Counterfactual Evaluation of Temporal Observation Protocols

Xizhe Zhang ZHANGXIZHE@GMAIL.COM
ORCID: 0000-0002-8684-4591

Abstract

We study counterfactual protocol evaluation: whether data collected under a realised observation protocol determine the predictive value of alternatives that were never deployed. Protocol value is the population R^2 of the Bayes-optimal predictor of a fixed trajectory-level target from the measurements an alternative would collect. We show that even infinite benchmark data need not determine this value: distinct latent covariance structures can induce the same benchmark measurement–target law while assigning different values to the same alternative. We develop a value-specific identification theory in which only latent ambiguity that changes the alternative's value matters. For linear targets, invisible covariance directions certify non-identification, while targeted measurements can restore identification without recovering the full latent covariance; an exact permutation construction extends the result to nonlinear aggregate targets. With finite dense calibration data, uniform error bounds control protocol-selection regret and distinguishable value gaps. Exact marginal gains then support cost-constrained, target-aware observation design. Simulations and retrospective analyses of Sleep-EDF and Long-Term AF show that broad temporal-layout differences can be more reliably distinguished than fine placements selected from finite data. Together, these results connect identification, calibration resolution and observation design for undeployed protocols.

Keywords: identifiability, observational equivalence, observation design, temporal aggregates, calibration

1 Introduction

Predictive performance in temporal learning depends not only on the learner, but also on the observation protocol under which the data are collected. Standard benchmarks typically hold this protocol fixed and compare learners on the resulting data. This leaves a different question unanswered: what predictive performance could be achieved under an undeployed observation protocol? We refer to this problem as *counterfactual protocol evaluation*. The central question is whether the joint measurement–target distribution under the realised protocol determines the predictive value of an undeployed protocol.

This question is particularly important when the prediction target is defined over a substantially longer time horizon than the observed input. In such settings, the realised protocol may omit temporal information relevant to the target. Examples include predicting whole-night sleep-stage proportions from partially observed sleep records and long-horizon atrial-fibrillation burden from shorter monitoring windows. Weak predictive performance may therefore reflect limitations of the learner, limited sample size, or target-relevant information excluded by the observation protocol. Better models and larger datasets can address the first two limitations, but the third requires changing what is observed.

Several lines of work address related questions. Methods for incomplete or irregular temporal data learn prediction rules from irregularly or incompletely observed training data (Che et al., 2018; Rubanova et al., 2019; Kidger et al., 2020). Experimental-design and sensor-placement methods, including optimal design for longitudinal data, choose future measurements to optimise a criterion

under an assumed or estimated model (Chaloner and Verdinelli, 1995; Krause et al., 2008; Ji and Müller, 2017). Bayes-error analyses characterise predictive limits for a given feature–target distribution (Fukunaga and Hummels, 1987; Ishida et al., 2023). Most closely related, active feature acquisition performance evaluation studies how to evaluate a new acquisition policy from data collected under a different retrospective acquisition process, under identification assumptions that include sufficient support for the acquisitions required by the target policy (von Kleist et al., 2025; Precup et al., 2000; Kallus and Uehara, 2020). In our setting, an alternative protocol may instead require measurements that were never collected under the realised protocol. This leaves a different identification question: whether the joint measurement–target distribution generated by one realised observation protocol determines the predictive value of another protocol whose measurements were never collected.

We develop a theory of counterfactual protocol evaluation centred on the information needed to determine the value of an undeployed protocol. The value of an alternative protocol is identified precisely when it is constant across all latent models that induce the same realised measurement–target law (Koopmans and Reiersøl, 1950; Rothenberg, 1971). This criterion is value-specific: the value may be identified even when the latent covariance is not, provided that the remaining ambiguity does not change the value being evaluated. Under a Gaussian latent-trajectory model with noisy linear measurements, we show that this condition can fail even with infinite data under the realised protocol, because distinct latent covariances can generate the same joint law of the realised measurements and target while assigning different values to the same alternative. For linear targets, we characterise the unresolved covariance variation through perturbations invisible to the realised law; nonzero directional sensitivity of the alternative's value along such a perturbation certifies local non-identification. Targeted augmentation can remove the value-changing ambiguity and identify the value of a specified alternative without recovering the full latent covariance. A stationary construction gives the sharp minimal failure case in the stationary, standardised, one-observation setting with a mean target, while a permutation construction establishes exact non-identification for nonlinear aggregate targets.

Targeted augmentation may be sufficient when the value of a specified alternative is sought. For an entire candidate family, densely observed calibration trajectories provide a stronger, family-wide route: at the population level, their law identifies the latent covariance used to evaluate all candidates. With only finitely many such trajectories, two questions remain: how accurately candidate values can be distinguished and how those distinctions should guide design. We derive uniform bounds that propagate covariance-estimation error to protocol-value error over the candidate family. The resulting bounds yield selection-regret guarantees and a calibration resolution below which value differences cannot be distinguished reliably. They also link the useful granularity of the candidate class to the amount of calibration data: finer classes can improve the attainable value, but their increasingly small differences support reliable selection only when the data can resolve them. At that resolution, we formulate cost-constrained, target-aware observation design and derive exact rank-one marginal gains for adding measurements under constraints on timing, support, repetition, noise, and cost. Simulations examine calibration error, protocol selection, and design under known data-generating laws. Retrospective analyses of Sleep-EDF and Long-Term AF annotations compare alternative temporal layouts at matched observation budgets (Kemp et al., 2000; Mourtazaev et al., 1995; Goldberger et al., 2000; Petrutiu et al., 2007); the pooled analyses show clearer differences between pre-specified dispersed and contiguous layouts than among learned Sleep supports, whose locations and held-out advantages vary across targets and resamples.

Taken together, these results place identification before optimisation in temporal observation design. The information required to evaluate a protocol is not generally the information required to recover the full latent dependence structure; it is the information needed to eliminate ambiguity that changes the protocol values under consideration. Finite calibration then determines the resolution at which those values can be compared, and the granularity of the design problem should be matched to that resolution. This value-directed view links the choice of additional measurements, the amount of calibration data, and the level of detail at which candidate protocols can be compared and optimised.

The remainder of the paper is organised as follows. Section 2 defines temporal aggregate targets, observation protocols, and protocol value. Section 3 develops value-specific identification and targeted augmentation. Section 4 studies finite calibration and protocol comparison, and Section 5 develops target-aware observation design. Section 6 reports the simulations and retrospective analyses. Section 7 reviews related work, and Sections 8 and 9 present the discussion and conclusion.

2 Problem Formulation and Protocol Value

Consider a benchmark collected under one observation protocol and a future study that may use another. The benchmark contains the measurements produced by the realised protocol together with a trajectory-level target used for prediction. The full latent trajectory need not be observed. An alternative protocol may include measurements that were not collected in the benchmark. We first define the target and protocol measurements in continuous time and then introduce the discrete Gaussian model used in the theoretical analysis.

An *observation protocol* is a finite collection of acquisition actions. An *observation design* is the decision problem of choosing a protocol from a feasible family. The *realised protocol* is the one that generated the benchmark data; an *alternative protocol* is a feasible protocol whose predictive value is to be evaluated even though the benchmark does not contain the complete measurement vector that it would produce.

2.1 Temporal Aggregate Targets and Observation Protocols

For unit i, let $Z_i = \{Z_i(t) : t \in [0, T]\}$ be a latent scalar trajectory. The supervised target is one number obtained from the whole trajectory. Specifically, we first transform the state at each time by a function g and then average the transformed values over the observation horizon:

$$\Theta_{i,g} = \int_0^T \omega_T(t)\, g\{Z_i(t)\}\, \mathrm{d}t, \qquad \omega_T(t) \geq 0, \qquad \int_0^T \omega_T(t)\, \mathrm{d}t = 1. \tag{1}$$

The weight function ω_T specifies which parts of the horizon contribute to the target. Uniform weights, $\omega_T(t) = 1/T$, give an ordinary time average; non-uniform weights can emphasise particular periods.

Two choices of g recur throughout the paper. If $g(z) = z$, then $\Theta_{i,g}$ is the average level of the latent process, such as average electricity demand over a day. If $g_c(z) = \mathbf{1}\{z > c\}$, then Θ_{i,g_c} is the weighted fraction of time for which the process exceeds c, and therefore represents exceedance-time burden. Sleep-stage proportions, atrial-fibrillation burden and time spent in a symptomatic state are examples of occupation-type targets. The target is fixed when protocols are compared: changing the protocol changes what is measured, not the quantity to be predicted.

One acquisition action a returns a noisy linear measurement of the trajectory, $Y_{i,a} = L_a Z_i + \varepsilon_{i,a}$, where L_a may represent a point measurement or an average over a time window. We assume $\varepsilon_{i,a} \sim \mathcal{N}(0, r_a)$, independently across actions and units and independently of Z_i. Each action has

cost c_a. For a cost budget $\mathcal{B} > 0$, a protocol is a finite set $S = \{a_1, \dots, a_D\}$ satisfying $\sum_{a \in S} c_a \le \mathcal{B}$, and $\Pi_{\mathcal{B}}$ denotes the collection of all such feasible protocols. Identification and estimation use this action representation; Section 5 specifies how actions encode timing, support, repetition, noise and cost.

2.2 Discrete Gaussian Model

The theoretical analysis uses a grid $t_1 < \cdots < t_p$ on $[0, T]$. We collect the latent states on this grid in $Z_i = (Z_i(t_1), \dots, Z_i(t_p))^\top$ and assume

$$Z_i \sim \mathcal{N}(0, K), \qquad \operatorname{diag}(K) = \mathbf{1}_p. \tag{2}$$

Thus K records the dependence between times, while every marginal state has mean zero and variance one. This standardisation fixes the scale on which g is defined. Here $\operatorname{diag}(M)$ is the vector of diagonal entries of M, whereas $\operatorname{Diag}(v)$ is the diagonal matrix with diagonal v.

Let $\omega = (\omega_1, \dots, \omega_p)^\top$ be non-negative quadrature weights with $\mathbf{1}_p^\top \omega = 1$. The continuous target and the protocol measurements then become

$$\Theta_{i,g} = \sum_{j=1}^{p} \omega_j \, g(Z_{ij}), \qquad Y_{i,S} = A_S Z_i + \varepsilon_{i,S}, \qquad \varepsilon_{i,S} \sim \mathcal{N}(0, R_S). \tag{3}$$

Each row $\ell_a^\top$ of $A_S \in \mathbb{R}^{D \times p}$ describes one action. A point measurement at t_j has $\ell_a = e_j$. A window measurement has non-negative entries on the grid points covered by that window and is normalised so that $\mathbf{1}_p^\top \ell_a = 1$. With independent action noise, $R_S = \operatorname{Diag}(r_a)_{a \in S}$. In short, K describes the latent trajectory, ω describes the target, and (A_S, R_S) describes what protocol S observes.

The empty protocol has no observations and explains no target variation. We set

$$Q_\varnothing(K) = 0_{p \times p}, \qquad \mathcal{I}_g(\varnothing; K) = 0. \tag{4}$$

2.3 Protocol Value under the Gaussian Model

We value a protocol by the fraction of target variance explained by the best population predictor based on its measurements.

Definition 1 (Bayes value of an observation protocol)**.** Let $\Theta \in L^2$ be a scalar target with $\operatorname{Var}(\Theta) > 0$. The Bayes protocol value under squared loss is

$$\mathcal{I}_{\mathrm{Bayes}}(S) = \frac{\operatorname{Var}\{\mathbb{E}(\Theta \mid Y_S)\}}{\operatorname{Var}(\Theta)}. \tag{5}$$

It is the population R^2 of the optimal predictor based on Y_S and lies in $[0, 1]$.

The best-linear analogue restricts the predictor to be affine. With $\Sigma_S = \operatorname{Var}(Y_S)$ and $c_S = \operatorname{Cov}(Y_S, \Theta)$, it is

$$\mathcal{I}_{\mathrm{L}}(S) = \frac{c_S^\top \Sigma_S^+ c_S}{\operatorname{Var}(\Theta)}, \tag{6}$$

where Σ_S^+ is the Moore–Penrose inverse (Penrose, 1955). It depends only on second moments and satisfies $\mathcal{I}_{\mathrm{L}}(S) \le \mathcal{I}_{\mathrm{Bayes}}(S)$, with equality when $\mathbb{E}(\Theta \mid Y_S)$ is affine in Y_S.

Equations (5) and (6) define population quantities. The empirical analyses evaluate fitted predictors by pooled held-out cross-fitted R^2.

Under the discrete Gaussian model, the Bayes value can be calculated from K, g and the protocol matrices. The target transformation g converts a latent Gaussian correlation into a covariance as follows. For standard bivariate normal variables (U, V_r) with correlation r, define

$$C_g(r) = \mathrm{Cov}\{g(U), g(V_r)\} = \sum_{k\geq 1} \frac{a_k^2}{k!}\, r^k, \qquad g \in L^2(\phi), \tag{7}$$

where ϕ is the standard normal distribution and $g - \mathbb{E}g = \sum_{k\geq 1}(a_k/k!)H_k$ is the Hermite expansion of g in $L^2(\phi)$. In particular, $\mathrm{Cov}\{g(Z_j), g(Z_k)\} = C_g(K_{jk})$. All dependence on the possibly nonlinear target function enters the value calculation through C_g.

The protocol enters through the part of the latent trajectory recoverable from its measurements. Let $\Sigma_S(K) = A_S K A_S^\top + R_S$. For a non-empty protocol with nonsingular $\Sigma_S(K)$, Gaussian conditioning gives

$$Q_S(K) = K A_S^\top \Sigma_S(K)^{-1} A_S K = \mathrm{Cov}\{\mathbb{E}(Z \mid Y_S)\}. \tag{8}$$

Thus $Q_S(K)$ is the covariance of the part of the trajectory recovered from Y_S. The residual covariance used in the marginal-gain calculation is introduced in Section 5.

Combining this covariance transform with Gaussian conditioning yields the protocol-value identity below (Zhang, 2026).

Proposition 2 (Gaussian protocol-value identity; Zhang, 2026)**.** *Under* (2)–(3)*, let* $g \in L^2(\phi)$ *and assume* $V_g(K) > 0$*. Define*

$$V_g(K) = \sum_{j,k=1}^{p} \omega_j \omega_k\, C_g(K_{jk}), \qquad F_g(S;K) = \sum_{j,k=1}^{p} \omega_j \omega_k\, C_g\{Q_S(K)_{jk}\}. \tag{9}$$

Then $V_g(K) = \mathrm{Var}(\Theta_g)$ *and* $F_g(S;K) = \mathrm{Var}\{\mathbb{E}(\Theta_g \mid Y_S)\}$*, and*

$$\mathcal{I}_g(S;K) = \mathcal{I}_{\mathrm{Bayes}}(S) = \frac{F_g(S;K)}{V_g(K)}. \tag{10}$$

Thus $V_g(K)$ is total target variance, $F_g(S;K)$ is the part explained by protocol S, and $\mathcal{I}_g(S;K)$ is the corresponding Bayes-optimal population R^2 based on Y_S. A proof is given in Appendix A.

For two target functions, the covariance transform has a simple form. For the mean target $g(z) = z$, $C_g(r) = r$, so $V_g(K) = \omega^\top K \omega$ and $F_g(S;K) = \omega^\top Q_S(K)\omega$; equivalently, $\Theta = \omega^\top Z$. For the occupation target $g_c(z) = \mathbf{1}\{z > c\}$, Plackett's identity (Plackett, 1954) gives

$$C_{g_c}(r) = G_c(r) = \int_0^r \frac{\exp\{-c^2/(1+s)\}}{2\pi\sqrt{1-s^2}}\, \mathrm{d}s, \tag{11}$$

with $G_0(r) = \arcsin(r)/(2\pi)$.

3 Value-Specific Identification

Section 2 showed how to calculate protocol value when the latent covariance K is known. We now ask what can be learned when the only available population information is the joint law of the measurements and target under the realised protocol A. Throughout this section, A and B denote the measurement operators induced by the realised and alternative protocols. Sections 4–5 return to the action-set notation S, with measurement matrix A_S.

We begin with the linear target $\Theta = h^\top Z$. Then (Y_A, Θ) is jointly Gaussian, so its full law is determined by second moments. We abbreviate $\mathcal{I}(B; K) = \mathcal{I}_g(B; K)$ for $g(z) = z$; in this linear-Gaussian case it also equals $\mathcal{I}_{\mathrm{L}}(B)$. We index grid points by $0, \dots, p-1$. Nonlinear aggregates require more than moment matching; we treat them separately through an exact permutation argument in Proposition 6.

The law of (Y_A, Θ) may fail to identify K even when it determines the scalar value of the specified alternative B.

3.1 Identification Criterion

For a benchmark law $\mathbb{P}_A$, define its observational equivalence class by

$$\mathcal{K}_A(\mathbb{P}_A) = \{K : \mathbb{P}_K(Y_A, \Theta) = \mathbb{P}_A\}. \tag{12}$$

Every $K \in \mathcal{K}_A(\mathbb{P}_A)$ generates exactly the same distribution of the benchmark data. No amount of additional sampling under the same protocol A can distinguish them. Consequently, for any target and model class under consideration,

$$\boxed{\begin{array}{c} \mathcal{I}_g(B;\cdot) \text{ is identified from } \mathbb{P}_A \\ \iff \quad \mathcal{I}_g(B; K_1) = \mathcal{I}_g(B; K_2) \quad \text{for all } K_1, K_2 \in \mathcal{K}_A(\mathbb{P}_A). \end{array}} \tag{13}$$

This criterion is value-specific: it requires uniqueness only of the specified alternative's value, not of every component of K. The class $\mathcal{K}_A(\mathbb{P}_A)$ may contain many covariance structures while assigning a unique value to B. Conversely, two members of this class with different values establish non-identification. This is functional identification over an observational-equivalence class in the sense of Koopmans and Reiersøl (1950).

For the linear target $\Theta = h^\top Z$, the realised benchmark law reveals three covariance blocks. Take $h = \omega$, and let $Y_A = AZ + \varepsilon_A$ with $\varepsilon_A \sim \mathcal{N}(0, R_A)$ and R_A known. Then (Y_A, Θ) is jointly Gaussian with covariance

$$\begin{pmatrix} AKA^\top + R_A & AKh \\ h^\top KA^\top & h^\top Kh \end{pmatrix}. \tag{14}$$

The Gaussian benchmark law therefore depends on K through the three blocks

$$AKA^\top, \qquad AKh, \qquad h^\top Kh, \tag{15}$$

which constitute its complete covariance information about K. The first block is the covariance of the realised measurements after subtracting the known measurement noise. The second is their cross-covariance with the target, and the third is the target variance. Any change in K that leaves all three blocks unchanged is invisible to the benchmark.

3.2 Minimal Stationary Counterexample

We first ask when invisible dependence can arise in the simplest stationary setting. A one-point benchmark reveals no lag information through $\mathrm{Var}(Y_A)$, because the process is standardised. It reveals only two weighted sums of the lag correlations: one through $\mathrm{Cov}(Y_A, \Theta)$ and one through $\mathrm{Var}(\Theta)$. These two equations determine all unknown lags on grids of at most three points. On a four-point grid they leave one degree of freedom for the first time, and moving along that direction can change the value of a protocol that observes (Z_1, Z_2).

Theorem 3 (Sharp minimal stationary counterexample)**.** *Let $p \geq 2$ and let Z be stationary and standardised on the grid $\{0, 1, \ldots, p-1\}$ with correlations $\rho(1), \ldots, \rho(p-1)$, let $\Theta = p^{-1} \sum_j Z_j$, and let A observe Z_0 with known noise variance ν^2. Then*

(i) for $p \leq 3$ the map from $(\rho(1), \ldots, \rho(p-1))$ to the observable functionals is injective, so the stationary lag-correlation vector, and hence K within this model class, is identified;

(ii) for $p \geq 4$ the space of stationary invisible directions has dimension $p - 3 \geq 1$.

For $p = 4$, the kernel is spanned by the lag perturbation

$$\delta = (\delta(0), \delta(1), \delta(2), \delta(3)) = (0, 1, -2, 1). \tag{16}$$

For any positive-definite base correlation matrix K with lag vector ρ, let $K_\pm$ have lag vectors $\rho \pm \varepsilon\delta$, with lag 0 fixed at one. For all sufficiently small $\varepsilon > 0$, these matrices are admissible and observationally equivalent under A. If B observes (Z_1, Z_2) with independent measurement noise $R_B = \nu_B^2 I_2$, where $\nu_B^2 \geq 0$, then

$$\mathcal{I}(B; K_\pm) = \frac{2b^2}{\{1 + \nu_B^2 + \rho(1) \pm \varepsilon\}\ \mathrm{Var}(\Theta)}, \qquad b = \frac{1 + 2\rho(1) + \rho(2)}{4}, \tag{17}$$

so the two values differ whenever $1 + 2\rho(1) + \rho(2) \neq 0$.

Proof idea. The realised point and the target reveal two independent weighted sums of the $p - 1$ unknown lag correlations. Their Jacobian has rank $\min\{2, p-1\}$, leaving $p - 3$ free directions exactly when $p \geq 4$; at $p = 4$, solving the two equations gives $(1, -2, 1)$. Along this direction the target variance and its covariance with Z_1 and Z_2 remain fixed, whereas the covariance of (Z_1, Z_2) changes, which yields (17). The calculation is given in Appendix A. □

This four-point threshold is specific to the stationary, standardised, one-observation model with a uniform mean target. Figure 1 uses $p = 4$, $h = \frac{1}{4}\mathbf{1}_4$, the baseline profile $\rho_0(u) = \exp(-u)$ (that is, $\tau = 1$), and the raw lag perturbation $\varepsilon = 0.1321$ in (16). Both protocols are noiseless, with $A = e_0^\top$ and $B = (e_1, e_2)^\top$. The two Toeplitz correlation matrices $K_\pm$ preserve $\mathrm{Var}(Y_A) = 1$, $\mathrm{Cov}(Y_A, \Theta) = 0.388250$ and $\mathrm{Var}(\Theta) = 0.428012$ to a maximum discrepancy of 10^{-16}. Substitution into (17) gives $\mathcal{I}(B; K_+) = 0.6817$ and $\mathcal{I}(B; K_-) = 0.8274$.

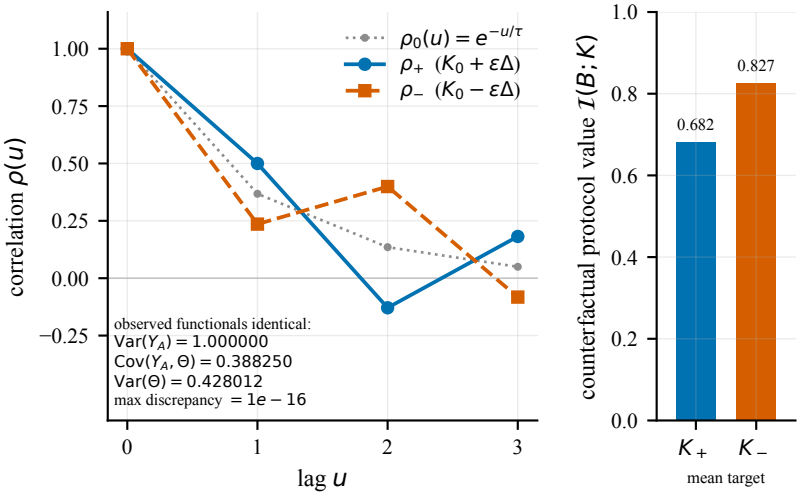

Figure 1: Four-point instance of Theorem 3. The profiles $\rho_\pm = \rho_0 \pm \varepsilon\delta$ are observationally equivalent under $A = \{Z_0\}$ but assign values 0.6817 and 0.8274 to $B = \{Z_1, Z_2\}$.

3.3 Invisible Directions

The vector $(0, 1, -2, 1)$ is one instance of a more general object: a covariance perturbation that changes latent dependence while preserving every covariance block visible under A.

Definition 4 (Invisible direction)**.** A symmetric $\Delta \in \mathbb{R}^{p\times p}$ with $\mathrm{diag}(\Delta) = 0$ is *invisible* to (A, h) if

$$A\Delta A^\top = 0, \qquad A\Delta h = 0, \qquad h^\top \Delta h = 0. \tag{18}$$

The three equalities preserve, respectively, the realised-measurement covariance, the realised-measurement–target cross-covariance and the target variance. The zero-diagonal requirement preserves the unit diagonal. For positive-definite K, sufficiently small perturbations $K \pm \varepsilon\Delta$ remain correlation matrices. We write $\mathcal{D}(A, h)$ for the linear space of directions invisible to (A, h).

Theorem 5 (Value-changing invisible directions imply non-identification)**.** *Let $0 \neq \Delta \in \mathcal{D}(A, h)$ and let $K_0 \succ 0$ be a correlation matrix. For all sufficiently small $|\varepsilon|$, $K_\varepsilon = K_0 + \varepsilon\Delta$ is a correlation matrix and has the same benchmark law as K_0; in particular,*

$$\mathbb{P}_{K_\varepsilon}(Y_A, \Theta) = \mathbb{P}_{K_0}(Y_A, \Theta). \tag{19}$$

Let B be an alternative protocol with noise $R_B \succeq 0$ such that $\Sigma_B(K_0) = BK_0B^\top + R_B \succ 0$, and set $W = \Sigma_B(K_0)^{-1}$. Its directional derivative is

$$\begin{aligned} D_B(\Delta; K_0) &:= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathcal{I}(B; K_0 + \varepsilon\Delta)\Big|_{\varepsilon=0} \\ &= \frac{2h^\top \Delta B^\top W B K_0 h - h^\top K_0 B^\top W (B\Delta B^\top) W B K_0 h}{h^\top K_0 h}. \end{aligned} \tag{20}$$

If $D_B(\Delta; K_0) \neq 0$, then the value functional $\mathcal{I}(B; \cdot)$ is not locally identified at K_0 within the model class: for every sufficiently small $\varepsilon > 0$, the observationally equivalent matrices $K_0 \pm \varepsilon\Delta$ assign different values to B.

Proof idea. The three invisibility equations preserve every covariance block of the Gaussian benchmark law. Differentiating the alternative-protocol value with the inverse rule gives (20); a non-zero derivative then separates the two compatible paths $K_0 \pm \varepsilon\Delta$ to first order. The full matrix calculation appears in Appendix A. □

Thus no estimator based only on (Y_A, Θ) can be consistent at both local alternatives, and collecting more units under the same protocol does not resolve the ambiguity.

If A has d rows, rank–nullity gives a lower bound on the dimension of the invisible space. A standardised K has $p(p-1)/2$ free entries whereas (15) supplies at most $d(d+1)/2 + d + 1$ linear constraints. Rank–nullity therefore gives

$$\dim \mathcal{D}(A, h) \geq \frac{p(p-1)}{2} - \frac{d(d+1)}{2} - d - 1, \qquad \text{when the right-hand side is positive.} \tag{21}$$

Near a positive-definite K_0, the equivalence class contains the corresponding open piece of the affine space $K_0 + \mathcal{D}(A, h)$ (Rothenberg, 1971). This count only measures ambiguity in K: value non-identification still requires sensitivity in (20). With fixed d, the contrast is nevertheless stark—the benchmark contributes $O(d^2)$ constraints to $O(p^2)$ covariance parameters.

3.4 Nonlinear Targets

The invisible-direction theorem uses joint Gaussianity of (Y_A, Θ) and therefore applies directly to a linear target. With a nonlinear aggregate, equality of the first two moments does not imply equality of the full benchmark law. The following symmetry argument instead constructs two exactly equivalent benchmark laws. The condition $A\Xi = A$ says that the permutation leaves every realised measurement unchanged, while $\Xi^\top \omega = \omega$ says that it preserves the aggregate weights.

Proposition 6 (Permutation equivalence for arbitrary targets)**.** *Let Ξ be a permutation matrix with $A\Xi = A$ and $\Xi^\top \omega = \omega$, and put $K' = \Xi K \Xi^\top$. For every $g \in L^2(\phi)$ with $V_g(K) > 0$, the joint laws of (Y_A, Θ_g) under K and under K' coincide, and*

$$\mathcal{I}_g(B; K') = \mathcal{I}_g(B\Xi; K) \qquad \textit{for every alternative protocol } B. \tag{22}$$

Consequently, whenever $\mathcal{I}_g(B\Xi; K) \neq \mathcal{I}_g(B; K)$, the value of B is not identified from the benchmark law.

Here $B\Xi$ is the algebraically permuted measurement operator used to compare values, with the same observation-noise covariance R_B as B; it need not belong to the original candidate family.

Proof. Let $Z \sim \mathcal{N}(0, K)$ and $Z' = \Xi Z \sim \mathcal{N}(0, K')$. Then $AZ' = A\Xi Z = AZ$, and since g acts entrywise $g(\Xi Z) = \Xi g(Z)$, so $\Theta_g(Z') = \omega^\top \Xi g(Z) = (\Xi^\top \omega)^\top g(Z) = \Theta_g(Z)$; in particular $V_g(K') = V_g(K) > 0$. The measurement noise is independent with the same law under both, so the joint laws agree. Applying the same substitution to B gives (22). □

Non-identification follows whenever the permutation changes the value of B. For example, no change occurs when $\Xi K \Xi^\top = K$. The next construction shows that three grid points suffice for every nonconstant target under uniform aggregation.

Example 7 (Three points suffice under uniform aggregation)**.** Take $p = 3$, $A = e_0^\top$, uniform ω, let Ξ transpose coordinates 1 and 2, and let $B = e_1^\top$ observe Z_1 with noise ν_B^2, writing $s = 1 + \nu_B^2$. For $0 < a < 1$ put

$$K_a = \begin{pmatrix} 1 & a & 0 \\ a & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad K_a' = \Xi K_a \Xi^\top. \tag{23}$$

Both are correlation matrices, $A\Xi = A$ and $\Xi^\top \omega = \omega$, so the benchmark laws coincide. Nevertheless, $\mathcal{I}_g(B; K_a) > \mathcal{I}_g(B; K_a')$ for every nonconstant $g \in L^2(\phi)$, every $a \in (0, 1)$ and every noise level. The explicit calculation, which reduces the gap to positive values of C_g, is given in Appendix A.

Without stationarity, three points suffice for arbitrary nonconstant square-integrable transformations under uniform aggregation; within the stationary standardised model with a one-point benchmark and a mean target, the sharp minimum is four.

3.5 Identification by Augmentation

Suppose A has already been run and a further acquisition A' is contemplated on the same units. Let A_{aug} stack the rows of A and A'. Their joint measurement reveals the covariance within A', its cross-covariance with A, and the corresponding target cross-covariances. Collect the revealed linear directions in

$$\bar{A} = \begin{bmatrix} A_{\mathrm{aug}} \\ h^\top \end{bmatrix}. \tag{24}$$

Let

$$\mathcal{D}(A_{\mathrm{aug}}, h) = \left\{ \Delta = \Delta^\top : \mathrm{diag}(\Delta) = 0,\ A_{\mathrm{aug}} \Delta A_{\mathrm{aug}}^\top = 0,\ A_{\mathrm{aug}} \Delta h = 0,\ h^\top \Delta h = 0 \right\} \tag{25}$$

be the invisible space of the augmented benchmark.

Proposition 8 (Exact sufficiency and a first-order failure certificate)**.** *Fix $K_0 \succ 0$ with $\mathrm{diag}(K_0) = \mathbf{1}$ and an alternative protocol B with known $R_B \succeq 0$ such that $BK_0B^\top + R_B \succ 0$.*

(i) Exact sufficiency. *If $\mathrm{row}(B) \subseteq \mathrm{row}(\bar{A})$, then the joint Gaussian law of (Y_B, Θ), and hence $\mathcal{I}(B; K)$, is identified exactly from the augmented benchmark, for every admissible K.*

(ii) First-order failure certificate. *If $D_B(\Delta; K_0) \neq 0$ for some $\Delta \in \mathcal{D}(A_{\mathrm{aug}}, h)$, then $\mathcal{I}(B; \cdot)$ is not identified near K_0: for every sufficiently small $\varepsilon > 0$ the correlation matrices $K_0 \pm \varepsilon\Delta$ induce the same law of $(Y_{A_{\mathrm{aug}}}, \Theta)$ and different values of B.*

The proof, including the row-space factorisation used in part (i), is given in Appendix A.

By contraposition, local identification requires $D_B(\cdot; K_0)$ to vanish on $\mathcal{D}(A_{\mathrm{aug}}, h)$. This condition need not be sufficient, just as the row-space condition is sufficient but need not be necessary. Both checks are linear algebra: compute a row-space inclusion and the kernel in (25), then restrict the linear functional D_B to that kernel.

The number of additional measurements required depends on the covariance restrictions. In the four-point stationary model of Theorem 3, augmenting $\{Z_0\}$ with Z_1 reduces the stationary invisible space from dimension 1 to 0 even though the span condition for $B = \{Z_1, Z_2\}$ fails: identification there comes from stationarity, not from Proposition 8(i). Without that restriction the invisible dimensions for $\{Z_0\}$, $\{Z_0, Z_1\}$ and $\{Z_0, Z_1, Z_2\}$ are 4, 2 and 0: observing Z_1 leaves a two-dimensional invisible space, whereas observing both Z_1 and Z_2 removes all invisible directions. Under stationarity, observing Z_1 already suffices. For an affine structured covariance class with direction space $\mathcal{T}$, the same local test applies on $\mathcal{D}(A_{\mathrm{aug}}, h) \cap \mathcal{T}$; stationarity is the Toeplitz instance.

A calibration protocol of full column rank satisfies the span condition and removes every invisible direction in the unrestricted model. Full column rank is stronger than is needed to identify one specified alternative, but it gives a simple route to evaluating an entire candidate family. The next question is how accurately that route works when dense calibration is available for only m units.

4 Finite Calibration and Protocol Comparison

Calibration measurements constrain covariance directions that are invisible under the realised protocol. We consider m independent units observed densely according to

$$W^{(i)} = Z^{(i)} + \eta^{(i)}, \qquad Z^{(i)} \sim \mathcal{N}(0, K), \qquad \eta^{(i)} \sim \mathcal{N}(0, R_0), \qquad i = 1, \dots, m. \tag{26}$$

The pairs $(Z^{(i)}, \eta^{(i)})$ are independent across i, and $\eta^{(i)}$ is independent of $Z^{(i)}$ for every i. Consequently, $\mathrm{Cov}(W^{(i)}) = K + R_0$. The finite-sample question is how error in the covariance recovered from these data propagates to comparisons among protocols. Appendix B gives the proofs, including covariance-repair and resolvent constants.

4.1 Covariance Calibration

Let $\widehat{\Sigma} = m^{-1} \sum_i W^{(i)} W^{(i)\top}$, and set $\widehat{R}_0 = R_0$ when the calibration-noise covariance is known; otherwise estimate R_0 from an independent or replicated calibration component. For a symmetric matrix M, define

$$\begin{aligned} \widetilde{M}_\tau &= \mathrm{proj}_{\{H = H^\top : H \succeq \tau I_p\}}(M), \\ D_\tau(M) &= \mathrm{Diag}\{\mathrm{diag}(\widetilde{M}_\tau)\}, \qquad \mathcal{R}_\tau(M) = D_\tau(M)^{-1/2} \widetilde{M}_\tau D_\tau(M)^{-1/2}. \end{aligned} \tag{27}$$

Thus $\mathcal{R}_\tau$ floors the eigenvalues and then rescales the result to a correlation matrix. The plug-in estimator is

$$\widehat{K} = \mathcal{R}_{\tau_m}(\widehat{\Sigma} - \widehat{R}_0), \qquad \widehat{\mathcal{I}}_g(S) = \frac{F_g(S; \widehat{K})}{V_g(\widehat{K})}, \tag{28}$$

where $\tau_m \downarrow 0$. The projection in (27) is the Frobenius-norm eigenvalue-clipping map (Higham, 1988). The asymptotic results require only $\tau_m \downarrow 0$.

The target enters the error rate through the modulus of continuity of its Gaussian covariance transform C_g.

Proposition 9 (Regularity of the target covariance transform)**.** *For each of the following cases there is a finite constant L such that, on the stated range,*

$$|C_g(r_2) - C_g(r_1)| \le L |r_2 - r_1|^\beta. \tag{29}$$

(i) If g belongs to the first-order Gaussian Sobolev space $W^{1,2}(\phi)$, then (29) *holds on* $[-1,1]$ *with* $\beta = 1$.

(ii) If $g_c(z) = \mathbf{1}\{z > c\}$, then it holds on $[-1,1]$ *with* $\beta = 1/2$. *This exponent cannot be improved uniformly, because*

$$C_{g_c}(1) - C_{g_c}(1-\delta) \sim \frac{e^{-c^2/2}}{2\pi}\sqrt{2\delta}, \qquad \delta \downarrow 0. \tag{30}$$

(iii) For the same threshold target and every $0 \le r_{\max} < 1$, (29) *holds on* $[-r_{\max}, r_{\max}]$ *with* $\beta = 1$.

Smooth targets therefore transmit covariance error linearly. Threshold targets do so linearly at a fixed model whose relevant correlations are separated from ± 1, but only through a square-root envelope when boundary correlations are allowed.

4.2 Uniform Value Error

We now propagate covariance-calibration error through the protocol-value calculation, $\widehat{K} - K \mapsto Q_S(\widehat{K}) - Q_S(K) \mapsto \widehat{\mathcal{I}}_g(S) - \mathcal{I}_g(S)$. Two quantities can destabilise this chain and therefore appear explicitly in the assumptions below: a vanishing target variance and nearly singular protocol covariances.

Theorem 10 (Uniform stability of protocol values)**.** *Let $V_g(K) \ge v_0 > 0$, and suppose that for a feasible family $\Pi_{\mathcal{B}}$,*

$$\sup_{\substack{S \in \Pi_{\mathcal{B}} \\ S \neq \varnothing}} \|A_S\|_{\mathrm{op}}^2 \le a < \infty, \qquad \inf_{\substack{S \in \Pi_{\mathcal{B}} \\ S \neq \varnothing}} \lambda_{\min}(A_S K A_S^\top + R_S) \ge \lambda > 0. \tag{31}$$

If the modulus (29) *with exponent β is valid for all arguments of C_g arising from entries of K, $\widehat{K}$, $Q_S(K)$ and $Q_S(\widehat{K})$, then, for all sufficiently small $e = \left\|\widehat{K} - K\right\|_{\mathrm{op}}$,*

$$\sup_{S \in \Pi_{\mathcal{B}}} \left|\widehat{\mathcal{I}}_g(S) - \mathcal{I}_g(S)\right| \le C e^{\beta}. \tag{32}$$

Here C depends only on the target modulus, v_0, bounds on K and $\widehat{K}$, and the family bounds (a, λ), not on S. Consequently $\beta = 1$ for smooth targets, $\beta = 1/2$ globally for threshold targets, and $\beta = 1$ for threshold targets at an interior model.

For a threshold target, the interior condition invoked in the theorem can be checked at the fixed model through

$$r_0 = \max\left\{ \max_{j \neq k} |K_{jk}|,\ \sup_{S \in \Pi_{\mathcal{B}}} \max_{j,k} |Q_S(K)_{jk}| \right\} < 1. \tag{33}$$

Uniform continuity of the resolvent then places the corresponding arguments for every sufficiently close $\widehat{K}$ in a common compact subset of $(-1, 1)$.

4.3 Calibration Rates

Theorem 11 (Fixed-model calibration rates)**.** *Fix p and a finite feasible family $\Pi_{\mathcal{B}}$, and suppose that $K \succ 0$, $V_g(K) > 0$ and the conditioning bound in* (31) *holds. If $\left\|\widehat{R}_0 - R_0\right\|_{\mathrm{op}} = O_p(m^{-1/2})$ and $\tau_m \downarrow 0$, then the estimator in* (28) *satisfies $\left\|\widehat{K} - K\right\|_{\mathrm{op}} = O_p(m^{-1/2})$. Hence*

$$\sup_{S \in \Pi_{\mathcal{B}}} |\widehat{\mathcal{I}}_g(S) - \mathcal{I}_g(S)| = \begin{cases} O_p(m^{-1/2}), & g \in W^{1,2}(\phi), \\ O_p(m^{-1/2}), & g = g_c \text{ and the interior condition (33) holds}, \\ O_p(m^{-1/4}), & g = g_c \text{ under the global threshold envelope}. \end{cases} \tag{34}$$

The global $m^{-1/4}$ rate is a worst-case guarantee over correlation matrices; the ordinary root-m rate applies to every fixed threshold model satisfying (33).

4.4 Selection Regret and Distinguishable Value Gaps

Uniform value error translates directly into a decision guarantee.

Corollary 12 (Selection regret and distinguishable value gaps)**.** *Let $\varepsilon_m = \sup_{S \in \Pi_{\mathcal{B}}} |\widehat{\mathcal{I}}_g(S) - \mathcal{I}_g(S)|$, let S^* maximise $\mathcal{I}_g$ over $\Pi_{\mathcal{B}}$, and let $\widetilde{S}$ be a feasible protocol with empirical optimisation gap*

$$\eta_m = \max_{S \in \Pi_{\mathcal{B}}} \widehat{\mathcal{I}}_g(S) - \widehat{\mathcal{I}}_g(\widetilde{S}). \tag{35}$$

Then

$$\mathcal{I}_g(S^*) - \mathcal{I}_g(\widetilde{S}) \le 2\varepsilon_m + \eta_m. \tag{36}$$

Exact empirical maximisation is the special case $\eta_m = 0$.

For nested classes $\Pi^{(1)} \subseteq \cdots \subseteq \Pi^{(L)} = \Pi_{\mathcal{B}}$, let $\widehat{S}_\ell$ maximise $\widehat{\mathcal{I}}_g$ over $\Pi^{(\ell)}$. On the event $\sup_{S \in \Pi^{(\ell)}} |\widehat{\mathcal{I}}_g(S) - \mathcal{I}_g(S)| \le \varepsilon_\ell$,

$$\mathcal{I}_g(S^*) - \mathcal{I}_g(\widehat{S}_\ell) \le \underbrace{\mathcal{I}_g(S^*) - \max_{S \in \Pi^{(\ell)}} \mathcal{I}_g(S)}_{\textit{class restriction}} + \underbrace{2\varepsilon_\ell}_{\textit{calibration error}}. \tag{37}$$

An estimated gap larger than $2\varepsilon_m$ certifies the same ordering at the population level, whereas a population gap larger than $2\varepsilon_m$ cannot be reversed by the plug-in estimates. Hence protocol-value differences below this scale cannot be distinguished uniformly from calibration error. Equation (37) separates this uncertainty from the approximation loss incurred by restricting the search to $\Pi^{(\ell)}$.

Remark (Best-linear calibration). The same perturbation and regret arguments apply to $\mathcal{I}_{\mathrm{L}}$. If the protocol covariances are uniformly nonsingular and their covariance and target-cross-covariance blocks, together with the target variance, are estimated uniformly at the root-m rate, then $\sup_{S \in \Pi_{\mathcal{B}}} |\widehat{\mathcal{I}}_{\mathrm{L}}(S) - \mathcal{I}_{\mathrm{L}}(S)| = O_p(m^{-1/2})$. Appendix B.4 states sufficient conditions and proves this claim; Corollary 12 then applies with $\mathcal{I}_g$ replaced by $\mathcal{I}_{\mathrm{L}}$.

5 Target-Aware Observation Design

Once candidate values have been identified and estimated accurately enough to distinguish them, the remaining task is to choose a feasible protocol. Because the target variance $V_g(K)$ does not depend on the protocol, maximising the normalised value $\mathcal{I}_g(S;K) = F_g(S;K)/V_g(K)$ is equivalent to maximising $F_g(S;K)$. Exact rank-one marginal gains support forward selection with one-swap refinement, and exhaustive search measures optimisation error when the catalogue is small enough. Appendix C gives the derivations, algorithmic details and comparator formulas.

5.1 Feasible Protocols and Design Objective

For design, an acquisition action is represented by

$$a = (\ell_a,\ r_a,\ c_a), \tag{38}$$

where $\ell_a \in \mathbb{R}^p$ specifies the linear measurement, $r_a \geq 0$ is its effective noise variance, and $c_a > 0$ is its cost. The observation is

$$Y_{i,a} = \ell_a^\top Z_i + \varepsilon_{i,a}, \qquad \varepsilon_{i,a} \sim \mathcal{N}(0, r_a), \tag{39}$$

with noises independent across actions and independent of Z_i. Timing and support are encoded in ℓ_a: a point measurement uses a coordinate vector, whereas a window average uses normalised weights over its support. Repetition is encoded through r_a; for example, averaging M_a independent measurements with per-measurement variance ν_a^2 gives $r_a = \nu_a^2/M_a$. The cost may include all repetitions and other acquisition burdens.

For a finite catalogue $\mathcal{V}$ and cost budget $\mathcal{B}$, the feasible family is

$$\Pi_{\mathcal{B}} = \Big\{ S \subseteq \mathcal{V} :\ \textstyle\sum_{a \in S} c_a \leq \mathcal{B} \Big\}. \tag{40}$$

Application-specific restrictions, such as choosing at most one precision variant on a given support, can be imposed on the catalogue or by taking a subfamily of $\Pi_{\mathcal{B}}$. The results below apply to any such subfamily. We restrict attention to feasible protocols for which every non-empty $\Sigma_S(K)$ is nonsingular; this is automatic when every selected action has $r_a > 0$.

At the population level, observation design is therefore the optimisation problem

$$\max_{S \in \Pi_{\mathcal{B}}} \ F_g(S;K), \tag{41}$$

where the candidate set $\mathcal{V}$ ranges over acquisition times, window lengths, repetition counts and noise levels. Given calibration data, the empirical version replaces K by $\widehat{K}$ and maximises $F_g(S;\widehat{K})$; Corollary 12 bounds what that substitution costs.

5.2 Exact Marginal Gains

Suppressing the argument K in $Q_S(K)$, write

$$P_S := K - Q_S(K) = \mathrm{Cov}(Z \mid Y_S). \tag{42}$$

Thus $v_{a|S} = P_S \ell_a$ is the covariance between a new measurement and the residual trajectory, and $s_{a|S}$ is its conditional variance, including measurement noise.

Proposition 13 (Rank-one marginal gains)**.** *Let $S = \varnothing$, or assume that $\Sigma_S(K)$ is nonsingular. Let $a \notin S$ satisfy $S \cup \{a\} \in \Pi_{\mathcal{B}}$ and $s_{a|S} > 0$, and put*

$$v_{a|S} = P_S \ell_a, \qquad s_{a|S} = \ell_a^\top P_S \ell_a + r_a. \tag{43}$$

Then $Q_{S\cup\{a\}} = Q_S + v_{a|S} v_{a|S}^\top / s_{a|S}$ and $P_{S\cup\{a\}} = P_S - v_{a|S} v_{a|S}^\top / s_{a|S}$, so that the exact marginal gain is

$$\Delta_g(a \mid S) = \sum_{j,k} \omega_j \omega_k \left[C_g\left\{ Q_{S,jk} + \frac{v_{a|S,j} v_{a|S,k}}{s_{a|S}} \right\} - C_g(Q_{S,jk}) \right]. \tag{44}$$

For the mean target this collapses to the closed form

$$\Delta_{\mathrm{mean}}(a \mid S) = \frac{(\omega^\top P_S \ell_a)^2}{\ell_a^\top P_S \ell_a + r_a}. \tag{45}$$

Equation (44) is the exact objective increment rather than a surrogate. The rank-one form lets search update the residual covariance without re-forming the full observation-covariance inverse.

The marginal-gain objective is monotone under nested protocols, although it need not exhibit diminishing returns.

Lemma 14 (Monotonicity under nested protocols)**.** *For every $g \in L^2(\phi)$, every correlation matrix K with nonsingular measurement covariance for each non-empty protocol concerned, and every $S \subseteq S'$, $F_g(S; K) \leq F_g(S'; K)$.*

Adding measurements cannot decrease F_g. Monotonicity, however, does not imply diminishing returns: even the mean target reduces to R^2 subset selection, which is not submodular in general (Das and Kempe, 2018).

For noiseless point measurements and a linear integral target, the objective reduces to kernel quadrature (Bach, 2017; Briol et al., 2019). Measurement noise and nonlinear targets place the general problem outside that noiseless formulation.

5.3 Search and Comparators

Starting from $S = \varnothing$ and $P_S = K$, forward selection repeatedly adds the feasible action with largest exact gain in (44), using gain per unit cost when costs differ, and then applies the rank-one update. A one-swap refinement makes feasible selected–unselected exchanges whenever the exchange improves the objective. Because monotonicity does not imply submodularity, exhaustive search is used as an optimisation benchmark when the catalogue is small enough. Pseudocode and operation counts are given in Appendix C.2.

To isolate which information drives a schedule, we compare the target-aware objective with several alternatives. Latent-state mutual information maximises $I(Z; Y_S)$ and is target-free. Integrated posterior variance minimises $\sum_j \omega_j (P_S)_{jj}$, so it uses temporal weights but not the target transformation. Linear-target design retains the actual weights and noise but sets $g(z) = z$, thereby maximising $\omega^\top Q_S(K) \omega$. Noiseless kernel quadrature uses the same linear criterion with $R_S = 0$. These comparisons separate target nonlinearity from the noise model and from target-free coverage criteria; their exact objectives and marginal gains are recorded in Appendix C.3.

6 Empirical Evaluation

6.1 Experimental Overview

The empirical evaluation addresses three questions that follow from the calibration and design results in Sections 4–5. First, with only finitely many dense calibration trajectories, how accurately can protocol values be estimated across a candidate family, and at what resolution can competing protocols be reliably distinguished and selected? Second, once those values are estimable, does designing measurements for the predictive value of the specified target improve on target-free or target-mismatched objectives, and how closely can practical search approach the exact target-specific optimum? Third, in fully annotated temporal data, are broad matched-budget differences in observation layout reproducible, and is there sufficient information to support the finer, data-dependent choice of exact observation locations?

The first two questions are studied under known data-generating laws. The calibration experiments compare plug-in values with the exact population values $\mathcal{I}_g(S; K)$ and evaluate family-uniform error, selection regret and nested candidate classes. The design experiments enumerate every feasible protocol in moderate catalogues, so the target-specific optimum and the optimisation error of forward selection with one-swap refinement are known exactly. The third question is studied retrospectively in Sleep-EDF and Long-Term AF by reconstructing partial observation protocols from complete annotations. Because population protocol values are not observed in these data, performance is measured by pooled held-out cross-fitted R^2; support selection and predictor tuning are confined to the corresponding outer training folds. Simulation specifications are given in Appendix D, and the real-data processing, cross-fitting and resampling procedures are given in Appendix E.

6.2 Finite Calibration and Protocol Comparison

Figure 2 examines three consequences of finite calibration. Panel (a) asks how accurately protocol values can be estimated over an entire candidate family. Panel (b) asks whether the remaining estimation error is large enough to affect selection of a near-optimal protocol. Panels (c,d) then examine how calibration accuracy limits the useful granularity of the candidate family. In all experiments, protocol values estimated from the calibrated covariance $\widehat{K}$ are compared with their exact population values under the known generating covariance. Full simulation specifications are given in Appendix D.

Panel (a) evaluates the family-uniform error $\varepsilon_m = \max_{S \in \Pi_{\mathcal{B}}} |\widehat{\mathcal{I}}_g(S) - \mathcal{I}_g(S; K)|$, the largest protocol-value error within the candidate family. This error decreases steadily as the number of calibration trajectories increases. The full-range log–log slopes are -0.4134 and -0.4142; over the three largest calibration sizes they are -0.4618 and -0.4637, approaching the fixed-model exponent $-1/2$ derived in Section 4.

In this experiment, panel (b) shows that selecting a near-optimal protocol can require substantially less accuracy than estimating every protocol value uniformly. Across the six covariance–target cells, the mean cellwise log–log slopes are -0.94 for realised selection regret and -0.49 for the uniform-error envelope. The median and maximum ratios of regret to this envelope are 0.024 and 0.37. The difference reflects the two criteria being controlled: ε_m is determined by the largest estimation error anywhere in the family, whereas near-optimal selection depends primarily on correctly ordering the leading candidates.

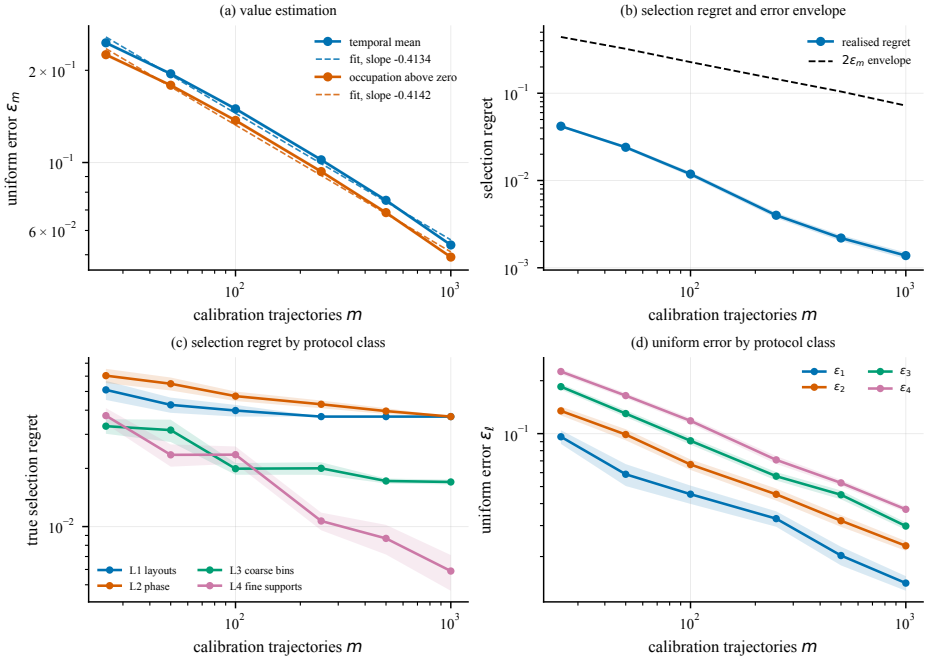

Figure 2: Finite-calibration error, selection regret and candidate-family resolution. (a) Family-uniform protocol-value error ε_m for the temporal-mean and occupation-at-zero targets over all 495 four-action protocols under a stationary OU Gaussian model on a $p = 128$ grid. (b) True regret of the protocol selected from estimated values, together with the $2\varepsilon_m$ envelope from Section 4, for four-action selection from a 14-action catalogue, averaged over two covariance models and three targets: the temporal mean and occupation above thresholds zero and one. (c,d) Selection regret and family-uniform error for four nested candidate families of increasing granularity, with sizes 2, 5, 75 and 568, under a nonstationary $p = 64$ covariance. Calibration sizes are $m \in \{25, 50, 100, 250, 500, 1000\}$. Panels (a,b) use 200 replications and panels (c,d) use 30; curves show replication means and bands show standard errors.

Panels (c,d) show how this distinction affects the useful resolution of the candidate family. At $m = 25$, the finest class has selection regret 0.03749, compared with 0.0510 for the coarsest class, despite a family-uniform error of 0.2255. Thus, a richer candidate class can yield a better selected protocol even before all of its values are estimated accurately. By $m = 1000$, the uniform error of the finest class falls to 0.0371, whereas the coarsest class still incurs restriction regret 0.03706 because better protocols lie outside that class.

Together, the four panels distinguish uniform estimation resolution from selection resolution. Reliable selection does not require every candidate value to be estimated to the same precision; it requires sufficient calibration to resolve the value differences among the candidates that compete near the optimum. Candidate-family refinement can therefore be useful before uniform estimation over the refined family becomes accurate.

6.3 Target-Aware Design and Search

Table 1 examines two questions about observation design. First, how much is lost when protocols are chosen using target-free or target-mismatched objectives rather than the predictive value of the specified target? Second, when the target-aware objective is used, how closely does the proposed search procedure approach the exact optimum?

We evaluate 25 instances formed by five aggregate targets and five covariance–action settings: stationary OU, stationary Matérn, horizon-varying covariance, recency-weighted prediction and heterogeneous actions. The first four settings select four of 12 point actions; the heterogeneous setting chooses among 18 actions under a cost budget of four. Every feasible protocol is enumerated, providing the exact target-specific optimum against which both alternative design objectives and approximate search can be evaluated. Full simulation settings are given in Appendix D.

For a method returning protocol S, we report relative efficiency $\mathcal{I}_g(S)/\mathcal{I}_g(S^*)$, where S^* is the exhaustive optimum for the target being evaluated. An efficiency of one therefore indicates that the selected protocol attains the target-specific optimum. Latent-state mutual information and integrated posterior variance provide target-free comparators. The linear-target comparator retains the actual temporal weights and measurement noise but replaces the target transformation by $g(z) = z$, whereas kernel quadrature uses the corresponding conventional noiseless integration criterion.

Method	minimum	mean	median
Latent-state mutual information	0.601	0.864	0.852
Integrated posterior variance	0.802	0.947	0.987
Actual-noise linear target	0.767	0.982	1.000
Noiseless kernel quadrature	0.231	0.899	0.988
Target-aware greedy	0.912	0.992	1.000
Target-aware greedy plus one-swap	0.987	0.999	1.000

Table 1: Relative efficiency of alternative design objectives and search procedures against the exhaustive target-specific optimum over 25 kernel–action–target instances.

The choice of design objective matters most when temporal symmetry is broken. In the stationary, uniformly weighted settings, supports optimised for different targets transfer with little loss. Under nonstationarity, non-uniform temporal weighting or heterogeneous actions, the target-specific optimal supports become more distinct; the lowest cross-target efficiency in the heterogeneous setting is 0.765. The target-free comparators show the same general distinction: good coverage of the latent trajectory does not necessarily yield a protocol that is optimal for the specified aggregate target.

The proposed search procedure is nevertheless close to the exhaustive target-aware optimum. One-swap refinement raises the minimum efficiency in the heterogeneous setting from 0.912 to 1.000 and reaches the exhaustive optimum there and in the stationary OU and horizon-varying settings. In the remaining settings its minimum efficiencies are 0.994 and 0.987, while evaluating only 0.09–0.23 of the exhaustive sets.

6.4 Retrospective Protocol Comparison

The retrospective analyses use two fully annotated temporal data sets to reconstruct partial observation protocols while computing the prediction target from the complete analysable record.

Sleep-EDF Expanded (Kemp et al., 2000; Mourtazaev et al., 1995; Goldberger et al., 2000; Pollard et al., 2026) contains 30-second sleep-stage annotations from the Sleep Cassette (SC) and Sleep Telemetry (ST) studies. The analysed sample contains 197 recordings from 100 subjects and 3818 hours of valid annotations. Stages 3 and 4 are combined as N3 (Rechtschaffen and Kales, 1968; Iber et al., 2007). Each recording is mapped to $p = 128$ relative-time anchors. The targets are the exact full-interval REM, N3 and Wake proportions. The Long-Term AF Database (Petrutiu et al., 2007; Goldberger et al., 2000) contains 84 records and 1934 hours of reviewed rhythm annotations, with median analysable record length 24.0 hours. Analysis begins at the first rhythm marker, excluding any unannotated prefix. Each record is mapped to $p = 96$ equal relative-time bins, and the target is the exact fraction of analysable annotated time spent in atrial fibrillation.

Sleep protocols expose stage-specific binary indicators at selected relative-time anchors, with budgets $N \in \{4, 8, 16, 32, 64\}$ counting distinct scored 30-second epochs. AF protocols expose bin-level AF fractions, with budgets $N \in \{1, 2, 4, 8, 16, 32\}$ observing the fraction N/p of each analysable record. Both analyses use five outer folds and report R^2 from pooled held-out predictions. Sleep folds are subject-disjoint and stratified by source study, with subject-balanced evaluation weights. Study, valid-duration and treatment covariates enter the Sleep predictors unpenalised. AF uses record-level folds and equal record weights because repeated-subject identifiers are unavailable.

At each budget, we compare a centred contiguous block with observations dispersed as uniformly as possible over relative time. Sleep additionally learns target-aware and kernel-quadrature supports within each outer training fold. The target-aware criterion maximises a regularised best-linear estimate of the fraction of residual target variance explained by the selected anchors. The kernel-quadrature comparator uses the same repaired temporal covariance without target outcomes. Support selection, ridge tuning and predictor fitting are completed within the outer training fold. For fixed-template contrasts, conditional 2.5th–97.5th percentile ranges are obtained by resampling held-out subjects for Sleep and records for AF while keeping the fitted predictors fixed. Further details of the common-grid mapping, weighting, support selection and resampling are given in Appendix E.

For REM, Figure 3(a) shows cross-fitted R^2 values of +0.648, +0.659 and +0.863 for the centred-contiguous template at $N = 4, 16, 64$, compared with +0.682, +0.738 and +0.881 for uniform dispersion. Across the 15 Sleep target–budget cells, the dispersed-minus-contiguous percentile range has a positive lower endpoint in 9 cells. The pooled Sleep results favour dispersion at several budgets, but the source-specific contrasts are heterogeneous. In separate SC and ST analyses, 10 of 30 source–target–budget ranges lie above zero, 5 lie below zero and 15 include zero. Additional analyses are reported in Appendix E.3.

For AF, at $N = 4$ each template observes 4.17% of the analysable record, equivalent to 1 hour on a 24-hour record. The contiguous and dispersed templates attain cross-fitted R^2 values +0.696 and +0.971, respectively, with difference +0.274 and conditional paired percentile range [+0.143, +0.458]. At $N = 16$, the corresponding values are +0.851 and +0.998, with difference +0.148 and range [+0.076, +0.249]. The paired percentile range has a positive lower endpoint at every AF budget containing at least two windows. At $N = 1$, the difference is -0.061 with range [-0.299, +0.164]. Because only one window is observed, this comparison reflects window location rather than dispersion. Outcome-defined cohort and grid-discretisation analyses are reported in Appendix E.3.

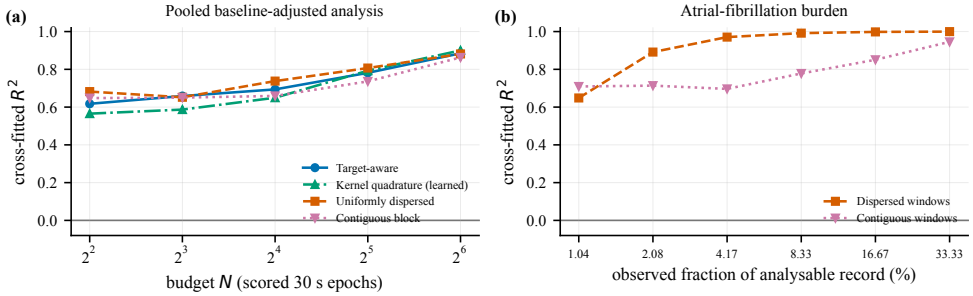

Figure 3: Held-out prediction from reconstructed Sleep and AF observation protocols. (a) Pooled baseline-adjusted cross-fitted R^2 for the Sleep targets under target-aware, learned kernel-quadrature, uniformly dispersed and contiguous supports. (b) Pooled cross-fitted R^2 for AF burden under dispersed and contiguous window layouts.

We next examine the stability of the learned Sleep supports. At $N = 16$, the target-aware REM support attains baseline-adjusted cross-fitted $R^2 = +0.694$, compared with +0.650 for learned kernel quadrature and +0.738 for fixed dispersion. The fold-averaged pairwise Jaccard overlap among the REM-, N3- and Wake-specific supports is 0.061–0.091. The fitted supports therefore differ substantially across targets, although this difference may also reflect finite-sample variation in support selection.

Figure 4(a) varies the number of subjects used for support selection from 20 to the complete outer-training-fold sample while keeping predictor fitting on the full training fold. The target-aware difference changes from -0.006 to +0.044 relative to learned kernel quadrature and from -0.004 to -0.044 relative to fixed dispersion. Panel (b) repeats the complete selection and evaluation pipeline on 1000 study-stratified subsamples containing 80% of subjects. The target-aware difference relative to kernel quadrature has median -0.014 and 2.5th–97.5th percentile range $[-0.102, +0.078]$. Relative to fixed dispersion, the corresponding values are -0.012 and $[-0.126, +0.097]$. Both ranges include zero.

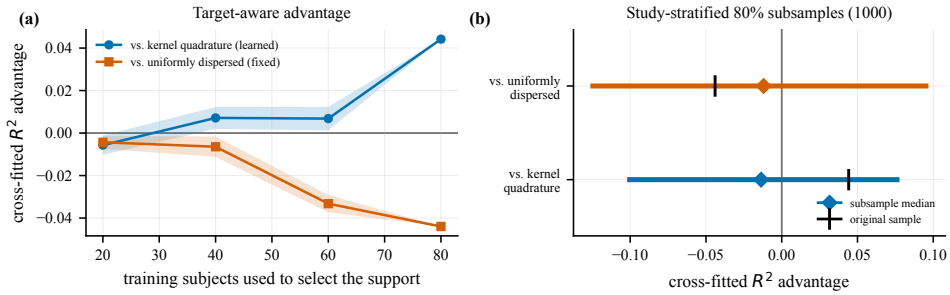

Figure 4: Stability of the learned REM support at $N = 16$. (a) Target-aware held-out advantage as the support-selection sample increases. (b) Target-aware advantage over learned kernel quadrature and fixed dispersion across 1000 study-stratified 80% subject subsamples.

The real-data results distinguish broad temporal-layout contrasts more clearly than exact learned support locations. Dispersion performs better than contiguous observation in several pooled Sleep comparisons and across the multi-window AF budgets, although the Sleep contrasts vary between source studies. The learned Sleep supports are less stable across targets and subject subsamples, and their paired held-out R^2 differences include zero. This agrees with the resolution pattern in Figure 2: coarse protocol differences can be easier to distinguish than fine support locations with finite data.

7 Related Work

Counterfactual protocol evaluation intersects four literatures: learning from incomplete temporal records, measurement design under an assumed process model, evaluation across acquisition regimes, and statistical identification of observation schemes. The main distinctions are whether the measurements required by an alternative occur in the observed data and whether its value is computable under a specified model or must be identified from the realised measurement–target law.

7.1 Learning from Incomplete Temporal Trajectories

Functional data analysis estimates population mean and covariance functions from sparse or irregular measurements under sampling and smoothness assumptions (Ramsay and Silverman, 2005; Yao et al., 2005; Hall et al., 2006; Li and Hsing, 2010). Longitudinal methods also allow observation times to depend on latent or previously observed outcomes (Pullenayegum and Lim, 2016; Weaver et al., 2023). These estimators pool within-subject pairs across units, so recovery of a covariance region requires the sampling design to supply informative pairs there, together with the assumptions imposed by the estimator. Under a single fixed protocol, times or supports unique to an alternative may occur for no unit. The corresponding covariance blocks are then not determined by the measurements alone; in the linear-Gaussian analysis, the aggregate target adds the cross-covariance and variance constraints characterised in Section 3.

Machine-learning methods for irregular time series instead learn prediction or interpolation rules directly from incomplete records. GRU-D models missingness indicators and elapsed times, whereas latent ODEs and neural controlled differential equations represent irregular observations in continuous time (Che et al., 2018; Rubanova et al., 2019; Kidger et al., 2020; Shukla and Marlin, 2020). Early-prediction and active feature acquisition methods determine when an observed prefix is sufficient or which feature to acquire next (Xing et al., 2009; Shim et al., 2018). Multiple-instance and label-proportion methods also use aggregate labels, but retain a fixed observation scheme (Dietterich et al., 1997; Quadrianto et al., 2009; Zhou, 2018). These methods operate on observation patterns represented in the data. By contrast, an alternative protocol here may require measurements absent for every unit in the realised sample; its joint law with the target is therefore not available as a marginal of the realised law.

7.2 Measurement Design under an Assumed Model

Classical optimal design allocates measurements to estimate parameters or linear functionals efficiently under a statistical model (Elfving, 1952; Pukelsheim, 2006). Bayesian and goal-oriented design optimise expected information or posterior uncertainty about a quantity of interest under a model or prior (Lindley, 1956; Chaloner and Verdinelli, 1995; Attia et al., 2018; Alexanderian, 2021; Rainforth et al., 2024). Active learning, sensor selection, Gaussian-process sensor placement, and

kernel quadrature likewise optimise uncertainty, information gain, estimation error, or integration accuracy under a fitted model, covariance, or kernel (MacKay, 1992; Settles, 2012; Joshi and Boyd, 2009; Krause et al., 2008; Bach, 2017; Briol et al., 2019).

Optimal design for longitudinal and functional data is especially close to the present temporal setting: pilot data can estimate dependence and guide future observation times, including designs for predicting a scalar response (Ji and Müller, 2017). Such procedures ask which measurements to collect once a model makes the criterion computable. Counterfactual protocol evaluation asks the preceding question: whether the value in Definition 1 is determined by the realised law at all. Once the required dependence has been identified or calibrated, Section 5 becomes a target-aware sensor-selection problem. Its objective is the population R^2 of the Bayes-optimal predictor of a fixed temporal aggregate, not trajectory-reconstruction error or target-free information gain. For noiseless point measurements and a linear aggregate it coincides with a kernel-quadrature criterion; measurement noise and nonlinear aggregates move beyond that case.

7.3 Evaluation across Acquisition Regimes

Active feature acquisition performance evaluation (AFAPE) evaluates a new acquisition policy from retrospective data generated by another acquisition process (von Kleist et al., 2025). Depending on its assumptions, AFAPE uses missing-data or offline-policy-evaluation methods, including direct, inverse-probability-weighted, and doubly robust estimators. More generally, off-policy evaluation requires overlap or related support conditions between the behaviour and target regimes (Precup et al., 2000; Kallus and Uehara, 2020; Sachdeva et al., 2020).

A fixed observation protocol can be represented as a deterministic acquisition policy, but an alternative considered here may assign positive demand to measurements having zero acquisition probability under the realised protocol. Reweighting observed acquisition trajectories cannot identify their distribution or predictive value. Identification must instead come from restrictions linking observed and unobserved components of the same latent trajectory to the aggregate target. Theorem 3 and Theorem 5 show that these links need not determine the alternative's value, whereas Proposition 8 gives conditions under which additional measurements remove the relevant ambiguity. Graphical missing-data theory provides a related account of recoverability from systematically incomplete observations (Mohan and Pearl, 2021; Nabi et al., 2020). Here, however, the acquisition rule is fixed and known; the unresolved object is the latent temporal dependence between measurements that were and were not collected.

7.4 Identification, Comparison, and Predictive Limits of Observation Schemes

Statistical identification asks whether a functional is constant over all latent models inducing the same observed law; partial identification retains the resulting set of compatible values (Koopmans and Reiersøl, 1950; Rothenberg, 1971; Manski, 2003; Tamer, 2010). We apply this principle to the predictive value of an undeployed protocol. Definition 4 exposes the observational-equivalence geometry in the linear-Gaussian model, Theorem 5 gives a local non-identification certificate, and Proposition 6 gives an exact construction for nonlinear aggregates. Proposition 8 conversely shows that a specified protocol value can be identified without recovering the complete latent covariance.

Bayes-error analyses characterise predictive limits for a fixed observed feature–target law (Fukunaga and Hummels, 1987; Ishida et al., 2023). The present question is whether that law determines the predictive limit under a different feature-generating protocol. Blackwell's comparison of experiments

and its extensions order observation schemes across classes of decision problems (Blackwell, 1953; Le Cam and Yang, 1990; Torgersen, 1991); protocol value is instead a scalar criterion for one fixed target and loss. Transportability and data-fusion methods infer quantities outside a realised regime through causal or invariance assumptions across populations or data sources (Pearl and Bareinboim, 2014; Bareinboim and Pearl, 2016). Here the population and target remain fixed, and the regime change concerns which measurements of the same latent trajectory are collected.

The Gaussian identity used to evaluate a fixed protocol in Proposition 2 was derived by Zhang (2026). The present work studies the identification of an undeployed protocol's value, measurement augmentation that restores it, and finite-calibration comparison and design.

8 Discussion

The observation protocol is part of a predictive system, not merely a fixed preprocessing choice. When the target summarises a longer trajectory than the measurements supplied to the learner, performance reflects both the learner and the information admitted by the protocol. More training units and more flexible predictors can reduce error under the realised protocol, but they do not reveal the value of measurements that were never collected. Even the population measurement–target law can leave dependence unresolved that is decisive for an alternative protocol.

The practical meaning of "more data" therefore depends on how those data are collected. If the ambiguity is invisible to the realised protocol, enlarging the sample under that protocol reproduces the same constraints. A smaller sample observed more extensively may be more useful. Targeted augmentation removes ambiguity affecting one specified alternative, whereas dense calibration supports a broader family. This motivates a large routine cohort accompanied by a smaller, intensively observed calibration subset chosen with future protocol decisions in mind.

Finite calibration determines how finely protocols can be compared. The precision of an optimiser's output should not be confused with the precision supported by the data. Figure 2 shows that richer candidate classes can reduce selection regret before all candidate values are estimated accurately, because selection depends mainly on ordering protocols near the optimum. Useful granularity is set by the value gaps among competitive protocols and the information available to resolve them.

The retrospective analyses illustrate this distinction. In Sleep, broad dispersed-versus-contiguous contrasts are more reproducible than exact learned supports, whose locations and held-out advantages vary across targets, source studies and subsamples (Figures 3 and 4). In AF, distributing a matched budget across the record is substantially more predictive of full-record burden than concentrating it in one block at the multi-window budgets considered. This is not a general rule in favour of dispersion; equal observation time need not carry equal information for a target defined over a longer horizon. Protocol quality is also target-dependent: Table 1 shows that observing the process well and observing it well for a specified aggregate can lead to different designs.

The same problem arises beyond temporal sampling. Short ECG recordings may not reveal the cross-day dependence needed to evaluate dispersed monitoring; an overlapping multimodal subset may be needed before imaging, proteomic or wearable panels can be valued; and current environmental sensors may not reveal the spatial dependence governing the value of new locations. These are not ordinary feature-selection problems, because the proposed measurements may never have been observed jointly with the target. Calibration precedes the usual optimisation problem in

longitudinal and sensor design (Chaloner and Verdinelli, 1995; Ji and Müller, 2017; Krause et al., 2008; Joshi and Boyd, 2009).

When point identification fails, the range of compatible values can still support partial comparisons and robust design based on worst-case value or minimax regret (Manski, 2003; Tamer, 2010). The Gaussian analysis exposes this ambiguity through covariance directions, but the principle is broader: an alternative is evaluable whenever its value is constant over the latent laws compatible with the realised data. In more general models, the corresponding task is to characterise that compatible set and collect measurements that shrink the decision-relevant value range.

9 Conclusion

Counterfactual protocol evaluation asks a question prior to observation design: whether data collected under one protocol determine what could be achieved under another. The results show that this can fail even with unlimited data under the realised protocol, and that resolving the failure requires new kinds of observation rather than more samples of the same kind. Finite calibration then sets the scale at which candidate protocols can be compared reliably. Observation design should therefore be optimised only after its value is identified, and only at a resolution supported by the available data. Although developed for temporal protocols, the same principle applies to alternative measurement systems more generally.

Code and Data Availability

Code and data are available at `https://github.com/zxzok/cpv-explainer`. An interactive explainer is available at `https://cpv.xizhe.net`.

Acknowledgments and Disclosure of Funding

The author declares no competing interests.

Appendix A. Proofs for Value-Specific Identification

A.1 Gaussian Protocol-Value Identity

Proposition 2 is due to Zhang (2026). Draw posterior replicates $Z^{(1)}, Z^{(2)}$ independently given Y_S; their cross-covariance is $Q_S(K)$. The Gaussian covariance identity and conditional independence give $\mathrm{Cov}\{\mathbb{E}[g(Z_j) \mid Y_S], \mathbb{E}[g(Z_k) \mid Y_S]\} = C_g\{Q_S(K)_{jk}\}$. Weighted summation yields $F_g(S;K)$, while total variance gives $V_g(K)$ and minimum Bayes error $V_g(K) - F_g(S;K)$, proving (10).

A.2 Minimal Stationary Counterexample

Under the assumptions of Theorem 3, stationarity and $\rho(0) = 1$ give the observable functionals

$$\begin{aligned} \mathrm{Var}(Y_A) &= 1 + \nu^2, \\ \mathrm{Cov}(Y_A, \Theta) &= \frac{1}{p}\left(1 + \sum_{l=1}^{p-1} \rho(l)\right), \\ \mathrm{Var}(\Theta) &= \frac{1}{p^2}\left(p + 2\sum_{l=1}^{p-1}(p-l)\rho(l)\right). \end{aligned} \tag{46}$$

The last two functionals have Jacobian rows proportional to $(1, \ldots, 1)$ and $(p-1, \ldots, 1)$. Hence $\operatorname{rank} J = \min\{2, p-1\}$ and $\dim\ker J = \max\{0, p-3\}$, proving parts (i)–(ii); at $p = 4$ the kernel is $\operatorname{span}\{(1, -2, 1)\}$. Positive definiteness is an open condition, so $K_\pm$ remain positive definite for all sufficiently small $\varepsilon > 0$.

The perturbation leaves $b = \mathrm{Cov}(Z_1, \Theta) = \mathrm{Cov}(Z_2, \Theta)$ and $\mathrm{Var}(\Theta)$ unchanged. With $R_B = \nu_B^2 I_2$, under $K_\pm$, $c = (b, b)^\top$ is an eigenvector of $\Sigma_B^\pm$ with eigenvalue $1 + \nu_B^2 + \rho(1) \pm \varepsilon$, giving (17). Subtraction yields

$$\mathcal{I}(B; K_-) - \mathcal{I}(B; K_+) = \frac{4b^2\varepsilon}{\mathrm{Var}(\Theta)\left[\{1 + \nu_B^2 + \rho(1)\}^2 - \varepsilon^2\right]}. \tag{47}$$

Admissibility makes the denominator positive, so the gap is strict exactly when $b \neq 0$.

A.3 General, Nonlinear and Augmented Protocols

For Theorem 5, positive definiteness of K_ε and of $\Sigma_B(K_\varepsilon)$ persists for sufficiently small $|\varepsilon|$, while $\mathrm{diag}(\Delta) = 0$ preserves the unit diagonal. Invisibility preserves the three Gaussian covariance blocks in (14), proving (19). Moreover, $h^\top \Delta h = 0$ makes the denominator of $\mathcal{I}(B; K_\varepsilon)$ constant. At zero, the inverse rule is $\mathrm{d}\Sigma_B(K_\varepsilon)^{-1}/\,\mathrm{d}\varepsilon = -W(B\Delta B^\top)W$. Applying the product rule to $h^\top K_\varepsilon B^\top \Sigma_B(K_\varepsilon)^{-1} B K_\varepsilon h$ gives (20). Since the difference between the values at $K_0 \pm \varepsilon\Delta$ is $2\varepsilon D_B(\Delta; K_0) + o(\varepsilon)$, a non-zero derivative gives local separation.

For Example 7, let $s = 1 + \nu_B^2$. Under K_a and K_a', respectively, the posterior-replica covariance matrices for the one-point alternative $B = e_1^\top$ are

$$Q_B(K_a) = \frac{1}{s}(a, 1, 0)^\top (a, 1, 0), \qquad Q_B(K_a') = \frac{1}{s}(0, 1, 0)^\top (0, 1, 0).$$

Because $C_g(0) = 0$, (9) gives

$$\mathcal{I}_g(B; K_a) - \mathcal{I}_g(B; K_a') = \frac{C_g(a^2/s) + 2C_g(a/s)}{3C_g(1) + 2C_g(a)}. \tag{48}$$

For nonconstant g, at least one nonconstant Hermite coefficient is non-zero, so (7) gives $C_g(r) > 0$ for every $r > 0$. The numerator and denominator in (48) are consequently positive for every $a \in (0, 1)$ and every finite noise variance, proving the universal strict inequality asserted in the example.

To prove Proposition 8, observe that the augmented law identifies $\bar{A}K\bar{A}^\top$. If $\mathrm{row}(B) \subseteq \mathrm{row}(\bar{A})$, there is a matrix C such that $B = C\bar{A}$. It then determines $BKB^\top = C(\bar{A}K\bar{A}^\top)C^\top$ and $BKh = C\bar{A}Kh$, and $h^\top Kh$ is already a block of the benchmark covariance. Together with known R_B, these quantities determine the joint Gaussian law of (Y_B, Θ) and hence its value, proving part (i). For part (ii), any $\Delta \in \mathcal{D}(A_{\mathrm{aug}}, h)$ preserves the augmented benchmark law, and Theorem 5 applied with realised operator A_{aug} gives the stated first-order failure certificate.

Appendix B. Proofs for Finite Calibration

B.1 Regularity of the Target Covariance Transform

For Proposition 9, if $g - \mathbb{E}g = \sum_{k\geq 1}(a_k/k!)H_k$, then $C_g(r) = \sum_{k\geq 1}(a_k^2/k!)r^k$. For $g \in W^{1,2}(\phi)$, termwise differentiation and the Gaussian Sobolev identity give

$$L_g := \sup_{|r|\leq 1} |C_g'(r)| \leq \sum_{k\geq 1} \frac{ka_k^2}{k!} = \mathbb{E}\{g'(U)^2\} < \infty. \tag{49}$$

The non-negative coefficients show that equality in the supremum is attained at $r = 1$, proving the Lipschitz claim in part (i).

For $g_c(z) = \mathbf{1}\{z > c\}$, Plackett's identity gives, for $-1 < r < 1$,

$$C_{g_c}'(r) = \frac{\exp\{-c^2/(1+r)\}}{2\pi\sqrt{1-r^2}}.$$

Together with the sharp arcsine inequality $|\arcsin b - \arcsin a| \leq (\pi/\sqrt{2})|b-a|^{1/2}$, this yields the global bound

$$|C_{g_c}(r_2) - C_{g_c}(r_1)| \leq \frac{1}{2\sqrt{2}}|r_2 - r_1|^{1/2}. \tag{50}$$

At $c = 0$ and $(r_1, r_2) = (-1, 1)$ the constant is attained. Expanding the same derivative as $r \uparrow 1$ and integrating over $[1-\delta, 1]$ gives (30), so no larger global Hölder exponent is possible. Finally, on $[-r_{\max}, r_{\max}]$ the derivative is bounded by

$$L_{g_c}(r_{\max}) = \frac{\exp\{-c^2/(1+r_{\max})\}}{2\pi\sqrt{1-r_{\max}^2}}, \tag{51}$$

which proves part (iii).

B.2 Covariance Repair and Resolvent Bounds

For covariance repair, put $\bar{K} = \widehat{\Sigma} - \widehat{R}_0$ and $e_0 = \left\|\widehat{\Sigma} - (K + R_0)\right\|_{\mathrm{op}} + \left\|\widehat{R}_0 - R_0\right\|_{\mathrm{op}}$. If $0 < \tau \leq \lambda_{\min}(K)/2$ and $e_0 \leq \min\{1/2, \lambda_{\min}(K)/2\}$, Weyl's inequality gives $\bar{K} \succeq \tau I_p$, so flooring is inactive. For $H = \mathrm{Diag}\{\mathrm{diag}(\bar{K})\}^{-1/2}$, the unit diagonal of K gives $\|H - I_p\|_{\mathrm{op}} \leq 2e_0$ and $\|H\|_{\mathrm{op}} \leq 2$. Expanding $H\bar{K}H - K = (H - I_p)\bar{K}H + (\bar{K} - K)H + K(H - I_p)$ yields

$$\left\|\mathcal{R}_\tau(\bar{K}) - K\right\|_{\mathrm{op}} \leq (6\|K\|_{\mathrm{op}} + 4)e_0 \leq 6(1 + \|K\|_{\mathrm{op}})e_0. \tag{52}$$

Thus diagonal rescaling preserves the raw rate at an interior K.

For the uniform resolvent bound, let $E = \widehat{K} - K$, $e = \|E\|_{\mathrm{op}}$, and $\kappa = \max\{\|K\|_{\mathrm{op}}, \left\|\widehat{K}\right\|_{\mathrm{op}}\}$. Under (31), if $ae \le \lambda/2$, then for every non-empty $S \in \Pi_{\mathcal{B}}$ the inverse and resolvent bounds are

$$\left\|(A_S \widehat{K} A_S^\top + R_S)^{-1}\right\|_{\mathrm{op}} \le 2/\lambda,$$
$$\left\|(A_S \widehat{K} A_S^\top + R_S)^{-1} - (A_S K A_S^\top + R_S)^{-1}\right\|_{\mathrm{op}} \le 2ae/\lambda^2.$$

Expanding the two outer factors and the inverse in Q_S then yields

$$\left\|Q_S(\widehat{K}) - Q_S(K)\right\|_{\mathrm{op}} \le L_Q e, \qquad L_Q = \frac{4\kappa a}{\lambda} + \frac{2\kappa^2 a^2}{\lambda^2}, \tag{53}$$

uniformly over the feasible family. Thus the constant worsens only through the latent covariance scale, measurement amplification a, and protocol conditioning λ.

B.3 Value Error, Calibration Rates and Regret

For Theorem 10, the empty protocol has value zero under both matrices. For non-empty S, non-negative weights summing to one, the target modulus and (53) imply

$$|F_g(S; \widehat{K}) - F_g(S; K)| \le L(L_Q e)^\beta, \qquad |V_g(\widehat{K}) - V_g(K)| \le L e^\beta.$$

For the local threshold case, the second bound uses that K and $\widehat{K}$ have the same unit diagonal, so only off-diagonal arguments vary. Since $0 \le F_g(S; K) \le V_g(K)$, the ratio identity gives the more explicit uniform bound

$$\sup_{S \in \Pi_{\mathcal{B}}} |\widehat{\mathcal{I}}_g(S) - \mathcal{I}_g(S)| \le \frac{L(L_Q^\beta + 1)e^\beta}{V_g(K) - L e^\beta}. \tag{54}$$

For sufficiently small e, the denominator is at least $v_0/2$ and κ is bounded by a fixed neighbourhood of K. Hence (54) is at most $2L(L_Q^\beta + 1)e^\beta / v_0$, which is (32) with exactly the dependencies stated there. For a threshold target satisfying (33), continuity of all finitely or uniformly conditioned resolvents keeps the varying correlations in a common compact subinterval of $(-1, 1)$; the local Lipschitz constant (51) then applies.

For Theorem 11, fixed-dimensional Gaussian covariance concentration gives

$$\left\|\widehat{\Sigma} - (K + R_0)\right\|_{\mathrm{op}} = O_p(m^{-1/2}).$$

The assumed rate for $\widehat{R}_0$ therefore makes $e_0 = O_p(m^{-1/2})$. Because $K \succ 0$ and $\tau_m \downarrow 0$, the conditions of (52) hold with probability tending to one, giving $\left\|\widehat{K} - K\right\|_{\mathrm{op}} = O_p(m^{-1/2})$.

Substitution into (54) gives root-m value error when $\beta = 1$ and $m^{-1/4}$ under the global threshold envelope $\beta = 1/2$. At a fixed threshold model satisfying (33), the local Lipschitz modulus has $\beta = 1$, restoring the root-m rate. The family is finite, so the conditioning and correlation neighbourhoods can be chosen uniformly over all $S \in \Pi_{\mathcal{B}}$, which gives the stated rate.

For Corollary 12, insert the empirical value on both sides of the comparison between S^* and $\widetilde{S}$:

$$\begin{aligned}\mathcal{I}_g(S^*) - \mathcal{I}_g(\widetilde{S}) &\le \{\mathcal{I}_g(S^*) - \widehat{\mathcal{I}}_g(S^*)\} + \{\widehat{\mathcal{I}}_g(S^*) - \widehat{\mathcal{I}}_g(\widetilde{S})\} + \{\widehat{\mathcal{I}}_g(\widetilde{S}) - \mathcal{I}_g(\widetilde{S})\} \\ &\le 2\varepsilon_m + \eta_m,\end{aligned}$$

which proves (36). Applying the same argument within $\Pi^{(\ell)}$ gives $\mathcal{I}_g(\widehat{S}_\ell) \ge \max_{S \in \Pi^{(\ell)}} \mathcal{I}_g(S) - 2\varepsilon_\ell$; subtracting this lower bound from the global optimum gives (37).

B.4 Best-Linear Calibration

Proposition (Uniform root-m error)**.** *For non-empty S, write $\Sigma_S = \mathrm{Var}(Y_S)$ and $c_S = \mathrm{Cov}(Y_S, \Theta)$, and let $v = \mathrm{Var}(\Theta)$. Suppose, uniformly over $S \in \Pi_{\mathcal{B}}$, that $v \geq v_0 > 0$, $\lambda_{\min}(\Sigma_S) \geq \lambda > 0$, $\|c_S\|_2 \leq C_c < \infty$, and*

$$\delta_m := \sup_{S \neq \varnothing} \{ \left\| \widehat{\Sigma}_S - \Sigma_S \right\|_{\mathrm{op}} + \|\widehat{c}_S - c_S\|_2 \} + |\widehat{v} - v| = O_p(m^{-1/2}).$$

Set both values of the empty protocol to zero; otherwise use $\mathcal{I}_{\mathrm{L}}(S) = c_S^\top \Sigma_S^{-1} c_S / v$ and its plug-in version. Then $\sup_{S \in \Pi_{\mathcal{B}}} |\widehat{\mathcal{I}}_{\mathrm{L}}(S) - \mathcal{I}_{\mathrm{L}}(S)| = O_p(m^{-1/2})$.

In the measurement model, where $\Sigma_S = A_S K A_S^\top + R_S$ and $c_S = A_S c$ for $c = \mathrm{Cov}(Z, \Theta)$, it suffices that $\sup_S \|A_S\|_{\mathrm{op}} < \infty$ and

$$\left\| \widehat{K} - K \right\|_{\mathrm{op}} + \|\widehat{c} - c\|_2 + |\widehat{v} - v| + \sup_S \left\| \widehat{R}_S - R_S \right\|_{\mathrm{op}} = O_p(m^{-1/2}),$$

using $\widehat{\Sigma}_S = A_S \widehat{K} A_S^\top + \widehat{R}_S$ and $\widehat{c}_S = A_S \widehat{c}$.

Proof. On the event $\delta_m \leq \min(\lambda/2, v_0/2)$, the resolvent identity and $q_S = c_S^\top \Sigma_S^{-1} c_S$ give, uniformly over non-empty S,

$$\left\| \widehat{\Sigma}_S^{-1} \right\|_{\mathrm{op}} \leq \frac{2}{\lambda}, \qquad \left\| \widehat{\Sigma}_S^{-1} - \Sigma_S^{-1} \right\|_{\mathrm{op}} \leq \frac{2\delta_m}{\lambda^2},$$
$$\sup_S |\widehat{q}_S - q_S| \leq \frac{2\delta_m}{\lambda}(2C_c + \delta_m) + \frac{2C_c^2 \delta_m}{\lambda^2} = O_p(m^{-1/2}).$$

Because $q_S \leq C_c^2/\lambda$ and $\widehat{v} \geq v_0/2$, division by $\widehat{v}$ and v proves the claim. The primitive rate implies the assumed one through $\left\| \widehat{\Sigma}_S - \Sigma_S \right\|_{\mathrm{op}} \leq \|A_S\|_{\mathrm{op}}^2 \left\| \widehat{K} - K \right\|_{\mathrm{op}} + \left\| \widehat{R}_S - R_S \right\|_{\mathrm{op}}$ and $\|\widehat{c}_S - c_S\|_2 \leq \|A_S\|_{\mathrm{op}} \|\widehat{c} - c\|_2$. □

Appendix C. Design Derivations and Algorithms

C.1 Rank-One Update

For a non-empty S, write $M = A_S K A_S^\top + R_S$, $b = A_S K \ell_a$ and $d = \ell_a^\top K \ell_a + r_a$. The Schur complement of $M_+ = \left(\begin{smallmatrix} M & b \\ b^\top & d \end{smallmatrix}\right)$ is $d - b^\top M^{-1} b = s_{a|S}$, so the block-inverse formula (Horn and Johnson, 2013, §0.8) gives

$$M_+^{-1} = \begin{pmatrix} M^{-1} + M^{-1} b s_{a|S}^{-1} b^\top M^{-1} & -M^{-1} b s_{a|S}^{-1} \\ -s_{a|S}^{-1} b^\top M^{-1} & s_{a|S}^{-1} \end{pmatrix}. \tag{55}$$

Substituting (55) into the definition of Q and collecting the correction as an outer product yields

$$Q_{S \cup \{a\}} - Q_S = \frac{(K\ell_a - K A_S^\top M^{-1} b)(K\ell_a - K A_S^\top M^{-1} b)^\top}{s_{a|S}} = \frac{P_S \ell_a \ell_a^\top P_S}{s_{a|S}}.$$

This is the stated Q update; subtracting it from K gives the P update. For $S = \varnothing$, the same formulas follow directly from $Q_\varnothing = 0$ and $P_\varnothing = K$. Substitution into (9) gives (44), and setting $C_g(r) = r$ gives (45).

A forward pass uses rank-one $O(p^2)$ scores and costs $O(d_{\mathcal{B}}|\mathcal{V}|p^2)$, where $d_{\mathcal{B}} = \max_{S\in\Pi_{\mathcal{B}}} |S|$. The implemented swap sweep directly recomputes at most $|S||\mathcal{V}|$ trial objectives, each in $O(p^2|S| + |S|^3)$, and therefore costs $O\{|S||\mathcal{V}|(p^2|S| + |S|^3)\}$.

For Lemma 14, $S \subseteq S'$ nests the two observation σ-fields. Conditional expectation is an L^2 projection, so projection contraction gives $F_g(S; K) \le F_g(S'; K)$.

C.2 Forward Selection and One-Swap Refinement

In exact-d mode, assume that $\Pi_{\mathcal{B}}$ contains a protocol of cardinality d. An addition $a \notin S$ is *completion-preserving* if there exists $S' \in \Pi_{\mathcal{B}}$ such that $S \cup \{a\} \subseteq S'$ and $|S'| = d$. We call an addition *admissible* if it is completion-preserving in exact-d mode, and if $S \cup \{a\} \in \Pi_{\mathcal{B}}$ in at-most-$\mathcal{B}$ mode.

Algorithm 1 Exact-gain forward selection with one-swap refinement

Require: K, g, ω, catalogue $\mathcal{V}$, family $\Pi_{\mathcal{B}}$; mode exact-d or at-most-$\mathcal{B}$

Initialise $S \leftarrow \varnothing$, $Q_S \leftarrow 0$, $P_S \leftarrow K$
while exact mode has $|S| < d$, or at-most mode has a positive-gain addition **do**
    choose the best admissible addition (gain per cost if costs differ), add it and update Q_S, P_S
end while
while a feasible improving swap exists and the swap cap is not reached **do**
    apply the first improving selected–unselected swap and update Q_S, P_S
end while
return S

Exact mode continues to $|S| = d$; at-most mode stops when no feasible addition has positive gain. Because F_g need not be submodular, a zero-gain action can enable a positive joint gain, so the latter rule is heuristic. Exhaustive optima are reported for small families.

C.3 Comparator Objectives

For the latent-state mutual-information comparator, when $R_S \succ 0$ the target-free objective and marginal gain are

$$F_{\mathrm{MI}}(S) = I(Z; Y_S) = \frac{1}{2} \log\det\Big(I + R_S^{-1/2} A_S K A_S^\top R_S^{-1/2}\Big), \tag{56}$$

$$\Delta_{\mathrm{MI}}(a \mid S) = \frac{1}{2} \log \frac{\ell_a^\top P_S \ell_a + r_a}{r_a}, \tag{57}$$

where the second line assumes $r_a > 0$.

The integrated-posterior-variance comparator minimises $\sum_j \omega_j (P_S)_{jj}$ and has marginal reduction

$$\Delta_{\mathrm{IPV}}(a \mid S) = \sum_j \omega_j \frac{(P_S \ell_a)_j^2}{s_{a|S}}. \tag{58}$$

The linear-target comparator replaces g by the identity and maximises $\omega^\top Q_S(K)\omega$. The kernel-quadrature comparator uses the same linear criterion with $R_S = 0$. Together, these objectives separate target nonlinearity, measurement noise and target-free latent coverage.

Appendix D. Simulation Settings

D.1 Models, Targets and Protocol Families

Unless stated otherwise, Z is a standardised Gaussian trajectory on a uniform grid. For lag u, the stationary correlations are OU, $\rho(u) = e^{-u/\tau}$; Matérn-3/2, $\rho(u) = (1 + 2u/\tau)e^{-2u/\tau}$, scaled so that $\int_0^\infty \rho(u)\,\mathrm{d}u = \tau$; a weighted two-scale OU mixture; and the damped-periodic kernel $e^{-u/6}\cos(2\pi u/3)$. A between-unit trait share α gives $K_{jk} = \alpha + (1-\alpha)\rho(|t_j - t_k|)$. The five targets in the design comparison are the temporal mean, occupation above thresholds 0 and 1.5, two-sided occupation outside $[-1.2, 1.2]$, and the logistic aggregate $g(z) = \{1 + e^{-2z}\}^{-1}$. The calibration experiments use the target subsets specified below.

Experiment	Latent and calibration model	Actions, family and repetitions
Uniform error, Figure 2(a)	$T = 20$, $p = 128$, OU $\tau = 1$, $\alpha = 0$; known calibration-noise variance 0.25, action-noise variance 0.5; mean and occupation-at-zero targets	12 equally spaced point actions, exactly four selected (495 protocols); $m \in \{25, 50, 100, 250, 500, 1000\}$ and 200 replications.
Selection regret, Figure 2(b)	$T = 20$, $p = 128$; trait–OU $(\alpha, \tau) = (0.2, 2)$ and two-scale OU $(\tau_1, \tau_2, w) = (0.5, 6, 0.5)$; noiseless calibration and action-noise variance 0.5; mean and occupation targets at thresholds zero and one	Points at $\{1, 3, \ldots, 19\}$, width-three windows centred at 5 and 15, and pre-specified three-replicate point packages at 2 and 18; exactly four of 14 actions; the same m grid and replication count as panel (a).
Nested protocol classes, Figure 2(c,d)	$T = 20$, $p = 64$; noiseless calibration; occupation-at-zero target; action-noise variance 0.4; $K_{jk} = 0.15 + 0.85\exp\{-\lvert t_j - t_k\rvert/[\tau(t_j)\tau(t_k)]^{1/2}\}$, with $\tau(t)$ linear from 0.3 to 3	Exactly four actions are selected. L1 compares contiguous with dispersed layouts; L2 adds early, middle, late and full-horizon phases; L3 uses supports on eight coarse bins; L4 adds 495 fine-grid four-point subsets. The cumulative, deduplicated class sizes are 2, 5, 75 and 568; the same m grid and 30 replications.

Table 2: Calibration and nested protocol-class configurations.

For the target-aware design comparison in Table 1, all settings use $T = 10$, $p = 128$ and $\alpha = 0$. The stationary OU and Matérn-3/2 settings use $\tau = 1$ and action-noise variance 0.4. The horizon-varying setting uses

$$K_{jk} = \exp\{-|t_j - t_k|/[\tau(t_j)\tau(t_k)]^{1/2}\}, \qquad \tau(t) = 0.5 + 3.5t/10,$$

with action-noise variance 0.15. The recency-weighted OU setting uses $\tau = 1$, target weights proportional to $e^{-(T-t)/6}$ and action-noise variance 0.4. The heterogeneous-action setting uses OU scale $\tau = 0.7$ and time-dependent action-noise variance $1 + 2t/10$.

The first four settings use 12 point actions at $\{0.5, 0.5 + 9/11, \ldots, 9.5\}$ and select exactly four actions. The heterogeneous family uses six centres $0.5 + 1.8k$, $k = 0, \ldots, 5$, widths $\{0, 1.5, 3.5\}$, cost $1 + 0.15w$ and a cost budget of four. All feasible protocols are enumerated to obtain the target-specific optimum. The linear comparator uses the actual action noise, kernel quadrature its standard noiseless criterion, and mutual information and integrated posterior variance the same covariance and action catalogue. Greedy search starts from the empty protocol; one-swap refinement takes the best feasible exchange until no improvement remains or 3 exchanges have been accepted.

D.2 Numerical Evaluation of Nonlinear Targets

For one-sided threshold targets, C_g is evaluated from the Plackett integral. If $G_c(r)$ denotes the corresponding one-sided covariance transform, then the two-sided occupation target uses

$$C_{\mathrm{two},c}(r) = 2\{G_c(r) + G_c(-r)\}.$$

The logistic transform is evaluated through a truncated Hermite expansion, with coefficients computed by Gauss–Hermite quadrature.

Appendix E. Sleep-EDF and Long-Term AF Analysis Details

E.1 Annotation Mapping and Estimands

Proportional binning maps unequal-length records to relative time. A Sleep action reads the distinct midpoint epoch of its bin, so adjacent anchors need not be consecutive raw epochs; an AF action averages one bin. With common-grid trajectory $X_i \in [0,1]^p$, protocol S observes $Y_{i,S} = A_S X_i$, while Θ_i is the exact state proportion over the complete analysable interval, not the anchor-grid mean.

Sleep retains pre- and post-sleep Wake and excludes movement and unknown epochs, leaving 3818 annotated hours (median 22.6 per record); night assignments follow `ST-subjects.xls`. The zero-cost baseline C_i contains SC/ST, valid-stage duration and placebo/temazepam, with treatment zero outside ST; neither it nor the validity mask is charged to the temporal budget. All stages use $w_i^{(0)} = 1/n_{s(i)}$ when subject $s(i)$ contributes $n_{s(i)}$ records. For AF, analysis starts at the first rhythm marker and excludes any unannotated prefix from the target and candidate windows. Templates average N bins, observing fraction $N/96$ despite record-specific elapsed time; records are treated as the analysis units and receive equal weights.

The four quantities serve different roles. $\mathcal{I}_g(S)$ is the population protocol-value estimand, and $J_{\tau|C}(S)$ is the foldwise criterion used to select Sleep supports. The pooled $\widehat{R}^2_{\mathrm{cf}}(S)$ reports held-out predictive performance, while the full-sample $\widehat{\mathcal{I}}_{\mathrm{L,p}}(S)$ is used only for the AF grid-sensitivity analysis.

E.2 Foldwise Selection and Held-Out Prediction

Within each Sleep outer-training fold, weighted least squares on an intercept and C gives residual anchors $X^{\perp}$, target $\Theta^{\perp}$ and moments

$$\widehat{\Sigma}_{0|C} = \mathrm{Cov}_w(X^{\perp}), \qquad \widehat{c}_{|C} = \mathrm{Cov}_w(X^{\perp}, \Theta^{\perp}), \qquad \widehat{v}_{|C} = \mathrm{Var}_w(\Theta^{\perp}). \tag{59}$$

Writing $\widehat{\Sigma}_{0|C} = U\,\mathrm{Diag}(\widehat{\lambda}_j)U^{\top}$, the fold-specific repair is

$$\widetilde{\Sigma}_{\tau} = U\,\mathrm{Diag}\{\max(\widehat{\lambda}_j, \tau)\}U^{\top}, \qquad \tau = q^{-1}\,\mathrm{tr}(\widehat{\Sigma}_{0|C})/p, \tag{60}$$

where q is the fold-specific number of independent training subjects (79–81; median 80); repeated nights do not increase it. Forward selection followed by at most 3 accepted one-swap updates searches for a high-scoring support under

$$J_{\tau|C}(S) = \frac{\widehat{c}_{|C}^{\top} A_S^{\top} (A_S \widetilde{\Sigma}_{\tau} A_S^{\top})^{-1} A_S \widehat{c}_{|C}}{\widehat{v}_{|C}}. \tag{61}$$

The target-agnostic learned comparator converts the same repaired residual covariance to correlations and applies greedy search to the noiseless kernel-quadrature criterion for the uniform mean of standardised, nonconstant anchors, without $\widehat{c}_{|C}$ or target outcomes. AF and fixed Sleep templates are pre-specified.

Given a support, outer-training ridge with an intercept selects among 25 log-spaced penalties on $[10^{-6}, 10^{2}]$ by three-fold inner cross-validation, grouped by subject for Sleep and record for AF. Centring and scaling use each inner-training split and are recomputed on the full outer-training fold; Sleep baselines are unpenalised, temporal features penalised, and predictions unclipped.

Predictions from five outer test folds are pooled. With weighted target mean $\bar{\Theta}_w$,

$$\widehat{R}^2_{\mathrm{cf}}(S) = 1 - \frac{\sum_i w_i \{\Theta_i - \widehat{\Theta}_i^{(-f(i))}\}^2}{\sum_i w_i (\Theta_i - \bar{\Theta}_w)^2}. \tag{62}$$

The statistic is computed once from all pooled held-out predictions rather than by averaging foldwise values; as an out-of-sample R^2, it can be negative.

E.3 Resampling and Sensitivity Analyses

Conditional 2.5–97.5 percentile ranges resample fixed-template held-out pairs 2000 times by Sleep subject and 2000 times by AF record. SC/ST-stratified Sleep draws give equal total weight per sampled subject. Because the fitted predictors remain fixed, these ranges describe variation in the realised held-out pairs conditional on those predictors. Pipeline stability is assessed by 1000 study-stratified samples of 80% of subjects without replacement. Every subsample reruns standardisation, moment estimation, support selection, inner tuning, predictor fitting and held-out evaluation, while keeping each subject within one outer and inner fold.

Alternative weighting and covariance-floor choices preserved the fixed-template ordering but changed the selected supports, consistent with the instability of fine support selection reported in the main analysis. Subject- versus recording-weighted moments preserve the fixed-template score sign in all 15 cells but yield support overlap as low as 0.14. Floor multipliers $\{0.5, 1, 2\}$ give maximum outer-fold mean contrast range 0.0074 and overlap 0.33; the fixed-template ordering is stable while the selected support varies.

SC-only and ST-only sensitivities rerun the full subject-grouped pipeline. At $N = 16$, Table 3 gives all target contrasts. For AF, two outcome-defined cohorts assess sensitivity to excluding records with AF burden at or near zero or one; Table 4 reports all fixed-template results at $N = 4$ and $N = 16$. The main text summarises all 30 Sleep source cells by positive, negative and unresolved ranges.

Target at $N = 16$	pooled adjusted	SC adjusted	ST adjusted
REM	$+0.079[-0.010, +0.168]$	$+0.221[-0.010, +0.428]$	$+0.161[-0.445, +0.650]$
N3	$+0.151[+0.058, +0.290]$	$-0.293[-0.430, -0.155]$	$+0.554[+0.286, +0.900]$
Wake	$+0.019[+0.005, +0.040]$	$+0.321[+0.088, +0.529]$	$-0.123[-0.624, +0.890]$

Table 3: Dispersed-minus-contiguous cross-fitted R^2 differences and conditional 2.5–97.5 percentile ranges at the middle Sleep budget.

		$N = 4$ (4.17%)		$N = 16$ (16.67%)	
Cohort	records	contiguous	dispersed	contiguous	dispersed
All records	84	$+0.696$	$+0.971$	$+0.851$	$+0.998$
$0 < \Theta_{\mathrm{AF}} < 1$	71	$+0.596$	$+0.958$	$+0.748$	$+0.998$
$0.05 \leq \Theta_{\mathrm{AF}} \leq 0.95$	33	$+0.180$	$+0.829$	$+0.430$	$+0.990$

Table 4: Cross-fitted R^2 in the primary AF cohort and two outcome-defined sensitivity cohorts.

The AF grid diagnostic complements held-out prediction by measuring discretisation sensitivity. For $p \in \{64, 128, 256\}$, let $F_i^{(p)}$ be record i's binwise raw AF fractions and $\Theta_i^{(p)} = p^{-1}\mathbf{1}^\top F_i^{(p)}$. On all 84 records, set $\Sigma_F = \mathrm{Cov}(F^{(p)})$, $c_F = \mathrm{Cov}(F^{(p)}, \Theta^{(p)})$, $v_F = \mathrm{Var}(\Theta^{(p)})$ and compute

$$\widehat{\mathcal{I}}_{\mathrm{L,p}}(S) = \frac{c_F^\top A_S^\top (A_S \Sigma_F A_S^\top)^+ A_S c_F}{v_F}.$$

Without flooring, singular values at most 10^{-10} times the largest are discarded and windows use exact fractional overlap between their continuous time intervals and the grid bins. The four diagnostics are a centred one-hour contiguous window, a centred six-hour contiguous window, four uniformly dispersed 15-minute windows and eight uniformly dispersed 15-minute windows. Across these templates, the largest within-protocol range is 0.0073. This diagnostic quantifies numerical grid sensitivity; Figure 3(b) and Table 4 provide the held-out and matched-budget comparisons.

References

A. Alexanderian. Optimal experimental design for infinite-dimensional Bayesian inverse problems governed by PDEs: A review. *Inverse Problems*, 37(4):043001, 2021.

A. Attia, A. Alexanderian, and A. K. Saibaba. Goal-oriented optimal design of experiments for large-scale Bayesian linear inverse problems. *Inverse Problems*, 34(9):095009, 2018.

F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.

E. Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016. doi: 10.1073/pnas.1510507113.

D. Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

F.-X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic. Probabilistic integration: A role in statistical computation? *Statistical Science*, 34(1):1–22, 2019.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8:6085, 2018. doi: 10.1038/s41598-018-24271-9.

A. Das and D. Kempe. Approximate submodularity and its applications: Subset selection, sparse approximation and dictionary selection. *Journal of Machine Learning Research*, 19(3):1–34, 2018.

T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71, 1997.

G. Elfving. Optimum allocation in linear regression theory. *The Annals of Mathematical Statistics*, 23(2): 255–262, 1952.

K. Fukunaga and D. M. Hummels. Bayes error estimation using Parzen and k-NN procedures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(5):634–643, 1987.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

P. Hall, H.-G. Müller, and J.-L. Wang. Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34(3):1493–1517, 2006. doi: 10.1214/009053606000000272.

N. J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118, 1988.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 2 edition, 2013.

C. Iber, S. Ancoli-Israel, A. L. Chesson, and S. F. Quan. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. American Academy of Sleep Medicine, Westchester, IL, 2007.

T. Ishida, I. Yamane, N. Charoenphakdee, G. Niu, and M. Sugiyama. Is the performance of my deep network too good to be true? a direct approach to estimating the Bayes error in binary classification. In *International Conference on Learning Representations*, 2023.

H. Ji and H.-G. Müller. Optimal designs for longitudinal and functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):859–876, 2017.

S. Joshi and S. Boyd. Sensor selection via convex optimization. *IEEE Transactions on Signal Processing*, 57 (2):451–462, 2009.

N. Kallus and M. Uehara. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

B. Kemp, A. H. Zwinderman, B. Tuk, H. A. C. Kamphuisen, and J. J. L. Oberyé. Analysis of a sleep-dependent neuronal feedback loop: The slow-wave microcontinuity of the EEG. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000.

P. Kidger, J. Morrill, J. Foster, and T. Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems 33*, pages 6696–6707. Curran Associates, Inc., 2020.

T. C. Koopmans and O. Reiersøl. The identification of structural characteristics. *The Annals of Mathematical Statistics*, 21(2):165–181, 1950.

A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008.

L. Le Cam and G. L. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, New York, 1990.

Y. Li and T. Hsing. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38(6):3321–3351, 2010. doi: 10.1214/10-AOS813.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956. doi: 10.1214/aoms/1177728069.

D. J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4): 590–604, 1992.

C. F. Manski. *Partial Identification of Probability Distributions*. Springer, New York, 2003.

K. Mohan and J. Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037, 2021. doi: 10.1080/01621459.2021.1874961.

M. S. Mourtazaev, B. Kemp, A. H. Zwinderman, and H. A. C. Kamphuisen. Age and gender affect different characteristics of slow waves in the sleep EEG. *Sleep*, 18(7):557–564, 1995.

R. Nabi, R. Bhattacharya, and I. Shpitser. Full law identification in graphical models of missing data: Completeness results. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7153–7163. PMLR, 2020.

J. Pearl and E. Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014. doi: 10.1214/14-STS486.

R. Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.

S. Petrutiu, A. V. Sahakian, and S. Swiryn. Abrupt changes in fibrillatory wave characteristics at the termination of paroxysmal atrial fibrillation in humans. *Europace*, 9(7):466–470, 2007.

R. L. Plackett. A reduction formula for normal multivariate integrals. *Biometrika*, 41(3/4):351–360, 1954.

T. Pollard, B. E. Moody, L.-w. Lehman, B. Gow, C. Fernandes, C. Xie, A. Johnson, R. G. Mark, and T. Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 2026. doi: 10.1038/s44360-026-00096-z.

D. Precup, R. S. Sutton, and S. P. Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 759–766. Morgan Kaufmann, 2000.

F. Pukelsheim. *Optimal Design of Experiments*. Number 50 in Classics in Applied Mathematics. SIAM, Philadelphia, PA, 2006.

E. M. Pullenayegum and L. S. H. Lim. Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. *Statistical Methods in Medical Research*, 25 (6):2992–3014, 2016. doi: 10.1177/0962280214536537.

N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10:2349–2374, 2009.

T. Rainforth, A. Foster, D. R. Ivanova, and F. Bickford Smith. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024. doi: 10.1214/23-STS915.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, 2 edition, 2005.

A. Rechtschaffen and A. Kales, editors. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects*. NIH Publication No. 204. U.S. Government Printing Office, Washington, DC, 1968.

T. J. Rothenberg. Identification in parametric models. *Econometrica*, 39(3):577–591, 1971.

Y. Rubanova, R. T. Q. Chen, and D. K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.

N. Sachdeva, Y. Su, and T. Joachims. Off-policy bandits with deficient support. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 965–975. ACM, 2020. doi: 10.1145/3394486.3403139.

B. Settles. *Active Learning*. Morgan and Claypool, 2012.

H. Shim, S. J. Hwang, and E. Yang. Joint active feature acquisition and classification with variable-size set encoding. In *Advances in Neural Information Processing Systems 31*, 2018.

S. N. Shukla and B. M. Marlin. A survey on principles, models and methods for learning from irregularly sampled time series, 2020. arXiv:2012.00168.

E. Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2:167–195, 2010.

E. Torgersen. *Comparison of Statistical Experiments*. Cambridge University Press, Cambridge, 1991.

H. von Kleist, A. Zamanian, I. Shpitser, and N. Ahmidi. Evaluation of active feature acquisition methods for time-varying feature settings. *Journal of Machine Learning Research*, 26(60):1–84, 2025.

C. Weaver, L. Xiao, and W. Lu. Functional data analysis for longitudinal data with informative observation times. *Biometrics*, 79(2):722–733, 2023. doi: 10.1111/biom.13646.

Z. Xing, J. Pei, and P. S. Yu. Early prediction on time series: A nearest neighbor approach. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 1297–1302. IJCAI Organization, 2009.

F. Yao, H.-G. Müller, and J.-L. Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005. doi: 10.1198/016214504000001745.

X. Zhang. The label defines the timescale: Trait–state limits of temporal-aggregate learning, 2026. arXiv:2608.01587.

Z.-H. Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018.