum_b \left[ \operatorname{tr}\Big( \{(K_b^{(\alpha)})^2 - (K_b^{(0)})^2\} Q_{ib} \Big) - 2c_b \operatorname{tr}(K_b^{(\alpha)} - K_b^{(0)}) \right]. \tag{134}$$

Table 8: Paired quadratic-score changes per scalar, averaged over three fixed Gaussian fits. C–O compares completed and original cyclic; U–S/U–O compare uniform with separated/original cyclic. Negative favors the first model. Source-preconditioned (P) and physical mV (V) coordinates define different losses.

| | C–O | | U–S | | U–O | |
|---|---|---|---|---|---|---|
| Cohort | P | V | P | V | P | V |
| PTB-XL | −188.33 | −87,322.52 | −0.19 | −159.77 | −187.63 | −87,287.29 |
| Chapman–Shaoxing/Ningbo | −1,138.69 | −618,024.02 | 91.57 | 7,058.17 | −1,094.19 | −615,240.45 |
| Georgia | −282.17 | −133,419.96 | 36.72 | 2,239.09 | −256.27 | −132,058.43 |

No evaluation covariance is fitted in this contrast. The three-fit means at $\alpha = 0, 1/2, 1$ are $0, -185.41, -188.33$ in the source-preconditioned coordinates. Acquired-marginal quadratic scores agree on every record. Under the population conditions above, the expectation is the joint Fisher-risk difference per scalar; averaging all 28 two-lead contexts gives the same conditional quadratic-score difference per missing scalar. The NLL experiment instead assigns one query to each record with balanced counts across queries. We compute the contrast in FP64 from all waveforms, check endpoint scores against the earlier full diagnostic, and check direct Fourier score norms and all 28 conditional query restrictions on three fixed audit records. Normalized discrepancies are below $3 \times 10^{-14}$. The 2,000 paired patient-bootstrap resamples use a fixed random stream, average the same three fits in each resample, and quantify patient sampling conditional on those fits. The paired data contrast complements the model divergence in (130). It is exploratory because these records were previously examined, although evaluation moments determine neither the completion operation nor its parameters.

### E.7 Fixed-marginal completion on external ECG cohorts

**Cohorts and frozen source models.** We use the Chapman–Shaoxing/Ningbo (CS/N) ECG release 1.0.0 (Zheng et al., 2022; 2020) and the public Georgia G12EC records in PhysioNet Challenge 2021 release 1.0.3 (Reyna et al., 2021; 2022). From 45,152 CS/N records, one header with ambiguous calibration is excluded and 87 exact duplicates are removed, leaving 45,064. All 10,344 Georgia file pairs are available; 52 five-second recordings fall outside the ten-second protocol and 91 exact duplicates are removed, leaving 10,201. No exact calibrated-waveform overlap is found between these two retained cohorts. Deduplication and format exclusions precede model scoring; no amplitude, diagnosis or model-error filter is applied.

Header gains, baselines and units determine mV signals, with independent WFDB decoding checks. Each of the same eight leads is resampled from 500 to 100 Hz using a fixed polyphase anti-aliasing filter (Kaiser parameter 5, constant padding); all 1,000 samples and the PTB-XL band preconditioner are retained. We evaluate fixed PTB-XL models without external training, selection or recalibration. Only the Challenge's public records are used, for waveform reconstruction rather than diagnostic classification. All three source repetitions are evaluated for original/completed cyclic Gaussian, covariance-matched fixed-$t_{10}$, and fitted-$t_{10}$ models, plus uniform and separated Gaussian fits: 24 configurations per cohort.

Each record receives one context from a balanced permutation of the 28 two-lead pairs, shared by all models. Exact random seeds are listed in the supplement. NLL is per missing scalar in mV, and point prediction uses conditional means. External summaries weight unique records equally. Without verified patient identifiers for clustering, we use 2,000 paired record-bootstrap resamples, conditional on the three source fits. Protocols, model configurations and cohort rules were fixed before the corresponding external predictive outcomes. The quadratic-score follow-up was specified after those outcomes were examined; its complete fixed-model comparisons use both source-preconditioned and physical coordinates, using a fixed bootstrap stream.

**Quadratic scores and acquisition designs.** Table 8 reports the completion contrast and both available Gaussian design contrasts on all three cohorts. The completion preserves every acquired marginal exactly and lowers quadratic risk in both coordinate systems and every source repeti-

Table 9: External completion: completed minus original, with all retained records and three specified families. Intervals below the estimates are paired record-bootstrap 95% intervals conditional on three fits. NLL is per missing mV scalar; MAE and RMSE changes are in $\mu$V.

| Cohort | Family | Δ NLL | Δ MAE | Δ RMSE |
|---|---|---|---|---|
| CS/N | Gaussian | −7.142<br>[−8.010, −6.404] | −0.374<br>[−0.408, −0.341] | 0.899<br>[−0.293, 2.425] |
| | Fixed $t_{10}$ | −0.440<br>[−0.442, −0.439] | −0.374<br>[−0.408, −0.341] | 0.899<br>[−0.293, 2.425] |
| | Fitted $t_{10}$ | −0.231<br>[−0.233, −0.230] | −0.166<br>[−0.192, −0.139] | 1.055<br>[−0.099, 2.598] |
| Georgia | Gaussian | −1.792<br>[−2.223, −1.438] | −0.436<br>[−0.487, −0.383] | −0.527<br>[−0.898, −0.200] |
| | Fixed $t_{10}$ | −0.411<br>[−0.414, −0.407] | −0.436<br>[−0.487, −0.383] | −0.527<br>[−0.898, −0.200] |
| | Fitted $t_{10}$ | −0.204<br>[−0.207, −0.201] | −0.172<br>[−0.204, −0.140] | −0.013<br>[−0.206, 0.215] |

tion. In preconditioned coordinates the external completion changes are $-1138.69$ (95% interval $[-1275.55, -1012.44]$) and $-282.17$ ($[-352.14, -230.11]$), for CS/N and Georgia respectively. These magnitudes depend on waveform amplitudes: the largest 1% of absolute record changes contribute 58% and 40% of the mean reductions, while 99.3% and 96.5% of records improve. All records remain included. Uniform–separated quadratic-score rankings reverse externally in both coordinate systems, showing that the preferred fitted design can change across cohorts. Decomposing this contrast over the eleven predefined frequency bands locates most of the external gap in the zero-frequency coefficient: it accounts for 86% of the mean preconditioned difference on CS/N and 97% on Georgia. This decomposition uses the fixed source models and cohort second moments. Summing the band contributions recovers every original contrast in both coordinate systems; the supplement provides all bands and source repetitions.

**Conditional density and point prediction.** Table 9 retains all three tail families. Completion lowers conditional NLL in every family and source repetition. The covariance-matched fixed-$t_{10}$ control has changes of $-0.440$ and $-0.411$ while preserving the corresponding Gaussian conditional means. Its largest 1% of absolute record changes contribute only about 2.2–2.3% of the total improvement, providing a less tail-sensitive check of the density effect. Point-prediction changes are small. Gaussian RMSE increases on CS/N, with an interval crossing zero. Separately, uniform Gaussian fits have lower MAE but higher NLL than separated fits on both external cohorts. The supplement gives the full grids, contrasts for each fit and intervals. Across these ECG populations, classical completion changes conditional density predictions while preserving every acquired marginal.

## F Additional related work

**Observed losses and joint targets.** Givens et al. (2025) approximate marginal score matching using importance weighting and variational inference. Our exact-marginal analysis isolates reconstruction from numerical marginalization, which remains necessary to evaluate the marginal scores. Ambient Diffusion learns conditional means by adding corruption to already corrupted records (Daras et al., 2023); its recovery result concerns that conditional-mean objective, and its proposed sampler uses an approximation. Conditional-mean recovery (Zhou et al., 2026), masked conditional-KL control via approximate tensorization (Li et al., 2024), and mixture-weight insensitivity (Cai & Shi, 2026) also concern different observed objectives. Basteri et al. (2026, Theorem 5.4) give a weighted marginal-KL guarantee for incomplete-data kernel density estimation. We study transfer from marginal discrepancies to joint spatial-score error.

**Interaction representations.** Low-order marginals identify log-density ANOVA models (Jeon, 2012). Geometric margins in Bayes spaces use a different projection (Genest et al., 2023). Our

density-ratio model uses ordinary marginalization to retain canonical components contained in each observed set.

**Statistical risk and efficiency.** Incomplete covariance estimation targets matrix error or graph recovery (Kolar & Xing, 2012; Cai & Zhang, 2016; Pavez & Ortega, 2021; Park et al., 2021). Our analysis uses classical information and projection principles (Gill & Levit, 1995; Dawid et al., 2016). Under full observation, score-matching efficiency is studied by Forbes & Lauritzen (2015); Koehler et al. (2023) and, through a Stein–Wasserstein decomposition, Nagai & Yano (2026). Proposition 14 weights the generator projection residual by inverse coverage. Related statistical bounds include full-data Fisher-risk convergence (Sriperumbudur et al., 2017) and finite-sample bounds for polynomial exponential families (Smedira et al., 2026). Regularity is necessary even with full observation (Lewis et al., 2025, Proposition 1). Proposition 11 varies joint-score sensitivity while fixing the informative marginal.