# BEYOND COVERAGE: WHICH OBSERVATIONS SHARE AN INSTANCE DECIDES WHAT A JOINT MODEL CAN LEARN

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

## ABSTRACT

Joint models learn from partially observed instances; designs are compared by coverage, how often each variable and pair is observed, not which observations share an instance. If three variables vary widely but their sum barely does, observing all three measures the sum's variance; observing pairs assembles it from far larger entries. For a Gaussian joint we define a design's information operator and prove that coverage fixes its trace, leaving its spectrum to which observations share an instance. The spectrum sets sample cost: at identical coverage, an isolated direction's variance costs a fixed number of instances when some instance observes all its variables and, for any estimator, one growing as the inverse square of that variance when none does. Moment and likelihood learners can rank coverage-matched designs oppositely. On budget-matched echocardiogram acquisition policies, our operator predicts the likelihood learner's covariance estimation error within $3.2\%$ with known truth, and a bound from a small pilot, before any fit, orders the policies, coverage-matched pairs included, as held-out likelihood does. A constructed domain realises the separation at four orders of magnitude. Coverage fixes how much a design observes; which observations share an instance decides what an unrestricted joint model can learn. Code: `https://anonymous.4open.science/r/beyond-coverage-code-3D5D`.

## 1 INTRODUCTION

Joint models are trained on partially observed instances: diffusion models on masked data (Daras et al., 2023; Ouyang et al., 2023), multimodal encoders on partially paired modalities (Girdhar et al., 2023), and planned-missingness studies (Graham et al., 2006; Rhemtulla et al., 2016). In an echocardiogram the ventricle's shapes at successive phases are coupled, and combinations that barely vary decide which contractions occur. An acquisition design, fixed in advance, decides which phases of which recording enter training.

Designs are compared by coverage: how often each variable is observed, and each pair together. Guarantees for masked generative models (Ouyang et al., 2023; Zhou et al., 2026) and covariance rates from incomplete data (Cai & Zhang, 2016) rest on it, and for a Gaussian joint a pair never observed together leaves its covariance undetermined (Vinci et al., 2019; Waghmare & Panaretos, 2022). Coverage does not record *co-observation*, the arrangement of measurements within instances (Mitra et al., 2023), so these guarantees cannot separate designs that differ only in co-observation.

Can two designs with identical coverage differ in what a joint model learns from them, by how much, and does the answer depend on the learner? (Q)

They can. Suppose three variables vary widely but their sum barely does. A fixed number of instances observing all three measures the sum's variance to fixed relative accuracy, however small that variance; instances observing pairs assemble it from separately estimated entries, each far larger. Design $A$ of Figure 1(a) observes all three on one instance and $B$ never; with a fourth variable, both make twelve observations on seven instances with identical coverage.

$A$ and $B$ differ in where information lies, not in how much. For a Gaussian joint, a design's Fisher information about the covariance is a linear map, its *information operator*. Coverage fixes its trace



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.

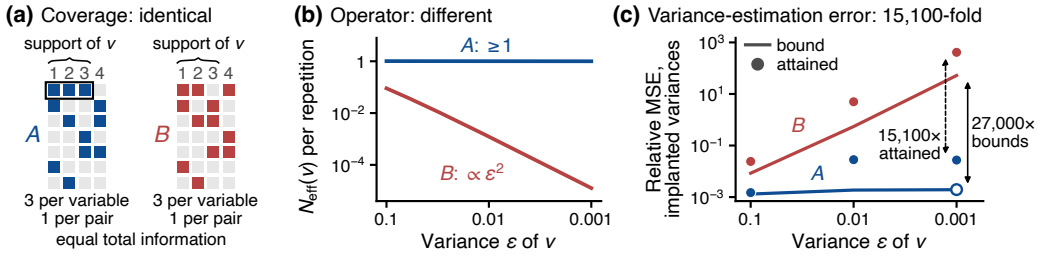



Figure 1: **Coverage ties designs $A$ and $B$; their information operators and the likelihood learner separate them.** (a) Instances (rows) by variables (columns), with coverage tallies below; $A$ observes the support of a low-variance direction $v$ on one instance (outlined), $B$ on none. (b) $N_{\mathrm{eff}}(v)$ per design repetition as the variance $\varepsilon$ of $v$ shrinks: at or above one under $A$, falling as $\varepsilon^2$ under $B$. (c) Relative mean squared error of the implanted variances in the 164-dimensional constructed domain under the same designs. Lines: Cramér–Rao bounds with every covariance parameter free. Markers: the likelihood learner (30 seeds; open marker: $A$ with its floor below $\varepsilon$, six seeds).

(Theorem 1); which observations share an instance shapes its spectrum, and with it the information along the sum. At identical coverage, the variance $\varepsilon$ along an isolated direction needs $O(1)$ instances for fixed relative accuracy when some instance observes all its variables, and $\Omega(\varepsilon^{-2})$ for any estimator when none does (Theorem 2).

The same example separates learners. Moment estimators combine separately estimated covariance entries; likelihood estimators jointly fit the observed-data model. They coincide under complete observation, respond differently when no instance observes every variable (*stitching*, a term from neural population recordings (Turaga et al., 2013)), and can rank coverage-matched designs oppositely (Theorem 3, Corollary 5).

A bound from the operator at a thirty-video pilot's scale orders all fifteen pairs of six budget-matched echocardiogram acquisition policies, coverage-matched ones included, as the likelihood learner's held-out negative log-likelihood does. With known truth, the operator predicts that learner's covariance estimation error within $3.2\%$. In a constructed domain, coverage-matched designs differ $15{,}100$-fold in that learner's relative mean squared error at the smallest implanted variance.

**Contributions.** Section 3 defines the information operator of a design, of which coverage fixes the trace and co-observation shapes the spectrum (Theorem 1). The resulting sample-cost separation at identical coverage holds on both sides with every covariance parameter free (Theorem 2 and its remark), and Section 4.4 realises it at four orders of magnitude. The likelihood learner's leading risk is $\frac{1}{2}\,\mathrm{tr}(\mathcal{M}^{-1})$, the moment learner pays a floor tax, and the two can rank coverage-matched designs oppositely (Theorem 3, Corollary 5). Section 4.2 turns the operator into a design diagnostic computed at a pilot's scale, before any model is fitted. Sections 4.5, 4.6 and 4.7 go beyond the unrestricted Gaussian learner: a structural prior fitted from the same observations recovers $91\%$ of the likelihood learner's stitching penalty, a segment-factorised neural model sees the design only through its segments, and a conditional Gaussian mixture keeps the coverage-matched order, with gaps that grow with its components.

## 2 Related Work

**Learning from masked data.** Guarantees rest on coverage or identification conditions: per-coordinate positive observation probability (Bora et al., 2018; Ouyang et al., 2023; Zhou et al., 2026), full-rank expected masking operator (Daras et al., 2023), identification from observed marginals (Givens et al., 2025; Vinci et al., 2019; Waghmare & Panaretos, 2022), true-conditional completions (Khan et al., 2026), or, in masked prediction, positive full-mask probability (Cai & Shi, 2026). At matched missing rate, independent dropout is harder than block observation (Li et al., 2019), and higher-order co-occurrence within an example decides identifiability (Liu et al., 2022). For corrupted images a few clean instances change the rate (Daras et al., 2025; Lu et al., 2025), as covering instances do in Theorem 2. EM-trained diffusion priors are likelihood-type neural learners (Rozet et al., 2024).

**Partially paired multimodal learning.** Pairwise objectives miss higher-order dependence (Saporta et al., 2024), and a hub binds modalities (Girdhar et al., 2023; Jeong et al., 2024; Xie et al., 2026). Which views are observed together decides which latent content is identifiable (Yao et al., 2024), non-Gaussian latents can be identified from unpaired marginals (Sturma et al., 2023), and unpaired second-modality samples add Fisher information along a direction primary-modality samples cannot (Gupta et al., 2026). Holding coverage fixed, we turn such identifiability and Fisher-information statements into a sample cost that co-observation sets. Adaptive acquisition chooses pairings or modalities per instance for a prediction (Zhang & Zhu, 2026; Rheude et al., 2025); we compare designs fixed in advance for learning the joint.

**Covariance estimation and planned missingness.** Minimax rates scale with the smallest pairwise count (Lounici, 2014; Cai & Zhang, 2016), and inverse-probability-weighted estimates under dependent missingness depend on marginal and pairwise observation probabilities alone, with a positive-definiteness correction that clips eigenvalues (Park et al., 2021). Stitching recovers the joint of partial neural recordings through latent structure (Turaga et al., 2013): the regime of Section 4.5. The fraction of missing information governs EM's convergence rate and asymptotic variance (Dempster et al., 1977; Meng & Rubin, 1991). Planned-missingness designs are built and judged by their marginal and pairwise counts: the three-form and balanced-incomplete-block designs equalise how many respondents answer each item set and pair of sets (Graham et al., 1996; 2006; Raghunathan & Grizzle, 1995; Rhemtulla et al., 2016). Their expected information is the pattern-weighted Fisher information, optimised by A- and D-criteria (Steel & Chipperfield, 2026; Yi et al., 2021; 2022); and split-questionnaire designs choose which blocks share a form from a pilot by an information-loss criterion (Adigüzel & Wedel, 2008). Chipperfield & Steel (2011) and Steneman & Vinci (2024) compute each observation pattern's contribution to estimator variance or information. Planned-missingness work sums pattern-specific Fisher information (Rhemtulla et al., 2016; Brandmaier et al., 2020; Jang et al., 2022); we whiten that sum so its trace becomes a coverage count, and add the coverage-matched rate of Theorem 2 and the learner-level account of Theorem 3.

**Classical design criteria.** Optimal design ranks experiments by spectral criteria (Kiefer, 1959; Pukelsheim, 2006); the coverage of reinforcement learning is a different quantity (Xie et al., 2023). The covariance guarantees and planned-missingness balance criteria above are functions of marginal and pairwise counts; we compare designs in which those counts are identical and measure how far their directional information and learners' risks separate (Appendix K).

## 3 Approach: the information operator of a design

The information operator describes how precisely each covariance perturbation can be estimated (Figure 2). The likelihood learner's risk reads off it; the moment learner adds the noise of separately pooled entries.

### 3.1 Designs, coverage and co-observation

A design decides which blocks of a Gaussian joint vector each instance reveals. The vector $a \in \mathbb{R}^D$ has $m$ blocks of $d$ coordinates, $D = md$, and $a \mid x \sim N(\beta^\top x, \Sigma)$ given covariates $x \in \mathbb{R}^q$, with $\Sigma \succ 0$. An instance with observation pattern $S$, a subset of blocks, reveals $z_S = H_S a$ through an observation map $H_S$ of full row rank $d|S|$.

A design is a multiset $\{(S, n_S)\}$: $n_S$ instances observe pattern $S$. Its budget $N_{\text{obs}} = \sum_S |S| n_S$ counts observed blocks.

Coverage is the marginal counts $n_j = \sum_{S \ni j} n_S$ and pairwise counts $n_{jk} = \sum_{S \supseteq \{j,k\}} n_S$; designs with equal budget and coverage are coverage-matched. Under complete co-observation every instance observes all $m$ blocks. A direction $v$'s support is the blocks where $v$ is nonzero.

Construction 1 is the three-measurement example of Section 1 as two coverage-matched designs (Figure 1(a)):

$$A = \{123, 14, 24, 3, 34, 1, 2\}, \qquad B = \{124, 13, 23, 4, 34, 1, 2\}. \tag{1}$$

Here $m = 4$, $d = 1$, $\Sigma = I - (1 - \varepsilon) v v^\top$ and $v = (1, 1, 1, 0)/\sqrt{3}$, so $v^\top a$ has variance $\varepsilon$ and every other direction variance one.

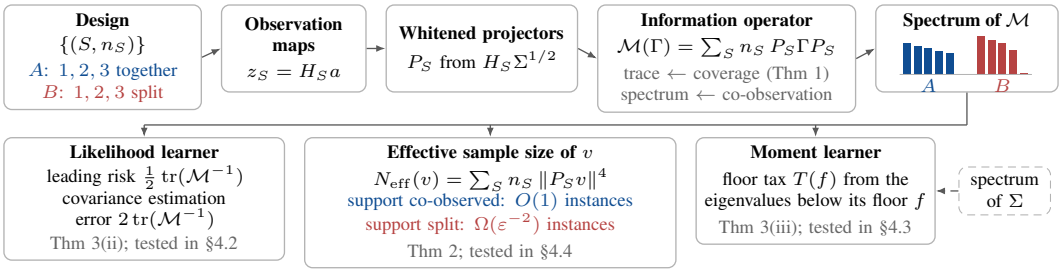



Figure 2: **A design reaches a learner through its information operator: coverage fixes its trace, co-observation shapes its spectrum.** Tags name each branch's result and its test.

### 3.2 Coverage fixes the trace; co-observation shapes the spectrum

With $G_S = H_S\Sigma^{1/2}$, the whitened projector $P_S = G_S^\top(G_S G_S^\top)^{-1}G_S$ projects onto the directions $S$ sees; summing over instances gives the operator. Part (i) of Theorem 1 whitens the pattern-weighted Gaussian information of planned-missingness designs (Rhemtulla et al., 2016; Jang et al., 2022); part (ii), the trace identity, ties coverage-matched designs.

**Theorem 1** (Coverage fixes the trace of the information operator)**.** Let instances be independent, let the design be fixed independently of the observed values, and let $\Gamma$ range over symmetric $D \times D$ matrices with the Frobenius inner product. **(i)** The Fisher information that the design carries about a whitened perturbation $\Sigma \mapsto \Sigma^{1/2}(I + \Gamma)\Sigma^{1/2}$ of the covariance is

$$\mathcal{I}(\Gamma, \Gamma) = \tfrac{1}{2}\sum_S n_S \|P_S\Gamma P_S\|_F^2 = \tfrac{1}{2}\langle \Gamma, \mathcal{M}\Gamma\rangle, \qquad \mathcal{M}(\Gamma) = \sum_S n_S P_S \Gamma P_S, \tag{2}$$

and the regression mean is Fisher-orthogonal to the covariance. **(ii)** With blocks of $d$ coordinates,

$$\operatorname{tr}\mathcal{M} = \tfrac{d(d+1)}{2}\sum_j n_j + d^2 \sum_{j<k} n_{jk}, \tag{3}$$

so coverage-matched designs have equal-trace operators.

Complete co-observation of $n$ instances gives $\mathcal{M} = nI$. Whitened at the truth, $\mathcal{M}/n$ is the covariance block of the complement of the expected fraction-of-missing-information matrix of expectation–maximisation (Dempster et al., 1977; Meng & Rubin, 1991): coverage fixes the mean fraction of information missing, and co-observation its spread. A sampling graph's spectrum, not its entry count, governs matrix completion in the same way (Bhojanapalli & Jain, 2014). Information directions, eigen-directions of $\mathcal{M}$, differ from covariance directions, eigen-directions of $\Sigma$. In Construction 1 both designs have trace 18. $B$'s information ellipsoid collapses along $vv^\top$, whose support no instance observes; $A$'s keeps at least one unit of information there. Proof idea: independent instances add information, and $\Gamma \mapsto P_S\Gamma P_S$ has trace $d|S|(d|S|+1)/2$. Proof in Appendix A.

### 3.3 Observing a low-variance direction's whole support sets its sample cost

Each instance gives a covariance direction $v$ the fourth power of its whitened projection length: one if it observes $v$'s support, small for low-variance $v$ otherwise. The sum is $v$'s effective sample size,

$$N_{\text{eff}}(v) = \sum_S n_S \|P_S v\|^4 \in \Big[0, \sum_S n_S\Big]. \tag{4}$$

**Theorem 2** (Observing a direction's whole support decides its sample cost)**.** Let $v$ be a unit vector with $\Sigma v = \lambda_v v$. With the other parameters known, the design's information about the variance $\lambda_v = v^\top\Sigma v$ along the fixed direction $v$ is $N_{\text{eff}}(v)/(2\lambda_v^2)$, so $2/N_{\text{eff}}(v)$ is the Cramér–Rao bound on the relative variance of its regular unbiased estimators. Let $\Sigma_\varepsilon = \varepsilon vv^\top + \Sigma_\perp$ with $\Sigma_\perp v = 0$ and $\Sigma_\perp$ positive definite on $v^\perp$, so that $\varepsilon$ is an isolated eigenvalue. Say that $S$ covers $v$ when $v$ lies in the row space of $H_S$ (for coordinate blocks: $S$ contains the support of $v$). **(i)** A covering instance carries information $1/(2\varepsilon^2)$ about $\varepsilon$. If $S$ does not cover $v$, then with $c_S = (H_S v)^\top(H_S\Sigma_\perp H_S^\top)^{-1}H_S v$, $\|P_S v\|^2 = \varepsilon c_S/(1+\varepsilon c_S)$ and $\mathcal{I}_S(\varepsilon,\varepsilon) = c_S^2/(2(1+\varepsilon c_S)^2) = O(1)$. **(ii)** Repeat the design $r$ times, with $v$ and the mean known. If some observed subset covers $v$, the mean square of the centred

covered contrast $v^\top a$ over the $rn_c$ covering instances ($n_c$ per repetition) estimates $\varepsilon$ with relative mean squared error $2/(rn_c)$ at every $\varepsilon$. If none does, every estimator whose relative error is below $1/3$ with probability at least $3/4$ at both $\varepsilon$ and $2\varepsilon$ needs $r \geq 2/(\varepsilon^2 \sum_S n_S c_S^2)$, which is $\Omega(\varepsilon^{-2})$ as $\varepsilon \to 0$ with $\Sigma_\perp$ and the design fixed; under $B$ of Construction 1, $\sum_S n_S c_S^2 = 51/4$ per repetition. Along $v$, a model carrying $2\varepsilon$ for $\varepsilon$ has predictive KL $\frac{1}{2}\log 2 - \frac{1}{4} = 0.0966$ nats at every $\varepsilon$.

In Construction 1, $N_{\text{eff}}(v)$ per repetition is at least one under $A$ and falls as $\varepsilon^2$ under $B$ (Figure 1(b)). At $\varepsilon = 0.001$, $B$'s scalar estimate of $\varepsilon$ has 42,300 times $A$'s relative mean squared error (Appendix C). Proof idea: covering instances see $v^\top a$, and others carry bounded information. A two-point test with only $\varepsilon$ unknown bounds $r$ and so binds every estimator of a larger model. Proof in Appendix A.

**Remark.** With $e_v = vv^\top$, $N_{\text{eff}}(v) = \langle e_v, \mathcal{M} e_v \rangle$. With every covariance parameter free and $\mathcal{M}$ invertible, the Cramér–Rao bound on the relative variance of the variance $v^\top \Sigma v$ along the fixed direction $v$ is $2\langle e_v, \mathcal{M}^{-1} e_v \rangle$, and $2/N_{\text{eff}}(v) \leq 2\langle e_v, \mathcal{M}^{-1} e_v \rangle \leq 2/n_c$ when $n_c$ instances cover $v$ (Appendix A), so the $O(1)$ cost holds with every parameter free (Section 4.4).

### 3.4 Moment and likelihood learners respond differently to stitching

Two learners turn identical observations into different covariance estimates and can rank designs oppositely. The moment learner is the inverse-probability-weighted estimator with eigenvalue projection of Park et al. (2021). It assembles blockwise available-case moments, which can be indefinite (Little & Rubin, 2019), and raises eigenvalues below a floor $f$ to $f$. The likelihood learner maximises the observed-data likelihood; its estimate is positive definite. Both select ridge and floor on held-out data. Part (i) of Theorem 3 rests on the complete-data equality of composite and full likelihood (Mardia et al., 2009), and part (ii) on the standard efficiency expansion, applied to the operator.

**Theorem 3** (The learners coincide under complete co-observation and differ short of it)**. (i) The learners coincide.** Under complete co-observation, with a common sample, a common mean fit, normalisation by $n$, a residual matrix of full column rank and inactive regularisation, the moment estimator is the Gaussian maximum-likelihood estimator. **(ii) The operator sets the likelihood risk.** Assume an identifiable design, $\mathcal{M} \succ 0$, repeated $r$ times, a fixed dimension and a correctly specified Gaussian model. Let $\widehat{\Sigma}$ be a regular efficient observed-data maximum-likelihood estimator with negligible penalties and the uniform integrability that turns asymptotic normality into expected risk. Then, to leading order, the covariance contributes $K_{\text{cov}} = \frac{1}{2}\operatorname{tr}(\mathcal{M}^{-1})$ to the expected predictive KL and the covariance estimation error $\left\| \Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} - I \right\|_F^2$ has expectation $2\operatorname{tr}(\mathcal{M}^{-1})$. A $q$-column regression mean adds a term $\frac{q}{2}\operatorname{tr}(\mathcal{M}_{\text{mean}}^{-1})$, with $\mathcal{M}_{\text{mean}} = \sum_S n_S P_S$, under the conditions in Appendix B. **(iii) Moment noise persists along partly observed low-variance directions and forces a floor.** Let the mean be known, and let $n_{ij} > 0$ instances observe coordinates $i$ and $j$ together, for every $i, j$; for a fixed design the inverse-probability weights equal $1/n_{ij}$. For a unit $v$, put $(V_t)_{ij} = v_i v_j \mathbf{1}\{i, j \text{ observed on instance } t\}/n_{ij}$. The unfloored moment estimate $\widehat{\Sigma}$ has $\mathbb{E}[v^\top \widehat{\Sigma} v] = v^\top \Sigma v$ and $\operatorname{Var}(v^\top \widehat{\Sigma} v) = 2\sum_t \|\Sigma^{1/2} V_t \Sigma^{1/2}\|_F^2$. The limit of this variance as $\varepsilon \to 0$ in the family $\Sigma_\varepsilon$ of Theorem 2 is positive whenever some instance observes part, but not all, of the support of $v$. Among covariances $\Sigma' \succeq fI$, the smallest $\mathrm{KL}(N(0, \Sigma) \,\|\, N(0, \Sigma'))/D$ is the *floor tax* $T(f)$:

$$T(f) = \frac{1}{2D} \sum_{\lambda_i < f} \Big[ \log \frac{f}{\lambda_i} - 1 + \frac{\lambda_i}{f} \Big], \tag{5}$$

attained by flooring the eigenvalues $\lambda_i$ of $\Sigma$ at $f$.

The least-informed directions dominate the likelihood learner's risk (Figure 2). Each moment entry pools every instance observing it. Along a partly observed direction the moment noise therefore keeps the coordinates' scale as the variance shrinks, even beside a covering instance. As $\varepsilon\sqrt{r} \to 0$ the estimate along $v$ turns negative with probability approaching one half, and the forced floor $f$ adds at least $T(f)$. The floor tax builds on the indefiniteness threshold of Jurczak & Rohde (2017) and the Gaussian KL of eigenvalue flooring of Tumminello et al. (2007). Proof in Appendix B.

**Corollary 4** (Complete co-observation is asymptotically optimal at a fixed budget)**.** Under equal block costs, continuous counts $n_S$, no minimum count for any block and enough instances to observe

completely, let $K_{\mathrm{cov}}$ be the efficient likelihood learner's leading covariance risk (Theorem 3(ii)), and let $p = D(D+1)/2$ be the number of covariance parameters. Among designs of budget $N_{\mathrm{obs}}$, for every $\Sigma \succ 0$ with an unrestricted covariance,

$$K_{\mathrm{cov}} = \tfrac{1}{2}\operatorname{tr}(\mathcal{M}^{-1}) \;\geq\; \frac{p^2}{2\operatorname{tr}\mathcal{M}} \;\geq\; \frac{mp}{2N_{\mathrm{obs}}}. \tag{6}$$

Equality holds throughout exactly for complete co-observation, with $\mathcal{M} = (N_{\mathrm{obs}}/m)I$.

Corollary 4 is classical A-optimality (Kiefer, 1959; Pukelsheim, 2006); Theorems 1 and 2 add why coverage ties designs and how fast co-observation separates them. Proof in Appendix A. Construction 2 realises opposite rankings by the two learners.

**Corollary 5** (The two learners can rank coverage-matched designs oppositely)**. Construction 2.** Over four variables, design $A$ observes each variable alone twice, each pair twice and all four together four times. Design $B$ observes each variable alone three times and each triple three times. Both use $24$ instances and $48$ measurements, $12$ per variable and $6$ per pair. For an equicorrelated Gaussian with correlation $0.9$, an unrestricted covariance, a known mean and $r$ repetitions, under the uniform integrability of Theorem 3(ii), the expected predictive KL of the whole four-variable joint is $K/r + o(1/r)$ nats. For the moment learner with an asymptotically inactive floor, $K_A = 19.09 > K_B = 18.22$. For the likelihood learner, $K = rK_{\mathrm{cov}} = \frac{1}{2}\operatorname{tr}(\mathcal{M}_1^{-1})$, $\mathcal{M}_1$ the operator of one repetition, and $K_A = 0.711 < K_B = 0.732$. The reversal is $4.8\%$ and $2.9\%$ of constants that differ 26-fold.

Both designs co-observe every pair, whose contrasts span the low-variance directions, so none is hidden from the likelihood learner; pairwise likelihood is efficient as correlation approaches one (Cox & Reid, 2004). At fixed coverage the moment learner's leading risk is affine in the three- and four-way co-observation counts, with weights set by $\Sigma$. $A$'s complete instances raise those counts and with them its moment risk; the likelihood learner gains from them. Proof in Appendix B. Designs must be compared for a specified learner.

## 4 Experiments, results and analysis

### 4.1 A randomised acquisition experiment on fully recorded videos

The six-policy acquisition study tests the operator's predicted error magnitudes and design order. Its thirty-video pilot is completely observed, so every covariance direction is identified under every policy; Section 4.4 tests Theorem 2's $\varepsilon^{-2}$ regime. EchoNet-Dynamic (Ouyang et al., 2020) supplies $7{,}242$ training videos (the pilot included), $64$ selection, $1{,}176$ evaluation and $1{,}244$ test videos. A block (Section 3.1) is a frame's segmenter-derived $41$-coordinate left-ventricular contour at interior phase $t \in \{0.2, 0.4, 0.6, 0.8\}$ between end-diastole ($t = 0$) and end-systole ($t = 1$). The joint vector $a$, the $164$ spline coefficients of the deformation from end-diastole, is Gaussian given $84$ covariates including the end-systolic deformation.

Each policy uses $14{,}400$ frames and observes each phase on $3{,}600$ of the $7{,}212$ videos outside the pilot (Table 1); an assignment, one random draw of which videos supply which phases, is the replication unit. windows observes four phases, triples three, pairs two; pairs+windows 1:1 and pairs+windows 3:2 split frames $50{:}50$ and $60{:}40$ between pairs and windows; singles+triples puts $10\%$ on single phases, $90\%$ on triples. Regularisers are selected on the selection videos from grids shared by every policy; the test videos are scored once, after every choice. Bootstrap intervals are crossed over assignments and videos; those of Section 4.3 come from the two-policy uncertainty study, windows against pairs at $14{,}424$ frames.

**Metrics.** Held-out conditional negative log-likelihood (NLL) and every KL are in nats per coefficient. With known truth, excess NLL is the NLL above the true law's and predictive KL its expectation; covariance-only KL leaves out the estimated mean. For the likelihood learner the operator's predicted excess is the leading-order predictive KL, and the covariance estimation error $\left\|\Sigma^{-1/2}\widehat{\Sigma}\,\Sigma^{-1/2} - I\right\|_F^2$ has leading-order expectation $2\operatorname{tr}(\mathcal{M}^{-1})$ (Theorem 3(ii)); for the moment learner the floor tax bounds covariance-only KL from below (Theorem 3(iii)).

Table 1: **At fixed budget and per-phase coverage, both learners' held-out NLLs and the covariance estimation error, truth known, follow the operator's order.** Rules bracket coverage-matched pairs. Predicted excess: the operator's covariance plus mean term from the cardiac reference covariance. Pairwise coverage in videos per phase pair; excess and NLL in nats per coefficient; NLL on 1,176 evaluation videos, mean (s.e.) over 30 assignments.

| | Predicted excess | Pairwise coverage | Held-out NLL, real videos | | Covariance estimation error, truth known | |
|---|---|---|---|---|---|---|
| Policy | | | Moment | Likelihood | Predicted | Measured |
| windows | 0.0229 | 3,600 | 0.4749 (0.0001) | 0.4774 (0.0001) | 7.45 | 7.60 |
| pairs+windows 1:1 | 0.0276 | 2,400 | 0.6837 (0.0071) | 0.4844 (0.0001) | 10.36 | 10.62 |
| triples | 0.0282 | 2,400 | 0.6859 (0.0051) | 0.4868 (0.0002) | 10.78 | 11.12 |
| pairs+windows 3:2 | 0.0294 | 2,160 | 0.7176 (0.0097) | 0.4874 (0.0001) | 11.41 | 11.70 |
| singles+triples | 0.0299 | 2,160 | 0.7180 (0.0113) | 0.4900 (0.0001) | 11.75 | 12.13 |
| pairs | 0.0522 | 1,200 | 0.9490 (0.0066) | 0.6297 (0.0036) | 25.79 | 25.45 |

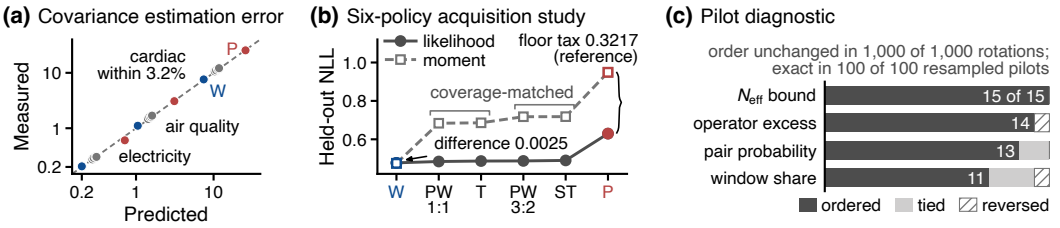



Figure 3: **The operator predicts the covariance estimation error, orders the policies, and orders them from a pilot.** (a) Likelihood learner's covariance estimation error, predicted by $2\,\mathrm{tr}(\mathcal{M}^{-1})$ and measured, truth known: within $3.2\%$ under every cardiac policy, $7.4\%$ under every air-quality policy and $3\%$ under five of six electricity policies; electricity pairs: $0.72$ predicted, $0.61$ measured. Dashed: equality. Covariance-only KL: Section 4.2. (b) Held-out NLL in the operator's order; brackets: coverage-matched pairs; brace: floor tax of the reference covariance. (c) Policy pairs, of fifteen, each pilot predictor orders as the scores do, ties or reverses; the resampled pilots are thirty-video pilots (Appendix D). W windows, PW pairs+windows, T triples, ST singles+triples, P pairs.

### 4.2 The operator predicts the covariance estimation error and orders the designs

With known truth and the real masks, the likelihood learner's covariance estimation error is within $3.2\%$ of its prediction under every policy (Figure 3(a)); in air quality (Zhang et al., 2017) within $7.4\%$ under every policy, and in electricity load (Trindade, 2015) within $3\%$ under five of six. Covariance-only KL carries higher-order terms, larger where the fitted covariance's smallest whitened eigenvalue is smaller: it is within 7–22% of its leading-order value under five policies and $2.9$ times it under pairs, where that eigenvalue is $0.08$ (Appendix E).

Both learners' held-out NLLs follow the operator's order (Spearman $1.0$; Figure 3(b)), as in both further domains (Spearman $0.94$ to $1.0$). The stitching penalty, pairs minus windows, compares designs differing in coverage and co-observation: pairs has one third of windows' pairwise coverage. It is $0.152$ for the likelihood learner, above its leading-order $0.029$ (Appendix E), and $0.474$ for the moment learner.

Within the coverage-matched pairs (pairs+windows 1:1, triples; pairs+windows 3:2, singles+triples), which coverage cannot separate, the likelihood learner resolves both contrasts in the operator's direction, second-named policy subtracted: $-0.0025$ each (crossed $95\%$ intervals $[-0.0030, -0.0020]$ and $[-0.0031, -0.0020]$). They replicate on the test videos ($-0.0023$ each). Like the stitching penalty, they exceed their leading-order sizes, $0.0006$ and $0.0005$. The moment learner's contrasts point the same way within its larger noise ($-0.0022$, s.e. $0.0085$; $-0.0004$, s.e. $0.0111$). In the further domains, coverage-matched contrasts lie within two standard errors of zero (Appendix F).

Before any model is fitted, the mean of $1/N_{\mathrm{eff}}$ over the pilot's covariance directions (Equation (4)) orders all fifteen policy pairs as both learners' held-out NLLs do, coverage-matched pairs included; two coverage summaries order 13 and 11 (Figure 3(c)). The order comes from the design operator,

not the pilot's covariance estimate: with an isotropic covariance at the pilot's scale the bound orders all fifteen in each of 1,000 random bases, and on the pilot's 29 determined directions alone it orders them the other way (Appendix D).

### 4.3 THE FLOOR TAX BOUNDS THE MOMENT LEARNER'S EXCESS FROM BELOW

Under pairs, the floor tax of Theorem 3(iii) bounds the moment learner's covariance-only KL from below. In the two-policy uncertainty study the learners differ by $0.3202$ ($95\%$ interval $[0.3025, 0.3386]$), beside a floor tax of $0.3217$ on the cardiac reference covariance, averaged over the selected floors ($0.3$ in 27 assignments, $0.1$ in three). The moment learner's KL is this difference plus the likelihood learner's own excess, $0.11$ with known truth (Appendix E). Under windows, where Theorem 3(i) predicts coincident learners, the implementations differ by $0.0025$ through their mean fits (Appendix B).

The floor trajectory is the mechanism (Table 5). The likelihood learner's floor lies below the spectrum: its floor tax is zero. Available-case noise forces the moment learner's floor: its held-out NLL falls from $493$ at $10^{-4}$ to a minimum of $0.950$ at $0.3$.

### 4.4 IN A CONSTRUCTED DOMAIN, COVERAGE-MATCHED DESIGNS DIFFER BY FOUR ORDERS OF MAGNITUDE

The constructed domain reaches Theorem 2's $\varepsilon^{-2}$ regime, which the completely observed pilot keeps out of the real-data study. Into the cardiac covariance it implants $41$ contrasts $v \otimes I_{41}$, $v = (1, -2, 1, 0)/\sqrt{6}$, of variance $\varepsilon \in \{0.1, 0.01, 0.001\}$, observed through the coverage-matched designs of Construction 1 with 1,034 instances per pattern, none complete. The implanted directions live on the first three phases, which no pattern of $B$ contains, so Lemma A.1 applies to $B$; at $\varepsilon = 0.001$, $N_{\mathrm{eff}}(v)$ is at least 1,034 under $A$ and median 17 under $B$.

At $\varepsilon = 0.001$ the likelihood learner's relative mean squared errors of the implanted variances separate the designs 15,100-fold (Figure 1(c)), and the Cramér–Rao bounds $2\langle e_v, \mathcal{M}^{-1} e_v\rangle$ with all 13,530 covariance parameters free separate them 27,000-fold (medians; Table 11). The learner's separation stays between 12,000- and 17,000-fold across an eightfold range of instance counts and holds on every floor grid; the exchange control, which swaps the roles of $A$ and $B$, swaps their covariance-only KL (Appendix G). Under $B$ the error grows at the predicted rate, with log–log slope $-2.1$ in $\varepsilon$. There the learner does not resolve the implanted variances: it overestimates them on average by $0.014$ to $0.022$ in absolute units at every $\varepsilon$.

Under $A$ the $O(1)$ regime tolerates leakage that is small against $\varepsilon$. When the implanted direction's squared mass outside the co-observed support is $0.15\varepsilon$, the relative mean squared error rises 2.5-fold; as that mass reaches $\varepsilon$ the error rises (59-fold at $1.7\varepsilon$). The moment learner offers a second, finite-sample route to Corollary 5's opposite rankings: at $\varepsilon = 0.001$ its selected floor reverses the order, and in covariance-only KL the likelihood learner prefers $A$ by $0.41$ nats per coefficient while the moment learner prefers $B$ by $0.83$ (Appendix G).

### 4.5 STRUCTURE RECOVERS MOST OF THE LIKELIHOOD LEARNER'S STITCHING PENALTY

When co-observation is unavailable, structure is the remedy. Theorem 2 binds estimators that leave a direction's variance free. The structural prior, fitted from the same observations, ties low-variance directions to well-observed quantities through each phase's covariance and a phase-distance correlation $\exp(-(\Delta t/\ell)^{\alpha})$. Shrunk toward it uniformly, the likelihood learner recovers $91\%$ of its stitching penalty (gain $0.1385$ nats per coefficient, $95\%$ interval $[0.131, 0.146]$); operator weighting adds $0.005$, within that interval (Figure 5(a); Appendix H).

With the floor also selected, the structural target replaces the moment learner's covariance in $149$ of $150$ fits outside windows. It scores $0.538$ nats per coefficient under pairs, against $0.949$ for the fitted moment learner and $0.092$ below the likelihood learner's $0.630$, so the target supplies two thirds of the $0.139$ gain.

### 4.6 A SEGMENT-FACTORISED NEURAL LEARNER SEES THE DESIGN ONLY THROUGH ITS SEGMENTS

The neural trajectory model, a stochastic interpolant (Chen et al., 2024), is the structured case of Section 4.5. Because its objective factorises over adjacent-phase segments, weighted by inverse inclusion probability (Islam et al., 2026), the model can respond to a design only through which segments it observes at both ends. Segment-matched dropout, which keeps each complete-data segment at the policy's inclusion rate, tests exactly this, with an equivalence margin of $\pm 0.10$ on the held-out complete-data objective fixed in advance. Over thirty seeds, pairs and triples are equivalent to their controls (paired differences $-0.013$ and $-0.011$, s.e. $0.035$ and $0.031$). pairs+windows 1:1 lies $0.052$ below its control ($95\%$ interval $[-0.111, +0.007]$), in the policy's favour, and no tested policy's interval admits a cost as large as the margin (Figure 5(b); Appendix I).

### 4.7 A MIXTURE LEARNER KEEPS THE ORDER, WITH GAPS THAT GROW WITH ITS COMPONENTS

A conditional Gaussian mixture with a full covariance per component, fitted through the same observed-data likelihood over all observed phases, carries the test beyond one Gaussian. Held-out selection keeps one component in every assignment, so the Gaussian analysis describes the selected model. With two or four components fixed, every assignment but one keeps the coverage-matched order, and the gaps grow with the number of components, to $-0.027$ and $-0.024$ nats per coefficient at four (paired s.e. $0.002$). The stitching penalty grows from $0.152$ nats per coefficient at one component to $0.50$ at four, through coverage and co-observation together (Appendix D).

The separation of Section 4.4 holds under a two-component truth: with implanted variances $\varepsilon/2$ and $2\varepsilon$, a two-component mixture learner recovers the variance above the selected floor, $2\varepsilon$, to $5\%$ under $A$ and overestimates it sixteenfold under $B$; its relative mean squared error of the implanted variances is $0.08$ under $A$ and $600$ under $B$ (Appendix G).

### 4.8 SCOPE

The analysis concerns observation designs fixed in advance and Gaussian joint distributions, with Gaussian mixtures as the non-Gaussian test. The information identities and the bounds of Theorem 2 hold at finite sample size; the rate concerns an isolated eigenvalue, with the shrinking-subspace version in Lemma A.1, and the risk expressions of Theorem 3 are leading order in the regular asymptotic regime. Acquisition is emulated on fully recorded EchoNet-Dynamic videos, with segmenter-derived contours as targets. In the constructed domain the KL comparisons move with the floor grid (Appendix G.8). The structural prior is one temporal family, and the neural learner's objective factorises over adjacent phases; neural learners with a joint objective over all observed phases are not measured here.

## 5 CONCLUSION AND BROADER IMPACT

Coverage fixes how much a design observes; which observations share an instance decides what an unrestricted joint model can learn. For a Gaussian joint, an isolated low-variance direction whose support no instance observes needs instances growing as the inverse square of its variance. The likelihood learner follows the operator's order and a non-Gaussian mixture keeps the coverage-matched order with larger gaps; the moment learner adds its floor tax and can reverse the order; a segment-factorised learner responds only to observed segments. Designs must be compared for a specified learner: compute the operator at a pilot's scale, co-observe low-variance supports, use existing structure, and report co-observation beside coverage in data releases and acquisition protocols.

### Reproducibility statement

Appendices A and B state the assumptions and prove Theorems 1, 2 and 3 and Corollaries 4 and 5; Appendix C checks both constructions at finite sample size; Appendices D–I give data processing, designs, learners, selected settings, controls and complete result tables; Appendix J gives the compute budget and the data licences. Regularisers are selected on a separate set of selection videos with the same grids under every policy, held-out scores are computed on the evaluation videos, and the test videos are scored once, after every selection choice is fixed. The equivalence margin of the neural comparison, $\pm 0.10$ on the held-out complete-data objective, was fixed before the runs (Appendix I). The released code, `https://anonymous.4open.science/r/beyond-coverage-code-3D5D`, reproduces every number in the paper.

### Ethics statement

EchoNet-Dynamic is public, de-identified echocardiography released by the Stanford University School of Medicine under a Research Use Agreement (`https://echonet.github.io/dynamic/index.html`) that permits personal, non-commercial research use only and prohibits commercial use, redistribution or publication of any part of the data or of its download link, reverse engineering or derivative works, any attempt to re-identify a data subject, and any clinical use; the data are provided without warranty and are not approved for diagnosis or patient care. This work uses the videos accordingly: every experiment is non-commercial research, the released code and result tables contain no video, frame or contour from the dataset, and no re-identification was attempted. The Beijing air-quality and electricity-load data are public research datasets. No patient was imaged and no new human-subject data were collected: acquisition is emulated by choosing which phases of already recorded videos enter training. The models are research artefacts for studying observation design, and the held-out cardiac scores measure agreement with contours from one fixed segmentation model. Neither is a clinical tool or a measure of individual patient predictions.

### AI use statement

Large language models assisted with code implementation, analysis scripting, literature retrieval and comparison, review of mathematical statements, figure construction and manuscript editing. Every reported number was produced by the released scripts and re-derived by separate verification scripts; responsibility for the manuscript rests with the authors.

## References

Feray Adigüzel and Michel Wedel. Split questionnaire design for massive surveys. *Journal of Marketing Research*, 45(5):608–617, 2008. doi: 10.1509/jmkr.45.5.608. URL `https://doi.org/10.1509/jmkr.45.5.608`.

Julian B. Axenfeld, Annelies G. Blom, Christian Bruch, and Christof Wolf. Split questionnaire designs for online surveys: The impact of module construction on imputation quality. *Journal of Survey Statistics and Methodology*, 10(5):1236–1262, 2022. doi: 10.1093/jssam/smab055. URL `https://doi.org/10.1093/jssam/smab055`.

Julian B. Axenfeld, Christian Bruch, and Christof Wolf. Composition of core modules and item allocation in split questionnaire designs: Impact on estimates from imputed data. *International Journal of Social Research Methodology*, 29(3):439–462, 2026. doi: 10.1080/13645579.2025.2561653. URL `https://doi.org/10.1080/13645579.2025.2561653`.

Martin Azizyan, Akshay Krishnamurthy, and Aarti Singh. Extreme compressive sampling for covariance estimation. *IEEE Transactions on Information Theory*, 64(12):7613–7635, 2018. doi: 10.1109/TIT.2018.2871077. URL `https://doi.org/10.1109/TIT.2018.2871077`.

R. A. Bailey and Peter J. Cameron. Combinatorics of optimal designs. In Sophie Huczynska, James D. Mitchell, and Colva M. Roney-Dougal (eds.), *Surveys in Combinatorics 2009*, volume 365 of *London Mathematical Society Lecture Note Series*, pp. 19–74. Cambridge University Press, 2009. doi: 10.1017/CBO9781107325975.003. URL `https://doi.org/10.1017/CBO9781107325975.003`.

Srinadh Bhojanapalli and Prateek Jain. Universal matrix completion. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1881–1889. PMLR, 2014. URL https://proceedings.mlr.press/v32/bhojanapalli14.html.

Ashish Bora, Eric Price, and Alexandros G. Dimakis. AmbientGAN: Generative models from lossy measurements. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hy7fDog0b.

Andreas M. Brandmaier, Paolo Ghisletta, and Timo von Oertzen. Optimal planned missing data design for linear latent growth curve models. *Behavior Research Methods*, 52(4):1445–1458, 2020. doi: 10.3758/s13428-019-01325-y. URL https://doi.org/10.3758/s13428-019-01325-y.

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. URL https://doi.org/10.1016/j.jmva.2016.05.002.

Yichao Cai and Javen Qinfeng Shi. An identifiability theory of masked prediction: Mode blindness and mask schedules, 2026. URL https://arxiv.org/abs/2608.01383.

Steven Cao, Percy Liang, and Gregory Valiant. One-sided matrix completion from two observations per row. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3599–3624. PMLR, 2023. URL https://proceedings.mlr.press/v202/cao23d.html.

Yifan Chen, Mark Goldstein, Mengjian Hua, Michael Samuel Albergo, Nicholas Matthew Boffi, and Eric Vanden-Eijnden. Probabilistic forecasting with stochastic interpolants and Föllmer processes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6728–6756. PMLR, 2024. URL https://proceedings.mlr.press/v235/chen24n.html.

Yilan Chen and Hongyun Liu. Planned missingness in intensive longitudinal studies: Extensions and comparisons of multiform designs. *Behavior Research Methods*, 58(8):215, 2026. doi: 10.3758/s13428-026-03035-8. URL https://doi.org/10.3758/s13428-026-03035-8.

James O. Chipperfield and David G. Steel. Efficiency of split questionnaire surveys. *Journal of Statistical Planning and Inference*, 141(5):1925–1932, 2011. doi: 10.1016/j.jspi.2010.12.003. URL https://doi.org/10.1016/j.jspi.2010.12.003.

Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. In *International Conference on Learning Representations*, pp. 42128–42149. OpenReview.net, 2025. URL https://openreview.net/forum?id=ftGnpZrW7P.

D. R. Cox and N. Reid. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737, 2004. doi: 10.1093/biomet/91.3.729. URL https://doi.org/10.1093/biomet/91.3.729.

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. Curran Associates, Inc., 2023. doi: 10.52202/075280-0015. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/012af729c5d14d279581fc8a5db975a1-Abstract-Conference.html.

Giannis Daras, Yeshwanth Cherapanamjeri, and Constantinos Daskalakis. How much is a noisy image worth? Data scaling laws for ambient diffusion. In *International Conference on Learning Representations*, pp. 28165–28193, 2025. URL https://openreview.net/forum?id=qZwtPEw2qN.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: 10.1111/j.2517-6161.1977.tb01600.x. URL https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Marcello D'Orazio, Marco Di Zio, and Mauro Scanu. *Statistical Matching: Theory and Practice*. Wiley Series in Survey Methodology. John Wiley & Sons, Chichester, 2006. ISBN 9780470023532. doi: 10.1002/0470023554. URL https://doi.org/10.1002/0470023554.

Yonina C. Eldar, Jerry Li, Cameron Musco, and Christopher Musco. Sample efficient Toeplitz covariance estimation. In *Proceedings of the 2020 ACM-SIAM Symposium on Discrete Algorithms*, pp. 378–397. SIAM, 2020. doi: 10.1137/1.9781611975994.23. URL https://doi.org/10.1137/1.9781611975994.23.

Shila Ghazanfar, Carolina Guibentif, and John C. Marioni. Stabilized mosaic single-cell data integration using unshared features. *Nature Biotechnology*, 42(2):284–292, 2024. doi: 10.1038/s41587-023-01766-z. URL https://doi.org/10.1038/s41587-023-01766-z.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01457. URL https://doi.org/10.1109/CVPR52729.2023.01457.

Josh Givens, Song Liu, and Henry 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. PMLR, 2025. URL https://proceedings.mlr.press/v267/givens25a.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

John W. Graham, Scott M. Hofer, and David P. MacKinnon. Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, 31(2):197–218, 1996. doi: 10.1207/s15327906mbr3102_3. URL https://doi.org/10.1207/s15327906mbr3102_3.

John W. Graham, Bonnie J. Taylor, Allison E. Olchowski, and Patricio E. Cumsille. Planned missing data designs in psychological research. *Psychological Methods*, 11(4):323–343, 2006. doi: 10.1037/1082-989X.11.4.323. URL https://doi.org/10.1037/1082-989X.11.4.323.

Robert Grone, Charles R. Johnson, Eduardo M. Sá, and Henry Wolkowicz. Positive definite completions of partial Hermitian matrices. *Linear Algebra and its Applications*, 58:109–124, 1984. doi: 10.1016/0024-3795(84)90207-6. URL https://doi.org/10.1016/0024-3795(84)90207-6.

Siyuan Guo, Jonas Wildberger, and Bernhard Schölkopf. Out-of-variable generalisation for discriminative models. In *International Conference on Learning Representations*, pp. 21264–21276. OpenReview.net, 2024. URL https://openreview.net/forum?id=zwMfg9PfPs.

Sharut Gupta, Shobhita Sundaram, Chenyu Wang, Stefanie Jegelka, and Phillip Isola. Better together: Leveraging unpaired multimodal data for stronger unimodal models. In *International Conference on Learning Representations*, pp. 73970–74025. OpenReview.net, 2026. URL https://openreview.net/forum?id=5OIgg5YkC3.

Ping-Han Huang and Ming-Hung Kao. New pilot-study design in functional data analysis, 2025. URL https://arxiv.org/abs/2508.00176.

Shahana Ibrahim and Xiao Fu. Recovering joint probability of discrete random variables from pairwise marginals. *IEEE Transactions on Signal Processing*, 69:4116–4131, 2021. doi: 10.1109/TSP.2021.3090960. URL https://doi.org/10.1109/TSP.2021.3090960.

Paul M. Imbriano and Trivellore E. Raghunathan. Three-form split questionnaire design for panel surveys. *Journal of Official Statistics*, 36(4):827–854, 2020. doi: 10.2478/jos-2020-0040. URL https://doi.org/10.2478/jos-2020-0040.

Mohammad Mohaiminul Islam, Thijs P. Kuipers, Sharvaree Vadgama, Coen de Vente, Afsana Khan, Clara I. Sánchez, and Erik J. Bekkers. Longitudinal flow matching for trajectory modeling. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 3871–3879. PMLR, 2026. URL https://proceedings.mlr.press/v300/islam26a.html.

Dae-Gyu Jang, Zhengyuan Zhu, and Cindy Yu. A-optimal split questionnaire designs for multivariate continuous variables, 2022. URL https://arxiv.org/abs/2203.04883.

Minoh Jeong, Zae Myung Kim, Min Namgung, Dongyeop Kang, Yao-Yi Chiang, and Alfred Hero. Anchors Aweigh! Sail for optimal unified multi-modal representations, 2024. URL https://arxiv.org/abs/2410.02086.

Hao Ji and Hans-Georg Müller. Optimal designs for longitudinal and functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):859–876, 2017. doi: 10.1111/rssb.12192. URL https://doi.org/10.1111/rssb.12192.

Kamil Jurczak and Angelika Rohde. Spectral analysis of high-dimensional sample covariance matrices with missing observations. *Bernoulli*, 23(4A):2466–2532, 2017. doi: 10.3150/16-BEJ815. URL https://doi.org/10.3150/16-BEJ815.

Nikos Kargas, Nicholas D. Sidiropoulos, and Xiao Fu. Tensors, learning, and "Kolmogorov extension" for finite-alphabet random vectors. *IEEE Transactions on Signal Processing*, 66(18): 4854–4868, 2018. doi: 10.1109/TSP.2018.2862383. URL https://doi.org/10.1109/TSP.2018.2862383.

Fairoz Nower Khan, Nabuat Zaman Nahim, and Peizhong Ju. Flow matching with missing data, 2026. URL https://arxiv.org/abs/2607.28698.

J. Kiefer. Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):272–304, 1959. doi: 10.1111/j.2517-6161.1959.tb00338.x. URL https://doi.org/10.1111/j.2517-6161.1959.tb00338.x.

Sayeri Lala and Niraj K. Jha. METRIK: Measurement-efficient randomized controlled trials using transformers with input masking, 2024. URL https://arxiv.org/abs/2406.16351.

Steven Cheng-Xian Li, Bo Jiang, and Benjamin M. Marlin. MisGAN: Learning from incomplete data with generative adversarial networks. In *International Conference on Learning Representations*. OpenReview.net, 2019. URL https://openreview.net/forum?id=S1lDV3RcKm.

Yan Li, Yunlong Deng, Yuewen Sun, Gongxu Luo, Kun Zhang, and Guangyi Chen. Multimodal LLMs under pairwise modalities, 2026. URL https://arxiv.org/abs/2605.21059.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, third edition, 2019. doi: 10.1002/9781119482260. URL https://doi.org/10.1002/9781119482260.

Bingbin Liu, Daniel J. Hsu, Pradeep Ravikumar, and Andrej Risteski. Masked prediction: A parameter identifiability view. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21241–21254. Curran Associates, Inc., 2022. doi: 10.52202/068431-1544. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/85dd09d356ca561169b2c03e43cf305e-Abstract-Conference.html.

Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012. doi: 10.1214/12-AOS1018. URL https://doi.org/10.1214/12-AOS1018.

Karim Lounici. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014. doi: 10.3150/12-BEJ487. URL https://doi.org/10.3150/12-BEJ487.

Haoye Lu, Qifan Wu, and Yaoliang Yu. Stochastic forward–backward deconvolution: Training diffusion models with finite noisy datasets. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40741–40768. PMLR, 2025. URL https://proceedings.mlr.press/v267/lu25n.html.

Kanti V. Mardia, John T. Kent, Gareth Hughes, and Charles C. Taylor. Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika*, 96(4):975–982, 2009. doi: 10.1093/biomet/asp056. URL https://doi.org/10.1093/biomet/asp056.

Leopold Mareis. Optimizing experimental design for causal effect estimation with partial measurements. In Xiao-Hua Zhou and Jinzhu Jia (eds.), *Causal Inference: 6th Pacific Causal Inference Conference, PCIC 2024, Shanghai, China, July 5–6, 2024, Revised Selected Papers*, volume 2200 of *Communications in Computer and Information Science*, pp. 74–85. Springer Nature Singapore, Singapore, 2025. doi: 10.1007/978-981-97-7812-6_7. URL https://doi.org/10.1007/978-981-97-7812-6_7.

Xiao-Li Meng and Donald B. Rubin. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86(416):899–909, 1991. doi: 10.1080/01621459.1991.10475130. URL https://doi.org/10.1080/01621459.1991.10475130.

Robin Mitra, Sarah F. McGough, Tapabrata Chakraborti, Chris Holmes, Ryan Copping, Niels Hagenbuch, Stefanie Biedermann, Jack Noonan, Brieuc Lehmann, Aditi Shenvi, Xuan Vinh Doan, David Leslie, Ginestra Bianconi, Ruben Sanchez-Garcia, Alisha Davies, Maxine Mackintosh, Eleni-Rosalina Andrinopoulou, Anahid Basiri, Chris Harbron, and Ben D. MacArthur. Learning from data with structured missingness. *Nature Machine Intelligence*, 5(1):13–23, 2023. doi: 10.1038/s42256-022-00596-z. URL https://doi.org/10.1038/s42256-022-00596-z.

Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 4348–4380. PMLR, 2023. URL https://proceedings.mlr.press/v206/nakada23a.html.

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020. doi: 10.1038/s41586-020-2145-8. URL https://doi.org/10.1038/s41586-020-2145-8.

Yidong Ouyang, Liyan Xie, Chongxuan Li, and Guang Cheng. MissDiff: Training diffusion models on tabular data with missing values. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023. URL https://arxiv.org/abs/2307.00467.

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. URL https://doi.org/10.1214/21-EJS1892.

Eduardo Pavez and Antonio Ortega. Active covariance estimation by random sub-sampling of variables. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4034–4038. IEEE, 2018. doi: 10.1109/ICASSP.2018.8462134. URL https://doi.org/10.1109/ICASSP.2018.8462134.

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. URL https://doi.org/10.1109/TIT.2020.3039118.

Friedrich Pukelsheim. *Optimal Design of Experiments*, volume 50 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2006. doi: 10.1137/1.9780898719109. URL https://doi.org/10.1137/1.9780898719109.

Trivellore E. Raghunathan and James E. Grizzle. A split questionnaire survey design. *Journal of the American Statistical Association*, 90(429):54–63, 1995. doi: 10.1080/01621459.1995.10476488. URL https://doi.org/10.1080/01621459.1995.10476488.

Susanne Rässler. *Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, volume 168 of *Lecture Notes in Statistics*. Springer, New York, 2002. doi: 10.1007/978-1-4613-0053-3. URL https://doi.org/10.1007/978-1-4613-0053-3.

Mijke Rhemtulla, Victoria Savalei, and Todd D. Little. On the asymptotic relative efficiency of planned missingness designs. *Psychometrika*, 81(1):60–89, 2016. doi: 10.1007/s11336-014-9422-0. URL https://doi.org/10.1007/s11336-014-9422-0.

Tillmann Rheude, Roland Eils, and Benjamin Wild. Cohort-based active modality acquisition, 2025. URL https://arxiv.org/abs/2505.16791.

François Rozet, Gérôme Andry, François Lanusse, and Gilles Louppe. Learning diffusion priors from observations by expectation maximization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87647–87682. Curran Associates, Inc., 2024. doi: 10.52202/079017-2783. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9f94298bac4668db4dc77ddb0a244301-Abstract-Conference.html.

Adriel Saporta, Aahlad Puli, Mark Goldstein, and Rajesh Ranganath. Contrasting with Symile: Simple model-agnostic representation learning for unlimited modalities. In *Advances in Neural Information Processing Systems*, volume 37, pp. 56919–56957. Curran Associates, Inc., 2024. doi: 10.52202/079017-1814. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6828259348d99d5e8994028bfdf15d09-Abstract-Conference.html.

David G. Steel and James Chipperfield. The efficiency of missing at random planned missing designs. *Mathematics*, 14(2):355, 2026. doi: 10.3390/math14020355. URL https://doi.org/10.3390/math14020355.

Joseph Steneman and Giuseppe Vinci. Covariance matrix completion via auxiliary information. Joint Statistical Meetings 2024, session New Methods for Networks, Graphical Models and Covariances, American Statistical Association, Portland, Oregon, August 7, 2024. URL https://ww2.amstat.org/meetings/jsm/2024/onlineprogram/abstract.cfm?sid=2878&tid=2878-p4.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34465–34492. Curran Associates, Inc., 2023. doi: 10.52202/075280-1496. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6c8985579293e0209bdaa4f21bb1d237-Abstract-Conference.html.

Julián Tachella, Dongdong Chen, and Mike Davies. Sensing theorems for unsupervised learning in linear inverse problems. *Journal of Machine Learning Research*, 24(39):39:1–39:45, 2023. URL https://jmlr.org/papers/v24/22-0315.html.

Ran Tao, Donglin Zeng, and Dan-Yu Lin. Optimal designs of two-phase studies. *Journal of the American Statistical Association*, 115(532):1946–1959, 2020. doi: 10.1080/01621459.2019.1671200. URL https://doi.org/10.1080/01621459.2019.1671200.

Subash Timilsina, Sagar Shrestha, and Xiao Fu. Identifiable shared component analysis of unpaired multimodal mixtures. In *Advances in Neural Information Processing Systems*, volume 37, pp. 127799–127833. Curran Associates, Inc., 2024. doi: 10.52202/079017-4058. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e6c967a0e4168df9614a81ed4195ba78-Abstract-Conference.html.

Artur Trindade. ElectricityLoadDiagrams20112014. UCI Machine Learning Repository, 2015. URL https://doi.org/10.24432/C58C86.

Michele Tumminello, Fabrizio Lillo, and Rosario N. Mantegna. Kullback–Leibler distance as a measure of the information filtered from multivariate data. *Physical Review E*, 76(3):031123, 2007. doi: 10.1103/PhysRevE.76.031123. URL https://doi.org/10.1103/PhysRevE.76.031123.

Srini Turaga, Lars Buesing, Adam M. Packer, Henry Dalgleish, Noah Pettit, Michael Hausser, and Jakob H. Macke. Inferring neural population dynamics from multiple partial recordings of the same neural circuit. In *Advances in Neural Information Processing Systems*, volume 26, pp. 539–547. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/hash/01386bd6d8e091c2ab4c7c7de644d37b-Abstract.html.

Giuseppe Vinci, Gautam Dasarathy, and Genevera I. Allen. Graph quilting: Graphical model selection from partially observed covariances, 2019. URL https://arxiv.org/abs/1912.05573.

N. N. Vorob'ev. Consistent families of measures and their extensions. *Theory of Probability & Its Applications*, 7(2):147–163, 1962. doi: 10.1137/1107014. URL https://doi.org/10.1137/1107014.

Kartik G. Waghmare and Victor M. Panaretos. The completion of covariance kernels. *The Annals of Statistics*, 50(6):3281–3306, 2022. doi: 10.1214/22-AOS2228. URL https://doi.org/10.1214/22-AOS2228.

Zehan Wang, Yang Zhao, Xize Cheng, Haifeng Huang, Jiageng Liu, Aoxiong Yin, Li Tang, Linjun Li, Yongqi Wang, Ziang Zhang, and Zhou Zhao. Connecting multi-modal contrastive representations. In *Advances in Neural Information Processing Systems*, volume 36, pp. 22099–22114. Curran Associates, Inc., 2023. doi: 10.52202/075280-0970. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/46362971bfc3a97e6a271f2eb90fba17-Abstract-Conference.html.

Jincheng Xie, Xingchen Xiao, Runheng Liu, Zhongyi Huang, Yu Zheng, and Heyan Huang. EmergentBridge: Improving zero-shot cross-modal transfer in unified multimodal embedding models. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 5708–5719. ACM, 2026. doi: 10.1145/3770855.3818172. URL https://doi.org/10.1145/3770855.3818172.

Tengyang Xie, Dylan J. Foster, Yu Bai, Nan Jiang, and Sham M. Kakade. The role of coverage in online reinforcement learning. In *International Conference on Learning Representations*. OpenReview.net, 2023. URL https://openreview.net/forum?id=LQIjzPdDt3q.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *International Conference on Learning Representations*, pp. 34817–34848. OpenReview.net, 2024. URL https://openreview.net/forum?id=OGtnhKQJms.

Si-Yu Yi, Yong-Dao Zhou, and Wei Zheng. Optimal designs for mean–covariance models with missing observations. *Journal of Statistical Planning and Inference*, 219:85–97, 2022. doi: 10.1016/j.jspi.2021.12.004. URL https://doi.org/10.1016/j.jspi.2021.12.004.

Siyu Yi, Yongdao Zhou, and Jianxin Pan. D-optimal designs of mean-covariance models for longitudinal data. *Biometrical Journal*, 63(5):1072–1085, 2021. doi: 10.1002/bimj.202000129. URL https://doi.org/10.1002/bimj.202000129.

Jiancheng Zhang and Yinglun Zhu. Towards multimodal active learning: Efficient learning with limited paired data. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=xMLajoct78.

Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in Beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205):20170457, 2017. doi: 10.1098/rspa.2017.0457. URL https://doi.org/10.1098/rspa.2017.0457.

Zihan Zhou, Chenguang Wang, Hongyi Ye, Yongtao Guan, and Tianshu Yu. Incomplete data, complete dynamics: A diffusion approach. In *International Conference on Learning Representations*, pp. 10918–10952. OpenReview.net, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/123a18dfd821c8b440f42a00a27648d6-Abstract-Conference.html.

Ziwei Zhu, Tengyao Wang, and Richard J. Samworth. High-dimensional principal component analysis with heterogeneous missingness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):2000–2031, 2022. doi: 10.1111/rssb.12550. URL https://doi.org/10.1111/rssb.12550.

# A Proofs for the information operator

This appendix proves Theorem 1, Corollary 4, Theorem 2 and the Remark that follows it, extends the $\Omega(\varepsilon^{-2})$ bound to a shrinking subspace (Lemma A.1), and computes Construction 1 and a constant-observation-size version of it exactly.

## A.1 Gaussian information and the geometry of an observation

The space $\mathrm{Sym}_D$ carries the Frobenius inner product and an orthonormal basis whose off-diagonal elements are $(e_ie_j^\top + e_je_i^\top)/\sqrt{2}$; unscaled upper-triangular entries would change the coordinate matrix of information. For one observation, put $C = H\Sigma H^\top = GG^\top$ and $C'_\Gamma = G\Gamma G^\top$. The Gaussian covariance score has bilinear information

$$\mathcal{I}_H(\Gamma,\Delta) = \tfrac{1}{2}\operatorname{tr}(C^{-1}G\Gamma G^\top C^{-1}G\Delta G^\top) \tag{7}$$

$$= \tfrac{1}{2}\operatorname{tr}(P\Gamma P\Delta), \qquad P = G^\top(GG^\top)^{-1}G. \tag{8}$$

Since $P = P^\top = P^2$, $\operatorname{tr}(P\Gamma P\Gamma) = \|P\Gamma P\|_F^2$. Summing independent observations proves Theorem 1. The mean score is linear in the centred observation and the covariance score is centred quadratic. Their product has zero expectation because centred Gaussian third moments vanish. The covariance block therefore remains unchanged when the regression mean is unknown; mean estimation contributes its own, separate predictive risk.

## A.2 Trace and complete-co-observation optimality

A projector of rank $r$ takes the form $P = \operatorname{diag}(I_r, 0)$ in a suitable orthonormal coordinate system. The map $\Gamma \mapsto P\Gamma P$ is the identity on the $r(r+1)/2$-dimensional symmetric leading block and zero on its orthogonal complement. Its trace is therefore $r(r+1)/2$. For equal-size blocks, $r = d|S|$, and

$$\operatorname{tr}\mathcal{M} = \sum_S n_S \frac{d|S|(d|S|+1)}{2} \tag{9}$$

$$= \frac{d(d+1)}{2}\sum_j n_j + d^2\sum_{j<k} n_{jk}. \tag{10}$$

The argument uses the ranks of the observations and nothing else, so $H_S$ need not be coordinate-aligned. It covers instance-specific operators $H_i$, with a sum over instances instead of identical subsets.

For any nonempty subset, $|S| \le m$ and $d|S|(d|S|+1)/2 \le |S|d(D+1)/2$. Thus $\operatorname{tr}\mathcal{M} \le N_{\mathrm{obs}}p/m$. If $\mu_1,\dots,\mu_p > 0$ are the eigenvalues of $\mathcal{M}$, Cauchy–Schwarz gives

$$\left(\sum_k \mu_k\right)\left(\sum_k \mu_k^{-1}\right) \ge p^2. \tag{11}$$

The efficient leading covariance KL is consequently at least $mp/(2N_{\mathrm{obs}})$. A feasible complete design gives $\mathcal{M} = (N_{\mathrm{obs}}/m)I$ and attains that value. Equality in the harmonic inequality requires equal information eigenvalues; equality in the budget trace inequality requires complete nonempty observations. Empty observations cost nothing and contribute zero information. When $\mathcal{M}$ is singular, the unrestricted local covariance risk is infinite under this approximation.

### A.3 Directional information and the isolated-eigenvalue bound

For a unit vector $v$, $e_v = vv^\top$ is a unit element of $\mathrm{Sym}_D$ and $\langle e_v, P_S e_v P_S\rangle = (v^\top P_S v)^2 = \|P_S v\|^4$, so $\langle e_v, \mathcal{M} e_v\rangle = N_{\mathrm{eff}}(v)$. If $\Sigma v = \lambda_v v$, the variance $\lambda_v = v^\top \Sigma v$ along the fixed direction $v$ moves in the family $\Sigma + t\, vv^\top$, whose whitened derivative is $\Sigma^{-1/2} vv^\top \Sigma^{-1/2} = e_v/\lambda_v$. By Equation (2) the information about $\lambda_v$ is $N_{\mathrm{eff}}(v)/(2\lambda_v^2)$, and its reciprocal $2/N_{\mathrm{eff}}(v)$ bounds the relative variance of regular unbiased estimators of $\lambda_v$ with the other parameters known.

**Directional information with every parameter free.** With every covariance parameter free and $\mathcal{M}$ invertible, the information is $\frac{1}{2}\mathcal{M}$ in whitened coordinates (Equation (2)). For an eigen-direction $v$, $\Sigma^{1/2} vv^\top \Sigma^{1/2} = \lambda_v e_v$, so the relative change of the variance $v^\top \Sigma v$ along the fixed direction $v$ under $\Sigma \mapsto \Sigma^{1/2}(I+\Gamma)\Sigma^{1/2}$ is the linear functional $\Gamma \mapsto \langle e_v, \Gamma\rangle$, and the Cramér–Rao bound on its relative variance is $2\langle e_v, \mathcal{M}^{-1} e_v\rangle$; the regression mean, Fisher-orthogonal to the covariance, leaves it unchanged. Cauchy–Schwarz gives $1 = \langle \mathcal{M}^{1/2} e_v, \mathcal{M}^{-1/2} e_v\rangle^2 \le \langle e_v, \mathcal{M} e_v\rangle \langle e_v, \mathcal{M}^{-1} e_v\rangle$, hence $2\langle e_v, \mathcal{M}^{-1} e_v\rangle \ge 2/N_{\mathrm{eff}}(v)$, with equality exactly when $e_v$ is an information direction. For the upper bound $2/n_c$ of the remark after Theorem 2, a covering instance has $v = H_S^\top w$ for some $w$, so $v = \lambda_v^{-1/2}\Sigma^{1/2} v = \lambda_v^{-1/2} G_S^\top w$ lies in the row space of $G_S$ and $P_S v = v$; hence $\mathcal{M} \succeq \mathcal{M}_c$, where $\mathcal{M}_c$ sums $\Gamma \mapsto P_S \Gamma P_S$ over the $n_c$ covering instances and $\mathcal{M}_c e_v = n_c e_v$. Every $y = te_v + y_\perp$ with $y_\perp \perp e_v$ has $\langle y, \mathcal{M} y\rangle \ge \langle y, \mathcal{M}_c y\rangle \ge n_c t^2$, so the variational form $\langle e_v, \mathcal{M}^{-1} e_v\rangle = \sup_y \{2\langle e_v, y\rangle - \langle y, \mathcal{M} y\rangle\}$ is at most $\sup_t\{2t - n_c t^2\} = 1/n_c$. In the constructed domain (Appendix G) the identity $N_{\mathrm{eff}}(v) = \langle e_v, \mathcal{M} e_v\rangle$ holds numerically to relative $10^{-12}$.

Now let $\Sigma_\varepsilon = \varepsilon vv^\top + \Sigma_\perp$, $\Sigma_\perp v = 0$, with positive definiteness on $v^\perp$. If $v \in \mathrm{row}(H)$, an observed linear contrast equals $v^\top a$. Also $v \in \mathrm{row}(H\Sigma_\varepsilon^{1/2})$, hence $Pv = v$ and the observation supplies $1/(2\varepsilon^2)$ information. If $v \notin \mathrm{row}(H)$, $C_0 = H\Sigma_\perp H^\top \succ 0$: a nonzero null vector would map through $H^\top$ to a nonzero vector in $\mathrm{span}(v)$, contradicting the row-space condition. Writing $b = Hv$, Sherman–Morrison gives

$$b^\top (C_0 + \varepsilon bb^\top)^{-1} b = \frac{c}{1+\varepsilon c}, \quad c = b^\top C_0^{-1} b, \tag{12}$$

and $v^\top P v = \|Pv\|^2 = \varepsilon c/(1+\varepsilon c)$. Since the information about $\varepsilon$ is $\|Pv\|^4/(2\varepsilon^2)$, this gives both identities of Theorem 2(i):

$$\|P_S v\|^2 = \frac{\varepsilon c}{1+\varepsilon c}, \qquad \mathcal{I}_S(\varepsilon,\varepsilon) = \frac{c^2}{2(1+\varepsilon c)^2} = O(1). \tag{13}$$

The rate lower bound is a two-point test between $\varepsilon$ and $2\varepsilon$ in the submodel where only $\varepsilon$ is unknown, so it binds every estimator of a larger model. With the mean known, a non-covering subset $S$ observes $N(0, C_\varepsilon)$ with $C_\varepsilon = C_0 + \varepsilon bb^\top$ and $C_{2\varepsilon} = C_\varepsilon + \varepsilon bb^\top$. The matrix $C_\varepsilon^{-1/2} C_{2\varepsilon} C_\varepsilon^{-1/2}$ has one eigenvalue $1+x$, with $x = \varepsilon b^\top C_\varepsilon^{-1} b = \varepsilon c/(1+\varepsilon c) \le \varepsilon c$, and all others equal to one, so

$$\mathrm{KL}\big(N(0, C_\varepsilon) \,\|\, N(0, C_{2\varepsilon})\big) = g(x) = \tfrac{1}{2}\Big[\log(1+x) - \frac{x}{1+x}\Big] \le \frac{x^2}{4} \le \frac{\varepsilon^2 c^2}{4}, \tag{14}$$

since $g(0) = 0$ and $g'(x) = x/(2(1+x)^2) \le x/2$ for $x \ge 0$. Independent instances add KL, so when no subset covers $v$ the laws $P_\varepsilon$ and $P_{2\varepsilon}$ of all observations of $r$ repetitions satisfy $\mathrm{KL}(P_\varepsilon \| P_{2\varepsilon}) \le r\varepsilon^2 \sum_S n_S c_S^2/4$. If an estimator $\hat{\varepsilon}$ has relative error below $1/3$ with probability at least $3/4$ at $\varepsilon$ and at $2\varepsilon$, the event $E = \{\hat{\varepsilon} \in (2\varepsilon/3, 4\varepsilon/3)\}$ has $P_\varepsilon(E) \ge 3/4$ and, being disjoint from $(4\varepsilon/3, 8\varepsilon/3)$, $P_{2\varepsilon}(E) \le 1/4$, so the total variation between the two laws is at least $1/2$. Pinsker's inequality, total variation at most $(\mathrm{KL}/2)^{1/2}$, gives $\mathrm{KL}(P_\varepsilon \| P_{2\varepsilon}) \ge 1/2$, hence $r \ge 2/(\varepsilon^2 \sum_S n_S c_S^2)$; with $\Sigma_\perp$ and the design fixed the $c_S$ are fixed, and $r = \Omega(\varepsilon^{-2})$. In Construction 1, $b$ has $k = |S \cap \{1,2,3\}|$ entries $1/\sqrt{3}$ and $C_0 = I - bb^\top$, so $c_S = k/(3-k)$; under $B$ three subsets have $k = 2$ and three have $k = 1$, and $\sum_S n_S c_S^2 = 3 \cdot 4 + 3 \cdot \frac{1}{4} = 51/4$ per repetition.

For the upper bound, with $v$ and the mean known, a covering observation supplies the centred contrast $v^\top(a-\beta^\top x)\sim N(0,\varepsilon)$. The mean square of these contrasts over the $rn_c$ covering instances ($n_c$ per repetition) is unbiased with variance $2\varepsilon^2/(rn_c)$, hence relative mean squared error $2/(rn_c)$ at every $\varepsilon$. Other directions have their own information, which can vanish with $\varepsilon$ under a covering design: in Construction 1, design $A$ gives the whitened cross-direction $(ve_4^\top+e_4v^\top)/\sqrt{2}$ information $3\varepsilon/[2(2+\varepsilon)]\to 0$.

**Lemma A.1** (A shrinking subspace)**.** Let $\Sigma_\varepsilon=\varepsilon UU^\top+\Sigma_\perp$ with $U\in\mathbb{R}^{D\times k}$, $U^\top U=I_k$, $\Sigma_\perp U=0$ and $\Sigma_\perp$ positive definite on the orthogonal complement of $\mathrm{range}(U)$. Suppose $\mathrm{row}(H_S)\cap\mathrm{range}(U)=\{0\}$ for every observed $S$, and put $W_S=H_SU$, $C_{0,S}=H_S\Sigma_\perp H_S^\top$ and $L_S=W_S^\top C_{0,S}^{-1}W_S$. Then

$$U^\top P_S U=\varepsilon L_S(I+\varepsilon L_S)^{-1}, \tag{15}$$

so every unit $u\in\mathrm{range}(U)$ has $\|P_Su\|^2\le\varepsilon\|L_S\|=O(\varepsilon)$, and every estimator of $\varepsilon$ whose relative error is below $1/3$ with probability at least $3/4$ at $\varepsilon$ and $2\varepsilon$ needs $r\ge 2/(\varepsilon^2\sum_S n_S\|L_S\|_F^2)$ repetitions of the design, which is $\Omega(\varepsilon^{-2})$ as $\varepsilon\to 0$ with $\Sigma_\perp$ and the design fixed.

*Proof.* Since $\Sigma_\perp U=0$ and $\Sigma_\perp\succeq 0$, $\Sigma_\perp^{1/2}U=0$, so $\Sigma_\varepsilon^{1/2}=\varepsilon^{1/2}UU^\top+\Sigma_\perp^{1/2}$ and $G_SU=\varepsilon^{1/2}W_S$. If $C_{0,S}y=0$, then $\Sigma_\perp^{1/2}H_S^\top y=0$, so $H_S^\top y\in\mathrm{row}(H_S)\cap\mathrm{range}(U)=\{0\}$ and $y=0$ by full row rank; hence $C_{0,S}\succ 0$. With $C_{\varepsilon,S}=H_S\Sigma_\varepsilon H_S^\top=C_{0,S}+\varepsilon W_SW_S^\top$, the identity $C_{\varepsilon,S}C_{0,S}^{-1}W_S=W_S(I+\varepsilon L_S)$ gives $C_{\varepsilon,S}^{-1}W_S=C_{0,S}^{-1}W_S(I+\varepsilon L_S)^{-1}$, hence $U^\top P_SU=(G_SU)^\top C_{\varepsilon,S}^{-1}G_SU=\varepsilon W_S^\top C_{\varepsilon,S}^{-1}W_S=\varepsilon L_S(I+\varepsilon L_S)^{-1}$. Its eigenvalues $\varepsilon\ell_i/(1+\varepsilon\ell_i)$, with $\ell_i$ those of $L_S$, are at most $\varepsilon\ell_i$, so a unit $u=Uw$ has $\|P_Su\|^2=w^\top U^\top P_SUw\le\varepsilon\|L_S\|$. For the rate, $C_{2\varepsilon,S}=C_{\varepsilon,S}+\varepsilon W_SW_S^\top$, so the nonzero eigenvalues of $C_{\varepsilon,S}^{-1/2}C_{2\varepsilon,S}C_{\varepsilon,S}^{-1/2}-I$ are those of $\varepsilon W_S^\top C_{\varepsilon,S}^{-1}W_S$, the positive $x_i=\varepsilon\ell_i/(1+\varepsilon\ell_i)\le\varepsilon\ell_i$. One observation's KL is $\sum_i g(x_i)\le\sum_i x_i^2/4\le\varepsilon^2\|L_S\|_F^2/4$, and the two-point test and Pinsker's inequality above give the bound on $r$. □

The condition concerns the whole subspace, not each direction in it: with $\Sigma=\varepsilon I_2$ and $H=[1,1]$, $e_1\notin\mathrm{row}(H)$, yet $\mathrm{row}(H)$ meets $\mathrm{range}(U)=\mathbb{R}^2$ and $\|Pe_1\|^2=1/2$ at every $\varepsilon$, against the $O(\varepsilon)$ of Equation (13).

### A.4 Exact constructions

For Construction 1, with the designs $A$ and $B$ of Equation (1), the observed covariance of a subset $S$ depends on $k=|S\cap\{1,2,3\}|$: it has one $\varepsilon$-dependent eigenvalue $(3-k+k\varepsilon)/3$; the other eigenvalues are constant. Its information about $\varepsilon$ is

$$\frac{k^2}{2(3-k+k\varepsilon)^2}. \tag{16}$$

Counting $k$ in the two pattern lists gives each design's information about $\varepsilon$ as a function of $\varepsilon$. Coordinate exchange sends one design to the other. Expanding the observed KL from $\varepsilon$ to $2\varepsilon$ under $B$ gives $(51/16)\varepsilon^2+o(\varepsilon^2)$, the two-point bound $\varepsilon^2\sum_S n_Sc_S^2/4$ to leading order; the finite-$\varepsilon$ coefficient is $3.17$ at $0.001$. Under $A$, the covering contrast contributes exactly $\frac{1}{2}\log 2-\frac{1}{4}$, with further nonnegative terms of order $\varepsilon^2$. A covariance perturbation can therefore have vanishing Frobenius norm and nonvanishing relative predictive KL.

The constant-observation-size extension uses six variables and $v=(1,1,1,0,0,0)/\sqrt{3}$:

$$A_3=\{123,145,246,356,156,245,346\}, \tag{17}$$
$$B_3=\{124,135,236,456,156,245,346\}. \tag{18}$$

The first four triples in each list have the same marginal counts and pairwise counts; the remaining triples are identical. Every pair is covered, every observation has size three, and the inverse-size-weighted concurrence $\sum_{S\supseteq\{j,k\}}n_S/|S|$ therefore also agrees. Counting overlap with $\{1,2,3\}$ yields the same information about $\varepsilon$ as Construction 1. The separation follows from these analytical identities, which the released code checks numerically for both constructions.

# B PROOFS FOR THE LEARNER MECHANISMS

This appendix proves Theorem 3 and Corollary 5, and specifies how the implemented learners' mean fits and the regression mean enter the measured risk.

## B.1 COMPLETE-DATA EQUALITY

For complete observations collected into $Y$ and full-rank regression matrix $X$, unregularised Gaussian likelihood gives

$$\widehat{\beta} = (X^\top X)^{-1} X^\top Y, \qquad \widehat{\Sigma} = \frac{1}{n}(Y - X\widehat{\beta})^\top (Y - X\widehat{\beta}). \tag{19}$$

These are the ordinary regression moment estimates with normalisation by $n$. The covariance is the interior positive definite maximiser when the residual matrix has full column rank, and inactive ridge and floor leave both estimates unchanged. With a fixed known mean, the same argument applies directly to the centred scatter. Complete-data relationships between full and suitable composite likelihoods are treated by Mardia et al. (2009).

Under windows the implemented learners differ in two specifications of their mean fits. The moment learner combines the covariate Gram matrix of all videos with inverse-probability-weighted outcomes, whereas the likelihood learner fits the completely observed videos alone, with the Gram matrix of those videos and a ridge penalty multiplied by the ratio of all videos to completely observed videos, so that one ridge value shrinks the two fits by different amounts. Together these specifications separate the implemented learners by $0.0025$ nats per coefficient in the moment learner's favour ($95\%$ interval $[0.0021, 0.0029]$ in the two-policy uncertainty study of Appendix D).

## B.2 EFFICIENT RISK AND THE REGRESSION MEAN

A fixed design is repeated $r$ times, with $\mathcal{M} = r\mathcal{M}_1$. The argument assumes fixed dimension, correct Gaussian specification, an identifiable design ($\mathcal{M}_1 \succ 0$), a regular consistent interior maximum-likelihood estimator, and penalties negligible at the statistical scale. Standard likelihood asymptotic normality gives covariance $(r\mathcal{I}_1)^{-1}$ in whitened symmetric coordinates. For a local whitened covariance deviation $\Gamma$,

$$\mathrm{KL} = \tfrac{1}{2}\{\mathrm{tr}((I+\Gamma)^{-1}) - D + \log\det(I+\Gamma)\} \tag{20}$$

$$= \tfrac{1}{4}\,\|\Gamma\|_F^2 + o(\|\Gamma\|_F^2). \tag{21}$$

Under the uniform integrability that turns asymptotic normality into expected risk, the efficient risk and the covariance estimation error expand locally as

$$\mathbb{E}\,\mathrm{KL} = \frac{1}{2r}\,\mathrm{tr}(\mathcal{M}_1^{-1}) + o(r^{-1}), \quad \mathbb{E}\,\|\Gamma\|_F^2 = \frac{2}{r}\,\mathrm{tr}(\mathcal{M}_1^{-1}) + o(r^{-1}), \tag{22}$$

so $K_{\mathrm{cov}} = \frac{1}{2}\,\mathrm{tr}(\mathcal{M}^{-1})$ and $rK_{\mathrm{cov}} = \frac{1}{2}\,\mathrm{tr}(\mathcal{M}_1^{-1})$.

For the mean, the relevant covariate object is the second moment, not the centred covariance, since the regression includes an intercept. With fixed $x_i$, observation projector $P_i$ and an evaluation covariate distribution $R$, define

$$J = \sum_i (x_i x_i^\top) \otimes P_i, \qquad Q_R = \mathbb{E}_R[xx^\top]. \tag{23}$$

The leading mean risk under $R$ is $\frac{1}{2}\,\mathrm{tr}[(Q_R \otimes I_D)J^{-1}]$. When assignment is independent of the covariates, the observation geometry is common within each pattern, and training and evaluation share a nonsingular covariate second moment, a $q$-column regression mean adds $\frac{q}{2}\,\mathrm{tr}(\mathcal{M}_{\mathrm{mean}}^{-1})$ with $\mathcal{M}_{\mathrm{mean}} = \sum_S n_S P_S$. Under these conditions the large-sample factorisation $J \simeq Q \otimes \mathcal{M}_{\mathrm{mean}}$ gives that term. In the known-truth simulation (Appendix E) the mean contributes $0.0070$ per coefficient under windows; the factorised value is $0.0116$, the difference arising from the video-specific phase times (Appendix D) and finite-sample covariate imbalance that the factorisation omits. The covariance term is unchanged.

### B.3 AVAILABLE-CASE NOISE AND INDEFINITENESS

For fixed observation patterns and known zero mean, the available-case estimate along $v$ is a quadratic form, $\widehat{q} = v^\top \widehat{\Sigma} v = \sum_t a_t^\top V_t a_t$, with $a_t$ the joint vector of instance $t$. For a fixed design the inverse-probability weights are the inverse counts $1/n_{ij}$, which sum to one over the instances observing each entry, so $\mathbb{E}\widehat{q} = v^\top \Sigma v$. Isserlis' identity gives

$$\mathrm{Cov}(a_i a_j, a_k a_l) = \Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk}, \tag{24}$$

and independence between instances implies

$$\mathrm{Var}(\widehat{q}) = 2\sum_t \mathrm{tr}(\Sigma V_t \Sigma V_t) = 2\sum_t \left\| \Sigma^{1/2} V_t \Sigma^{1/2} \right\|_F^2. \tag{25}$$

Write $\Sigma = \varepsilon v v^\top + \Sigma_\perp$ as in Section 3, so that $v^\top \Sigma v = \varepsilon$. The singular-limit variance, the limit of $\mathrm{Var}(\widehat{q})$ as $\Sigma \to \Sigma_\perp$, i.e. as $v^\top \Sigma v \to 0$, is $2\sum_t \left\| \Sigma_\perp^{1/2} V_t \Sigma_\perp^{1/2} \right\|_F^2$; it is positive precisely when at least one $\Sigma_\perp^{1/2} V_t \Sigma_\perp^{1/2}$ is nonzero. This is a condition on the estimator's pooling weights, distinct from whether the arrangement contains a covering subset. It holds whenever some instance $t$ observes a support index $j$ and misses another support index $i$, in particular whenever no instance contains the support of $v$. Because $\Sigma_\perp$ is positive definite on $v^\perp$, $\Sigma_\perp^{1/2} V_t \Sigma_\perp^{1/2} = 0$ would give $V_t = vu^\top + uv^\top$ for some $u$; row $i$ of $V_t$ is zero, so $u$ is a multiple of $v$ and $V_t = c\, vv^\top$ with $c\, v_i^2 = (V_t)_{ii} = 0$, which contradicts $(V_t)_{jj} = v_j^2/n_{jj} > 0$. Substituting the counts of Construction 1 (Equation (1)) for the design repeated $r$ times gives

$$\mathrm{Var}_A(\widehat{q}) = \frac{298\varepsilon^2 + 16\varepsilon + 64}{243\, r}, \qquad \mathrm{Var}_B(\widehat{q}) = \frac{130\varepsilon^2 + 136\varepsilon + 112}{243\, r}, \tag{26}$$

which agree at $\varepsilon = 1$, where $\Sigma = I$. As $\varepsilon \to 0$ they tend to $64/(243r)$ and $112/(243r)$: in $A$, pooling each covariance entry over all available instances discards the coherent cancellation present in its triple.

For the design repeated $r$ times, $\sqrt{r}(\widehat{q} - \varepsilon)$ converges to a centred normal law with nonzero limiting variance when some $\Sigma_\perp^{1/2} V_t \Sigma_\perp^{1/2}$ is nonzero. Along $\varepsilon\sqrt{r} \to 0$, the probability that $\widehat{q} < 0$ converges to one half. A negative quadratic form implies an indefinite covariance. For comparison, $v = (1, -1, 0, 0)/\sqrt{2}$ with observations $\{1, 2, 3\}$ and $\{1, 2, 4\}$ preserves the complete cancellation and variance $2\varepsilon^2/n$. Asymptotic indefiniteness thresholds for Bernoulli masks are due to Jurczak & Rohde (2017); the variance identity above supplies the finite-sample accounting for a fixed block design.

### B.4 PROOF OF THE FLOOR TAX

For any positive definite estimate $S$, covariance KL is

$$\mathrm{KL}(N(0, \Sigma) \| N(0, S)) = \tfrac{1}{2}\{\mathrm{tr}(S^{-1}\Sigma) - D + \log\det S - \log\det \Sigma\}. \tag{27}$$

In precision coordinates $C = S^{-1}$ the constraint is $0 \prec C \preceq f^{-1} I$. Diagonalise $\Sigma$. Replacing $C$ by its diagonal preserves $\mathrm{tr}(C\Sigma)$, preserves feasibility, and does not decrease its determinant by Hadamard's inequality. It therefore cannot increase the objective $\mathrm{tr}(C\Sigma) - \log\det C$. The remaining scalar minimisation gives $C_{ii} = \min\{\lambda_i^{-1}, f^{-1}\}$, or $S_{ii} = \max\{\lambda_i, f\}$. Dividing the resulting KL by $D$ proves Equation (5). Gaussian KL as an information measure for covariance filtering is established in Tumminello et al. (2007).

Misestimated eigenvectors, unfloored eigenvalues and the mean add predictive KL to $T(f)$, which Equation (5) gives at whichever floor validation selects.

### B.5 DERIVATION OF THE RANKING REVERSAL

Construction 2 is the pair of coverage-matched 24-instance designs of Section 3.4. For it, use free upper-triangular covariance coordinates $\theta$ and their symmetric basis matrices $D_a$. The full-observation Fisher matrix has entries $I_{\mathrm{full,ab}} = \frac{1}{2}\,\mathrm{tr}(\Sigma^{-1} D_a \Sigma^{-1} D_b)$, and the design information is

the sum of the analogous observed-submatrix forms. For available-case moments, the shared-record covariance identity of Steneman & Vinci (2024, Lemma 3.1), specialised using Isserlis' identity, gives

$$C_{(ij),(kl)} = \frac{n_{ijkl}}{n_{ij}n_{kl}}(\Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk}), \tag{28}$$

where $n_{ijkl}$ counts instances containing the union of the indices. Consequently the leading total joint-KL coefficients are

$$K_{\text{moment}} = \tfrac{1}{2}\operatorname{tr}(I_{\text{full}}C), \qquad K_{\text{likelihood}} = \tfrac{1}{2}\operatorname{tr}(I_{\text{full}}I_{\text{design}}^{-1}). \tag{29}$$

For the moment estimator, impose a fixed positive eigenvalue floor below the population minimum; it makes finite-sample KL well defined and is asymptotically inactive. The trace $\operatorname{tr}(I_{\text{full}}I_{\text{design}}^{-1})$ does not depend on the coordinates, and in whitened orthonormal coordinates $I_{\text{full}} = \frac{1}{2}I$ and $I_{\text{design}} = \frac{1}{2}\mathcal{M}_1$, so $K_{\text{likelihood}} = \frac{1}{2}\operatorname{tr}(\mathcal{M}_1^{-1}) = rK_{\text{cov}}$. At fixed coverage the denominators $n_{ij}n_{kl}$ and the counts $n_{ijkl}$ of one- and two-element unions are fixed, so $K_{\text{moment}}$ is affine in the counts of three- and four-element unions, with weights set by $\Sigma$; in Construction 2 these are $4$ for every triple and for the quadruple under $A$, whose four complete instances contain them, and $3$ for every triple and $0$ for the quadruple under $B$. Inserting the two pattern lists and $\Sigma_{ii} = 1$, $\Sigma_{ij} = 0.9$ gives the leading coefficients $K = \lim_{r\to\infty} r\,\mathbb{E}\,\mathrm{KL}$, in nats for the whole four-variable joint: $K_{\text{moment}} = 19.0923$ under $A$ and $18.2246$ under $B$, and $K_{\text{likelihood}} = 0.7113$ and $0.7317$. These constants are finite sums and matrix inverses; the released code computes them and confirms them in exact rational arithmetic. The near-null eigenspace in equicorrelation is spanned by pairwise contrasts, so the reversal here concerns the covariance of pooled entry estimates, whereas Construction 1 concerns a single isolated low-variance direction.

## C Finite-sample checks of the two constructions

This appendix confirms both constructions at finite sample size: Construction 1's separation, attributed to co-observation by an exchange control and a symmetric control, and Construction 2's opposite rankings.

The check of Theorem 2 on Construction 1 uses the one-parameter family ($v$ known) and $2{,}000$ replications. The estimator maximises the scalar likelihood of $\varepsilon$ over a bounded interval, using all observed subsets. The reference $2/r$ is the exact relative MSE of the mean square of the centred covering contrast; the likelihood estimates under $A$ stay within $6\%$ of it at every tested $\varepsilon$ and $r$. At $\varepsilon = 0.1$ their relative MSEs are $0.0420$, $0.0101$ and $0.00254$ at $r = 50, 200, 800$, and at $\varepsilon = 0.001$ they are $0.0377$, $0.00964$ and $0.00235$. Across these $r$, the $B/A$ ratio is 10–11 at $\varepsilon = 0.1$, 450–530 at $0.01$ and 42,000–48,000 at $0.001$ (42,300 at $r = 200$; Figure 4(a)). Under $B$ the median relative error of the estimate is near one for $\varepsilon \le 0.01$.

The separation could instead come from the particular variables each design observes. Exchanging variables 3 and 4 in the population exchanges the designs' information exactly and, up to Monte Carlo variation, their simulated relative MSEs. In the symmetric control, $v = (1, 1, 1, 1)/2$, whose support neither design co-observes, the $B/A$ ratios range 0.94–1.03.

Construction 2 shows the opposite rankings of Corollary 5 at finite $r$ (Table 2). The intervals are independent-design normal approximations, $\pm 1.96\sqrt{\text{s.e.}_A^2 + \text{s.e.}_B^2}$, over $2{,}000$ replications per design. With $60{,}000$ replications per design and intervals computed the same way, the moment learner's $A - B$ intervals are $[1.08, 1.47]$ at $r = 800$ and $[0.85, 1.14]$ at $r = 10{,}000$, in the same total $r$ KL units.

## D Cardiac setup, learners and diagnostics

This appendix specifies the cardiac representation, designs and learners, traces the pilot diagnostic's order to the design operator, reports the mixture learner in full, and gives the uncertainty study and spectral diagnostics behind Sections 4.2 and 4.3.

Table 2: The opposite rankings of Corollary 5 hold at finite sample size. The moment learner favours $B$ (positive values), with 95% intervals excluding zero from $r = 800$ on, and at $r = 800$ and 10,000 with 60,000 replications (see the text); the likelihood learner favours $A$ (negative values), with intervals excluding zero at $r = 200$ and 3,200. Construction 2 at finite sample size: design $A - B$ in total $r\,\mathrm{KL}$ (equivalently $Dr$ times excess per coefficient), with independent-design 95% Monte Carlo intervals.

| $r$ | Moment | Likelihood |
|---|---|---|
| 200 | $+0.11\ [-0.74, +0.96]$ | $-0.037\ [-0.059, -0.015]$ |
| 800 | $+1.52\ [+0.46, +2.58]$ | $-0.020\ [-0.042, +0.001]$ |
| 3,200 | $+1.00\ [+0.15, +1.85]$ | $-0.032\ [-0.053, -0.010]$ |

### D.1 Representation, conditioning and observation

Each video carries expert end-diastolic (ED) and end-systolic (ES) tracings; the larger cavity defines ED and the smaller ES. Eligible videos have the ED frame before the ES frame, so the ED-to-ES interval does not wrap around the clip. Geometry and preprocessing are fitted on training data. The contour representation keeps the long-axis length and two radii at 20 axial locations. Deformations are relative to ED and normalised by the ED long-axis length: the deformation at phase $t$ is $\delta(t) = tE + F(t)a$, with $E$ the ES deformation and $F(t)$ the endpoint-zero cubic modes below, so that $\delta(0) = 0$ and $\delta(1) = E$ hold exactly.

The cubic B-spline knot vector is $[0, 0, 0, 0, 1/3, 2/3, 1, 1, 1, 1]$; dropping its first and last basis functions leaves four endpoint-zero modes per coordinate. The joint vector is standardised by $s \in \mathbb{R}^{164}$, the standard deviations of the thirty pilot videos' coefficients, floored at $10^{-4}$; the pilot covariance is fitted with ridge 1 and eigenvalue floor 0.01, and selection and evaluation videos enter no fitted scale. An observed subset $S$ of phases exposes $H_S = (B_S \otimes I_{41})\,\mathrm{diag}(s)$, where $B_S$ evaluates the four modes at the phases in $S$. Fitted models evaluate each video's basis at its actual phase times, rounded to frames; the information operator uses the nominal phases $(0.2, 0.4, 0.6, 0.8)$, and the known-truth simulation (Section 4.2) tests its predictions with this difference present. The regression vector is $x = [1, C_{\mathrm{ED}}, E] \in \mathbb{R}^{84}$ with ED features $C_{\mathrm{ED}} \in \mathbb{R}^{42}$. The recorded ES deformation enters both the anchor $\delta(1) = E$ and the regressor $x$, so the joint model is conditional on both endpoints.

### D.2 Data splits, budget and learners

The 30-video pilot is completely observed and disjoint from the 7,212 acquisition videos; it adds the same complete observations to every policy, and every acquisition video a policy does not select has an empty observation pattern. Videos per pattern: windows 3,600 with all four phases; pairs+windows 1:1 600 per pair and 1,800 with all four; triples 1,200 per triple; pairs+windows 3:2 720 per pair and 1,440 with all four; singles+triples 360 per single phase and 1,080 per triple; pairs 1,200 per pair. The model family, ED/ES conditioning, video splits, frozen contour targets and selection videos are identical across policies. Regularisers are selected on the selection videos, every reported held-out NLL is computed on the evaluation videos, and the test videos are scored once, after every choice on the selection videos is fixed (Table 4).

The moment learner forms inverse-probability-weighted blockwise moments, solves for the Gaussian parameters and applies an eigenvalue floor; the likelihood learner maximises the observed-data Gaussian likelihood by accelerated expectation-maximisation (EM) with objective tolerance $10^{-10}$, parameter tolerance $10^{-8}$ and patience 5. A development-scale fit on 252 training videos under pairs specifies convergence at these tolerances: two starts converge in 13,081 and 16,081 iterations, and 40,000 further iterations at tolerances $10^{-13}$ and $10^{-11}$ change the objective by $5 \times 10^{-10}$ and the covariance by $4.6 \times 10^{-5}$ in relative Frobenius norm (KL $2 \times 10^{-6}$). Both learners choose a ridge from $\{10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$ and an eigenvalue floor from $\{10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.3, 1, 3\}$, the same grid under every policy, and share preprocessing, videos, observation patterns and selection videos.

**The pilot diagnostic.** The pilot diagnostic's order comes from the design operator, not from the pilot's covariance estimate. The thirty pilot videos give a covariance of rank 29, floored in 135 direc-

tions. Against both learners' scores the four pilot predictors of Figure 3(c) reach Spearman $1.0$ (the effective-sample-size bound, $15$ of $15$ pairs ordered), $0.943$ (the predicted excess, $14$, one adjacent swap of triples and pairs+windows 3:2), $0.971$ (pair probability, $13$ and $2$ ties) and $0.820$ (window share, $11$ and $3$ ties). The floored directions, which the floor makes isotropic, carry the order. On the $29$ determined directions the bound reverses it (Spearman $-0.89$; $2$ of $15$ pairs ordered): partial observation captures these directions almost completely, and on them $7{,}200$ acquisition videos observing two phases give a larger effective sample size than $3{,}600$ observing four. Rotating the floored directions $1{,}000$ times moves each policy's bound by at most $3.6\%$ and never changes the order.

Table 3 replaces the pilot covariance. The structural target fitted from the pilot videos and the cardiac reference covariance order all fifteen pairs with both predictors. Because the bound averages over the covariance's directions, an isotropic covariance leaves the basis free: in random bases the bound orders all fifteen pairs, and in the coordinate basis it orders $13$ and neither coverage-matched pair, as does the diagonal of the pilot covariance.

The order is stable across resampled pilots. One hundred resampled thirty-video pilots give the exact order in all $100$ ($99$ and $100$ at floors $0.001$ and $0.1$), sixty-video pilots in all $100$, and $120$-video pilots order at least fourteen pairs in all $100$ and all fifteen in $31$. The fifteenth pair, triples against pairs+windows 3:2, has the smallest measured gap of the fifteen for the likelihood learner, $0.0006$ nats per coefficient (standard errors $0.00016$ and $0.00014$); for the moment learner the gap is $0.032$. The predicted excess, computed exactly for ten of the resampled thirty-video pilots, orders fourteen pairs in each, with the same adjacent swap.

Table 3: The pilot diagnostic with the pilot covariance replaced: of the fifteen pairs of policies, the number the effective-sample-size bound orders as both learners' scores do, whether the bound orders the two coverage-matched pairs, and the number the operator's predicted excess orders. The isotropic covariance uses the pilot for the per-coefficient scale alone.

| Covariance | Bound: pairs ordered | Bound: coverage-matched pairs | Excess: pairs ordered |
|---|---|---|---|
| Pilot covariance, rank 29 | 15 | both | 14 |
| Structural target from the pilot videos | 15 | both | 15 |
| Cardiac reference covariance | 15 | both | 15 |
| Isotropic, coordinate basis | 13 | neither | 14 |
| Isotropic, random basis, 1,000 rotations | 15 in 1,000 of 1,000 | both, in every rotation | 14 |
| Diagonal of the pilot covariance | 13 | neither | 14 |

### D.3 The mixture learner

The mixture learner tests whether the coverage-matched order survives beyond one Gaussian, with every other part of the likelihood learner held fixed. It is a mixture of Gaussians conditional on the covariates, each component with its own regression and full covariance, fitted by EM through the same observed-data likelihood, observation maps, masks, ridge and eigenvalue floor as the likelihood learner. With one component it is the likelihood learner itself: its rescoring reproduces the study's per-video scores to $8 \times 10^{-15}$. The two-component fit starts from the fitted Gaussian split along its leading eigen-directions, and the four-component fit starts from the two-component fit split the same way. Every fit stops by the likelihood learner's convergence rule; all $360$ mixture fits on the videos converge, no component collapses and no objective increases. The number of components, one, two or four, is chosen on the $64$ selection videos by held-out conditional NLL; the evaluation videos are scored for every number of components, and the test videos after that choice. Held-out selection keeps one component in every assignment under every policy, so the Gaussian analysis describes the selected model. The fixed two- and four-component fits carry the test beyond one Gaussian: on the evaluation videos the coverage-matched order holds in every assignment but one, and the gaps and the stitching penalty grow with the number of components (Table 4).

The moment learner's coverage-matched contrasts are $-0.0022$ (s.e. $0.0085$) and $-0.0004$ (s.e. $0.0111$) on the evaluation videos and $-0.002$ (s.e. $0.009$) and $-0.003$ (s.e. $0.011$) on the test videos, and its stitching penalty is $0.474$ and $0.475$.

Table 4: The mixture learner keeps the coverage-matched order, with gaps and a stitching penalty that grow with the number of components. On the evaluation videos every assignment orders both coverage-matched pairs as the Gaussian learners do at one and two components, and every assignment but one does at four. Six-policy acquisition study, thirty assignments, held-out NLL in nats per coefficient. Held-out selection on the 64 selection videos keeps one component in 30 of 30 assignments under every policy. Upper block: evaluation videos, mean over assignments (s.e.). Lower blocks: paired contrasts, a negative value favouring the first-named policy, with the paired s.e. and the number of the thirty assignments whose sign is opposite to the mean, and beneath each contrast its 95% bootstrap interval crossed over assignments and videos (4,000 replications). The stitching penalty is pairs minus windows. One component is the likelihood learner.

| | One component | Two components | Four components |
|---|---|---|---|
| Held-out NLL, evaluation videos (s.e. over assignments) | | | |
| windows | .477 (.0001) | .485 (.0002) | .530 (.0004) |
| pairs+windows 1:1 | .484 (.0001) | .505 (.0003) | .582 (.0015) |
| triples | .487 (.0002) | .520 (.0007) | .610 (.0015) |
| pairs+windows 3:2 | .487 (.0001) | .514 (.0003) | .605 (.0013) |
| singles+triples | .490 (.0001) | .528 (.0005) | .629 (.0019) |
| pairs | .630 (.0036) | .742 (.0039) | 1.027 (.0049) |
| Paired contrasts, evaluation videos (paired s.e.; assignments of 30 with the opposite sign) | | | |
| pairs+windows 1:1 minus triples | −.0025 (.0002; 0) | −.0156 (.0007; 0) | −.0273 (.0021; 0) |
| crossed 95% interval | [−.0030, −.0020] | [−.0194, −.0126] | [−.0329, −.0223] |
| pairs+windows 3:2 minus singles+triples | −.0025 (.0002; 0) | −.0135 (.0005; 0) | −.0241 (.0022; 1) |
| crossed 95% interval | [−.0031, −.0020] | [−.0169, −.0106] | [−.0291, −.0195] |
| Stitching penalty | .152 (.004; 0) | .256 (.004; 0) | .497 (.005; 0) |
| crossed 95% interval | [.141, .165] | [.236, .279] | [.469, .529] |
| Paired contrasts, test videos (paired s.e.; assignments of 30 with the opposite sign) | | | |
| pairs+windows 1:1 minus triples | −.0023 (.0002; 0) | −.0138 (.0007; 0) | −.0256 (.0021; 0) |
| crossed 95% interval | [−.0028, −.0018] | [−.0161, −.0118] | [−.0300, −.0212] |
| pairs+windows 3:2 minus singles+triples | −.0023 (.0002; 1) | −.0121 (.0005; 0) | −.0226 (.0022; 1) |
| crossed 95% interval | [−.0029, −.0017] | [−.0142, −.0102] | [−.0274, −.0183] |
| Stitching penalty | .146 (.004; 0) | .246 (.004; 0) | .477 (.005; 0) |
| crossed 95% interval | [.136, .156] | [.230, .262] | [.455, .500] |

### D.4 The two-policy uncertainty study

The two-policy uncertainty study supplies the intervals for the learner comparison and the floor tax of Section 4.3. It compares pairs with windows at 14,424 frames on the videos outside the pilot, with crossed bootstrap intervals for the two-policy contrasts (Table 7, Real column); the six-policy contrasts carry the crossed intervals of Table 4. Its mean evaluation-video NLLs, 0.9491 (moment learner) and 0.6289 (likelihood learner) under pairs and 0.4748 and 0.4774 under windows, lie within 0.001 of the six-policy acquisition study's (Table 1). The moment learner's excess under pairs could instead come from how it weights the blockwise moments. The study therefore also fits variants of the moment learner that weight the blockwise moments differently or smooth each video before forming moments, and an estimator that uses complete videos alone: four learners under windows and five under pairs. Their mean evaluation-video NLLs spread over 0.003 nats per coefficient under windows and 0.694 under pairs, and under pairs the moment learner's weighting variants lie within 0.005 of one another.

### D.5 Spectrum and calibration diagnostics

These diagnostics give the spectrum on which the floor tax is evaluated, the indefiniteness that forces the moment learner's floor, and both learners' calibration. The cardiac reference covariance of Section 4.2, on which the floor tax $T(f)$ of Equation (5) is evaluated, is the mean of the 30 moment estimates under windows, each fitted on 3,630 complete videos. Its minimum, mean and maximum eigenvalues are 0.010, 0.86 and 23.5; 79 eigenvalues lie below 0.1, 110 below 0.3 and 143 below 1. Under pairs the moment learner floors 111.5 directions on average; before flooring, its estimate in the first assignment has minimum eigenvalue −10.7 and ten negative eigenvalues. Available-case second moments of Gaussian data drawn from the reference covariance, with each policy's observa-

tion patterns and the mean known, have minimum eigenvalue $-0.258$ and five negative eigenvalues under pairs and are positive definite under the other five policies.

**The floor trajectory.** At the selected floor $0.3$ the moment learner lies $0.321$ nats per coefficient above the likelihood learner under pairs; given the likelihood learner's mean it scores $0.938$ (Table 5), so $0.308$ of the difference lies in the covariance and $0.012$ in the mean fit, beside the floor tax $T(0.3) = 0.343$.

Table 5: The floor trajectory under pairs, thirty assignments, mean fit fixed: held-out NLL of the moment learner in nats per coefficient at each floor of the grid, beside the floor tax $T(f)$ of the cardiac reference spectrum (Equation (5)), which is zero at floors below its smallest eigenvalue $0.010$. The last three rows exchange the mean fits: the moment learner at floor $0.3$ with the likelihood learner's mean, and the likelihood learner with its own mean and with the moment learner's.

| Floor $f$ | $T(f)$ | Held-out NLL |
|---|---|---|
| $10^{-4}$ | 0 | 493 |
| $10^{-3}$ | 0 | 49.8 |
| $10^{-2}$ | 0 | 5.43 |
| 0.1 | .131 | 1.077 |
| 0.3 | .343 | .950 |
| 1 | .704 | 1.194 |
| 3 | 1.134 | 1.592 |
| Moment learner, $f = 0.3$, likelihood learner's mean | .343 | .938 |
| Likelihood learner, own mean | 0 | .630 |
| Likelihood learner, moment learner's mean | 0 | .653 |

The likelihood learner's mean held-out Mahalanobis statistic exceeds the Gaussian dimension $164$ under every policy, from $175.8$ under windows through $182.2$–$185.9$ under the four intermediate policies to $247.2$ under pairs; the moment learner's is $150.5$ under pairs. The smallest effective sample size over the $164$ directions of the cardiac reference covariance is $3{,}630$ under windows and $541$ under pairs.

## E THE KNOWN-TRUTH SIMULATION: COVARIANCE ESTIMATION ERROR AND REPRODUCTION OF THE REAL PENALTIES

This appendix gives the exact account of where the leading-order predictions of Theorem 3(ii) describe the likelihood learner and where, and through which mechanism, they deviate; it also establishes that a simulation with resampled real residuals reproduces $91\%$ of the real likelihood stitching penalty and $94\%$ of the real moment stitching penalty (Table 7).

**Specification.** The known-truth simulation keeps the real covariates, each video's basis evaluated at its measured phase times, and the actual observation masks of all six policies. It draws the responses from a reference truth: the mean of the thirty windows estimates of the regression coefficients and of the covariance (the cardiac reference covariance). The residuals come in two versions: Gaussian draws with the cardiac reference covariance, and real residuals, resampled and linearly transformed to have that covariance. The resampled version carries the empirical non-Gaussian residual structure, fourth moments included, together with the relation between residuals and covariates. Both learners use the estimators of the six-policy acquisition study with smaller candidate sets than its common $7 \times 7$ grid: the moment learner selects its ridge from $\{0.01, 0.1, 1, 10, 100, 1000\}$ and its floor from $\{10^{-4}, 10^{-3}, 0.01, 0.1, 0.3, 1\}$, and the likelihood learner uses ridge $0.01$ and selects its floor from $\{0.001, 0.01, 0.1\}$. All $1{,}080$ likelihood fits converged, with a median of $43$ EM iterations under Gaussian truth.

**Covariance estimation error and covariance-only KL.** With the truth known, $2\,\mathrm{tr}(\mathcal{M}^{-1})$ predicts the likelihood learner's covariance estimation error within $3.2\%$ under every policy (Figure 3(a)). Covariance-only KL also carries higher-order terms, and their size tracks the smallest

relative eigenvalue of $\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}$ (Table 6). Under the five policies other than pairs that eigenvalue stays at $0.49$ or above, and the measured covariance KL lies within 7–22% of its local approximation. Under pairs the eigenvalue reaches $0.08$, and the measured covariance KL is $2.9$ times the local approximation.

Table 6: Known-truth simulation, Gaussian version, likelihood learner: the measured covariance KL is within 7–22% of its local approximation while the smallest relative eigenvalue of $\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}$ stays at $0.49$ or above, and $2.9$ times it under pairs, where that eigenvalue is $0.08$. KL is in nats per coefficient; the eigenvalue is dimensionless.

| Policy | KL predicted | KL measured | Ratio | Smallest relative eigenvalue |
|---|---|---|---|---|
| windows | .0114 | .0122 | 1.07 | .63 |
| pairs+windows 1:1 | .0158 | .0181 | 1.14 | .55 |
| triples | .0164 | .0195 | 1.19 | .52 |
| pairs+windows 3:2 | .0174 | .0204 | 1.17 | .53 |
| singles+triples | .0179 | .0218 | 1.22 | .49 |
| pairs | .0393 | .1124 | 2.86 | .08 |

**The mean term.** Under windows the operator's predicted excess, $0.0229$ nats per coefficient (Table 1), splits into $0.0114$ for the covariance, the local approximation of Table 6, and $0.0116$ for the mean. The measured mean excess is $0.0070$ under windows (Appendix B) and $0.0096$ under pairs.

**Reproduction of the real penalties.** Table 7 separates the stitching penalties into their sources. The local theory gives the leading-order term; the Gaussian version adds the higher-order terms under a Gaussian truth; the resampled version adds the empirical fourth moments and the relation between residuals and covariates; the real column is the measured penalty. The likelihood learner's penalty exceeds its local term 3.5-fold with the truth known (Gaussian column) and fivefold on the videos (Real column).

Table 7: The resampled-residual simulation reproduces $91\%$ of the real likelihood stitching penalty ($0.1374$ against $0.1516$ nats per coefficient) and $94\%$ of the real moment stitching penalty ($0.4450$ against $0.4743$), the Gaussian version $68\%$ and $90\%$. Reproduction of the real penalties, nats per coefficient. Stitching penalties are pairs minus windows, so a positive value favours windows. Local theory: the operator's predicted difference (Theorem 3), the pairs total minus the windows total of Table 1, and zero under windows by Theorem 3(i). The Gaussian and resampled columns are exact differences in population total KL; the real column holds the held-out evaluation-video NLL differences of the two-policy uncertainty study with $95\%$ bootstrap intervals crossed over assignments and videos. An em dash marks an entry without a value, not a zero.

| Contrast | Local theory | Gaussian | Resampled | Real |
|---|---|---|---|---|
| Likelihood stitching penalty | .0293 | .1028 | .1374 | .1516 [.1399, .1653] |
| Moment stitching penalty | — | .4292 | .4450 | .4743 [.4581, .4917] |
| pairs: moment minus likelihood | — | .3265 | .3080 | .3202 [.3025, .3386] |
| windows: learner difference, magnitude | 0 | .0002 | — | .0025 [.0021, .0029] |

Under Gaussian truth the floor tax of the true spectrum at the floors the moment learner selects is $0.2722$ nats per coefficient, $83\%$ of the Gaussian version's moment-minus-likelihood difference under pairs. Across the six policies of the Gaussian version, triples closes $92.8\%$ of the gap from pairs to windows for the likelihood learner and $60.7\%$ for the moment learner; on the real data it closes $93.8\%$ and $55.5\%$.

## F THE TWO FURTHER DOMAINS: AIR QUALITY AND ELECTRICITY LOAD

This appendix establishes that the operator's predictions carry to two further domains: with the truth known it predicts the likelihood learner's covariance estimation error, and on the real data its

predicted excess orders the six policies for both learners; the coverage-matched contrasts are not resolved in either domain.

**Specification.** In both domains an instance is one forecast from one station or client, and instances start every 72 hours: 24 hours of context followed by 24 hours of targets. The Beijing Multi-Site Air-Quality data (Zhang et al., 2017) supply instances from 12 stations, each with a joint vector of six pollutants at four future hours (one block of six per hour). The UCI ElectricityLoadDiagrams 2011–2014 data (Trindade, 2015) supply instances from 370 clients, each with the next 24 hours in four blocks of six. Both joint vectors have dimension 24. Air-quality training instances precede 1 March 2015 and validation instances 1 March 2016, with the data running to 1 March 2017; electricity training instances precede 1 January 2013 and validation instances 1 January 2014, with the data running to 1 January 2015. The validation period is divided into alternating seven-day blocks of selection and evaluation instances, and the target scale is fitted on the first thirty training instances. Every policy starts from a complete 30-instance pilot common to all policies and spends its budget of observed blocks (one block of one instance, the analogue of a cardiac frame) on the remaining training instances. Beijing has 2,172 training instances besides the pilot and a budget of 2,160; electricity has 12,000 training instances, the pilot included, and a budget of 11,880. Both learners select ridge and floor on the selection instances. Electricity targets are divided by a blended output scale that mixes, with weight 0.5, the load level of the instance's context and that of its client; in the original load units the same fits score 4.48–4.92.

Table 8: In both further domains the operator's predicted excess ranks the six policies at Spearman 0.943 for both learners on Beijing and at 1.0 (moment) and 0.943 (likelihood) on electricity, each swap exchanging the coverage-matched pair pairs+windows 3:2 and singles+triples, whose scores differ by 0.004 or less. The two further domains, real data: conditional NLL in nats per coefficient on held-out evaluation instances, mean (s.e.) over thirty assignments. Predicted excess: the operator's predicted excess per coefficient, covariance plus mean term (Theorem 3(ii)), from the complete training covariance. Rows follow the operator's cardiac order.

| Policy | Moment | Likelihood | Predicted excess |
|---|---|---|---|
| Beijing air quality | | | |
| windows | .874 (.007) | .850 (.005) | .062 |
| pairs+windows 1:1 | .976 (.004) | .875 (.006) | .067 |
| triples | .979 (.004) | .875 (.006) | .068 |
| pairs+windows 3:2 | .984 (.004) | .881 (.007) | .069 |
| singles+triples | .982 (.004) | .877 (.006) | .070 |
| pairs | 1.009 (.004) | .934 (.004) | .090 |
| Electricity load, blended output scale | | | |
| windows | .328 (.006) | .2549 (.0022) | .0071 |
| pairs+windows 1:1 | .533 (.017) | .2598 (.0019) | .0080 |
| triples | .563 (.017) | .2615 (.0018) | .0080 |
| pairs+windows 3:2 | .587 (.016) | .2644 (.0013) | .0083 |
| singles+triples | .588 (.016) | .2640 (.0016) | .0083 |
| pairs | .690 (.002) | .2818 (.0022) | .0130 |

**Known truth.** Beside the real-data scores of Table 8, the known-truth simulation draws Gaussian outcomes from each domain's fitted mean and complete training covariance. On Beijing the operator predicts the likelihood learner's covariance estimation error $\left\|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - I\right\|_F^2$ within 7.4% under every policy: 3.12 predicted against 3.08 measured under pairs, 1.05 against 1.11 under windows, and 1.49 against 1.60 under triples, the largest deviation. The predicted excess orders the likelihood learner's scores exactly and the moment learner's at Spearman 0.943, and Gaussian finite-sample effects reproduce 127% of the moment learner's real stitching penalty and 78% of the likelihood learner's.

On electricity the likelihood learner selects ridge 0.001. Its covariance estimation error is 0.200 predicted against 0.205 measured under windows (within 2.4%, the regular efficient regime of Theorem 3(ii)), 0.27–0.31 under the intermediate policies, each within 3% of its prediction, and 0.72

predicted against 0.61 measured under pairs. Ridge selection matters in this domain because the mean signal is strong: the true regression coefficients have squared norm 131 after whitening by the covariance. With a ridge fixed at 0.01, the closed-form Gaussian maximum-likelihood fit on the complete instances has covariance estimation error 0.387, against 0.19 with zero ridge, and the likelihood learner measures 0.39.

**Coverage-matched contrasts.** Neither further domain resolves its coverage-matched contrasts: every contrast lies within two standard errors of zero. Paired over the thirty assignments, in held-out NLL per coefficient, pairs+windows 1:1 minus triples and pairs+windows 3:2 minus singles+triples are, on air quality, $-0.0001$ (s.e. 0.0071) and $+0.0036$ (s.e. 0.0048) for the likelihood learner and $-0.0030$ (s.e. 0.0044) and $+0.0017$ (s.e. 0.0046) for the moment learner; on electricity load, $-0.0017$ (s.e. 0.0019) and $+0.0004$ (s.e. 0.0017) for the likelihood learner and $-0.030$ (s.e. 0.017) and $-0.0015$ (s.e. 0.018) for the moment learner.

## G THE CONSTRUCTED DOMAIN: SUPPORT, CONTROLS AND REGULARISATION

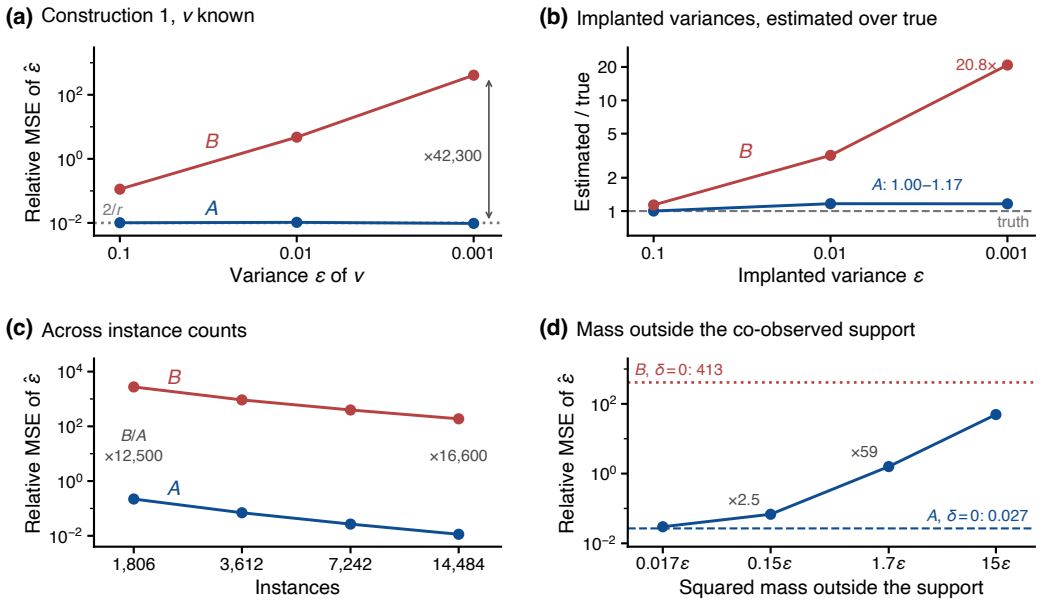



Figure 4: **Co-observing a direction's support separates coverage-matched designs by four orders of magnitude at $\varepsilon = 0.001$ across the tested sample sizes, within a quantified tolerance to mass off that support.** Blue: $A$, which co-observes each direction's support; red: $B$, which does not. (a) Construction 1 with $v$ known and the designs repeated $r = 200$ times: relative mean squared error of the scalar likelihood estimate of $\varepsilon$, within 4% of $2/r$ under $A$ and 42,300 times $A$'s under $B$ at $\varepsilon = 0.001$; across $r = 50$ to 800, within 6% and 42,000 to 48,000. (b) to (d) Constructed domain, likelihood learner. (b) Estimate over truth for the 41 implanted variances, means over 30 seeds with floors from $\{10^{-4}, 10^{-3}, 10^{-2}\}$: 1.00 to 1.17 under $A$ and 20.8 under $B$ at $\varepsilon = 0.001$. (c) Relative mean squared error at $\varepsilon = 0.001$ against the instance count (mean $\pm$ s.e., ten seeds): across an eightfold range the ratio of $B$ to $A$ stays between 12,000 and 17,000. (d) Design $A$ at $\varepsilon = 0.001$ (ten seeds) with a component $\delta = 0.01$, 0.03, 0.1 or 0.3 of the implanted direction on the fourth phase, $v = (1, -2, 1, \delta)/\sqrt{6 + \delta^2}$, against the direction's squared mass outside the co-observed support, $\delta^2/(6+\delta^2)$, as a multiple of $\varepsilon$. The relative mean squared error stays near its value at $\delta = 0$ (dashed, 0.027) while that mass is small against $\varepsilon$ (a factor 2.5 at $0.15\varepsilon$) and rises as the mass reaches $\varepsilon$ (a factor 59 at $1.7\varepsilon$).

This appendix specifies the constructed domain of Section 4.4 and establishes that its separation comes from co-observation: the exchange control swaps the designs' errors, the negative control

shows how small the background's own contrast is, and the separation of the implanted variances holds across instance counts, under a two-component truth and on every floor grid.

### G.1 Population and summaries of the implanted variances

Phase space is the observation space ($H = I$). The background $C_{\mathrm{b}}$ is the covariance of the 41 contour coordinates at the four interior phases implied by the cardiac reference covariance, standardised to unit variance. Let $U = v \otimes I_{41} \in \mathbb{R}^{164\times 41}$, $v = (1, -2, 1, 0)/\sqrt{6}$, be the orthonormal basis of the implanted contrasts (Section 4.4), and $Q = I - UU^{\top}$. The population $C_\varepsilon = QC_{\mathrm{b}}Q + \varepsilon UU^{\top}$ has 41 eigenvalues exactly equal to $\varepsilon$, at $\varepsilon \in \{0.1, 0.01, 0.001\}$; outside $\mathrm{range}(U)$ it has 75 eigenvalues below 0.1, minimum 0.0219, and maximum 57.6. Both learners fit the unrestricted joint, including an intercept, from the two coverage-matched designs, and select their floors from $\{10^{-4}, 10^{-3}, 10^{-2}\}$ (the primary floor grid) on 700 selection instances masked by the same design. The reported KL is the covariance-only population KL to $C_\varepsilon$, which leaves out the estimated mean; it differs from the conditional negative log-likelihood of the six-policy acquisition study and from the total population KL of the known-truth simulation.

The factor 15,100 is a ratio of mean relative MSEs, each the average over 30 seeds of the per-seed relative MSE over the 41 implanted variances (Table 9). At the two smaller $\varepsilon$ the selected floor moves $A$'s estimates, so $A$'s relative MSE exceeds $2/r$, the relative MSE of the direct sample variance over the $r = 1{,}034$ triples of $A$ (Table 12).

Table 9: Coverage-matched designs differ 15,100-fold in the likelihood learner's relative MSE of the implanted variances at $\varepsilon = 0.001$ (0.027 under $A$ against 413 under $B$, 30 seeds). Constructed domain, likelihood learner, primary floor grid. Ratio: estimated over true implanted variance, mean over seeds (seed s.e.). Abs. rel.: mean absolute relative error. Rel. MSE: relative mean squared error. Both are averaged over all 41 implanted variances.

| $\varepsilon$ | Ratio $A$ | Ratio $B$ | Abs. rel. $A/B$ | Rel. MSE $A$ | Rel. MSE $B$ | $B/A$ |
|---|---|---|---|---|---|---|
| .1 | 1.001 (.001) | 1.135 (.003) | .031/.138 | .0015 | .024 | 16 |
| .01 | 1.166 (.001) | 3.184 (.029) | .166/2.18 | .028 | 4.99 | 176 |
| .001 | 1.163 (.001) | 20.83 (.27) | .163/19.8 | .027 | 413 | 15,100 |

Under $B$ the estimated implanted variances exceed $\varepsilon$ by 0.0135, 0.0218, and 0.0198 on average: an upward offset of 0.014 to 0.022 in absolute units that does not shrink with $\varepsilon$. As $\varepsilon$ falls, $B$'s relative MSE grows by factors 206 and 83 per decade (log–log slope $-2.1$ over two decades). Under $A$ the direct sample variance of the contrast co-observed by the triple has relative MSE $2/r = 0.001934$ in theory and 0.002026 (s.e. 0.000083) observed.

Table 10: Constructed domain, likelihood learner, primary floor grid: covariance estimation error $\left\| C_\varepsilon^{-1/2} \widehat{C}\, C_\varepsilon^{-1/2} - I \right\|_F^2$ and covariance-only KL per coefficient. Seed s.e. in parentheses; the s.e. of the contrast is paired.

| $\varepsilon$ | Estimation error $A$ | Estimation error $B$ | KL $A$ | KL $B$ | KL $B - A$ |
|---|---|---|---|---|---|
| .1 | 32.6 (.1) | 53.2 (.7) | .1370 (.0012) | .1365 (.0008) | $-.0005$ (.0013) |
| .01 | 23.9 (.1) | 1,709 (61) | .0431 (.0002) | .0983 (.0005) | $+.0552$ (.0005) |
| .001 | 26.0 (.1) | 140,998 (5,316) | .0469 (.0003) | .4566 (.0058) | $+.4097$ (.0058) |

Table 10 gives the likelihood learner's covariance estimation error and covariance-only KL. The moment learner ranks the designs the other way at the smallest $\varepsilon$. Its covariance KL for $A/B$ is $0.3595/0.3794$, $0.2534/0.2664$, and $1.9927/1.1622$ at the three $\varepsilon$ values; at 0.001 its $B - A$ contrast is $-0.8305$ (s.e. 0.1927), and its two fits overestimate the implanted variances by factors 30.6 ($A$) and 38.8 ($B$). This reversal arises at finite sample size through the floor the moment learner selects from its grid; Construction 2 (Appendix B) reverses the ranking asymptotically, with the floor inactive.

### G.2 SUBSPACE INFORMATION

Lemma A.1 applies with $\Sigma_\perp = QC_bQ$ to every subset of either design that does not contain phases 1, 2 and 3 together, giving each implanted direction $u$ a squared projection $\|P_S u\|^2 = O(\varepsilon)$. $B$ has no other subsets, so the lemma's $\Omega(\varepsilon^{-2})$ bound holds under $B$, while the triple of $A$ observes all implanted contrasts, $P_S U = U$. Under $A$ the effective sample size of the implanted directions is 1,034 to 1,038 at $\varepsilon = 0.001$; under $B$ its median over those directions is 2,013, 450, and 17 at $\varepsilon = 0.1$, 0.01, and 0.001. At $\varepsilon \leq 0.01$ some eigenvalues of $\mathcal{M}$ fall below one, and there the selected regulariser shapes the fitted covariance (Appendix G.8).

### G.3 CRAMÉR–RAO BOUNDS WITH EVERY COVARIANCE PARAMETER FREE

The bound $2/N_{\text{eff}}(v)$ holds the other covariance parameters fixed; with all 13,530 free it is $2\langle e_v, \mathcal{M}^{-1} e_v\rangle$ (Appendix A). Under $A$ the two agree within 0.2% at $\varepsilon = 0.001$, 7% at 0.01 and 38% at 0.1. Under $B$ the free bound grows as $\varepsilon^{-1.9}$ (factors 65 and 95 per decade), and at $\varepsilon = 0.001$ it is 450 times $2/N_{\text{eff}}(v)$ as a ratio of medians (417 as the median of per-direction ratios). The free bounds separate $B$ from $A$ by factors 6.4, 292 and 27,000 as $\varepsilon$ falls. The likelihood learner's relative MSE (Table 9) is 2.9, 9.1 and 7.9 times the free bound under $B$; under $A$ it is 1.14 times the free bound at $\varepsilon = 0.1$ and 14 to 15 times at the two smaller $\varepsilon$, through the selected floor.

Table 11: The Cramér–Rao bounds with every parameter free separate $B$ from $A$ 27,000-fold at $\varepsilon = 0.001$; the likelihood learner attains 15,100-fold. Constructed domain: Cramér–Rao bounds on the relative variance of an implanted variance, medians over the 41 implanted directions. $2/N_{\text{eff}}$: other covariance parameters known. Free: every covariance parameter free, $2\langle e_v, \mathcal{M}^{-1} e_v\rangle$; under $B$ at $\varepsilon = 0.001$ it ranges from 28 to 104 over the directions.

| $\varepsilon$ | $2/N_{\text{eff}}$ $A$ | Free $A$ | $2/N_{\text{eff}}$ $B$ | Free $B$ |
|---|---|---|---|---|
| .1 | .00095 | .00131 | .00099 | .0084 |
| .01 | .00175 | .00187 | .00445 | .55 |
| .001 | .00193 | .00193 | .116 | 52 |

### G.4 BIAS AND VARIANCE OF THE IMPLANTED-VARIANCE ERROR

Under $B$ the likelihood learner's error in the implanted variances is almost entirely squared bias, and under $A$ with the floor below $\varepsilon$ it attains the free-parameter bound. Table 12 splits the relative MSE of the implanted variances over seeds into squared bias and variance.

Table 12: Under $A$ at $\varepsilon = 0.001$ with the floor below $\varepsilon$ the likelihood learner's relative MSE of the implanted variances is 0.00195, of which 0.0004 is squared bias (20%), at the free-parameter bound 0.0019; under $B$ it is 396 with 98% squared bias, the squared offset of a twentyfold overestimate. Constructed domain: relative MSE of the implanted variances split into squared bias and variance over seeds, per implanted direction (ten seeds; six for the floor below $\varepsilon$). Bias share: squared bias over relative MSE. Variance is the unbiased variance over seeds, so squared bias plus $(n-1)/n$ times the variance equals the relative MSE.

| Design | $\varepsilon$ | Learner | Floor | Rel. MSE | Squared bias | Variance | Bias share |
|---|---|---|---|---|---|---|---|
| $A$ | .1 | likelihood | selected | .0013 | .0002 | .0013 | 13% |
| $A$ | .001 | likelihood | selected | .0268 | .0262 | .0007 | 98% |
| $A$ | .001 | likelihood | below $\varepsilon$ | .00195 | .0004 | .0019 | 20% |
| $B$ | .001 | likelihood | selected | 396 | 389 | 8.4 | 98% |
| $A$ | .001 | moment | selected | 998 | 880 | 131 | 88% |
| $B$ | .001 | moment | selected | 1,504 | 1,370 | 149 | 91% |

### G.5 Instance count and support tolerance

The separation is stable across sample sizes. Across the instance counts of Figure 4(c), $1{,}806$ to $14{,}484$ (258, 516, $1{,}034$ and $2{,}069$ per pattern; ten seeds each), the relative MSE falls from $0.221$ to $0.011$ under $A$ and from $2{,}748$ to $188$ under $B$; at $1{,}034$ per pattern the ratio is $14{,}793$-fold over ten seeds and $15{,}100$-fold over thirty. The support-tolerance sweep of Figure 4(d) quantifies how much of the implanted direction may fall outside the co-observed support before the relative mean squared error under $A$ rises. Over the support components $\delta = 0$ to $0.3$, the median $N_{\text{eff}}$ under $A$ falls from $1{,}036$ to $248$ and the smallest eigenvalue of $\mathcal{M}$ from $0.062$ to $0.0010$.

### G.6 A two-component truth

The separation holds when the truth is a mixture. At $\varepsilon = 0.001$ the population has two equiprobable components with means $\pm 3\sqrt{\lambda_1}\, u_1$, three standard deviations along the leading eigen-direction $u_1$ of the background covariance ($\lambda_1 = 57.6$), a marginal feature both designs observe equally, and implanted contrast variances $\varepsilon/2$ and $2\varepsilon$ (pooled $1.25\varepsilon$). The mixture learner of Appendix D.3 is fitted with one, two and four components under both designs, ten seeds each (Table 13). With two components the variance $\varepsilon/2$ lies below the selected floor $0.001$, which constrains its estimate under $A$, and the covariance estimation error of the recovered component is $63$ under $A$ and $87{,}200$ under $B$. Held-out selection keeps one component in every seed under both designs: at this sample size a second component's gain per coefficient lies below the held-out NLL its estimation adds. The moment learner's excess held-out NLL ranks the designs the other way, as its covariance-only KL does.

Table 13: Constructed domain with a two-component truth, $\varepsilon = 0.001$, ten seeds: with two components the mixture learner recovers the variance above the floor to $5\%$ under $A$ and separates the designs $7{,}600$-fold in the pooled relative MSE of the implanted variances. Excess NLL: held-out NLL over the truth's, nats per coefficient, mean (seed s.e.). Relative MSE of the implanted variances: pooled over the two components (true pooled variance $1.25\varepsilon$), and for the component with variance $2\varepsilon$, above the selected floor $0.001$, and the component with variance $\varepsilon/2$, below it; after $\times$, the mean ratio of estimated to true variance. An em dash marks an entry the learner does not define: the per-component columns need fitted components that correspond one to one with the truth's.

| | Excess NLL | | Relative MSE of the implanted variances | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Pooled | | Above the floor | | Below the floor | |
| Learner | $A$ | $B$ | $A$ | $B$ | $A$ | $B$ | $A$ | $B$ |
| One component | .105 (.001) | .553 (.006) | .0089 | 259 | — | — | — | — |
| Two components | .177 (.004) | .835 (.008) | .079 | 601 | .0063 ($\times$1.05) | 242 ($\times$16) | 1.41 ($\times$2.19) | 3,722 ($\times$61) |
| Four components | .927 (.016) | 1.680 (.010) | .175 | 1,473 | — | — | — | — |
| Moment learner | 4.57 (1.03) | .959 (.038) | 2,841 | 5,200 | — | — | — | — |

### G.7 Controls

Three controls attribute the separation to co-observation.

**Exchange control.** The separation could instead come from which phases the implant uses. Permuting phases 3 and 4 in the population exchanges the roles of $A$ and $B$, and in covariance KL all twelve checks (two learners, three values of $\varepsilon$, two designs) lie within $1.4$ standard errors of the corresponding swapped result. At $\varepsilon = 0.001$ the likelihood learner's KL is $0.4501$ under exchanged $A$ against $0.4566$ under $B$ in the main population, and $0.0468$ under exchanged $B$ against $0.0469$ under $A$; the moment learner's $B - A$ becomes $+0.8435$ (s.e. $0.1727$). With the two-component truth, the two-component learner's excess NLL is $0.826$ under exchanged $A$ against $0.835$ under $B$, and $0.186$ under exchanged $B$ against $0.177$ under $A$. The permuted population is sampled through its own eigendecomposition, with a different random basis, so the exchange holds in distribution.

**Symmetric control.** The designs could also differ through features of the population other than which design co-observes the implant's support. The symmetric control makes the background and

the implant invariant to the exchange, and declares the designs equivalent when the paired 95% interval of the KL contrast lies within $\pm 0.005$ (Table 14). The criterion holds at $\varepsilon = 0.1$ and $0.01$ for both learners. At $0.001$ the interval half-widths are $0.010$ and $0.014$, wider than the margin. At that level the likelihood learner's implanted-variance ratios are $18.84$ under $A$ and $18.91$ under $B$.

Table 14: Constructed domain, symmetric control: covariance KL $B - A$ per coefficient (paired seed s.e.). The equivalence margin is $\pm 0.005$.

| $\varepsilon$ | Likelihood | Moment | Interval within margin |
|---|---|---|---|
| .1 | $+.0000$ $(.0007)$ | $-.0003$ $(.0014)$ | yes |
| .01 | $-.0003$ $(.0004)$ | $-.0008$ $(.0006)$ | yes |
| .001 | $-.0004$ $(.0049)$ | $-.0061$ $(.0070)$ | no |

**Negative control.** The separation could come from the cardiac background alone. The negative control is the unmodified cardiac background, with no implant. There the designs differ by $B - A = +0.0031$ (s.e. $0.0009$) for the likelihood learner and $+0.055$ (s.e. $0.013$) for the moment learner. The background is not exchange-symmetric, and the information operator predicts the contrast's direction, with a positive local contrast of $0.0013$, and predicts the likelihood learner's covariance estimation error as $28.96/29.78$ under $A/B$, against $29.64/29.80$ measured, as Theorem 3(ii) expects where the operator's eigenvalues are large. The implant multiplies the likelihood learner's separation by factors $18$ and $130$ at the two smaller $\varepsilon$ values. With the two-component truth and no implant, the mean shift kept, $B$ minus $A$ in excess NLL lies within $0.01$ in $9$ of $10$ seeds at one and two components and in $8$ of $10$ at four, where one further seed lies $0.015$ below. The remaining seed is common to all three: there design $B$ selects the smallest floor on a near-tie between floors and scores $11.2$ (two components) and $25.9$ (one component).

### G.8 Floor sensitivity

The primary floor grid contains $\varepsilon$ at $\varepsilon = 0.01$ and $0.001$, and every seed selects that value there, so the separation could depend on the grid. The sensitivity analyses replace the primary floor grid with grids that bracket $\varepsilon$, each on six matched seeds (the primary grid uses 30). The separation in the recovered implanted variances (the ratio and relative MSE of Table 9) holds under every grid; the KL levels, and at $\varepsilon = 0.001$ the KL contrast, move with the floor. At $\varepsilon = 0.01$, the bracketing grid $\{3 \times 10^{-3}, 3 \times 10^{-2}\}$ gives the likelihood learner's $B - A = 0.052$, against $0.055$ with the primary floor grid. At $0.001$, the grid $\{5 \times 10^{-4}, 2 \times 10^{-3}\}$ gives $0.130$ (s.e. $0.007$), and the grid $\{3 \times 10^{-4}, 3 \times 10^{-3}\}$ gives $A = 0.386$ and $B = 0.407$ with contrast s.e. $0.13$. Under this last grid every seed selects the floor $3 \times 10^{-4}$, below $\varepsilon$, under $A$ (the row of Table 12 with the floor below $\varepsilon$), and the likelihood learner's relative MSE of the implanted variances under $B$ is $412$.

Splitting the KL of one seed along the eigenvectors of $C_\varepsilon^{-1/2} \widehat{C}\, C_\varepsilon^{-1/2}$ attributes $0.002$–$0.05$ of $A$'s KL and $0.13$–$0.14$ of $B$'s to the implanted subspace $\mathrm{range}(U)$ across floors; the remaining directions have small operator eigenvalues. The held-out likelihood on the selection instances separates the candidate floors by $0.2$–$1.1$ under $A$ and $0.001$–$0.46$ under $B$.

## H The structural prior: target and weighting

This appendix establishes that the target, not the weighting, carries the recovery of Section 4.5. The structural target $\Sigma_{\mathrm{s}}$ keeps the within-phase residual covariances, estimated with ridge $1\%$ of the mean diagonal from the videos observing each phase (3,630 per phase and 1,230 per pair of phases under pairs, pilot included). After block standardisation it fits $\exp[-(\Delta t/\ell)^\alpha]$ to the pairwise temporal correlations, takes the separable product and maps the result back to coefficient space; the fitted $\ell$ lies in $0.11$–$0.14$ with $\alpha = 0.4$. The observed correlations under pairs at phase distances $0.2$, $0.4$ and $0.6$ average $0.30$, $0.20$ and $0.15$ in the likelihood learner's fits and $0.32$, $0.22$ and $0.17$ in the moment learner's. The target has 3,446 free parameters (the four within-phase covariances, $\ell$ and $\alpha$) against 13,530 for the unrestricted covariance.

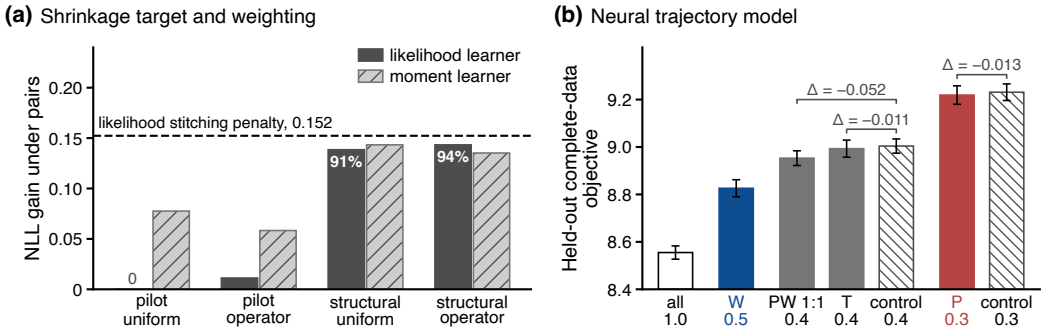



Figure 5: **A structural prior fitted from the same observations recovers most of the likelihood learner's stitching penalty, and no tested policy's interval admits a neural-model cost of $0.10$, the pre-specified margin, over segment-matched dropout.** (a) Held-out NLL gain under pairs ($30$ assignments) from shrinking each fitted covariance toward the pilot covariance or a structural prior, uniformly or with operator weighting, which is stronger along the least-informed information directions; dashed: the likelihood learner's stitching penalty, $0.152$, of which uniform shrinkage toward the structural prior recovers $91\%$ ($95\%$ interval of the gain $[0.131, 0.146]$); operator weighting adds $0.005$, within that interval. (b) Held-out complete-data objective of the neural trajectory model (lower is better; mean $\pm$ s.e., thirty seeds) with complete data (all), four budget-matched policies, and beside each policy its segment-matched control, which keeps each complete-data segment at the policy's inclusion rate for it (hatched); segment shares under the bars. pairs and triples are equivalent to their controls within $\pm 0.10$ (paired differences $-0.013$ and $-0.011$, s.e. $0.035$ and $0.031$); pairs+windows 1:1 lies $0.052$ below its control (s.e. $0.029$), in the policy's favour. Labels as in Figure 3.

Each fitted covariance $\widehat{\Sigma}$ is shrunk toward a target $\Sigma_0$, either $\Sigma_\mathrm{s}$ or the covariance $\Sigma_\mathrm{p}$ of the thirty completely observed pilot videos. Uniform shrinkage is $(1-\rho)\Sigma_0 + \rho\widehat{\Sigma}$, with $\rho$ selected on the selection videos. Operator weighting expands the target-whitened deviation $\Sigma_0^{-1/2}\widehat{\Sigma}\,\Sigma_0^{-1/2} - I$ in the information directions of $\mathcal{M}$ evaluated with $\Sigma_0$ in place of $\Sigma$ in the projectors $P_S$ of Theorem 1, and multiplies coordinate $k$ by $\mu_k/(\mu_k+\kappa)$, where $\mu_k$ is the $k$-th eigenvalue of that operator and $\kappa$ is selected on the selection videos. The empirical-Bayes rule sets $\kappa$ without selection, transferring the signal variance of the best-informed directions to all others. Every shrunk model keeps the original fitted mean and smallest fitted eigenvalue, through the final eigenvalue floor that keeps it positive definite, so the moment learner's floor tax is held fixed.

Under every policy other than pairs, operator weighting changes the structural-target gain by at most $0.0011$ for either learner. The empirical-Bayes rule over-shrinks the likelihood learner's fits under the intermediate policies and under-shrinks the moment learner's fits.

### H.1 The pilot target

Where a fit departs from its target depends on the target. Under every policy other than windows, where every direction is equally informed, $59$–$81\%$ of the squared Frobenius norm of the moment learner's deviation from $\Sigma_\mathrm{p}$ lies in the quarter of the information directions with the smallest $\mu_k$, against $3$–$4\%$ of its deviation from $\Sigma_\mathrm{s}$. Under pairs the gains of operator-weighted shrinkage toward $\Sigma_\mathrm{p}$ (Table 15) have standard errors $0.005$ (moment) and $0.001$ (likelihood). Under the other policies the same shrinkage gains the moment learner $0.058$ nats per coefficient under pairs+windows 3:2, $0.039$ under pairs+windows 1:1, $0.019$ under singles+triples, $0.0075$ under triples and below $0.0001$ under windows, and the likelihood learner at most $0.0006$ under every policy other than pairs. The pairs models of the two-policy uncertainty study (Appendix D), fitted separately at a budget of $14{,}424$ frames, reproduce these gains within $0.0001$ for both learners.

Table 15: The structural target recovers $91\%$ of the likelihood learner's stitching penalty under pairs with uniform shrinkage and $94\%$ with operator weighting; the pilot target recovers at most $7\%$. Structural-prior study: gain in held-out negative log-likelihood per coefficient on the evaluation videos (fitted minus shrunk), mean over thirty assignments. Percentages are shares of the learner's gap under that policy (likelihood learner: to windows, $0.1523$ under pairs; moment learner: to the likelihood learner, $0.3194$ under pairs); windows gains are absolute. Brackets: paired $95\%$ Student-$t$ intervals over the assignments.

| | Likelihood gain | Moment gain |
|---|---|---|
| Structural target, uniform shrinkage | | |
| windows | .0022 | .0023 |
| pairs+windows 1:1 | .0052 (75%) | .0782 (39%) |
| triples | .0069 (73%) | .0776 (39%) |
| pairs+windows 3:2 | .0070 (70%) | .1066 (46%) |
| singles+triples | .0090 (72%) | .1040 (46%) |
| pairs | .1385 [.1311, .1460] (91%) | .1433 [.1269, .1597] (45%) |
| pairs, other targets and weightings | | |
| structural / operator | .1433 (94%) | .1352 (42%) |
| structural / empirical Bayes | .1334 (88%) | .0293 (9%) |
| pilot / uniform | .0000 (0%) | .0776 (24%) |
| pilot / operator | .0112 (7%) | .0583 (18%) |

### H.2 The structural target with the floor also selected

Selecting the floor together with the shrinkage lets the structural target replace the moment learner's covariance. With the floor held fixed, the shrunk moment models sit $0.331$ nats per coefficient above windows under pairs and $0.131$–$0.139$ above under the four intermediate policies, at the scale of the floor taxes $T(0.3) = 0.343$ and $T(0.1) = 0.131$ of Equation (5). With the floor also selected, from $0.01$ and the target's smallest eigenvalue, selection sets $\rho = 0$ in $149$ of the $150$ fits outside windows. The resulting negative log-likelihood under pairs, $0.538$, is $0.4115$ below the fitted moment model's (paired $95\%$ interval $[0.3979, 0.4250]$), and under the intermediate policies it is $0.523$–$0.529$. The target alone, with the likelihood learner's means and floor $0.01$, scores $0.517$–$0.519$: $0.04$ above the likelihood learner under windows ($0.4774$) and $0.11$ below the stitched likelihood learner.

## I The neural trajectory model: objective, controls and sample metrics

This appendix establishes that the neural trajectory model's objective is design-unbiased and factorises over adjacent-phase segments, so that a design reaches the model through the segments it observes at both ends, and that segment-matched dropout tests exactly this channel.

### I.1 An evaluable design-unbiased objective

The model's trajectory runs through the end-diastolic deformation $y_0 = 0$, the interior deformations $y_1, \dots, y_4$ and the end-systolic deformation $y_5 = E$, with $y_k$ at the fixed phase knot $\tau = k/5$ (Section 4.1). The interpolant's time $\tau$ is the cardiac phase and its base is a point mass at the zero end-diastolic deformation. For video $i$ the interpolant follows the piecewise-linear centreline through $y_{i0}, \dots, y_{i5}$ plus a Brownian perturbation whose scale falls linearly to zero at $\tau = 1$; a drift network conditioned on the end-diastolic features is regressed on the interpolant's velocity, and samples are drawn by Euler–Maruyama integration of the learned drift. The objective is built from the five adjacent segments $(k, k+1)$; it uses no pair of non-adjacent observed times and no interpolation across an unobserved intervening knot. For $\ell_{ik}(\theta)$ the drift network's squared deviation from the interpolant's velocity target, averaged over the interpolant noise and integrated over the phases $\tau \in [k/5, (k+1)/5]$ of segment $k$ of video $i$, and $\pi_{ik}$ the joint inclusion probability of that

segment's two endpoints, define

$$\widehat{L}(\theta) = \frac{1}{n} \sum_{i} \sum_{k=0}^{4} \frac{\mathbf{1}\{y_{ik}, y_{i,k+1} \text{ observed}\}}{\pi_{ik}} \ell_{ik}(\theta). \tag{30}$$

With positive known joint inclusion probabilities and outcome-independent assignment, the expectation of $\widehat{L}$ over the design is the complete-data objective $L(\theta) = \frac{1}{n} \sum_i \sum_{k=0}^{4} \ell_{ik}(\theta)$. Because the end-diastolic and end-systolic frames are always available, the inclusion probabilities of the first and last segments are the marginal observation probabilities of $y_1$ and $y_4$. The implementation samples the interpolation time uniformly and applies the weight of the corresponding segment.

A single observed frame does not determine a velocity target. Let $B(t) \in \mathbb{R}^{1\times 4}$ hold the values at phase $t$ of the four endpoint-zero cubic modes of one contour coordinate (the nonzero entries of that coordinate's row of $F(t)$, Appendix D). A unit coefficient perturbation $b \in \mathbb{R}^4$ can satisfy $B(0.4)b = 0$ while $B'(0.4)b = 2.44$, with unchanged endpoints. Adjacent observed endpoints instead define the piecewise-linear segment and its derivative.

The network receives the $42$ end-diastolic geometry features, not the recorded end-systolic deformation, which enters as the endpoint $y_5$ of the training and reference contour trajectories. The Gaussian models of the six-policy acquisition study instead condition on it through $x$, so the neural model's scores are on their own scale and are compared across acquisition policies. The drift network is a multilayer perceptron on the state, a Fourier embedding of the phase and the end-diastolic features, trained for $20{,}000$ updates with model selection by the complete-data objective on the selection videos; its layer and optimiser settings are shared across conditions and released with the code. Sample metrics use $16$ Euler–Maruyama samples per held-out evaluation video, read at the four interior phases, in standardised units: each coordinate is divided by the training standard deviation of the end-systolic deformation, so the point mass stays at zero. For a video, let $S_s$ stack the $164$ coordinates of sample $s$ at the four interior phases and $T$ the observed deformations there. The energy score (Gneiting & Raftery, 2007) is $\frac{1}{16}\sum_s \|S_s - T\| - \frac{1}{2\cdot 16^2}\sum_{s,s'} \|S_s - S_{s'}\|$ (lower is better), and the sample-mean RMSE is the root mean square over the $164$ coordinates of the difference between the mean of the $16$ samples and $T$. Both are averaged over the evaluation videos.

### I.2 The design reaches the model through its usable segments

Because the objective sums over adjacent-phase segments and the sampler is Markov in phase, the learned joint factorises over adjacent phases, and a design reaches the model only through the segments it observes at both ends. Two random-dropout controls, each with the complete-data objective as its expected objective, test this channel. The segment-matched control keeps segment $k$ of the complete data independently with probability $\pi_k$, the policy's own inclusion rate for that segment, and weights it by $1/\pi_k$, so coverage matches segment by segment. The equivalence margin, $\pm 0.10$ on the objective, is pre-specified. The energy score and sample-mean RMSE of each policy agree with those of its segment-matched control within $1.1$ standard errors in every comparison.

The uniform-share control is a secondary check that matches the policy's overall usable-segment share rather than each segment's rate: it keeps every segment of the complete data with probability equal to that share and weights it by the inverse share. At ten seeds each policy minus it is $-0.064$ (s.e. $0.072$), $-0.055$ ($0.042$) and $-0.058$ ($0.052$) under pairs, pairs+windows 1:1 and triples. Those fits place pairs $0.397$ (s.e. $0.057$) above windows, and their sample-quality metrics differ from the control's by less than one standard error.

## J Reproducibility and compute

This appendix establishes the compute budget of each study and the licences under which the data are used. All Gaussian experiments run on $48$-core CPU nodes; the constructed domain takes $56.5$ hours of eight-thread computation, and the structural-prior study $32.5$ hours plus $14.7$ for the structural target with the floor also selected. The neural trajectory model trains on A100 GPUs, with ten seeds per condition for the uniform-share control and thirty for the segment-matched comparison. The code is released with the paper; EchoNet-Dynamic (Ouyang et al., 2020) is used under its non-commercial research-use agreement, and the Beijing air-quality (Zhang et al., 2017) and electricity-

Table 16: Over thirty seeds pairs and triples are equivalent to their segment-matched controls within the pre-specified margin of $\pm 0.10$, pairs+windows 1:1 lies $0.052$ below its control, in the policy's favour, and no tested policy's interval admits a cost as large as the margin. Neural trajectory model: held-out complete-data objective on the evaluation videos, mean (standard error) over thirty seeds, lower is better, and each policy minus its segment-matched control, paired by seed. Budgeted policies use $14{,}400$ interior frames on the acquisition videos; complete data uses all $28{,}968$ frames. The control rows are the segment-matched controls, which keep segment $k$ of the complete data at the stated rate and weight it by the inverse rate.

| Condition | Objective | Minus segment-matched control (s.e.) [paired 95% interval] |
|---|---|---|
| complete data | $8.555$ $(.028)$ | |
| windows | $8.826$ $(.036)$ | |
| pairs+windows 1:1 | $8.953$ $(.031)$ | $-.052$ $(.029)$ $[-.111, +.007]$ |
| triples | $8.993$ $(.036)$ | $-.011$ $(.031)$ $[-.075, +.052]$ |
| pairs | $9.219$ $(.039)$ | $-.013$ $(.035)$ $[-.084, +.059]$ |
| control at rates $(0.50, 0.33, 0.33, 0.33, 0.50)$ | $9.004$ $(.030)$ | |
| control at rates $(0.50, 0.17, 0.17, 0.17, 0.50)$ | $9.231$ $(.035)$ | |

load (Trindade, 2015) data under the Creative Commons Attribution 4.0 licence of the UCI Machine Learning Repository.

## K Further distinctions from prior work

Design criteria in survey, longitudinal and two-phase studies score one parameter at a time or equalise counts. At matched cost, split-questionnaire forms estimate a between-set covariance less efficiently than complete forms below a correlation near $0.6$ (Rhemtulla et al., 2016). Multiform designs over occasions are compared by the simulated bias and power of single parameters (Chen & Liu, 2026). A-optimal splits for means improve on random splits by at most a factor of two (Jang et al., 2022), while for a small covariance eigenvalue Theorem 2 makes the factor between coverage-matched designs unbounded. Topic-spreading modules give better bivariate estimates after imputation than single-topic modules (Axenfeld et al., 2022; 2026), though their pairwise counts change. Concurrence is a sufficient design statistic for treatment contrasts with fixed block effects (Bailey & Cameron, 2009), and the six-variable extension of Construction 1 (Appendix A.4) matches it and keeps the separation of Theorem 2. Pilot, longitudinal and three-form designs choose which times or item sets are observed (Huang & Kao, 2025; Ji & Müller, 2017; Imbriano & Raghunathan, 2020). Two-phase and complete-partial designs minimise one estimand's asymptotic variance (Tao et al., 2020; Mareis, 2025). Bayesian D-optimal longitudinal designs trade subjects against measurements per subject (Yi et al., 2021). Budgeted covariance sampling chooses each variable's observation probability (Pavez & Ortega, 2018). METRIK fits an input mask to one learner (Lala & Jha, 2024). This paper holds these counts fixed, varies how co-observations are grouped into instances, and scores the joint's leading predictive risk $\frac{1}{2}\operatorname{tr}(\mathcal{M}^{-1})$ (Theorem 3(ii)), which complete co-observation minimises at a fixed budget for every covariance (Corollary 4).

Estimation from incomplete or corrupted observations is analysed through observation probabilities, operator counts or structure. Covariance estimation from $m$ of $d$ coordinates has a minimax effective sample size of order $m^2/d^2$, set by subset size, not arrangement (Azizyan et al., 2018). Bounds for unbiased estimators under non-uniform, dependent observation depend only on marginal and pairwise observation probabilities (Pavez & Ortega, 2021). Missing-covariate regression uses corrected matrices that can be indefinite (Loh & Wainwright, 2012), and along partly observed low-variance directions this forces the moment learner's floor (Theorem 3(iii)). Inverse-probability-weighted principal components degrade under heterogeneous missingness where an iterative refit does not (Zhu et al., 2022), which Corollary 5 states for the whole covariance. The model dimension and the number of distinct measurement operators decide whether a signal model can be learned from incomplete measurements (Tachella et al., 2023). Toeplitz and low-rank structure make pairwise co-observation sufficient (Eldar et al., 2020; Cao et al., 2023), the regime of Section 4.5. Theorem 2 gives the rate at which identifiable designs sharing these probabilities separate under an unrestricted covariance.

Identification of a joint from overlapping margins rests on structure. Chordal completion and the marginal problem characterise when partial Hermitian matrices have positive definite completions and consistent marginals extend to a joint (Grone et al., 1984; Vorob'ev, 1962), guaranteeing existence, not uniqueness. Statistical matching studies uncertainty about unobserved cross-file associations (D'Orazio et al., 2006; Rässler, 2002). Low-rank tensor and coupled factor models identify discrete joints from selected lower-order marginals (Kargas et al., 2018; Ibrahim & Fu, 2021). Shared components of linear mixtures are identifiable from unpaired samples (Timilsina et al., 2024). Each modality's shared latent block is subspace-identifiable from pairwise-only data on a connected modality graph (Li et al., 2026). Mosaic single-cell integration needs a bridge dataset when datasets share no features (Ghazanfar et al., 2024). Contrastive spaces transfer through a shared modality to the never-paired pair (Wang et al., 2023). Multimodal contrastive learning gains from unpaired data once hidden pairs are recovered (Nakada et al., 2023). Gram-volume alignment needs every modality on the instance (Cicchetti et al., 2025). Out-of-variable prediction reuses marginal information from overlapping variable sets (Guo et al., 2024). These works establish identification or build methods; this paper fixes the learner and coverage and compares stitched and co-observed designs for a Gaussian joint at finite sample size.