WHEN CAN A TRAJECTORY BE ALIGNED? AN INFORMATION BOUND FOR AMPLITUDE–PHASE SEPARATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Elastic functional data analysis separates trajectories into amplitude and phase by warping each one onto a common template and taking the warps as the phase, and recent work makes the warp a neural network that integrates a positive velocity field. Developed for densely sampled curves, it is now applied to trajectories with very few samples: contrast-enhanced MRI with three post-contrast frames, or ICU chemistry panels drawn three times in two days. We compute the Fisher information a set of measurement times carries about a warp parameter. Under the within-exam baseline normalisation these pipelines apply, it depends on the centred derivative of the response and the tangent direction of the warp family, neither of which involves the estimator, so the bound follows from acquisition metadata alone and holds for a network of arbitrary capacity. It implies that a warp pinning both ends of the time axis, the construction these differentiable pipelines are built from, cannot represent a rigid arrival delay at all; that a response with constant derivative carries no timing information once its onset precedes the window, at any sample size or SNR; and that fitting an unknown amplitude costs $3.7\times$ on a near-monotone response. Across six warp families and three encoders, in simulation and on real acquisitions, no configuration exceeds the bound, and the endpoint-pinned family recovers 0.595 of the arrival variation against 1.03 for free-endpoint ones: impaired on real scans rather than disabled as it is in simulation. On $3{,}303$ prostate DCE-MRI examinations across two protocols, thinning a dense acquisition degrades a measured arrival estimate from 0.25 s at forty samples to 20.1 s at three, and the bound tracks that degradation across a 119-fold range ($r = 0.96$). Sample count alone orders those schedules equally well; holding it at six and moving only where the six fall, the bound still orders the measured error. Applied to PhysioNet 2012 timestamps alone, the same calculation spans $10.6\times$ across 25 ICU variables.

1 INTRODUCTION

A contrast-enhanced MRI examination injects a bolus of contrast agent and images the same tissue repeatedly as the agent arrives, accumulates and washes out, producing one intensity trajectory per region of interest. Trajectories from different patients vary in magnitude and in timing, and the two carry different information: the magnitude reflects how much agent the tissue takes up, while the timing reflects how quickly blood reaches it, so that a delayed arrival indicates impaired perfusion. Estimating the timing separately from the magnitude is a standard objective here and in any setting where a shared response is observed on different clocks in different subjects.

Elastic functional data analysis provides the standard decomposition. Each trajectory is warped onto a common template by a monotone reparameterisation of time, leaving amplitude in the aligned trajectories and phase in the warps (Srivastava et al., 2011; Srivastava & Klassen, 2016). Recent work replaces the dynamic-programming solver with a neural network that produces the warp by integrating a positive velocity field (Xiong et al., 2026), which makes the decomposition differentiable and allows structure to be shared across trajectories.

The decomposition was developed for functions observed on dense grids, and is increasingly applied to sequences of a few samples. In that regime an alignment procedure remains well defined and

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

returns a warp for every trajectory, while its output gives no indication of whether the estimated warp reflects timing present in the data or an arbitrary selection from a set of warps that fit the observations equally well. This is what the identifiability results following nonlinear ICA ask of latent-variable models (Khemakhem et al., 2020; Locatello et al., 2019) – what a class of models structurally cannot recover – asked of the sampling design rather than the population: identifiability here is a property of the acquisition, and is not diagnosable from a fitted model.

This paper contributes the following.

- A closed-form Fisher information for a warp parameter under the within-exam baseline normalisation these pipelines apply. It involves the response only through its derivative and the family only through its tangent direction, so it is computable from a protocol before anything is acquired, and it bounds any unbiased estimate, a network of arbitrary capacity included.
- Three consequences, two falling on constructions now in use. A response rising at a constant rate from before the window carries no timing information at any sample size or SNR; a warp family pinning both ends of the time axis, which is how the differentiable pipelines of Section 2 are built, cannot represent a rigid delay at all; and fitting an unknown amplitude costs a factor set by the response shape and the measurement times.
- Validation on 3,303 prostate examinations across two protocols, a public breast cohort and an ICU benchmark, where the bound tracks the precision lost to coarser sampling and no learned model of six warp families and three encoders exceeds it.

2 Related work

Amplitude and phase variation. Decomposing a sample of curves into magnitude and timing variation is long-standing in functional data analysis. Kneip & Gasser (1992) introduced structural averaging so that landmark timing is not smeared by pointwise means, and Wang & Gasser (1997) formalised alignment through dynamic time warping. The square-root velocity function framework (Srivastava et al., 2011; Srivastava & Klassen, 2016) placed the problem on a metric footing, carrying the reparametrisation-invariant Fisher–Rao metric to an ordinary L^2 distance so that alignment becomes a distance computation. It extends to generative models in which amplitude and phase are drawn jointly (Tucker et al., 2013); Marron et al. (2015) survey the area.

Learned warps and models for irregular sampling. Recent alignment models parametrise the warp with a neural network, integrating a positive velocity field (Xiong et al., 2026) with the neural ODE machinery of Chen et al. (2018). These sit within a broader literature on learning from irregularly

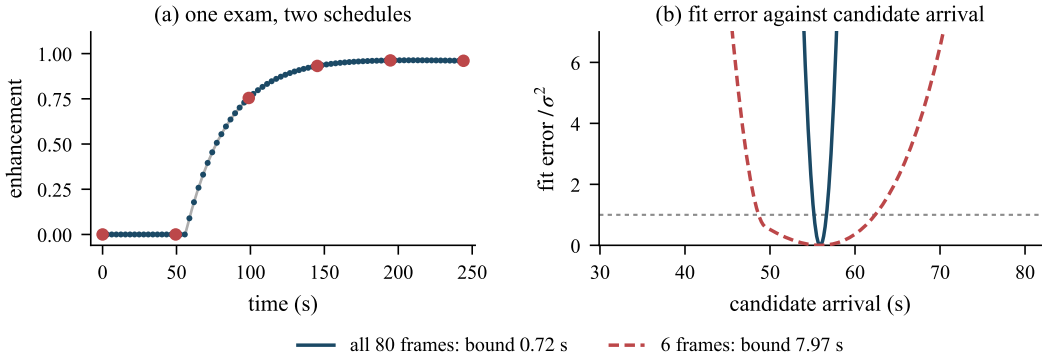

Figure 1: One examination from the 80-frame protocol, at 3.09 s frame spacing and SNR 20. **(a)** The full acquisition and the six frames a sparse schedule keeps. **(b)** The fit error each schedule gives as a function of the candidate arrival, divided by the noise variance, which makes it the log-likelihood ratio against the best fit; the dashed line marks 1. Eighty frames put the arrival within a second, six within eight. The bound in the legend is computed from the measurement times alone, so it is known before the exam is acquired.

Table 1: Ways to ask whether an alignment is identifiable (✓ yes, × no, – not applicable). The first two need the data in hand; the third concerns a model class rather than a schedule; the fourth bounds kinetic parameters under no within-exam normalisation.

	before acquisition	estimator-free	per-schedule number	warp parameter	baseline centring
reconstruction error of a fit	×	×	✓	✓	×
profile likelihood (Raue et al., 2009)	×	×	✓	–	×
nonlinear-ICA identifiability	✓	✓	×	–	×
Fisher design, tracer kinetics	✓	✓	✓	×	×
this paper	✓	✓	✓	✓	✓

sampled series (Shukla & Marlin, 2020): hidden-state decay (Che et al., 2018), continuous-time latent dynamics (Rubanova et al., 2019; Kidger et al., 2020), learned temporal embeddings with attention (Shukla & Marlin, 2021), and set-valued encoders over time-value pairs (Horn et al., 2020).

Identifiability. Nonlinear independent component analysis and the identifiability results following it establish that latent factors may be unrecoverable without auxiliary structure (Hyvärinen & Morioka, 2016; Khemakhem et al., 2020; Locatello et al., 2019), at the population level where the failure persists at infinite sample size. Raue et al. (2009) separate structural from practical non-identifiability using the profile likelihood, for data already in hand.

Tracer kinetics and experiment design. Tracer-kinetic modelling estimates physiological parameters from enhancement curves (Tofts et al., 1999; Sourbron & Buckley, 2013), and the temporal resolution they require is long studied: Henderson et al. (1998) showed for breast DCE-MRI that they degrade sharply as sampling coarsens. Fisher information is likewise established for selecting MR measurements, in diffusion imaging (Alexander, 2008) and parallel imaging (Bouhrara et al., 2018), and the derivation in Section 3 follows standard theory (Lehmann & Casella, 1998). That work asks which kinetic parameters a design determines, for a signal model carrying no within-exam normalisation.

3 The bound

3.1 Model and notation

A trajectory y is observed at n measurement times $t_1 < \cdots < t_n$, where n is the number of frames the protocol acquires. We model it as a fixed response shape f, scaled by a per-trajectory amplitude a, with its time axis deformed by a warp γ_θ drawn from a family indexed by a single scalar θ:

$$y(t_i) \;=\; a\, f(\gamma_\theta(t_i)) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

The warp family determines what θ means. Throughout we take $\gamma_\theta(t) = t + \theta$, a *rigid delay* that translates the whole trajectory in time without changing its shape, so that θ is the bolus arrival time and is the quantity to be estimated. Section 3.4 treats other families.

Pipelines in this area normalise each trajectory against itself, by subtracting or dividing by a measurement taken at t_1, so that every trajectory begins at zero and differences in scanner gain or patient size are removed. Under that normalisation (1) becomes

$$y(t_i) \;=\; a\,[f(\gamma_\theta(t_i)) - f(\gamma_\theta(t_1))] + \varepsilon_i. \tag{2}$$

3.2 Fisher information under baseline normalisation

For a scalar parameter θ observed under additive Gaussian noise of variance σ^2, the Fisher information is the expected squared sensitivity of the log-likelihood,

$$I(\theta) \;=\; \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(\frac{\partial\, \mathbb{E}[y(t_i)]}{\partial \theta} \right)^2, \tag{3}$$

and the Cramér–Rao bound $I(\theta)^{-1/2}$ lower-bounds the standard deviation of any unbiased estimate of θ. Writing $v = \partial_\theta \gamma_\theta$ for the tangent direction of the warp family, differentiating (2) and substituting into (3) gives

$$I(\theta) \;=\; \frac{a^2}{\sigma^2} \sum_{i=1}^{n} \Big[\underbrace{f'(\gamma(t_i))\, v(t_i) - f'(\gamma(t_1))\, v(t_1)}_{\text{centred}} \Big]^2. \tag{4}$$

The response enters only through its derivative f' and the warp family only through v. The difference of the two terms is the trace of the baseline normalisation, which measures every sample against the first. Since (4) depends on no property of the estimator, it can be evaluated from a set of intended measurement times and an assumed response shape in advance of acquisition. Three consequences are derived below and verified numerically in Appendix K.

Equation (3) assumes independent noise. For a covariance Σ the general form is $I(\theta) = s^\top \Sigma^{-1} s$ with $s_i = \partial\, \mathbb{E}[y(t_i)]/\partial\theta$, and Section 5.1 measures the residuals against a first-order autoregressive process, AR(1), in which each sample's noise carries a fraction ρ of the previous sample's. For a schedule retaining frames at original indices $k_1 < \cdots < k_n$ that gives

$$\Sigma_{ij} \;=\; \sigma^2 \rho^{|k_i - k_j|}, \qquad I(\theta) \;=\; s^\top \Sigma^{-1} s, \tag{5}$$

with (3) the case $\rho = 0$. A schedule's gaps enter through the exponent, so decimation decorrelates the samples it keeps.

3.3 Constant-derivative responses

For a rigid delay $v \equiv 1$, so (4) reduces to a sum over the differences $f'(t_i + \theta) - f'(t_1 + \theta)$.

Proposition 1 (Constant-derivative responses)**.** *Let $\gamma_\theta(t) = t + \theta$ and suppose f' is constant on $[t_1 + \theta,\, t_n + \theta]$. Then $I(\theta) = 0$ for every sample size n, every noise level σ and every set of measurement times, so no unbiased estimator of θ has finite variance.*

Every difference vanishes term by term; Appendix G gives the argument in full.

This holds only once the response is already rising at the first measurement. An onset *inside* the window puts a discontinuity in f' there, and that discontinuity is informative: translating the trajectory moves the onset relative to the samples. Once the onset precedes t_1 the observed trajectory is an unbroken ramp and the information is zero.

3.4 Pinned endpoints

The warp families used in this literature hold both ends of the time axis fixed, with $\gamma(t_1) = t_1$ and $\gamma(t_n) = t_n$, whether constructed as a softplus integral or a neural ODE flow (Srivastava et al., 2011; Xiong et al., 2026).

Proposition 2 (Pinned endpoints exclude a delay)**.** *Let $\{\gamma_\theta\}$ satisfy $\gamma_\theta(t_1) = t_1$ and $\gamma_\theta(t_n) = t_n$ for all θ. Then every tangent direction obeys $v(t_1) = v(t_n) = 0$, and the rigid delay $t \mapsto t + \theta$ with $\theta \neq 0$ is not a member of the family, for any capacity or sample size.*

It can approximate one on the interior, with the residual concentrated at the boundary (Figure 2a; Appendix K). Equation (4) is not specific to a delay: for the power family $\gamma_\theta(t) = t^\theta$ of Section 5.3, $v = t^\theta \ln t$. A family is suited to a deformation when its tangent directions align with the centred derivative, and blind to one when they do not.

3.5 Null space

Equation (4) is the squared norm of a vector obtained by applying a linear map to the tangent direction,

$$A : v \;\longmapsto\; C[f'(t)\, v(t)]\,, \qquad Cu = u - u(t_1),$$

in which C performs the baseline subtraction. The unidentifiable directions are those A maps to zero, namely $\{v : f'(t)v(t) \equiv \text{const}\}$, and this set is never empty: for any response shape the deformation

$$v_{\text{null}}(t) \;\propto\; 1/f'(t)$$

carries no information. Since v_{null} is inversely proportional to the derivative, it is concentrated wherever the response is flattest.

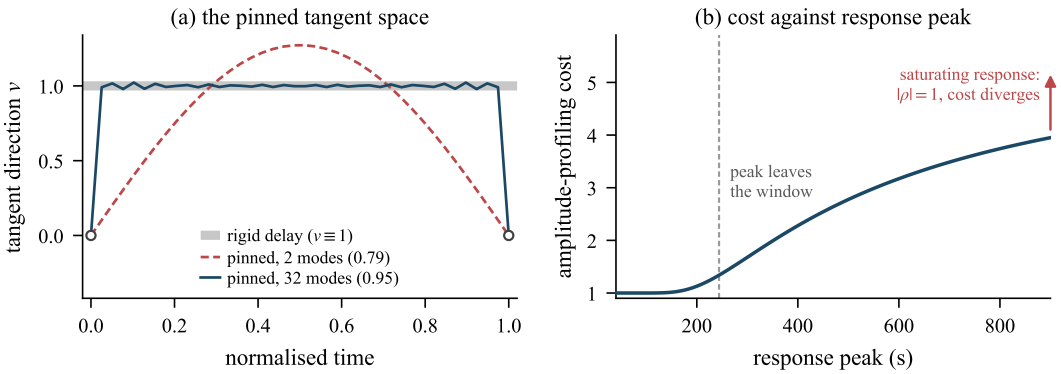

Figure 2: A pinned warp family on normalised time and a gamma-variate response against a 244 s window. **(a)** Pinning both ends of the time axis forces every tangent direction to vanish there (circles), so $v \equiv 1$ is not in the family and capacity buys interior accuracy, not the endpoints. **(b)** The cost of (6) as the response peak leaves the window: while the response is still rising the amplitude and timing scores point nearly the same way, and for a saturating response they are exactly parallel.

Exact degeneracies. In practice the amplitude is unknown and is fitted alongside θ, and the two can be confounded to a degree that depends on the similarity of their effects on the observations. Writing $C[f]$ for the effect of perturbing the amplitude and $C[f'v]$ for the effect of perturbing θ:

Proposition 3 (Amplitude–timing confound)**.** *With the amplitude unknown and profiled out, the bound on θ is inflated by $(1-\rho^2)^{-1/2}$, where ρ is the cosine between the two score vectors below, and the cost diverges exactly when those two are parallel.*

$$\text{efficiency cost} \;=\; \big(1-\rho^2\big)^{-1/2}, \qquad \rho \;=\; \frac{\langle C[f],\, C[f'v]\rangle}{\|C[f]\|\,\|C[f'v]\|}. \tag{6}$$

Here ρ is a cosine rather than a Pearson correlation, and Figure 2b gives the cost it implies. The saturating exponential $f(t) = 1 - e^{-kt}$ makes the two scores exactly parallel, since $C[f'] = -k\,C[f]$ holds identically for any t_1 (Appendix G), so profiling the amplitude removes the timing entirely and neither more samples nor higher SNR repairs it. That is the persistent-enhancement curve, one of the three standard kinetic types in contrast imaging. The constant-derivative response of Section 3.3 is not a case of $|\rho| = 1$: there $C[f'v]$ vanishes identically, so ρ is undefined and Proposition 1 has already disposed of it.

Scope of the kernel. Carrying no first-order information and being undetermined by the data are distinct properties, so the kernel does not license projecting v_{null} out of a learned warp: a deformation along it stays visible at second order, and measurably so (Appendix G). The bounds below concern a warp parameter in a parametric family, where the first-order statement is the applicable one.

4 Experimental Setup

Prostate DCE-MRI. An internal cohort of 4,936 contrast-enhanced prostate examinations on GE and Siemens scanners, of which 2,279 acquire 35 frames, 1,578 acquire 80 and 592 acquire 50. Section 5.1 uses the 80-frame protocol, which samples at a median of 3.1 s, and repeats the analysis on the 35-frame protocol, which samples at a median of 9.1 s and supplies 1,832 further exams on a disjoint set of patients. Filtering to exams with a physical, uniform acquisition clock and one exam per patient leaves 1,473 with an extracted trajectory, and 1,832 of the 35-frame exams (Appendix H). The 50-frame protocol is excluded, 591 of its 592 exams carrying an irregular clock, so reconstructing their sampling times would fabricate the schedule under test. The cohort has no lesion segmentations, so trajectories come from the decile of gland voxels with the largest enhancement; Section 6 measures what that substitution costs. This cohort cannot be redistributed, so we release its per-exam measurement times and fitted arrival estimates, which are the inputs to every number in Figure 3.

MAMA-MIA. A public breast DCE-MRI dataset (Garrucho et al., 2025) pooling four trial collections: ISPY2 (980 curves), DUKE (291), ISPY1 (171) and NACT (64), with medians of 6, 4, 3 and 3 frames at intervals of 90, 117, 341 and 376 s. Expert tumour segmentations ship with the dataset, and we use them both for the trajectories and to measure the prostate region-of-interest substitution.

PhysioNet Challenge 2012. A public ICU benchmark (Silva et al., 2012) of 12,000 ICU stays, 4,000 of them in training set A, carrying 37 time-series variables measured at irregular times. We read the first 900 records of set A and keep the 25 variables recorded for at least 400 of those patients, using at most 250 patients per variable. The three limits are computational. The bound is evaluated per variable, so a larger subset would tighten each estimate and leave the ordering between variables unchanged. Section 5.4 reads only the measurement timestamps and none of the recorded values.

Subsampling design. No dataset here records a true arrival time, so Section 5.1 substitutes a more precise measurement for it. Comparing the width of a fitted confidence interval against the predicted bound would not serve, since both are computed from the same measurement times through the same response shape and are coupled independently of the data: permuting trajectories across patients reproduces the agreement in full. Throughout, (4) is evaluated at $\theta = 0$, which places the response onset at the first sample and is the quantity a protocol determines before any exam exists. The 80-frame protocol determines the arrival to a predicted 0.72 s, so we estimate it on the full acquisition, retain that as the reference, then discard samples from the same acquisition, refit, and record the displacement. Density is therefore manipulated within patient, so tissue, injection, scanner, reconstruction and region of interest are identical across schedules. Schedules are specified relative to each acquisition's cadence, since the per-exam median interval ranges from 1.77 to 6.08 s and a fixed schedule would truncate the window for half the cohort. Every retained sample is acquired; nothing is interpolated. Errors are summarised by the median absolute value: a minority of scans land in a second, wrong solution and dominate a squared average, which at twenty samples reads 3.42 s against a median of 0.55.

Warp families and encoders. Section 5.3 sweeps six families: an affine warp, a monotone piecewise-linear warp with a free offset, a direct delay, the softplus-integral construction of Xiong et al. (2026) with endpoints pinned and again with a free offset and span, and a neural ODE warp integrated by fixed-step fourth-order Runge–Kutta (Chen et al., 2018). Two tasks run against each: under a power warp $\gamma(t) = t^e$ the ends of the time axis are preserved and every family can express the transformation, while under a rigid delay $\gamma(t) = t + \theta$ only families with free endpoints can. The sweep is simulated because a bound on attainment needs a planted truth, which Section 5.1's estimated arrival is not. It is then repeated on the 1,473 dense acquisitions, with only the delay task and with the arrival estimated on the full acquisition as the target. Both runs fix the template to the fitted response.

Simulated trajectories. Quantities *computed from the derivation* evaluate (4) or (6) at a chosen response and warp direction, establishing that the algebra and its implementation agree. Quantities *measured on simulated trajectories* generate curves, add noise and run an estimator. Quantities *measured on data* use the cohorts above.

Simulated trajectories use a gamma-variate response $f(t) = t^{\alpha} e^{-t/\beta}$ with β placing the peak at a chosen time and $\alpha > 1$, below which the derivative at the origin diverges and (4) is infinite for every design. The noise level is set to a/SNR at a stated SNR; Appendix A shows why the fitted residual cannot serve. Each cell uses 2000 trajectories unless stated, and every result below states which of the three it is.

5 Results

5.1 Real acquisitions

The subsampling covers all 1,473 scans of the 80-frame protocol, where the error grows monotonically as samples are removed. On the 35-frame protocol it does not: six samples spanning the window reach 2.74 s against 3.49 s for eight (Figure 3a). Appendix I gives the per-schedule values.

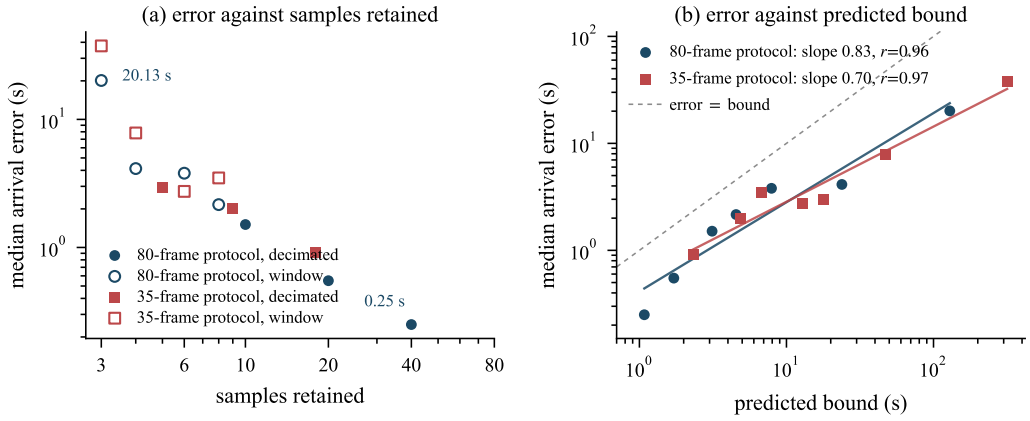

Figure 3: Arrival-time estimation on 3,303 prostate examinations across two protocols, both panels logarithmic. **(a)** Median absolute error against samples retained; filled markers decimate the full acquisition, open markers keep a contiguous window. **(b)** The same errors against the bound computed from the measurement times alone, with a least-squares fit per protocol.

Arrival times vary between patients by 18.3 s, recovered from the 25.9 s a mismatched pairing gives at the dense schedules, where the estimate's own error is negligible. Twenty samples recover an arrival to 0.55 s, a fortieth of that spread, and six to 3.8 s. **Three samples give 20.1 s, larger than the spread itself**, so such an estimate says no more about an individual patient than the population mean does. Two of the four MAMA-MIA collections acquire three samples.

The bound (Figure 3b) spans $119\times$ across these schedules while the measured error spans $81\times$, and a regression of $\log$ error on $\log$ bound has slope 0.834 at $r = 0.96$, so the error follows the bound over two orders of magnitude. The slope falls short of one because each sparse schedule is a subset of the reference's own samples, so the two fits share data and their difference understates the sparse estimator's own error.

The design-time evaluation puts the measured error below the bound in all seven schedules, against five of seven at the arrival each exam turns out to have had (Appendix F).

Breaking the pairing between scan and reference inflates the error 2.9 to 36-fold wherever the estimate is informative, and $1.3\times$ at three samples, where little is left to carry. Appendix F adds the per-exam correlations and, pooling the protocols, separates the bound reading the schedule from its reading cadence alone at partial rank correlation 0.515. Residuals are autocorrelated rather than independent, at lag-1 0.89 and 0.68 on the two protocols, and recomputing the bound under the AR(1) model of (5) leaves the ordering identical and moves the slope by at most 0.025 (Appendix J).

5.2 Where the samples fall

Every schedule above has a different number of samples, so the count orders them as well as the bound does, both at rank correlation 1.000 on the 80-frame protocol. Holding the count at six separates the two.

Over the thirteen placements of Figure 4 the bound orders the measured error at rank correlation 0.654, and 0.621 on the 35-frame protocol, against a permutation threshold of 0.560. Evaluating it at the fitted arrival rather than at $\theta = 0$ gives 0.577 and 0.599, so on this design the two evaluation points perform comparably, unlike the schedules above where the arrival matters more.

The windows that open after the bolus has arrived observe a saturating curve, which Section 3.5 shows is exactly degenerate once the amplitude is profiled; the failure reproduces against a known arrival in simulation (Appendix C).

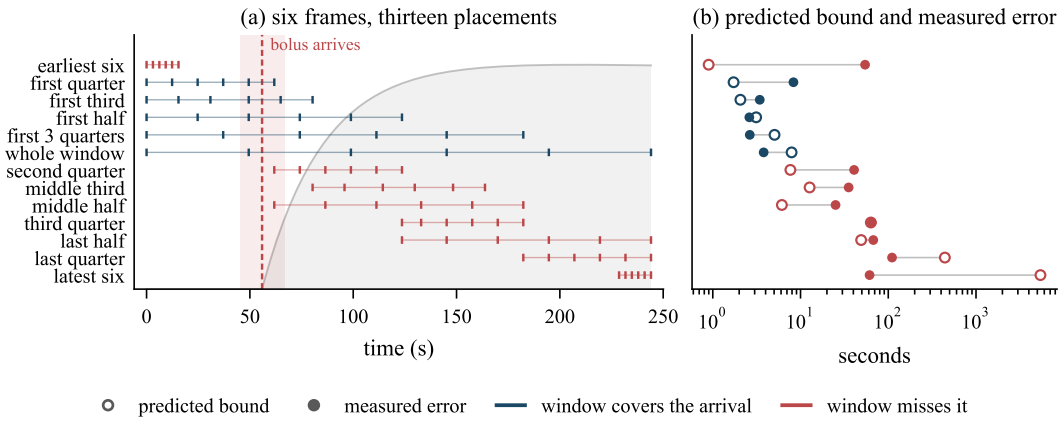

Figure 4: Thirteen placements of six real frames on the 80-frame protocol. Every row keeps six samples, so the count explains none of the variation and only the placement differs. **(a)** Each placement against the fitted response (shaded), coloured by whether its window covers the median arrival, the band being its spread across the cohort. **(b)** The bound predicted from the measurement times alone (open) and the error measured on real exams (filled). The bound separates the windows covering the arrival from those opening after it, and rates the earliest six best of all thirteen while they recover nothing: at $\theta = 0$ those frames are assumed to straddle an onset they precede by 40 s. Values in Appendix I.

Table 2: Warp families against the bound. Recovery is the median fraction of the true warp recovered, where 1.00 is exact and a negative value means the fitted warp moves the wrong way; efficiency is the bound over the calibrated error, where a value above one would violate the bound. Bold marks the best efficiency in each arm. *Delay?* is whether the family can express a rigid delay at all, which Proposition 2 settles from its endpoints and not from any fit. No cell in any arm exceeds the bound.

			simulated, 84 cells			real, template fixed		real, template learned	
warp family	ends	delay?	power	delay	eff. ↑	delay	eff. ↑	delay	eff. ↑
direct delay	free	yes	0.895	1.006	**0.939**	0.974	**0.922**	0.114	0.632
affine	free	yes	1.059	0.988	0.505	1.188	0.357	0.023	0.635
softplus integral	free	yes	0.830	1.020	0.420	1.022	0.329	0.155	0.534
neural ODE (RK4)	free	yes	0.837	1.014	0.427	1.127	0.288	0.280	0.670
monotone spline	free	yes	0.875	0.980	0.934	0.546	0.317	0.417	**0.720**
softplus integral	pinned	no	0.918	−0.116	0.799	0.595	0.733	0.337	0.610

5.3 Learned warp families

Only the endpoint-pinned family fails the delay task, and it fails in the direction Proposition 2 predicts, recovering a negative fraction, while matching the free families on a power warp its tangent space can express. It still returns a warp and still lowers its reconstruction error, fitting the sparsest real trajectories at $R^2 = 1.000$. The separation is milder on real scans than in simulation: the arrival spread is 18.3 s inside a 243 s window, which leaves a pinned warp interior room to imitate a delay of that size.

Learning the template jointly with the per-curve warps collapses recovery in every family and reverses the ordering between them: template and warp trade against each other, so the model explains a delay by moving the template.

An MLP and a GRU recover 0.85 to 1.06 of either transformation at every density; a permutation-invariant encoder is erratic at median 0.045, which we attribute to training instability (Appendix B).

5.4 PHYSIONET 2012

Heart rate arrives hourly for a median of 55 measurements per patient, the chemistry panels every 15 hours for a median of 3. Converted into the timing-relevant quantity, the bound spans 0.81 to 8.62 min across the 25 variables for hour-scale events, a factor of 10.6, while models are trained across all of it at once and scored with one pooled error. For half-day events it spans only 3.9. Measuring how well a model *recovers* timing here would need a known arrival time, which PhysioNet does not record.

5.5 ATTAINMENT AND NUISANCE PARAMETERS

Given a gamma-variate response, the shape known exactly, the amplitude known and a noiseless baseline, a maximum-likelihood estimate never beats the bound across 72 simulated settings spanning 6 to 80 samples and SNR 5 to 50: efficiency exceeds one in none of them and peaks at 0.962. Median efficiency is 0.858 on a near-monotone shape and 0.370 on a peaked one, whose likelihood carries a second solution that a few trajectories fall into. Appendix L gives the sweep, the shape dependence and the per-setting spread.

Adding one missing quantity at a time to that estimator, the two large costs are fitting an amplitude ($2.77\times$ pooled) and taking the baseline from one noisy sample ($1.75\times$). The amplitude cost is strongly shape dependent, reaching $3.71\times$ on a near-monotone response while a peaked one instead gains, its efficiency rising from 0.424 to 0.524. Equation (6) predicts $4.00\times$ and $1.00\times$ from the response shape and the measurement times. The two agree only where the estimator is efficient (Appendix L).

5.6 RESPONSE-SHAPE MISSPECIFICATION

Across seventeen wrong response shapes the estimator acquires a bias of up to 166 s, and a Cramér–Rao bound constrains the variance of unbiased estimates, so it then bounds nothing: ranking designs by total error inverts, at a median rank correlation of -0.79, while the ordering against the spread of the estimates survives at 0.71. Coverage of a nominal 95% interval falls from 0.947 to a median of 0.038 without the interval widening. Appendix E gives the sweep and Appendix D carries it to clinical labels, where such a feature clears a conventional significance threshold.

6 DISCUSSION AND CONCLUSIONS

Whether a set of trajectories admits an amplitude–phase separation is a property of when they were sampled, and (4) settles it from the measurement times alone. A protocol can be audited before it is run, and a negative verdict is not a modelling failure to be answered with capacity: an endpoint-pinned family cannot represent a rigid delay at any width or depth, and a response already rising at the first frame carries no timing information at any SNR. Both constructions are in current use.

No configuration in Table 2 exceeded the bound, and the endpoint-pinned family failed the delay task where Proposition 2 says it must, its reconstruction error giving no sign. The same calculation applies to any irregular series, and on PhysioNet 2012 the single pooled error these models are scored on mixes variables whose phase is recoverable with variables whose phase is not.

Fixing a target resolution instead of a schedule turns (4) around (Appendix C). The amplitude has to be profiled, roughly doubling every sample count, and the schedule anchored to the expected arrival rather than the start of acquisition, which on this cohort recovers nothing.

A design chosen this way yields estimates whose spread is what the bound predicts and whose bias is whatever the assumed response shape leaves behind, so calling one design more accurate needs that shape to be right, and a fitted warp does not report whether it is. The bound is also first-order and single-parameter; the second-order and matrix versions are the natural next steps (Appendix G). The prostate trajectories come from an enhancement-selected gland decile, mostly non-tumour against MAMA-MIA's expert segmentations (precision 0.222) though reproducing the expert shape at correlation 0.981, so the bound describes a protocol and not an individual scan; all three datasets are clinical time series untested beyond that setting. Separating a trajectory's size from its timing nonetheless has an information limit computable from the measurement times alone, and nothing we tried exceeds it.

REPRODUCIBILITY STATEMENT

Every number in this paper is produced by a script in the accompanying code release, `https://anonymous.4open.science/r/dce-response-manifold-56A5/`, and each script carries a self-test that asserts the properties its results depend on. Sections 5.1 and 5.4 are measured on data; Section 5.3 reports a simulated sweep and a repetition of it on real acquisitions; Section 5.5 and Appendix E are measured on simulated trajectories; Appendix K is computed from the derivation. The synthetic results (Sections 5.5, C and 5.3) run from a fixed seed on any machine with NumPy and, for Section 5.3, PyTorch; they need no data and no GPU. Nuisance sources on the ladder draw from independent random streams, so a rung cannot borrow another rung's noise draw.

MAMA-MIA (Garrucho et al., 2025) and PhysioNet Challenge 2012 (Silva et al., 2012) are public, so Section 5.4, the trajectory extraction and the substitution measurement of Section 6 reproduce directly. The prostate cohort is not redistributable, and Section 5.1 does rest on it alone: it is the one cohort in reach with an acquisition dense enough to serve as its own reference. We therefore release that section's per-exam acquisition times and fitted arrival estimates, which are the inputs every number in Figure 3 is computed from, so the analysis can be rerun without the images even though the extraction cannot.

Three points of practice affect whether the numbers reproduce as reported. Summary statistics are taken over all cases, never over the cases whose fit succeeded, since conditioning on success inverts the ordering; errors are summarised by the median absolute value rather than the RMSE, because a minority of trajectories land in a secondary likelihood mode and dominate a squared average. Every reported comparison on real data is accompanied by a null in which the pairing between case and reference is broken, and that null is computed by the same code path as the result. The gamma-variate response is normalised at a fixed abscissa rather than by its maximum over the caller's grid, which otherwise makes the template depend on the sampling being evaluated. Its shape parameter is required to exceed one, below which the derivative at the origin diverges and every design scores as identifiable.

AI USE STATEMENT

Generative AI tools were used in preparing this work: to write and revise analysis code, to draft and revise this manuscript, and in proposing candidate experiments. They were not used to fabricate data or results. Every number attributed to a cohort derives from that cohort, and Section 4 states which results are computed from the derivation, which are measured on simulated trajectories and which are measured on data. Each analysis script carries a self-test asserting the properties its results depend on, and every table was transcribed from the stored result files and checked back against them by hand. The authors take responsibility for the final content of this work, including all text, claims and artefacts.

REFERENCES

I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark. Predicting in-hospital mortality of ICU patients: the PhysioNet/Computing in Cardiology Challenge 2012. In *Computing in Cardiology*, volume 39, pages 245–248, 2012.

A. Srivastava, W. Wu, S. Kurtek, E. Klassen, and J. S. Marron. Registration of functional data using Fisher–Rao metric. *arXiv:1103.3817*, 2011.

A. Srivastava and E. P. Klassen. *Functional and Shape Data Analysis*. Springer, 2016.

X. Xiong, S. Jiang, and P. Zeng. NeuralFLoC: Neural flow-based joint registration and clustering of functional data. *arXiv:2602.03169*, 2026.

L. Garrucho et al. A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations. *Scientific Data*, 12:453, 2025.

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.

Y. Rubanova, R. T. Q. Chen, and D. Duvenaud. Latent ODEs for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, pages 5321–5331, 2019.

P. Kidger, J. Morrill, J. Foster, and T. Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

M. Horn, M. Moor, C. Bock, B. Rieck, and K. Borgwardt. Set functions for time series. In *International Conference on Machine Learning*, pages 4353–4363, 2020.

S. N. Shukla and B. M. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021.

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

E. Henderson, B. K. Rutt, and T.-Y. Lee. Temporal sampling requirements for the tracer kinetics modeling of breast disease. *Magnetic Resonance Imaging*, 16(9):1057–1073, 1998.

D. C. Alexander. A general framework for experiment design in diffusion MRI and its application in measuring direct tissue-microstructure features. *Magnetic Resonance in Medicine*, 60(2):439–448, 2008.

M. Bouhrara and R. G. Spencer. Fisher information and Cramér–Rao lower bound for experimental design in parallel imaging. *Magnetic Resonance in Medicine*, 79(6):3249–3255, 2018.

A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

J. D. Tucker, W. Wu, and A. Srivastava. Generative models for functional data using phase and amplitude separation. *Computational Statistics & Data Analysis*, 61:50–66, 2013.

N. M. Hylton et al. Neoadjuvant chemotherapy for breast cancer: functional tumor volume by MR imaging predicts recurrence-free survival. *Radiology*, 279(1):44–55, 2016.

A. Kneip and T. Gasser. Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, 20(3):1266–1305, 1992.

K. Wang and T. Gasser. Alignment of curves by dynamic time warping. *The Annals of Statistics*, 25(3):1251–1276, 1997.

J. S. Marron, J. O. Ramsay, L. M. Sangalli, and A. Srivastava. Functional data analysis of amplitude and phase variation. *Statistical Science*, 30(4):468–484, 2015.

R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural ordinary differential equations. In *NeurIPS*, 2018.

A. Hyvärinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *NeurIPS*, 2016.

I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: a unifying framework. In *AISTATS*, 2020.

F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2019.

P. S. Tofts et al. Estimating kinetic parameters from dynamic contrast-enhanced T1-weighted MRI of a diffusable tracer: standardized quantities and symbols. *Journal of Magnetic Resonance Imaging*, 10(3):223–232, 1999.

S. P. Sourbron and D. L. Buckley. Classic models for dynamic contrast-enhanced MRI. *NMR in Biomedicine*, 26(8):1004–1027, 2013.

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 2nd edition, 1998.

Table 3: Fewest uniformly spaced samples meeting a target resolution, with the total acquisition time in parentheses. Response peaking at 90 s, SNR 20, and the *amplitude-profiled* bound of (6) required below half the target, since an analysis fits the amplitude rather than knowing it. Selecting on the amplitude-known bound instead halves every count in this table. No entry has two samples: two observations determine an amplitude and a shift exactly, so profiling the amplitude removes the timing entirely and (6) diverges. Dashes mark targets no sample count reaches at that interval. The schedule must also begin at the expected arrival; Section 5.1 measures what happens when it does not.

	sampling interval (s)					
target	3	5	10	30	60	120
5 s	16 (48)	12 (60)	8 (80)	5 (150)	—	—
10 s	11 (33)	8 (40)	5 (50)	3 (90)	3 (180)	—
20 s	8 (24)	6 (30)	4 (40)	3 (90)	3 (180)	—

A Choice of noise level

The fitted residual $\mathrm{SSE}/(n-2)$ cannot serve as the noise estimate here. Two parameters are fitted, so a three-sample curve leaves one degree of freedom and the model very nearly interpolates it. Estimated per collection on MAMA-MIA the residual runs monotone in the degrees of freedom, 0.042 for ISPY2 at six samples, 0.033 for DUKE at four, 0.012 and 0.008 for ISPY1 and NACT at three. The collections predicted least identifiable would receive the tightest acceptance thresholds, and the ordering would invert on degrees of freedom rather than on phase. Pooling one residual across all collections conceals this rather than removing it, since the pooled value is dominated by whichever collection contributes most cases. We therefore set $\sigma_i = a_i/\mathrm{SNR}$ at a stated SNR, identically for every collection and matching the assumption the predicted bound is computed under, and report the sweep above rather than a single value.

The fraction of unbounded intervals is the reported statistic. Median width over the subset with bounded intervals is conditioned on the fit having succeeded and inverts the ordering: NACT, which fails in 77% of cases, shows a narrower median than ISPY2, which fails in 29%, because the cases that succeed there are the easiest ones.

B The permutation-invariant encoder

Holding the warp family fixed, an MLP and a GRU over (t_i, y_i) pairs both recover between 0.85 and 1.06 of either transformation at every sampling density. A permutation-invariant encoder does not: it recovers 0.98 of a power warp at twenty samples and essentially nothing at six or eighty, with the same spread on the shift. Raising its width from 32 to 128 and its budget from $1{,}500$ to $6{,}000$ epochs does not stabilise it.

At twenty samples alone the permutation-invariant encoder reads as recovering shape but not delay, which would match the warp-family result; across the full sweep it fails on the power warp as well, where no structural obstruction exists. Its behaviour therefore supports no claim about representation. The warp-family separation holds at every sampling density and both SNRs.

C Study design implied by the bound

Fixing the arrival-time difference to be detected, (4) returns the schedules that deliver it. Table 3 does this for a gamma-variate response peaking at 90 s at SNR 20, requiring the amplitude-profiled bound of (6) to sit at half the target effect. Profiling is not a small correction here: at these short spans the response is still rising, so (6) costs a factor of 3.6 to 6.4 and the required sample counts roughly double.

Coverage dominates density (Figure 5). Four samples at a 5 s interval span 15 s and give a bound of 22.9 s; four at 40 s span 120 s and give 2.44 s. The finer schedule is nine times worse because all of its samples land where the centred derivative is still small, and there the amplitude confound is worst

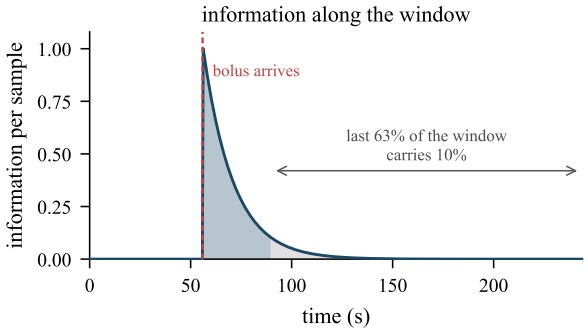

Figure 5: Information per sample along the window, on the response fitted to this cohort. The last 63% of the window carries 10% of it, which is why coverage of the informative part dominates sampling density above.

as well. Density helps only once the window is fixed: thirteen samples across the same 120 s reach 1.34 s.

Placement matters more than either. Six samples confined to $[0, 100]$ s give a bound of 2.07 s. The same six spread evenly over 0–600 s give 31.5 s, and the same six placed at $[270, 370]$ s give 1653 s, a spread of eight hundredfold at fixed count. A protocol that extends into the washout tail gains almost nothing, because that is where f' is flat and the centred derivative of Section 3 is near zero. This is the mechanism behind the MAMA-MIA NACT collection's schedule: its median samples, at 0, 376 and 752 s, fall almost entirely in the uninformative region, and no increase in SNR rescues them.

Schedules anchored at the start of imaging. Built on this cohort's own acquisitions but anchored at the start of imaging rather than at the expected arrival, the schedules above recover nothing: none of the nine testable rows meets its target, and the median error is 34 to 55 s against targets of 5, 10 and 20 s. The reason is placement rather than count. This cohort's bolus arrives 55.9 s into the acquisition while eight of the nine schedules close before 31 s, so they observe pre-contrast baseline and 98 to 100% of exams carry no timing information at all. The one row reaching past the arrival, three samples over 61 s, is also the only one with a materially better error, 34 s against about 54 s elsewhere. Anchoring the same counts and intervals to the expected arrival returns seven of the nine to within their targets; the two that remain outside are the tight-target, coarse-interval rows where (6) alone costs 2.9 and $7.1\times$.

The saturating degeneracy across the two cohorts. Across the thirteen placements of Figure 4, (6) puts the amplitude-profiling cost at 134 to $380\times$ on the windows that open after the arrival, against 1.1 to $1.8\times$ on those opening before it. The failure reproduces on simulated curves against a known arrival: six samples over the second quarter of the window carry a predicted bound of 1.18 s and a measured median error of 12.8 s.

Over a 243 s window the wash-out constant is not estimable, so the response is effectively saturating, which Section 3.5 shows is exactly degenerate. The degeneracy needs the arrival to precede the window: with the arrival corner inside it two shifts are not amplitude rescalings of one another (residual 27.7%), and with arrival long before it they are, to 0.000%. A 3-second acquisition samples that corner densely and three breast samples at 376 s intervals do not.

D A DOWNSTREAM TEST ON TREATMENT-RESPONSE LABELS

MAMA-MIA ships pathological complete response labels, so a misspecified alignment can be scored against a clinical endpoint. We fitted a shift to every labelled trajectory and scored it for pCR within each collection, never across them, with peak enhancement as a positive control.

Amplitude separates pCR in none of the four collections; all four intervals contain 0.5. Single-visit region-averaged curves are not expected to predict response, since the established imaging predictors are functional tumour volume and change across treatment (Hylton et al., 2016). The experiment

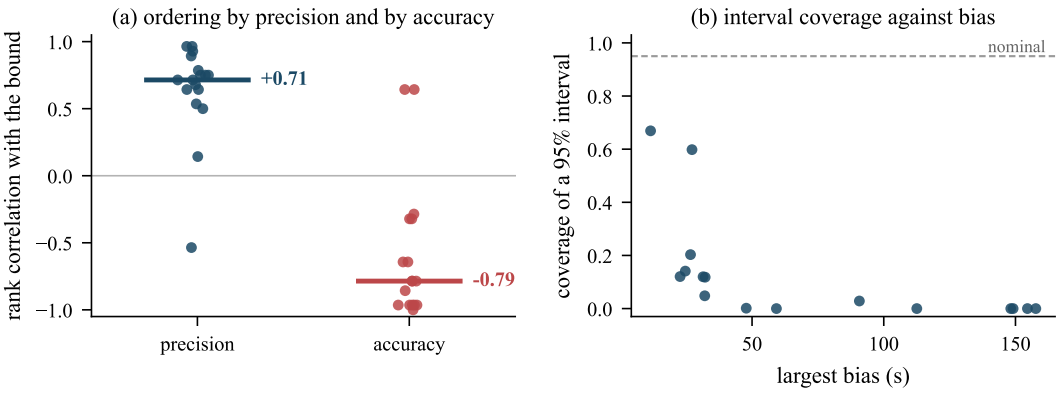

Figure 6: Design ordering and interval coverage under seventeen misspecified response shapes. Each point is one scenario. **(a)** Rank correlation between the bound and the estimator's spread (left) and between the bound and its total error (right); bars are medians. **(b)** Coverage of a nominal 95% profile interval against the largest bias in that scenario: the intervals do not widen as the bias grows, so nothing in the fit reports the failure.

therefore cannot test whether identifiability governs downstream usefulness: a null phase result would be equally consistent with no signal of any kind being present.

In all four collections the fitted shift tracks the time to peak read directly off the curve with no model fitted, at rank correlation -0.67 to -0.87. The assumed template peaks at 90 s and the median fitted shift in ISPY2 is 78 s, which is squarely the regime of Appendix E: the estimator absorbs shape into the warp because the shape it was given is wrong.

In ISPY2 the shift appears to predict pCR, at AUC 0.588 (95% CI 0.550–0.627), and it does not survive decomposition. An indicator of whether the fit reached the edge of its search range predicts pCR by itself at 0.565 (0.538–0.594); with the non-converged fits removed the shift scores 0.537 (0.491–0.580), an interval containing chance. The effect is carried by estimation failure and curve shape rather than by phase.

A quantity reported as a timing measurement clears a conventional significance threshold on real clinical labels while tracking curve shape and estimation failure. The diagnostics that would catch it are the fraction of fits that fail and the correlation with a model-free shape statistic, and neither is standard practice.

E Response-shape misspecification

The analyst always assumes a gamma-variate response peaking at 90 s with $\alpha = 3$. The truth ranges over peak times from 45 to 180 s, shape parameters from 2 to 4.5, and a persistent enhancement curve with no washout, a different family rather than a different parameter and one of the three standard kinetic types in contrast imaging. Seventeen misspecified scenarios, seven acquisition designs each, 400 replicates per cell. The cost of assuming the wrong family is defined against a truth only a simulation has.

Bias and the RMSE ordering. The estimator acquires a bias of up to 166 s. Ordering designs by RMSE against the predicted bound then gives a median rank correlation of -0.79 across the seventeen scenarios. A dense acquisition fits the wrong template confidently and inherits the full bias, while a sparse one carries too little information to be pulled anywhere, so the design the bound calls best is the one misspecification damages most.

The variance ordering. Against the estimator's spread, the quantity a variance bound constrains, the median rank correlation is 0.71, and it inverts in one scenario of seventeen (-0.54, true peak 130 s at $\alpha = 2$). A wrong template moves the likelihood surface without flattening it, and flatness is

what (4) predicts. The bound therefore orders designs by precision and, under misspecification, not by accuracy.

Interval coverage. Coverage of the nominal 95% profile interval falls from 0.947 under correct specification to a median of 0.038 across the misspecified scenarios, with a maximum of 0.669. On the densest design the interval stays finite in every replicate and contains the truth in none of them. The saturating, persistent-enhancement case is the most damaging, which Section 3.5 predicts: it is one of the two exact degeneracies, so profiling the amplitude removes the phase entirely and no amount of data repairs it.

F CASE-LEVEL INFORMATION IN THE BOUND

Section 5.1 establishes that the bound predicts the magnitude of the degradation across designs. A stronger claim would be that it predicts which individual *exam* is estimated worse, at a fixed design. It does, weakly.

Per-exam rank correlations. Within a single schedule every exam has the same sample count and differs only in when those samples fell. Correlating the bound against the observed error across the $1{,}473$ exams gives rank correlations of 0.240, 0.215, 0.138, 0.258, 0.016, 0.310 and 0.445 for the seven sparse schedules. A permutation null that breaks the exam-to-bound pairing while leaving both marginals intact has a 95th percentile near 0.04 throughout, so six of the seven are outside it. The exception is six samples spanning the window, at $\rho = 0.016$, for which we have no mechanism.

Scale of the correlations. The achievable rank correlation in this test is itself small. Within a schedule the bound varies only 1.18 to 2.44-fold between exams (ninetieth over tenth percentile), and an estimator whose error sat exactly at the bound would still scatter about it. Drawing errors as $|\mathcal{N}(0, \mathrm{CRB}_i)|$ and correlating gives 0.079, 0.098, 0.139, 0.171, 0.232, 0.361 and 0.352, the ceiling for this test. The observed values reach it in four of the seven schedules and sit at it in a fifth (0.138 against 0.139); the six- and four-sample spanning designs fall short, at 0.07 and 0.86 of their ceilings. Departures in either direction reflect the error distribution not being half-normal rather than any property of the bound, and ρ is not interpretable without the ceiling beside it. The median ratio across the seven is 1.26.

Collinearity of window and cadence. In an 80-frame protocol the final sample lies at 79 cadences, so the window and the cadence are collinear, at $\mathrm{corr} = 1.000$ across the $1{,}571$ exams, with $t_{\text{last}}/(79\,\Delta t)$ having standard deviation 8×10^{-5}. Within one protocol the acquisition is a one-parameter family, and no partial correlation can separate "the bound reads the sampling design" from "the bound reads one scalar that happens to be cadence".

Pooling the two protocols supplies the second parameter, because the same cadence buys a 79-frame window in one and a 34-frame window in the other. In the cadence band where both exist, 3.06 to $6.08\,\mathrm{s}$, $\mathrm{corr}(\Delta t, \text{window})$ falls to -0.004 over $6{,}405$ exam-schedule pairs, and both arms there are the same vendor and pass the same uniform-clock audit, so the protocol is not standing in for the scanner. Restricting to the four schedules that keep the same sample count in both protocols, the bound predicts the per-exam error at partial rank correlation 0.452 to 0.548, median 0.515, controlling for cadence. The bound reads more than one scalar.

The design is unbalanced: the 35-frame arm contributes 73 of the 915 exams per schedule in the overlap band, so a null here would have been weak evidence and only the positive direction is informative. And the three decimation schedules are excluded rather than reported, because they keep a fixed *fraction* of frames and so pool 40 samples against 18; there the sample count is a proxy for the protocol and the protocol for the window, and their partial correlations, -0.24 to 0.06, measure that confound rather than the bound.

The error-to-bound ratio. Dividing the observed error by the predicted bound gives 0.13 to 0.48. That ratio is not a bound check. Each sparse schedule is a *subset* of the reference's samples, so the two estimates share data and their difference has spread $\sqrt{\mathrm{var}_{\text{sparse}} - \mathrm{var}_{\text{ref}}}$ rather than the sum, confirmed in simulation, where the nested identity predicts 2.169 and the independent one 2.394

against a measured 2.185. The measured error therefore understates the sparse estimator's own error by construction. On top of that, declaring $\sigma = a/\mathrm{SNR}$ fixes an arbitrary scale, worth about ten per cent even on clean synthetic data. Ordering is the only thing this design can test.

Where the bound is evaluated. Equation (4) is evaluated at $\theta = 0$ throughout, which places the response onset at the first sample and is the quantity a protocol fixes before any exam exists. Evaluating it instead at the arrival each exam turns out to have had gives a different reading, since the bolus arrives $55.9\,\mathrm{s}$ into the window rather than at its start: the span narrows to $18.5\times$ and the slope rises to 1.470 at $r = 0.977$, or $14.4\times$ and 1.514 from a single population arrival, one constant per protocol and so still available in advance. The design-time evaluation is the conservative one, putting the measured error below the bound in all seven schedules against five of seven at the arrival.

G Proofs

Throughout, $Cu = u - u(t_1)$ is the baseline subtraction of (2), $v = \partial_\theta \gamma_\theta$ is the tangent direction of the warp family, and $I(\theta)$ is the Fisher information (4). Appendix K checks each statement numerically against the code that implements it.

Proposition 1. For $\gamma_\theta(t) = t + \theta$ the tangent is $v \equiv 1$, so the centred term of (4) is $f'(t_i + \theta) - f'(t_1 + \theta)$. If f' is constant on $[t_1 + \theta,\, t_n + \theta]$, every $t_i + \theta$ lies in that interval and each difference is zero term by term. Hence $I(\theta) = 0$ regardless of n, of σ and of where the samples fall, and the Cramér–Rao bound $I(\theta)^{-1/2}$ is infinite. A vanishing Fisher information means the log-likelihood is flat in θ to first order, so no unbiased estimator has finite variance. The hypothesis is what fails once the onset lies inside the window: an onset at $t^* \in (t_1 + \theta,\, t_n + \theta)$ puts a discontinuity in f' there, the differences no longer vanish, and the information is positive. □

Proposition 2. Differentiating $\gamma_\theta(t_1) = t_1$ with respect to θ gives $v(t_1) = \partial_\theta \gamma_\theta(t_1) = 0$, and the same at t_n, so every admissible perturbation vanishes at both endpoints. The rigid delay has $\gamma_\theta(t_1) = t_1 + \theta$, which differs from t_1 whenever $\theta \neq 0$, so it is not a member of the family; its tangent $v \equiv 1$ violates $v(t_1) = 0$. Neither statement involves the number of samples, the noise level or the capacity of whatever produces the warp. □

Proposition 3. Write $u = C[f'v]$ and $w = C[f]$, so that under (2) the mean is $\mu(\theta, a) = a\, C[f \circ \gamma_\theta]$ with $\partial_\theta \mu = a\, u$ and $\partial_a \mu = w$. For a Gaussian model with known σ the Fisher matrix in (θ, a) is

$$I \;=\; \frac{1}{\sigma^2} \begin{pmatrix} a^2\|u\|^2 & a\langle u, w\rangle \\ a\langle u, w\rangle & \|w\|^2 \end{pmatrix}, \qquad \det I \;=\; \frac{a^2}{\sigma^4}\Big(\|u\|^2\|w\|^2 - \langle u, w\rangle^2\Big).$$

Inverting and reading the (θ, θ) entry,

$$\big(I^{-1}\big)_{\theta\theta} \;=\; \frac{\|w\|^2/\sigma^2}{\det I} \;=\; \frac{\sigma^2}{a^2\|u\|^2} \cdot \frac{1}{1 - \rho^2}, \qquad \rho = \frac{\langle u, w\rangle}{\|u\|\,\|w\|},$$

where $\sigma^2/(a^2\|u\|^2)$ is exactly the amplitude-known variance bound of (4). Profiling the amplitude therefore multiplies the variance by $(1 - \rho^2)^{-1}$ and the standard deviation by $(1 - \rho^2)^{-1/2}$, which is (6). By Cauchy–Schwarz $|\rho| \leq 1$ with equality exactly when u and w are parallel, and there $\det I = 0$: the two parameters have proportional effects on the observations and neither is separately identifiable. Both named cases are of that kind. A constant-derivative response gives $u = C[f'v] = 0$, so the information in θ is already zero by Proposition 1. For $f(t) = 1 - e^{-kt}$ with $v \equiv 1$,

$$C[f'](t) \;=\; ke^{-kt} - k \;=\; -k\left(1 - e^{-kt}\right) \;=\; -k\, C[f](t)$$

holds identically in t and for any t_1, so $u = -k\, w$ exactly, $|\rho| = 1$, and profiling the amplitude removes the timing entirely. Neither more samples nor higher SNR repairs this, since the degeneracy is in the direction of the score rather than in its magnitude. □

The kernel of the score map. Equation (4) is $\|Av\|^2$ for the linear map $A : v \mapsto C[f'v]$. Then $Av = 0$ means $f'(t)v(t)$ is constant in t, so whenever f' does not vanish the kernel is one-dimensional and spanned by $v_{\text{null}} \propto 1/f'$. Section 3.5 states why this does not license projecting that direction out of a learned warp, and The paragraph below measures what happens if one does.

Table 4: Values behind Figure 4. The bound is (4) at $\theta = 0$. Railed fits are those whose profile search reached the edge of its range; they are counted separately and the median is taken over the rest, since a railed fit reports the search range rather than the data.

	covers the	seconds		
six frames placed over	arrival	predicted bound	median abs. error	railed
the earliest six	no	0.90	54.40	1%
the first quarter	yes	1.73	8.30	0%
the first third	yes	2.06	3.43	0%
the first half	yes	3.12	2.61	0%
the first 3 quarters	yes	5.06	2.64	0%
the whole window	yes	7.91	3.80	0%
the second quarter	no	7.67	40.78	0%
the middle third	no	12.69	35.36	1%
the middle half	no	6.13	25.07	0%
the third quarter	no	63.30	64.77	5%
the last half	no	49.08	67.53	4%
the last quarter	no	441.74	110.41	16%
the latest six	no	5427.21	61.22	20%

Second-order detectability of the null direction. Section 3.5 identifies $v_{\text{null}} \propto 1/f'$ as carrying exactly zero Fisher information, and notes that this does not license projecting it away.

Fisher information is a property of the likelihood's derivative at a point, so a direction in the kernel carries no first-order signal; a finite deformation along it is still visible at second order. Measured on the response used throughout, the data effect of the least visible direction grows as $\epsilon^{1.95}$ against $\epsilon^{1.00}$ for the most visible one, which matches the order the derivation predicts.

The curvature term is not small. The least visible direction is concentrated where f' is near zero, and that is exactly where f'' is large: flatness of the response and curvature of the response coincide. So the nominally invisible deformation becomes detectable above the noise at an amplitude of 0.030, while the most visible direction needs 0.069, despite a singular value ratio of $15{,}000$ between them. At realistic amplitudes the null direction is the easier of the two to detect.

We therefore do not propose projecting near-null components out of a learned warp, and the bounds this paper reports are unaffected: they concern a warp *parameter* in a parametric family, where the local statement is the one that applies and the Cramér–Rao inequality holds as usual.

H Cohort filtering

Of the $1{,}578$ exams on the 80-frame protocol, an automated audit confirms a physical and uniform acquisition clock for $1{,}571$; one exam per patient leaves $1{,}475$, of which $1{,}473$ have an extracted trajectory. The same filters leave $1{,}832$ of the 35-frame exams, on a disjoint set of patients. Two of those reach the edge of the reference search range, leaving no arrival to measure error against, so the 35-frame arm is analysed on $1{,}830$. Summary statistics throughout are taken over all retained cases and never over the cases whose fit succeeded, since conditioning on success inverts the ordering.

I Tabulated values behind the figures

Weighting the bound by the amplitude-profiling cost of (6) repairs the ordering on the 35-frame protocol, from 0.599 to 0.775, and does not repair it on the 80-frame one, from 0.544 to 0.495.

Per-schedule subsampling numbers. Table 5 holds the values behind Figure 3 and the wrong-reference column the figure does not carry.

Table 5: Recovering an arrival time from real prostate scans as samples are removed from the same curve, against the bound predicted from that schedule's measurement times alone. Within each protocol the first three rows keep every nth frame, so they thin the sampling and also shorten the window, while the last four spread a fixed number of frames across the full window, which is how sparse clinical protocols place them. Count and placement therefore vary separately. Each protocol's own full-frame fit is the reference, so its error is zero by construction and is not shown. The last column pairs each sparse estimate with a *different* patient's reference, as a ratio of median absolute errors: a comparison carrying information about this patient must get worse when that pairing is broken.

		seconds		
schedule	samples	predicted bound	median abs. error	wrong reference
80-*frame protocol,* 1,473 *exams*				
every 2nd	40	1.08	0.25	36.4×
every 4th	20	1.71	0.55	16.4×
every 8th	10	3.12	1.51	6.1×
8 spanning the window	8	4.54	2.16	5.0×
6 spanning the window	6	7.91	3.80	2.9×
4 spanning the window	4	23.82	4.13	3.0×
3 spanning the window	3	129.03	20.13	1.3×
35-*frame protocol,* 1,830 *exams*				
every 2nd	18	2.33	0.91	11.5×
every 4th	9	4.89	2.00	5.9×
every 8th	5	17.78	2.96	4.1×
8 spanning the window	8	6.80	3.49	3.7×
6 spanning the window	6	12.91	2.74	4.4×
4 spanning the window	4	47.51	7.83	2.1×
3 spanning the window	3	316.62	37.49	1.2×

The figure plots each protocol against its own bound, so it cannot compare them at matched sample count. Across the four window schedules the 35-frame protocol's bound is worse by 1.50, 1.63, 1.99 and 2.45 at eight, six, four and three samples, so the penalty grows as samples are removed. Its median cadence of 9.1 s over 35 frames spans roughly 310 s against 245 s for the 80-frame protocol, and the extra span falls in the washout, where the centred derivative is near zero. Spreading a fixed number of samples across the longer window therefore places fewer of them where the timing information is. This is the placement effect of Appendix C, computed there from a response shape, appearing between two real protocols.

J THE BOUND UNDER AR(1) NOISE

Equation (4) takes the noise on the normalised trajectory to be independent. Residuals from the fitted response do not support that: their lag-1 autocorrelation is 0.89 on the 80-frame protocol and 0.68 on the 35-frame one, decaying to 0.31 and 0.08 by lag six.

For a scalar parameter under Gaussian noise of known covariance, $I(\theta) = (\partial_\theta \mu)^\top \Sigma^{-1} (\partial_\theta \mu)$, and (4) is the case $\Sigma = \sigma^2 I$. Taking instead $\Sigma_{ij} = \sigma^2 \rho^{|k_i - k_j|}$ on the original frame indices k, so that a schedule's gaps enter the covariance, gives the bound an estimator faces when the noise is AR(1). Setting $\rho = 0$ recovers (4) exactly.

The correction is differential: decimating an AR(1) series decorrelates the samples it retains as ρ^k, so a sparse schedule is nearly white while the reference is strongly correlated. It therefore acts along the same axis the paper orders schedules on. Measured at the observed autocorrelation, the inflation is $1.34\times$ on the densest schedule and $1.00\times$ by four samples. The ordering is unchanged at rank correlation 1.000 against the white-noise bound in every arm, and the log-log slope moves from 0.834 to 0.856 on the 80-frame protocol and from 0.703 to 0.712 on the 35-frame one. The sub-unity slope is therefore not an artefact of the noise model.

The AR(1) bound is not monotone in ρ: it rises to a maximum near $\rho \approx 0.7$ and falls after it. Baseline centring forces the first entry of the score vector to zero, so the interior terms shrink as $(1-\rho)^2$ while

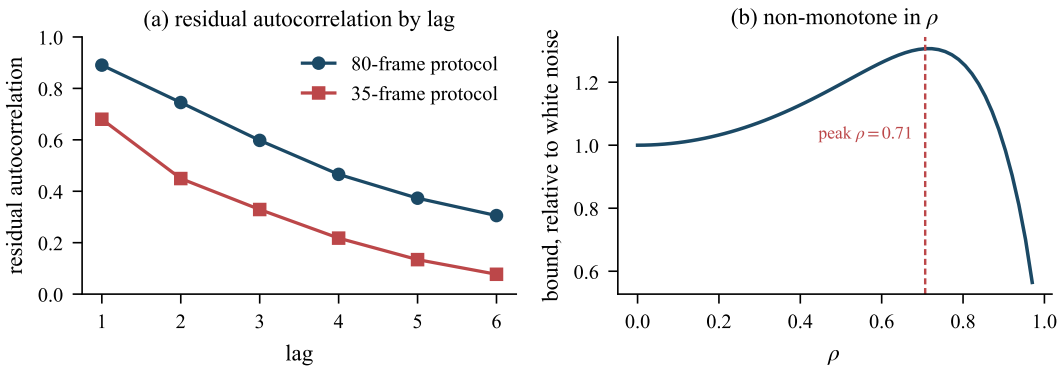

Figure 7: Residual autocorrelation and the bound under AR(1) noise. **(a)** Residual autocorrelation against lag on both protocols; independent noise would sit at zero from lag one. **(b)** The bound on the densest schedule as a function of ρ, relative to its white-noise value, recomputed with the function that produced the published bounds. It is not monotone: baseline centring forces the first score entry to zero, so the interior terms shrink as $(1-\rho)^2$ while the endpoint term carries $(1-\rho^2)^{-1}$, and strongly correlated noise is largely removable by differencing.

the endpoint term carries $(1-\rho^2)^{-1}$ and diverges. Strongly correlated noise is largely removable by differencing, and what remains of the signal difference is still informative.

K Numerical verification of the derivation

Every quantity in this section is computed by evaluating (4) or (6) at a chosen response and warp direction. No trajectories are generated, no noise is added and no estimator is run, so nothing here depends on a generating process we selected: these are checks that the algebra and the code implementing it agree. The claims on data are in Sections 5.1 and 5.4.

The one exception is the constant-derivative sweep at the end, which does generate trajectories and run an estimator, because whether an estimator beats a do-nothing baseline is not an algebraic question.

Null direction. Across five response shapes, $I(v_{\text{null}})$ evaluates to below 10^{-32} times the information carried by a random direction of equal norm. On the four with a non-constant derivative, 96 to 99.9% of its energy lies in the flatter half of the observation window, so flat regions of the response are uninformative pointwise and not merely in aggregate. The fifth is the constant-derivative ramp, for which the statistic is undefined rather than small: its derivative is the same everywhere, so which half counts as flatter is decided by numerical tie-breaking and the share it returns carries no meaning.

Pinned families. Projecting the constant direction onto the pinned tangent space of Section 3.4 recovers 0.79 of it with two modes and 0.95 with thirty-two, so imitating a delay requires increasingly high-frequency bending. That bending concentrates at the ends of the window: 96% of the residual falls in the 20% of samples nearest the boundary, and at t_1 the residual is a full unit. The two design choices are independent rather than compounding, which Proposition 2 already implies: a pinned family has $v(t_1) = 0$, so the centring term $f'(\gamma(t_1))v(t_1)$ of (4) vanishes identically and the baseline normalisation is an exact no-op there. Evaluated on the pinned tangent space the centred and uncentred information agree to machine precision at every mode count. A pipeline adopting both loses representation alone.

Amplitude–timing confound. On the templates and sampling grid of Table 6, equation (6) predicts efficiency costs of $4.00\times$ on the near-monotone response and $1.00\times$ on the peaked one, against measured costs of $3.71\times$ and $0.81\times$ in Section 5.5. The near-monotone prediction is accurate and the peaked one is not. (6) is a ratio of two Cramér–Rao bounds, while the ladder reports a ratio of achieved errors, and the two coincide only when the estimator is efficient at both rungs. It is not: the peaked oracle rung runs at 0.424 efficiency, so fitting the amplitude there recovers some of that

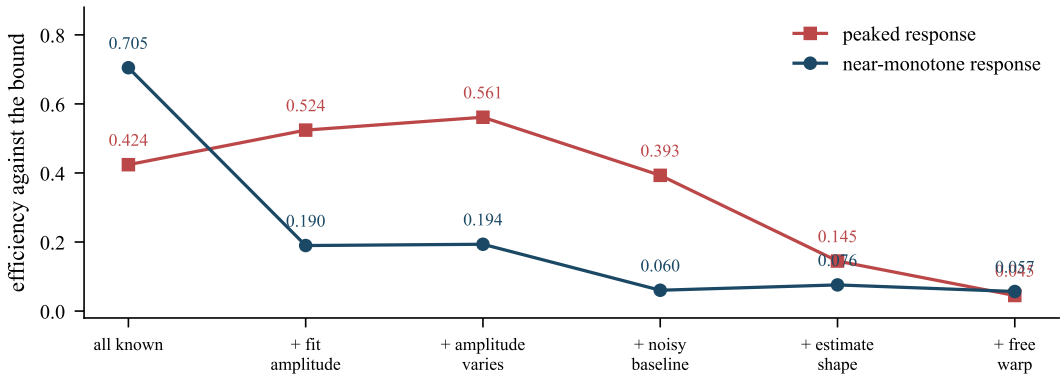

Figure 8: Efficiency as each oracle advantage is withdrawn, at 35 samples and SNR 20. The two response shapes cross at the amplitude rung: fitting an amplitude costs the near-monotone response a factor of 3.71 and *gains* on the peaked one, whose oracle rung is itself only 0.424 efficient. By the last rung, a pipeline with nothing given to it, both run near 0.05.

shortfall and the measured ratio comes out below one. The prediction also depends on where the first sample falls, not on the shape alone: holding the template fixed at 35 samples and moving t_1 from 0 to 0.02 of the window moves the peaked prediction from $1.00\times$ to $1.30\times$, because the grid begins exactly at the onset where f and f' both vanish.

Second-order detectability. The deformation along v_{null} rises above the noise at an amplitude of 0.030, against 0.069 for the direction carrying the most first-order signal, despite a ratio of 15,000 between their singular values. Appendix G gives the measurement.

Constant-derivative responses. Across 36 cells generated from a constant-derivative response, an estimator beats a do-nothing baseline in 5 cells when the onset precedes the window and in 21 when it falls inside. At SNR 5 the first count is zero in every setting.

L ESTIMATOR EFFICIENCY AGAINST THE BOUND

We use a gamma-variate response $f(t) = t^{\alpha} e^{-t/\beta}$, the conventional shape for contrast kinetics, with β set so the peak falls at a chosen time and $\alpha > 1$. Below $\alpha = 1$ the slope at the origin is infinite and (4) is infinite for every design.

Give an estimator every advantage: the response shape known exactly, the amplitude known, and a baseline measured without noise. Across 72 settings spanning 6 to 80 samples and SNR from 5 to 50, it never beats the bound. Efficiency, the bound divided by the achieved error, exceeds 1 in 0 settings and peaks at 0.962.

Median efficiency is 0.858 on a near-monotone response and 0.370 on a peaked one, so the bound is not reached and the shortfall depends on the response shape. The peaked case has a second, wrong solution that a few trajectories fall into, with errors many times the bound, and those few dominate a root-mean-square. Table 6's peaked column is measured against a baseline that is itself only half-efficient.

These figures need many trajectories per setting. At 150 the efficiency of a single setting has a relative spread near 70%, enough for 12 of the 72 to report a value above 1 that an unbiased estimator cannot achieve. At 2000 none do.

The nuisance ladder. A real analysis has none of the advantages Section 5.5 gave the estimator. One missing piece is added at a time. The true timings are held fixed across rows so the comparison is not contaminated by the draw, at 2000 trajectories per setting.

The two columns move in opposite directions at the amplitude row. Fitting an amplitude costs $3.71\times$ on a near-monotone response, while on a peaked one the efficiency rises from 0.424 to 0.524. On a

Table 6: Efficiency against the bound as realistic departures accumulate, at 35 samples and SNR 20. The peaked column's first row is itself only half-efficient (Section 5.5), so its ratios are a floor.

	efficiency against the bound	
what the estimator is given	peaked	near-monotone
response shape, amplitude and baseline known	0.424	0.705
+ amplitude fitted per trajectory	0.524	0.190
+ amplitude varies across trajectories	0.561	0.194
+ baseline from one noisy sample	0.393	0.060
+ response shape estimated from the data	0.145	0.076
+ free warp in place of a rigid delay	**0.045**	**0.057**

response that is still rising, scaling it up and shifting it earlier produce nearly the same measurements, so the two parameters trade against each other. This is a property of the response shape rather than of the estimator, and (6) predicts $4.00\times$ and $1.00\times$ on this grid. The peaked column moves the other way because its oracle rung is itself only 0.424 efficient, so the comparison there is between a bound ratio and an error ratio that the estimator's own inefficiency separates.

Pooled over all settings the rows cost: amplitude fitted $2.77\times$, amplitude varying $0.98\times$, noisy baseline $1.75\times$, response shape estimated $1.04\times$, free warp $1.08\times$. Both large costs consume information the trajectory does not have: an amplitude that trades against timing, and a baseline measured once under noise. The free warp's extra parameters buy local smoothness rather than response structure, and the effect is visible on real trajectories. With scattered interior samples held out, the free warp predicts them better than a one-parameter shift does, at 0.692 against 0.499 improvement over an amplitude-only baseline. With a contiguous tail held out instead, the ordering reverses and the shift is eight times more accurate. The extra degrees of freedom bend locally to fit whatever surrounds a gap, which interpolation rewards and extrapolation punishes; a shift, carrying one parameter tied to the response's shape, has nothing to bend. A free warp evaluated by random masking will therefore look like the better model while having learned less about the response, since random masking almost always leaves samples on both sides of the gap. Timing a bolus from a sparse acquisition is the extrapolation problem.