WHEN FISHER INFORMATION MISPREDICTS PRIOR SENSITIVITY: MODEL MISSPECIFICATION, NOT NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Physics-based inverse models resolve what the measurements cannot determine by adding a prior, so part of every recovered quantity comes from the prior rather than from the object measured. Fisher information is the standard tool for predicting how large that part is. We show that the prediction fails, characterise when, and give the mechanism. For a maximum a posteriori fit whose Fisher matrix has a near-null direction v, the omitted curvature term multiplies every parameter's prior sensitivity by a single scalar, $1 - v^\top C v/\lambda$, in which the parameter index cancels. The sign of that scalar is the correlation between the fit residual and the model's curvature along v. That correlation vanishes for noise independent of the model, which is what the standard justification assumes. On 4,608 clinical MRI exams it does not vanish, and two nulls built from the cohort's own residuals reject both halves of the assumption. The Fisher prediction is then too large by 29–38%, exceeds the measured sensitivity in 92.8% to 98.8% of exams, and loses the ordering across exams for T_1. Because the error occupies one eigendirection, it can be repaired without computing the curvature: shifting a single eigenvalue recovers 92–99% of the full-Hessian correction at two gradient evaluations, whatever the parameter count. Both the failure and the repair reproduce on public diffusion MRI from another laboratory and on outdoor photovoltaic modules, and both vanish on transit photometry, where the residual is genuinely noise. The same substitution underlies Laplace approximations, influence functions and Fisher-based continual learning. Measured on a neural network, with weight decay as the prior, Fisher over-predicts how far the weights follow a shifted prior by 2.0 to 3.4 times across seven widths, and the exact curvature recovers it to 0.5%. On the clinical images, self-supervised networks differing only in random seed disagree 2 to $3.2\times$ more along the unconstrained direction than across it. What the measurements cannot determine, the training run decides.

1 INTRODUCTION

Recovering physical parameters from measurements is ill-posed whenever the data cannot determine every parameter, which is the usual case in quantitative imaging: the MRI fits below recover four tissue properties per voxel from three images. Every practical method closes the gap with a prior, whether an explicit penalty toward nominal values or a training population supplying the same pull implicitly. Part of every estimate therefore comes from the prior rather than from the object measured, and interpreting it means knowing how large that part is.

Fisher information is the standard instrument for that question. For a maximum a posteriori (MAP) fit the prior's contribution has a closed form, and the same matrix that yields the Cramér–Rao bound also predicts it. That second prediction is used far less carefully than the first, and to our knowledge has not been checked directly against measurement. We measure it against the fits themselves, on clinical data. Figure 1 runs that check on one photovoltaic module, where the standard prediction is too large by a factor of 6.3.

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

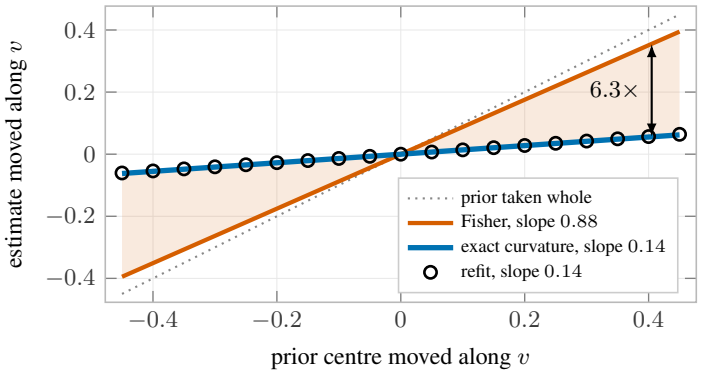


Figure 1: Prior sensitivity on one photovoltaic module, 407 days outdoors, at the median effect among the aged modules. The prior is Gaussian of unit width on the five log-parameters, centred on a datasheet-style estimate read off the curve. Moving its centre along the direction v that the measurements constrain least and refitting at each step gives the black points, whose slope is what Fisher information is asked to predict: Fisher gives 0.88 where the refit gives 0.14, so the shaded wedge is its error, and the exact curvature gives 0.14. Both predictions are straight by construction; the refit is not. Along v the shunt resistance can sit between 24 and $58\,\mathrm{k}\Omega$ for a 0.005σ change in the residual, against the 1.35σ of misfit the fit already carries.

Contributions.[1] We derive the effect of the omitted curvature term on prior sensitivity and find it collapses to a single scalar in which the parameter index cancels (§3.1). We identify what dropping that term requires, which is a non-correlation rather than a small residual, and reject it on the cohort against two nulls built from its own residuals (§3.2). We measure the resulting error on 4,608 clinical exams, including a loss of T_1 ordering that no scale correction repairs, and reproduce both the failure and its repair on public diffusion MRI, outdoor photovoltaic modules, three neural-network regressions and influence functions, against transit photometry as a null (§4). The repair needs two gradient evaluations rather than the curvature, and has a scope boundary the cohort locates (§4.6). For learned estimators on the same images, what the measurements cannot determine is decided by the training run, so self-supervised quantitative maps are not reproducible in exactly those directions.

2 Related work

Sloppy-model analysis (Gutenkunst et al., 2007; Transtrum et al., 2011) established that nonlinear least-squares problems have Fisher spectra spanning many decades, with predictions stiff in a few directions and insensitive in the rest. The identifiability literature approaches the same structure from the other side, asking which parameters the data determines at all (Raue et al., 2009). Transtrum and Sethna argue that the small-curvature approximation is safe when the model manifold's extrinsic curvature is small, and separate that from parameter-effects curvature, which a change of coordinates removes in many cases (Transtrum & Sethna, 2012).

Criticism of the Fisher matrix in parameter estimation is long established in gravitational-wave analysis, where its variance predictions are known to mislead outside the high-signal regime (Vallisneri, 2008). In quantitative MRI (Ma et al., 2013) the same matrix is the standard tool for choosing an acquisition before any patient is scanned (Zhang et al., 2022), and networks are now fitted to the same forward models, supervised against reference maps (Sun et al., 2023) or against the physics alone (van Lune et al., 2026). The empirical Fisher approximation is separately known to be a poor stand-in for curvature in optimisation (Kunstner et al., 2019).

Replacing the Hessian by the Fisher or Gauss–Newton matrix, justified by the residual term being small at a converged fit, is not peculiar to physics-based fitting. It is the standard construction in Laplace approximations for neural networks (MacKay, 1992; Ritter et al., 2018; Daxberger et al., 2021), in the natural-gradient and Kronecker-factored curvature family (Amari, 1998; Martens & Grosse, 2015; Martens, 2020), and in the matrix that influence functions (Koh & Liang, 2017; Bae

[1] Code: `https://anonymous.4open.science/r/mri-identifiability-ceiling-7BE0/`.

et al., 2022) and Fisher-based continual-learning penalties (Kirkpatrick et al., 2017) invert. Influence functions in deep networks are already known to be unreliable, which Basu et al. (2021) attribute to non-convexity, depth and regularisation.

3 METHODS

For one measurement site write the unknown parameters as x and the forward model as f, so that $y = f(x) + \varepsilon$. Let $J = \partial f/\partial x$ be the Jacobian and W the inverse noise covariance, diagonal throughout, so W_m is the weight on measurement m. We estimate x by MAP fitting against a Gaussian prior of mean x_0 and precision $\Lambda = \lambda I$. Writing $F = J^\top W J$ for the Fisher information matrix, and linearising the fit about the true parameter value,

$$\hat{x} - x_{\text{true}} \;\simeq\; \underbrace{(F+\Lambda)^{-1} J^\top W \varepsilon}_{\text{noise}} \;-\; \underbrace{(F+\Lambda)^{-1}\Lambda\,(x_{\text{true}} - x_0)}_{\text{prior contribution}}. \tag{1}$$

The first term averages away over repeated measurements; the second does not. It is a bias, pulling the estimate a fixed fraction of the way toward the prior mean . That fraction is the *leak matrix*, which Fisher information predicts as

$$M_F = (F+\Lambda)^{-1}\Lambda. \tag{2}$$

With $\Lambda = \lambda I$ its eigenvalues in the eigenbasis of F are $\lambda/(f_k + \lambda)$.

F is an approximation. The exact curvature of the fitting problem is

$$H = F - \sum_m W_m r_m\, \partial^2 f_m \;\equiv\; F + C, \tag{3}$$

with $r_m = y_m - f_m(x)$. Replacing H by F is justified on the grounds that C is proportional to the residuals and residuals are small at a good fit.

The leak measurement. The leak a fit actually has is the exact derivative $M_H = \mathrm{d}\hat{x}/\mathrm{d}x_0 = (H+\Lambda)^{-1}\Lambda$, which follows from the implicit function theorem applied to the MAP stationarity condition rather than from equation 2; appendix A gives the derivation, the conditions it needs, and the landscape of one fitted exam (figure 4). We measure M_H directly: move the prior's nominal value for one parameter by a known Δ and refit the same data, which moves that parameter's estimate by $M_H[j,j]\Delta$. Dividing that displacement by the Fisher prediction $M_F[j,j]\Delta$ gives a ratio that is 1 if Fisher is correct, and it is the quantity §3.1 derives in closed form.

3.1 EFFECT OF THE OMITTED CURVATURE TERM

Assumption. Let F have eigenvalues $f_1 \le f_2 \le \dots$ with eigenvectors $v = v^{(1)}, v^{(2)}, \dots$, and suppose

$$f_1 \;\ll\; \lambda \;\ll\; f_2. \tag{4}$$

The left inequality says the data does not constrain v, which holds exactly when there are fewer measurements than parameters. The right says the prior is weak against the directions the data does constrain, the regime a prior is chosen for. Both hold on the clinical cohort, and §4.2 exhibits a system where the left one fails and the effect vanishes with it.

Result. Write v_j for the jth component of $v = v^{(1)}$, and $R_j = \sum_{k\ge 2} v_j^{(k)2}/(f_k+\lambda)$ for the part of parameter j's leak carried by the directions the data does constrain. Expanding $(F+\Lambda+C)^{-1}$ and keeping R_j rather than discarding it,

$$\frac{M_H[j,j]}{M_F[j,j]} \;=\; \frac{v_j^2/(f_1 + v^\top C v + \lambda) + R_j}{v_j^2/(f_1+\lambda) + R_j}. \tag{5}$$

When $v_j^2 \gg \lambda R_j$ as well, both R_j terms drop and the parameter index cancels:

$$\frac{M_H[j,j]}{M_F[j,j]} \;\approx\; 1 - \frac{v^\top C v}{\lambda}. \tag{6}$$

Appendices B and C carry the steps, and §4.2 measures how well each form holds.

Consequences. *(i) Where $v_j^2/(f_1+\lambda) \gg R_j$ the correction is parameter-independent*: the leak is multiplied by $(f_1+\lambda)/(f_1+v^\top C v+\lambda)$, one scalar for every such parameter. *(ii) Where $v_j = 0$ the parameter is untouched*, since the correction lives in the $vv^\top$ block and the surviving λR_j is unchanged. Real parameters lie between, which is why (i) holds only to about 15% on the cohort. The correction must be applied in the eigenbasis, to the v–v term alone: rescaling the whole leak matrix by the scalar of (i) gives a median observed over predicted of 1.49 against 0.94 for leaving F alone, so it is worse than no correction. *(iii) The sign is an inner product.* With $g_m = v^\top \partial^2 f_m v$,

$$v^\top C v = -\sum_m W_m r_m \, g_m = -\langle Wr, g\rangle, \tag{7}$$

so the correction reduces the leak exactly when the residual is *negatively* correlated with the model's curvature along the direction the data cannot see.

3.2 The sign of the correction

If r were zero-mean noise, independent of g, then $\mathbb{E}\langle Wr, g\rangle = 0$ and the sign would be a coin flip: the correction would add variance to the prediction without biasing it. That independence is what the standard justification quietly assumes. It fails under model misspecification, where r carries a component that does not shrink with more data, and it fails through the fit itself, since g is evaluated at $\hat{x}$ and $\hat{x}$ depends on the data. A fit absorbs misspecification by moving, that displacement changes the model's curvature along v, and residual and curvature therefore covary whatever the sign of the misspecification itself.

The requirement. The argument for replacing H with F is that C is proportional to the residual and residuals are small at a good fit. The requirement is that the residual be *uncorrelated with the model's curvature along the unconstrained direction*, which smallness alone does not give. Noise satisfies this; a misspecified model need not. The failure is therefore structural rather than a small-sample artifact, and it is largest exactly in the directions that motivated the prior in the first place.

4 Results

Data. Five systems carry the evidence, with sizes and settings in table 1. The clinical arm is $4{,}608$ quantitative prostate MRI exams from a single centre, four tissue parameters against three measurements per voxel. Public diffusion MRI from another laboratory repeats that fit on a different acquisition and population. Outdoor photovoltaic modules put the same measurement outside medicine, where a single-diode fit to a weathered module is misspecified by construction. Three neural-network regressions put weight decay in the role of the prior, and influence functions on those networks test the substitution where it is most used. Transit photometry is the null arm, where the residual is genuinely instrument noise.

Table 1 summarises what follows: the condition holds on four systems and fails on one, and where it holds the Fisher prediction is too large and the exact curvature recovers the measurement.

4.1 Clinical MRI

A quantitative MRI exam supplies three images per voxel, from which four tissue parameters are inferred: proton density (PD), the relaxation times T_1 and T_2, and the diffusion coefficient D. Three measurements against four unknowns, so F is singular by construction. We use 4,608 clinical prostate exams from a single centre.

Estimates move 29–38$\%$ less than predicted (Table 2). The shortfall holds exam by exam, in 92.8% to 98.8% of exams depending on the parameter and in 94% to 100% under each of three measurement sets, where a coin flip would give 50%. Counting exams is a sign test, so no tail of extreme ratios can produce it.

Nulls from the cohort's own residuals. A contrasting cohort would mean controlling the misspecification, which a fixed clinical cohort cannot offer. Only the *pairing* of r and g carries the claim, so we measure the inner product per voxel and compare it against two nulls built from the

Table 1: Every arm, with its measured leak divided by each prediction. A ratio of 1 is agreement and the closest prediction is bold. Units are whatever the arm resamples: exams, sessions, curves, planets or admitted training runs. The yacht network nearly interpolates its data, so its curvature is indefinite at the solution and no leak is defined; transit photometry is the null arm, where $v^\top C v$ vanishes and the leak ratios are not interpretable. On the clinical cohort f_1 is exactly zero, since three measurements cannot determine four parameters.

	setting			condition			measured / predicted		
system	units	p	meas.	f_1/λ	$v^\top C v/\lambda$	z_{flip}	Fisher	rank-one	exact
clinical MRI	4,608 exams	4	3	–	$+0.655$	-37.8	0.636	0.960	**0.963**
IVIM diffusion	5 sessions	4	15	0.519	$+0.590$	-13.5	0.726	0.966	**1.002**
photovoltaic	40 curves	5	∼4,000	0.136	$+6.183$	–	0.158	1.006	**1.006**
network, airfoil	15 runs	71–281	1503	0.228	$+2.876$	–	0.335	1.106	**1.002**
network, wine	12 runs	131–521	1599	0.396	$+1.912$	–	0.567	1.295	**1.000**
network, yacht	0 of 8 admitted	–	308	–	–	–	–	–	–
transit photometry	4 planets	4	1,321	0.004	-0.011	-2.5	–	–	–

Table 2: Displacement when the prior's nominal value is raised by $1.25\times$, over 4,608 exams. Each column is a median over exams, so the first two do not divide into the third. Intervals resample patients.

	displacement (log units)		measured / predicted		below
	measured	Fisher	ratio (1 = agreement)	95% CI	prediction
PD	$+0.003$	$+0.004$	0.706	[0.700, 0.715]	92.8%
T_1	$+0.119$	$+0.212$	0.636	[0.633, 0.639]	93.9%
T_2	$+0.009$	$+0.017$	0.617	[0.607, 0.626]	98.8%

same residuals. Permuting voxels pairs each residual with a different voxel's curvature, which preserves both marginals exactly and destroys the correspondence alone, so a surviving signal cannot come from independence. Negating a random half of the per-voxel inner products leaves the statistic symmetric about zero whenever r is zero-mean noise, so a surviving signal cannot come from a zero-mean residual either.

On every measurement set $\langle Wr, g\rangle$ is negative in 99.9% of exams where a coin flip gives 50%; against the sign-flip null the median exam sits at $z = -37.84$ with every exam past $|z| = 2$, and against the permutation null at $z = +7.79$ with 89.5% past it. The permutation z is *positive* while the inner product is negative, so breaking the pairing makes it more negative still: what signs the correction is that the residual is not zero-mean, and the voxel-level correlation is separately significant but acts against it.

Ordering across exams. Figure 2 shows the T_1 prediction saturating: for half the cohort it returns nearly one answer while the measurement spans 0.019 to 0.159 between the 5th and 95th percentiles. Its rank correlation with the measurement is 0.419, rising to 0.915 once the curvature term is restored. PD and T_2 keep theirs, at 0.963 and 0.865, where Fisher costs magnitude alone. A median can be repaired by a scale factor; rank information cannot.

The corrected prediction. Recomputing with the exact curvature brings all three ratios within 6% of 1: 1.061, 0.963 and 1.018. None reaches 1, and the bootstrap intervals $[1.059, 1.062]$, $[0.962, 0.964]$ and $[1.016, 1.022]$ all exclude it, so a small residual survives.

A rank-one correction. Equation 6 says the error lives entirely in one eigendirection, which means it can be repaired without building H at all: take F's near-null eigenvector, replace f_1 by $f_1 + v^\top C v$, and leave the rest of the spectrum alone. The directional second derivative that needs costs two gradient evaluations whatever p is, against $2p$ for a finite-difference Hessian and $O(p^2)$ for a dense one. On the cohort it reproduces the full-Hessian correction: 1.034, 0.960 and 1.009 against 1.061, 0.963 and 1.018, recovering 92% to 99% of it, and leaving D at 1.000. We report the

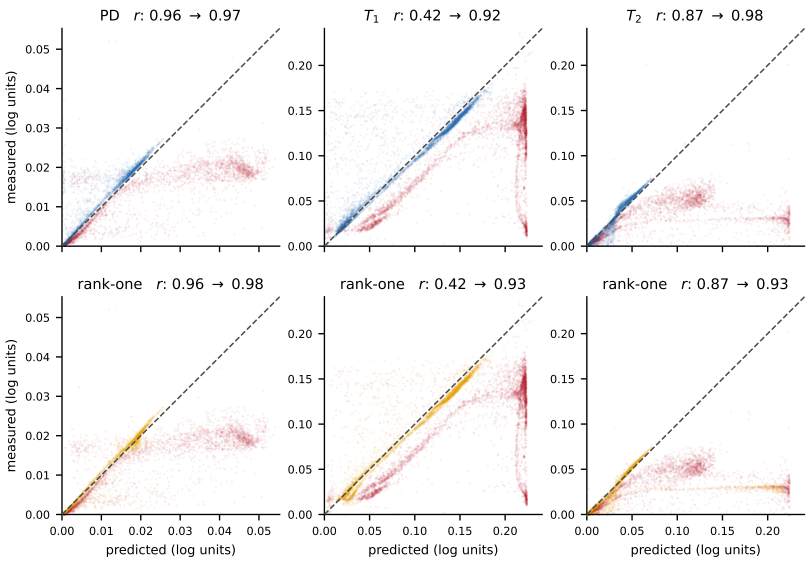


Figure 2: How far each estimate moved when the prior moved, against how far it was predicted to move, one point per exam over $4{,}608$ clinical MRI exams. Dashed is agreement. Red (top) is the Fisher prediction, which sits below the line and flattens to the right: for T_1 it predicts over 0.95 of the full shift on 50.9% of exams while the measurement does not follow. Blue is the same prediction with the omitted curvature term restored. Orange (bottom) is the rank-one repair on the same axes, which tracks the exact curvature except on a tail of T_2 exams whose leak is not carried by v. Panel titles give the rank correlation with the measurement.

saving as a count rather than a speedup: at $p = 4$ the per-voxel eigendecomposition dominates and the scalar route is the slower of the two.

The two consequences. Consequence (i) holds to leading order and no further: the three ratios agree to about 15%, with a median spread of 0.174 *within* a single exam. Appendix G reports consequence (ii), the robustness checks and the remaining predictions.

4.2 Diffusion, photometry and photovoltaics

The same measurement runs on three further systems, two that the theory says should show the correction and one that it says should not.

Diffusion MRI. Intravoxel incoherent motion (IVIM) (Le Bihan et al., 1988) fits $S(b)/S_0 = fe^{-b(D+D^*)} + (1-f)e^{-bD}$ per voxel against a set of b-shells. We take a multi-shell test–retest brain acquisition from OpenNeuro (ds008695, Philips 3T), $3{,}000$ voxels per session. The pseudo-diffusion coefficient D^* is the parameter at issue and is routinely held fixed in that literature, so the prior is established practice rather than our imposition. Fisher over-predicts its sensitivity on every session, every bootstrap interval excluding 1, and two of the five sessions are one subject a week apart, so the agreement survives a rescan.

Transit photometry. §3.2 requires the residual to correlate with the curvature along v, which a wrong model does not by itself produce. A fit whose residual is measurement noise should therefore show nothing. We fit a limb-darkened transit to $1{,}321$ events from four Kepler planets, where the residual is the photon noise to within 10–20% of the catalogue error, and the correction collapses. This is the noise-only comparison a fixed clinical cohort cannot supply.

Photovoltaic I–V curves. A solar module is neither MRI nor an imaging model. The single-diode model has five parameters and a residual that is model error rather than noise, since real modules depart from it. We fit 60 flash-test curves sampled evenly across exposure from a public

Table 3: The network arm: measured displacement along v divided by each prediction, over seven widths of the airfoil network. A ratio of 1 is agreement and the closest prediction is bold. All seven pass the admission tests of appendix D.

		diagnostics		measured / predicted		
h	p	$\cos(\Delta w, v)$	$v^\top C v/\lambda$	Fisher	rank-one	exact
10	71	$+0.999$	$+3.07$	0.294	1.059	**1.007**
15	106	$+0.995$	$+1.23$	0.492	1.061	**1.005**
20	141	$+0.975$	$+2.92$	0.487	1.218	**0.995**
25	176	$+0.993$	$+2.97$	0.295	1.083	**1.004**
30	211	$+0.940$	$+2.92$	0.292	1.106	**1.002**
40	281	$+0.987$	$+2.50$	0.332	1.084	**1.002**
50	351	$+0.972$	$+2.74$	0.331	1.100	**0.995**

Sandia cohort of 616 (Byford et al., 2025); each carries about four thousand points, so the near-null direction is a parameter degeneracy rather than a shortage of measurements. The spread is wide rather than tight, and ten of the forty aged curves sit where Fisher is already close to right. These modules carry outdoor exposure to 1,885 days, which varies the misspecification on a fixed cohort. Ageing moves the condition rather than the curvature term: median f_1/λ falls from 66.5 on unexposed modules to 0.13 beyond four hundred days while $v^\top C v/\lambda$ stays near $+6$. A new module has no near-null direction to speak of. Across these curves and their 300 well-conditioned parameter ratios, equation 5 predicts the measured ratio to a median of 1.0% and equation 6 to 62%. Appendix F gives the model and the caveats on that comparison.

The role of the residual, on the diffusion arm. Hold the data, the scanner and the noise fixed, and change only how many shells are fitted. At three shells the four parameters absorb everything and $v^\top C v/\lambda$ is -0.014, positive in 41% of voxels; at fifteen it is $+0.828$ and positive in 72%. A systematic residual is what the sign condition needs, and it arrives either from a prior strong enough to prevent an exact fit or from model error the fit cannot absorb. Underdetermination is one road to it, not the requirement itself.

4.3 Learned estimators

The analysis so far concerns a per-exam optimiser. Physics-informed networks are trained on the same objective, so the same directions should be undetermined for them, and the consequence is sharper: a MAP fit is unique given its objective, while a network is whatever its training run converged to. Self-supervised training (van Lune et al., 2026) minimises the same loss and needs no ground-truth maps, so it runs on the cohort. We train five networks differing only in random seed and split each pair's disagreement on held-out exams along v and across it.

Across 80 held-out exams the median disagreement is 0.078 log units along v against 0.034 across it, a ratio of 2.28, with 88.8% of exams following that ordering. Other training-set sizes, widths and lengths give 3.02 to 3.19 at 40,000 steps and six seeds, excluding one run whose output bound saturated, so the concentration is pinned down and its size is not. In the parameter itself 0.078 log units is 8.1%. Two runs of one architecture on one dataset therefore produce maps differing by about 8%, concentrated where the measurements are silent. For a literature that treats these maps as scanner-invariant tissue properties (Borges et al., 2023; Lee et al., 2025; Cai et al., 2023; van Lune et al., 2026), the map is in part a property of the training run.

4.4 A neural network

Laplace approximations, influence functions and Fisher-based continual-learning penalties all replace the Hessian by F, on the argument Section 3.1 has been testing. Weight decay toward a centre w_0 is a Gaussian prior with mean w_0 and precision λ, so a network admits the measurement the cohort answers: how far the estimate follows when w_0 moves.

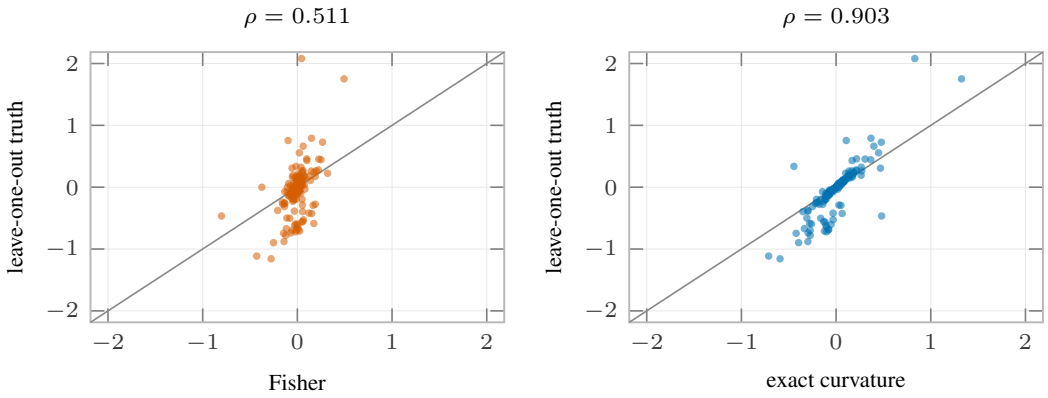


Figure 3: Influence estimates against the leave-one-out answer they approximate, for the configuration with the largest gap. Each point is one training example; the grey line is identity.

We fit a tanh network to the airfoil self-noise measurements (Brooks et al., 1989), 1,503 wind-tunnel records with five inputs, using the Levenberg-damped Gauss–Newton MAP estimator of Section 3, so the optimiser is shared with the other arms. The activation is tanh rather than ReLU because C is a curvature term and a piecewise-linear network has $C = 0$ almost everywhere. We move w_0 along v, refit, and project the displacement onto v, where equation 2 predicts $\lambda/(f_1 + \lambda)$ and equation 6 predicts $\lambda/(f_1 + v^\top C v + \lambda)$.

Refits are warm-started from the unshifted solution, since equation 12 describes how one stationary point deforms. Appendix D sets out the estimator, the three admission tests and the sweeps behind this table.

Across seven widths, $p = 71$ to 351 (table 3), Fisher over-predicts the prior's pull by 2.0 to 3.4 times, and the exact curvature recovers it to within 0.7%, median 0.5%. The quantity $v^\top C v$ is positive in every airfoil configuration, as on the clinical cohort and in IVIM. Five seeds at each of two widths give the same picture, with all ten admitted. Lowering λ through three decades leaves the exact correction between 0.99 and 1.02 across the admitted settings. The rank-one form tracks its own prediction where $f_1 \ll \lambda$ and departs from it where the spectral gap closes, which is the assumption of Section 3.1 failing where it says it will.

Competing approximations. The empirical Fisher matrix, the outer product of per-sample gradients, gives ratios of 0.26 to 0.30, slightly worse than F itself. That is consistent with Kunstner et al. (2019), though milder than their claim, which is that the empirical Fisher need capture no second-order information at all. The diagonal of F is the penalty matrix of elastic weight consolidation (Kirkpatrick et al., 2017), and it gives ratios between 8.1×10^3 and 2.8×10^4: it predicts almost no prior influence at all. Here v mixes coordinates, so each coordinate keeps a large diagonal entry and a diagonal approximation never sees that a mixed direction is near-null. Dropping the off-diagonals removes the effect rather than approximating it, and the diagonal Hessian behaves the same way.

4.5 Influence functions

Influence functions supply a quantity a practitioner computes together with a way of checking it, since the influence of a training point on the test loss approximates the effect of removing that point, and that effect can be had by retraining. Writing A for the curvature that gets inverted,

$$I(z_i) \;=\; \nabla L_{\text{test}}^\top A^{-1} \nabla L(z_i), \qquad \text{truth} \;=\; L_{\text{test}}(\hat{w}_{-i}) - L_{\text{test}}(\hat{w}), \tag{8}$$

with $A = F + \lambda I$ in standard practice and $A = H + \lambda I$ exactly. On the network of §4.4 we retrain without each of 200 training points and rank both approximations against that truth.

Figure 3 shows one configuration. Across 24 the median Spearman correlation is 0.628 using F and 0.881 using H, and H is closer in 23 of 24. Both understate the magnitude of the effect, with median slopes of 1.73 and 1.54, so what H buys is ordering rather than scale. One configuration fails for both, which is the fragility Basu et al. (2021) report.

Replacing f_1 by $f_1 + v^\top C v$ and leaving the rest of the spectrum alone recovers a median -0.02% of the distance from F to H. A test point's gradient is almost orthogonal to v, which carries 0.3% or less of the predictive variance, and the same geometry predicts the Laplace null we measure. Appendix E validates the implementation against a linear model whose leave-one-out answer is known in closed form.

4.6 Range of validity

Equation 6 assumes one unconstrained direction and the spectral gap of equation 4. Both can be varied on the cohort, by removing a measurement and by changing the prior width. The sharpest of those tests asks how much of the full-Hessian correction a *rank-one* repair recovers, which is a direct test of the one-direction assumption: 92–99% at null dimension one, against 31–60% at two and 34–48% at three. The step is between one and more than one, and the higher-codimension arms are indistinguishable from each other, which is what a rank-one correction should do. Lowering λ four decades leaves the ratio near 0.63 while tightening the within-exam spread, as the gap assumption implies. Appendix H reports both sweeps, and the reasons the measurement-set arms are not a controlled codimension series.

The derivative exists only where $H + \Lambda \succ 0$. All eight yacht networks fail that test, because a network that nearly interpolates has indefinite curvature at the solution, and any fit trained close to interpolation fails it the same way. The repair is also specific to prior sensitivity: quantities contracting the inverse curvature against data gradients are governed by the whole of C, which is why the rank-one shift does nothing for the influence functions of §4.5.

5 Discussion

Fisher information does two different jobs. Asking which parameters an acquisition can determine, before any data exists, is exact, and for protocol design F is both right and the only one to hand, since H needs the images and a completed fit. Asking how much of a fitted value came from the prior is the job that fails, and it fails in a way a practitioner can check: the deciding quantity is an inner product between the residual and the model's curvature along the unconstrained direction, two gradient evaluations from a completed fit. A prior sensitivity reported from F alone overstates how much of the estimate came from the prior. The repair needs no Hessian, but it must be applied in the eigenbasis: as a rescaling of the leak matrix it is worse than no correction. Magnitudes belong at a stated prior width, since the factor moves with λ (figure 6, appendix I). The direction survives every setting we tested; no ordering of systems by effect size does.

The same substitution sits inside Laplace approximations, influence functions and Fisher-based continual-learning penalties. Exact curvature raises influence-function agreement with retraining from 0.628 to 0.881, and the diagonal of F that elastic weight consolidation penalises with misses the effect by four orders of magnitude, because v mixes coordinates and a diagonal never sees that a mixed direction is near-null.

Whether a cheap correction exists for the quantities of §4.6, which the rank-one repair does not reach, is open.

Misspecification cannot be varied on a fixed cohort, so the structured residual is not decomposed by shortcut here. We do not explain why the corrected ratio lands at 0.963 rather than 1, nor why the rank-one correction beats the full Hessian for PD and T_2; finite-difference noise, the obvious explanation, is wrong. The learned arm's concentration appears in every configuration but its size does not hold still, 2.18 to 3.19, and averaging runs shrinks it at about the rate expected of seed noise.

6 Conclusion

Fisher information predicts the prior's contribution to a fitted parameter too large whenever the residual is misspecification rather than noise. Dropping the curvature term needs that residual uncorrelated with the model's curvature along v, and on 4,608 clinical exams it is not. The repair costs two gradient evaluations and holds at codimension one.

Reproducibility statement

The derivation of equation 6 and its assumptions are stated in §3.1, and the sign condition of equation 7 in §3.2. Every number carries a claim tag traced to the run that produced it. Code, tests and each analysis arm are at the anonymised repository cited in §1, which includes the transit, diffusion, photovoltaic and neural-network arms in full. Four public datasets are identified by accession. The diffusion cohort is OpenNeuro ds008695 and the transit photometry comes from the Kepler archive. The photovoltaic curves are Sandia PV-IV-EL (OpenEI submission 8378, CC BY 4.0), and the network arm uses three UCI regression sets, airfoil self-noise (Brooks et al., 1989), wine quality (Cortez et al., 2009) and yacht hydrodynamics (Gerritsma et al., 1981). The clinical cohort is not redistributable, so the results in §4.1 cannot be rerun from the repository. The processing steps are described there and the analysis code is included, so the pipeline is inspectable even where the images cannot be shared.

AI use statement

We used a generative AI coding assistant for software implementation, for exploratory data analysis, and for drafting and editing prose. All AI-assisted code was reviewed by the authors and is covered by the repository's test suite; every quantitative claim in this paper was produced by running that code and is traceable to a logged run rather than to model output. The research questions, experimental designs, choices of control and null experiment, and the interpretation of results are the authors' own. We did not use generative AI to produce or alter any measurement, and no text was included without author review. We take responsibility for the final content of this work, including all text, claims and artifacts.

References

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746.

Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger B. Grosse. If influence functions are the answer, then what is the question? In *Advances in Neural Information Processing Systems*, 2022. https://arxiv.org/abs/2209.05364.

Samyadeep Basu, Phillip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations*, 2021.

Pedro Borges, Virginia Fernandez, Petru-Daniel Tudosiu, Parashkev Nachev, Sébastien Ourselin, and M. Jorge Cardoso. Unsupervised heteromodal physics-informed representation of MRI data: Tackling data harmonisation, imputation and domain shift. In *Simulation and Synthesis in Medical Imaging (SASHIMI), MICCAI Workshops*, LNCS 14288, 2023.

Thomas Brooks, D. Pope, and Michael Marcolini. Airfoil self-noise. UCI Machine Learning Repository, 1989.

Brandon Byford, Jennifer Braid, Marios Theristis, Charles Robinson, William Snyder, Kevin Santistevan, Norman Jost, and Bruce King. Photovoltaic module current-voltage and electroluminescence image data (PV-IV-EL). Sandia National Laboratories, Open Energy Data Initiative, 2025. https://data.openei.org/submissions/8378.

Qingrui Cai, Liuhong Zhu, Jianjun Zhou, Chen Qian, Di Guo, and Xiaobo Qu. Bloch equation enables physics-informed neural network in parametric magnetic resonance imaging. *arXiv preprint arXiv:2309.11763*, 2023.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4): 547–553, 2009.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux – effortless Bayesian deep learning. In *Advances in Neural Information Processing Systems*, pp. 20089–20103, 2021.

J. Gerritsma, R. Onnink, and A. Versluis. Yacht hydrodynamics. UCI Machine Learning Repository, 1981.

Ryan N. Gutenkunst, Joshua J. Waterfall, Fergal P. Casey, Kevin S. Brown, Christopher R. Myers, and James P. Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLoS Computational Biology*, 3(10):e189, 2007. doi: 10.1371/journal.pcbi.0030189.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, 2017.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical Fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems*, 2019.

Denis Le Bihan, Eric Breton, Denis Lallemand, Marie-Louise Aubin, Jean Vignaud, and Maurice Laval-Jeantet. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. *Radiology*, 168(2):497–505, 1988.

Gawon Lee et al. A preliminary attempt to harmonize using physics-constrained deep neural networks for multisite and multiscanner MRI datasets (PhyCHarm). *NeuroImage*, 2025. doi:10.1016/j.neuroimage.2025.121361.

Dan Ma, Vikas Gulani, Nicole Seiberlich, Kecheng Liu, Jeffrey L. Sunshine, Jeffrey L. Duerk, and Mark A. Griswold. Magnetic resonance fingerprinting. *Nature*, 495(7440):187–192, 2013. doi: 10.1038/nature11971.

David J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *International Conference on Machine Learning*, 2015.

Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009. doi: 10.1093/bioinformatics/btp358.

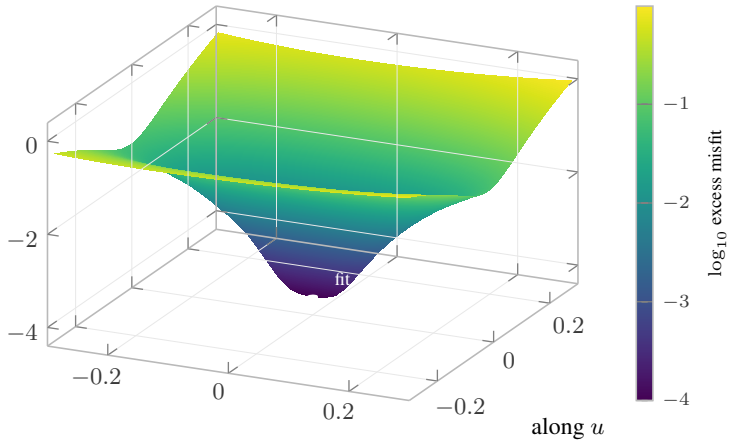
Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable Laplace approximation for neural networks. In *International Conference on Learning Representations*, 2018.

Haoran Sun, Lixia Wang, Timothy Daskivich, Shihan Qiu, Fei Han, Alessandro D'Agnolo, Rola Saouaf, Anthony G. Christodoulou, Hyung Kim, Debiao Li, and Yibin Xie. Retrospective T2 quantification from conventional weighted MRI of the prostate based on deep learning. *Frontiers in Radiology*, 2023. PMC10598780.

Mark K. Transtrum and James P. Sethna. Geodesic acceleration and the small-curvature approximation for nonlinear least squares. *arXiv preprint arXiv:1207.4999*, 2012.

Mark K. Transtrum, Benjamin B. Machta, and James P. Sethna. Geometry of nonlinear least squares with applications to sloppy models and optimization. *Physical Review E*, 83:036701, 2011.

Michele Vallisneri. Use and abuse of the Fisher information matrix in the assessment of gravitational-wave parameter-estimation prospects. *Physical Review D*, 77:042001, 2008.



displacement along v, the direction the data does not constrain

Figure 4: The data misfit of one fitted voxel from the clinical cohort, on the plane spanned by the two least-constrained eigendirections of F. Three measurements against four parameters make F rank-deficient, and the measured f_1/λ is 2×10^{-11} against f_2/λ of 1023: along v the measurements contribute nothing and the valley is flat. Where the estimate lands on that floor is decided by the prior, and how far the prior moves it is what this paper measures. The view is cropped to ±0.30 log units, the neighbourhood in which the leak is a derivative; further out the model's nonlinearity curves the valley.

Jelmer van Lune, Stefano Mandija, Oscar van der Heide, Matteo Maspero, Martin B. Schilder, Jan Willem Dankbaar, Cornelis A. T. van den Berg, and Alessandro Sbrizzi. Quantitative mapping from conventional MRI using self-supervised physics-guided deep learning: Applications to a large-scale, clinically heterogeneous dataset. *arXiv preprint arXiv:2601.05063*, 2026.

Xiaoxia Zhang, Quentin Duchemin, Kangning Liu, Cem Gultekin, Sebastian Flassbeck, Carlos Fernandez-Granda, and Jakob Assländer. Cramér–Rao bound-informed training of neural networks for quantitative MRI. *Magnetic Resonance in Medicine*, 88(1):436–448, 2022.

A Derivation of the leak matrix

Equation 2 is stated for the linearised problem, where the curvature is $F+\Lambda$ by construction. Substituting H for F inside it does not follow from that equation. The object this paper measures comes instead from the implicit function theorem applied to the exact nonlinear stationarity condition.

Write the MAP objective as

$$L(x,x_0) \;=\; \tfrac{1}{2}\big(y-f(x)\big)^\top W\big(y-f(x)\big) \;+\; \tfrac{1}{2}(x-x_0)^\top\Lambda(x-x_0), \tag{9}$$

whose stationarity condition at the fitted point $\hat{x}$ is

$$G(\hat{x},x_0) \;=\; -J(\hat{x})^\top W\big(y-f(\hat{x})\big) \;+\; \Lambda(\hat{x}-x_0) \;=\; 0. \tag{10}$$

The implicit function theorem gives the derivative of the solution with respect to the prior centre. Differentiating equation 10 in x_0,

$$\frac{\partial G}{\partial\hat{x}}\,\frac{\mathrm{d}\hat{x}}{\mathrm{d}x_0} \;+\; \frac{\partial G}{\partial x_0} \;=\; 0, \qquad \frac{\partial G}{\partial x_0} \;=\; -\Lambda, \tag{11}$$

and $\partial G/\partial\hat{x}$ is the exact posterior curvature $H+\Lambda$, where $H = J^\top WJ + C$ and $C = -\sum_m W_m r_m \partial^2 f_m$. Therefore

$$\frac{\mathrm{d}\hat{x}}{\mathrm{d}x_0} \;=\; (H+\Lambda)^{-1}\Lambda \;=\; M_H. \tag{12}$$

Conditions.

1. *f is twice continuously differentiable.* C is a curvature term, so a piecewise-linear model has $C = 0$ almost everywhere and the question is vacuous by construction. The network arm uses tanh for this reason.
2. *$H + \Lambda$ is invertible at $\hat{x}$.*
3. $H + \Lambda \succ 0$. Invertibility alone permits a saddle, and C is indefinite in general, so this does not follow from the previous condition. It is the guard the implementation carries, and the network arm rejects configurations that violate it. On the yacht dataset that is every run: the network nearly interpolates its training data and the posterior curvature is indefinite at the solution, so no leak is defined there.
4. *The solution branch does not change between x_0 and the shifted centre.* This concerns the finite shift used in the measurement rather than the derivative itself. The displacement is measured at two shift sizes for that reason, and the ratio drifts by up to 10% over a doubling, so the measurement is a path average of M rather than M evaluated at $\hat{x}$.

B Expansion of the leak ratio

The steps behind equation 6, which the body states as a result.

Write $A = F + \Lambda$ and expand the inverse to first order in C:

$$(A + C)^{-1} \;\approx\; A^{-1} - A^{-1} C A^{-1}. \tag{13}$$

Multiplying by λ and taking the jth diagonal entry, with $a_j = A^{-1} e_j$,

$$M_H[j,j] - M_F[j,j] \;\approx\; -\lambda\, a_j^\top C\, a_j. \tag{14}$$

Now resolve a_j in the eigenbasis of F. Since

$$A^{-1} \;=\; \sum_k \frac{1}{f_k + \lambda}\, v^{(k)} v^{(k)\top}, \tag{15}$$

the spectral gap of equation 4 makes the $k = 1$ term dominant: its coefficient is $1/(f_1 + \lambda) \approx 1/\lambda$, while every other is at most $1/(f_2 + \lambda) \ll 1/\lambda$. Keeping only that term,

$$a_j \;\approx\; \frac{v_j}{\lambda}\, v, \qquad v_j = v^\top e_j, \tag{16}$$

$$M_F[j,j] \;=\; \lambda\, (A^{-1})[j,j] \;\approx\; v_j^2. \tag{17}$$

Substituting equation 16 into equation 14 gives $M_H[j,j] - M_F[j,j] \approx -v_j^2\, v^\top C v/\lambda$. Dividing that by equation 17,

$$\boxed{\frac{M_H[j,j]}{M_F[j,j]} \;\approx\; 1 - \frac{v^\top C v}{\lambda}}$$

which is equation 6. The factor v_j^2 cancels from both sides, which is why the parameter index disappears.

The cancellation is the step that carries a condition. Equation 17 drops λR_j against v_j^2, so it needs $v_j^2 \gg \lambda R_j$, and a parameter whose leak is carried mostly by the constrained directions does not satisfy it. Appendix C keeps that term instead of dropping it.

C The closed form and its limits

Figure 5 shows the closed form against the measurement it predicts.

Let F have eigenpairs $(f_k, v^{(k)})$ with $v = v^{(1)}$, and write $A = F + \lambda I$ and $v_j = v^\top e_j$. Since $A^{-1} = \sum_k (f_k + \lambda)^{-1} v^{(k)} v^{(k)\top}$, the jth diagonal entry of the Fisher leak is exact:

$$M_F[j,j] \;=\; \lambda\, (A^{-1})[j,j] \;=\; \lambda \sum_k \frac{v_j^{(k)2}}{f_k + \lambda}. \tag{18}$$

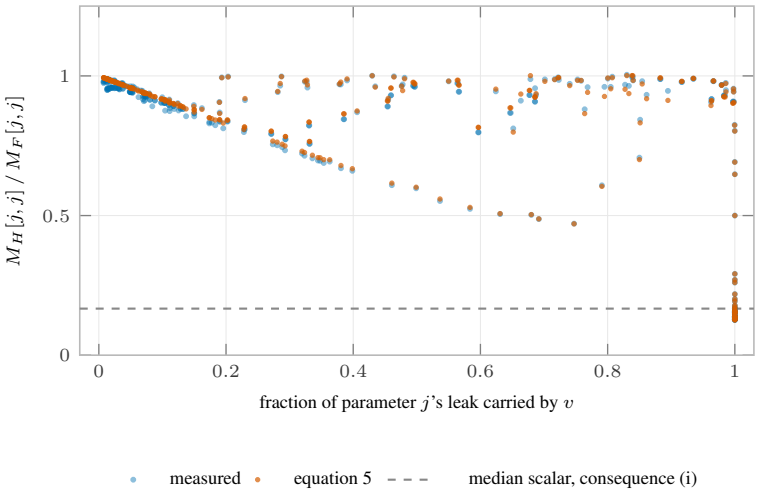


Figure 5: Every parameter of every photovoltaic curve: the measured ratio against how much of that parameter's leak the unconstrained direction carries. The dashed line is the median scalar of consequence (i), which the measurement meets only at the right-hand edge; consequence (ii) is the approach to 1 at the left. Equation 5 interpolates between them.

Separating the near-null term from the rest, and writing

$$R_j \;=\; \sum_{k\geq 2} \frac{v_j^{(k)2}}{f_k + \lambda} \tag{19}$$

for the part of parameter j's leak carried by the directions the data does constrain,

$$M_F[j,j] \;=\; \lambda \left[\frac{v_j^2}{f_1 + \lambda} \;+\; R_j \right], \tag{20}$$

$$M_H[j,j] \;=\; \lambda \left[\frac{v_j^2}{f_1 + v^\top C v + \lambda} \;+\; R_j \right]. \tag{21}$$

Equation 21 holds when C's action outside the $vv^\top$ block is negligible, which is the same suppression by λ/f_2 that equation 4 supplies: the correction shifts the near-null eigenvalue and leaves the rest of the spectrum alone. Dividing equation 21 by equation 20, the overall factor of λ cancels and equation 5 follows.

The v-dominated limit. When $v_j^2/(f_1+\lambda) \gg R_j$, both R_j terms are negligible and

$$\frac{M_H[j,j]}{M_F[j,j]} \;\longrightarrow\; \frac{f_1 + \lambda}{f_1 + v^\top C v + \lambda}, \tag{22}$$

which carries no index j. This is consequence (i). Expanding equation 22 for $v^\top C v \ll \lambda$, and using $f_1 \ll \lambda$, recovers equation 6:

$$\frac{f_1 + \lambda}{f_1 + v^\top C v + \lambda} \;=\; \left(1 + \frac{v^\top C v}{f_1 + \lambda}\right)^{-1} \;\approx\; 1 - \frac{v^\top C v}{\lambda}. \tag{23}$$

The orthogonal limit. When $v_j = 0$ the numerator and denominator of equation 5 both reduce to R_j, so

$$\frac{M_H[j,j]}{M_F[j,j]} \;=\; \frac{R_j}{R_j} \;=\; 1 \tag{24}$$

exactly. This is consequence (ii). The reason is structural rather than a matter of the leak being small: the correction lives in the $vv^\top$ block and contributes nothing to the jj entry when $v_j = 0$, while the surviving λR_j is untouched. Note that $M_F[j,j] = \lambda R_j$ is not itself zero, so the ratio is well defined.

Practical consequences. At the clinical cohort's $v^\top Cv/\lambda = 0.655$ the first-order and resummed forms give 0.345 and 0.604, and on the network arm, where $v^\top Cv/\lambda$ reaches $+3$, the first-order form is negative while the resummed one stays in $(0, 1)$. The correction also has to be applied in the eigenbasis. Rescaling the whole of M by the scalar of equation 22 applies the v-dominated answer to parameters that are not v-dominated, and against measured per-parameter displacement on photovoltaic curves it gives a median observed over predicted of 1.49, against 0.94 for leaving F uncorrected and 0.999 for equation 5.

D THE NETWORK ARM

The model is a single-hidden-layer tanh network fitted by the Levenberg-damped Gauss–Newton MAP estimator used in every other arm, so the optimiser is not a difference between arms. Refits after a prior shift are warm-started from the unshifted solution, because equation 12 describes how one stationary point deforms and says nothing about which point a cold restart would find. On this landscape a cold restart usually finds a different one: at $n_{\text{iter}} = 3000$, with every fit converged to a gradient norm near 10^{-4}, one configuration moved from objective 33.08 to 29.82.

A configuration is admitted when the displacement along F's stiffest eigendirection is below 1% of the leak along v, when the displacement is parallel to v with a positive cosine above 0.9, and when the fit is a minimum, $\lambda_{\min}(H) > 0$ with the measured leak in $(0, 1]$, as equation 12 requires. Over 77 runs the stiff-direction test rejected nothing, the cosine test rejected 18, and the minimum test rejected 24. The last is what separates airfoil and wine quality (Cortez et al., 2009), where the leak is defined, from yacht hydrodynamics (Gerritsma et al., 1981), where the network nearly interpolates the data and the curvature is indefinite at the solution.

E INFLUENCE FUNCTIONS AND THE LAPLACE APPROXIMATION

The influence implementation was validated against a problem with a closed-form answer. Replacing the network with a linear model, where leave-one-out has an exact Sherman–Morrison form, the code reproduces the analytic answer to machine precision. The reported regression slopes of about 1.06 in a single configuration are the leverage factor $1/(1 - \bar{h})$ and not a scale error; over 24 configurations the median slopes are 1.73 and 1.54.

The rank-one correction recovers a median of -0.02% of the distance from F to H in influence ranking, across those same 24 configurations. Substituting F for H in the Laplace predictive variance changes it by about 9%, stably across a sweep of the assumed noise scale, while the z-score spread moves by about 1%. The absolute coverage depends entirely on that noise scale and is not a calibration measurement. Both follow from the same geometry: a test point's gradient is nearly orthogonal to v, with median $|\cos|$ of order 10^{-4}.

F THE PHOTOVOLTAIC ARM

The single-diode model $I = I_L - I_0(e^{(V+IR_s)/a} - 1) - (V + IR_s)/R_{sh}$ is implicit in the current, so it is solved by Newton iteration and differentiated implicitly, which keeps the Jacobian free of the solver tolerance. The five parameters are fitted as logarithms. Each curve carries about four thousand points against those five, so the near-null direction is a parameter degeneracy rather than a shortage of measurements.

The relation between exposure and f_1 is a step rather than a gradient: alternative binnings are not monotone, the best cut sits near 235 days, and only two curves fall between 43 and 223 days. The rank correlation between exposure and the Fisher ratio has a bootstrap interval that includes zero. And f_1 tracks the fitted shunt resistance far more closely (Spearman -0.98) than it tracks exposure (-0.31), so the mechanism is the shunt branch carrying less information as the module degrades.

G ROBUSTNESS OF THE COHORT RESULT

Varying the prior width over two decades moves the Fisher-predicted T_1 ratio across 0.040 and the corrected ratio across 0.007; removing measurements one at a time moves them across 0.147 and

Table 4: Both assumptions varied on the cohort. The prior-width sweep behaves as the gap assumption implies: the spread tightens as λ falls and the prior becomes weak relative to the constrained directions. The measurement-set sweep does not separate codimensions. Removing the ADC anchor changes *which* parameter is unconstrained as well as how many, so these arms differ in more than one respect.

measurement set				prior width		
arm	null dim	T_1 ratio	within-exam spread	λ	T_1 ratio	within-exam spread
A1 (3 meas.)	1	0.636	0.174	1	0.653	0.204
A4 (2 meas.)	2	0.766	0.101	10^{-1}	0.625	0.220
A3 (1 meas.)	3	0.783	0.119	10^{-2}	0.636	0.174
				10^{-3}	0.634	0.129
				10^{-4}	0.613	0.126

0.005 respectively. All eight corrected values sit below 1, too regularly for sampling noise. Doubling the prior shift instead of using $1.25\times$ leaves the ratios at 0.777, 0.611 and 0.625. PD moves by 10% and T_1 by 4% over a doubling, so the measurement is a path average of M rather than M at the fitted point, with an error of that order.

Variation across instruments. Grouped by instrument, the T_1 ratio runs from 0.582 to 0.658 across five instruments of one vendor, a spread of 0.076 whose extremes fall on two of the largest groups rather than the smallest. Proton density and T_2 separate further. A correction calibrated on one instrument should not be assumed to transfer, which is what the mechanism implies: instruments misspecify the model differently, and the misspecification sets the size of the effect.

Further predictions. D should be untouched: the exam measures it directly, so it has no loading on v. Its ratio is 1.000. Acquisitions with less opportunity for misspecification should show a smaller shortfall. Removing contrasts gives T_1 ratios of 0.636, 0.766 and 0.783 for three, two and one contrast, monotone in the opportunity for inter-contrast inconsistency.

H Assumption sweeps on the cohort

Equation 6 assumes one unconstrained direction and the spectral gap of equation 4. Neither can be set freely on a fixed clinical cohort, but both can be varied: removing a measurement changes the codimension, and the prior width sets λ directly. Table 4 reports both sweeps.

Prior width. Lowering λ four decades tightens the within-exam spread from 0.220 to 0.126 while leaving the ratio near 0.63 throughout. A weaker prior means a cleaner separation between the unconstrained direction and the rest, which is the regime equation 4 describes and the regime a prior is normally chosen for. The effect is modest, and we report it as directional support rather than a tight confirmation.

Codimension. Removing measurements raises the ratio monotonically, 0.636 to 0.766 to 0.783, which is consistent with misspecification driving the shortfall: a single-contrast fit cannot have an inter-contrast inconsistency. The within-exam spread does not order with codimension, so that statistic does not separate the arms. A sharper one does. Asking how much of the full-Hessian correction a *rank-one* repair recovers is a direct test of the one-direction assumption, and it answers cleanly: 92–99% at null dimension one, against 31–60% at two and 34–48% at three. The step is between one and more than one, and the two- and three-dimensional arms are indistinguishable from each other, which is what a rank-one correction should do. This comparison is also less exposed to the arms' confound than the ratios are, since the scalar and Hessian corrections are compared within an arm on the same fits rather than across arms. It does not make these arms a controlled codimension series, since dropping the ADC anchor still changes which parameter the data fails to determine, but it does locate the boundary the derivation predicts.

I PRIOR WIDTH AND THE SIZE OF THE EFFECT

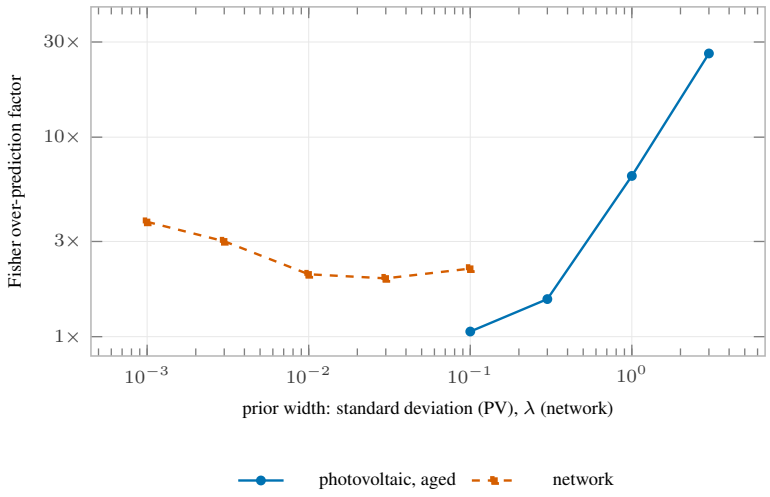


Figure 6: The over-prediction factor against the prior's width, for the aged photovoltaic curves and for the network arm. The factor is $(f_1 + v^\top C v + \lambda)/(f_1 + \lambda)$, so it moves with λ on every system. The direction of the error does not.

The over-prediction factor is $(f_1 + v^\top C v + \lambda)/(f_1 + \lambda)$, so it depends on where the prior's precision sits relative to f_1. On the aged photovoltaic curves the median factor is 1.06 at a prior standard deviation of 0.1, 1.54 at 0.3, 6.37 at 1.0 and 25.8 at 3.0. The network arm's own sweep moves the Fisher ratio from 0.266 to 0.600 across three decades of λ.

The direction survives every setting tested: $v^\top C v$ stays positive, Fisher over-predicts, and the exact curvature recovers the measurement. The magnitude does not, and neither does any ordering of the systems by effect size. Where this paper reports a factor, it reports it at a stated λ and does not compare it against another arm's.