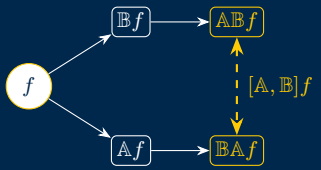
Statistical Commutators

Operator geometry for estimation, inference, and regularization tuning



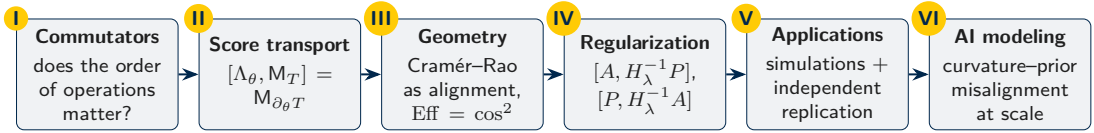
Ivo D. Dinov[1] **Eric Tatt Wei Ho**[2] **S. Ejaz Ahmed**[3]

[1]University of Michigan [2]Universiti Teknologi Petronas [3]Brock University

Statistics and Computing **36**, Article 228 (2026) doi:10.1007/s11222-026-10991-w

https://socr.umich.edu/ | https://DSPA3.predictive.space/

Outline


I
Commutators
does the order of operations matter?
II
Score transport
$[\Lambda_\theta, \mathsf{M}_T] = \mathsf{M}_{\partial_\theta T}$
III
Geometry
Cramér–Rao as alignment, $\mathrm{Eff} = \cos^2$
IV
Regularization
$[A, H_\lambda^{-1} P]$, $[P, H_\lambda^{-1} A]$
V
Applications
simulations + independent replication
VI
AI modeling
curvature–prior misalignment at scale

When does order matter?

Let D differentiate a smooth function and let M_x multiply it by x. For instance, let $f(x) = x^2$. Then, $D(\mathsf{M}_x f) = (x^3)' = 3x^2, \qquad \mathsf{M}_x(Df) = x(2x) = 2x^2.$

The commutator is the difference

For linear operators A, B on a common domain where both products make sense,

$$[A, B]f = A(Bf) - B(Af).$$

Here $[D, \mathsf{M}_x]f = (xf)' - xf' = f$ for every smooth f. Two operators *commute* if the difference is zero for every input in the domain.

$[A, B]$ measures the non-commutativity of 2 linear operators. In *physics*, a nonzero commutator signals that sequential *measurements* or transformations interfere, i.e., measuring or evolving along observable A then B alters the state differently than B then A. Canonical commutation relations, e.g., $[\hat{x}, \hat{p}] = i\hbar$ (position and momentum) or angular momentum algebras $[L_i, L_j] = i\hbar\epsilon_{ijk}L_k$, yield fundamental uncertainty bounds and determine whether observables share eigenstates or a generator leaves a quantity invariant.

Two statistical questions

Estimation. Suppose we take independent normal readings of the same mean, each with the same variance. Why does their average have the smallest variance among unbiased estimators?

- The likelihood score measures how the data respond to a change in the mean. The answer depends on how closely the estimator aligns with that score.

Regularization. Suppose nearby regression coefficients should be similar. How strongly should we penalize their differences?

- We compare the order of two matrix operations. The resulting score gives a candidate penalty strength that we can evaluate on held-out data.

The likelihood score

Let $f_\theta(x)$ be the density or probability mass of the observed data. The *score* is the slope of the log likelihood with the observation held fixed,

$$S_\theta(x) = \partial_\theta \log f_\theta(x).$$

A positive score means that a small increase in θ raises the log likelihood of this observation.

A regular model

Use a common reference measure, fixed support, and a differentiable density. Assume we can differentiate the expectations used below under the integral. Then $\mathbb{E}_\theta S_\theta = \partial_\theta \int f_\theta \, d\mu = 0$.

The *Fisher information* is the mean squared score, $I(\theta) = \mathbb{E}_\theta S_\theta^2$. We assume $0 < I(\theta) < \infty$.

How an expectation changes with the parameter

Inside $\mathbb{E}_\theta[g] = \int g(x,\theta) f_\theta(x)\, d\mu(x)$, changing θ can change both the quantity we average and the probabilities used to average it.

$$\frac{d}{d\theta}\mathbb{E}_\theta[g] = \underbrace{\mathbb{E}_\theta[\partial_\theta g]}_{\text{change in the quantity}} + \underbrace{\mathbb{E}_\theta[gS_\theta]}_{\text{change in the weights}},$$

by the product rule of differentiation, since for $f_\theta(x) > 0$, $S_\theta(x) = \partial_\theta \log f_\theta(x) = \frac{\partial_\theta f_\theta(x)}{f_\theta(x)}$, multiplying both sides by $f_\theta(x)$ gives $\partial_\theta f_\theta(x) = f_\theta(x) S_\theta(x)$.

The score-transport operator, Λ_θ

The product rule gives $\partial_\theta(g f_\theta) = f_\theta(\partial_\theta g + g S_\theta)$. Writing $\Lambda_\theta g = \partial_\theta g + g S_\theta$ makes the identity $\frac{d}{d\theta}\mathbb{E}_\theta[g] = \mathbb{E}_\theta[\Lambda_\theta g]$.

A sufficient differentiation condition is given in the backup slides.

What does the transport commutator measure?

Multiplication by a quantity T is the operator $\mathsf{M}_T g = Tg$. Compare multiplication with score transport in the two orders.

$$\begin{aligned} \Lambda_\theta(Tg) &= (\partial_\theta T)g + T\partial_\theta g + TgS_\theta, \\ T\Lambda_\theta g &= \qquad\qquad T\partial_\theta g + TgS_\theta. \end{aligned}$$

Subtracting leaves the explicit dependence on θ

$$[\Lambda_\theta, \mathsf{M}_T]g = (\partial_\theta T)g.$$

For an ordinary statistic $T(X)$, the difference is zero. Its mean can still change because $\frac{d}{d\theta}\mathbb{E}_\theta T = \mathrm{Cov}_\theta(T, S_\theta)$.

A sample mean and its centered error

Take n independent $N(\theta, \sigma^2)$ readings with known σ^2. Their joint score is $S_\theta = n(\bar{X} - \theta)/\sigma^2$.

	Sample mean $\bar{X}$	Centered error $\bar{X} - \theta$
Mean $\mathbb{E}_\theta T$	θ	0
Direct change $\mathbb{E}_\theta[\partial_\theta T]$	0	-1
Weight change $\mathbb{E}_\theta[T S_\theta]$	1	1
Total change $\frac{d}{d\theta}\mathbb{E}_\theta T$	$0 + 1 = 1$	$-1 + 1 = 0$

The centered error is a *pivot* here. Its distribution is $N(0, \sigma^2/n)$ for every θ. Its explicit derivative cancels the change in the weights.

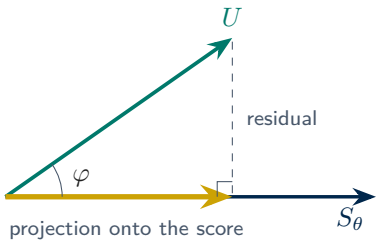
Cramér–Rao as a projection

View centered random variables as vectors. Their inner product is $\langle U, V \rangle = \mathbb{E}_\theta[UV]$. Set $U = T - \mathbb{E}_\theta T$.

$$\|U\|^2 = \mathrm{Var}_\theta(T), \qquad \|S_\theta\|^2 = I(\theta).$$

Cauchy–Schwarz gives

$$\mathrm{Cov}_\theta(T, S_\theta)^2 \le \mathrm{Var}_\theta(T) I(\theta).$$



Equality holds exactly when U is a multiple of the score, almost surely.

The angle φ measures alignment with the score. The perpendicular residual adds variance and has zero covariance with the score.

Why averaging all the readings is efficient

For an unbiased estimator of θ in a regular model, $\mathrm{Cov}_\theta(T, S_\theta) = \frac{d}{d\theta}\mathbb{E}_\theta T = 1$. Thus

$$\mathrm{Var}_\theta(T) \geq \frac{1}{I(\theta)}, \qquad \mathrm{Eff}_\theta(T) = \frac{1/I(\theta)}{\mathrm{Var}_\theta(T)} = \cos^2\varphi.$$

Normal mean estimator	Information in the sample	Variance	Efficiency
All readings, $T = \bar{X}$	n/σ^2	σ^2/n	1
One reading, $T = X_1$	n/σ^2	σ^2	$1/n$

Since $\bar{X} - \theta = (\sigma^2/n)S_\theta$, the mean lies exactly along the score. Both estimators have zero transport commutator.

Bias changes the variance bound

For a statistic $T(X)$ with mean $\psi(\theta)$ and bias $b(\theta) = \psi(\theta) - \theta$,

$$\mathrm{Var}_\theta(T) \geq \frac{\psi'(\theta)^2}{I(\theta)} = \frac{(1 + b'(\theta))^2}{I(\theta)}.$$

Shrinking the normal sample mean

Choose a fixed $0 < c < 1$ and report $T = c\bar{X}$. Its mean changes at rate c. It attains the biased variance bound, yet its mean squared error includes bias,

$$\mathrm{MSE}_\theta(c\bar{X}) = \underbrace{c^2\sigma^2/n}_{\text{variance}} + \underbrace{(1-c)^2\theta^2}_{\text{squared bias}}.$$

Shrinkage reduces error near $\theta = 0$. Farther from zero, the bias can outweigh the reduction in variance.

A moving endpoint changes the calculation

If $X \sim \mathrm{Uniform}(0, \theta)$, increasing θ adds possible observations. For a continuous, parameter-independent g,

$$\frac{d}{d\theta}\mathbb{E}_\theta[g] = \frac{d}{d\theta}\frac{1}{\theta}\int_0^\theta g(x)\,dx = -\frac{1}{\theta}\mathbb{E}_\theta[g(X)] + \underbrace{\frac{g(\theta)}{\theta}}_{\text{moving endpoint}}.$$

For $g(x) = x$, the two contributions are $-1/2$ and 1, giving $\frac{d}{d\theta}\mathbb{E}_\theta X = 1/2$.

With n independent observations, the maximum estimates the endpoint. Its unbiased correction $\widetilde{\theta} = (n+1)\max_i X_i/n$ has variance $\theta^2/[n(n+2)]$. The moving support explains why the regular Cramér–Rao argument does not apply.

A penalty for neighboring coefficients

Suppose columns of X are measurements at successive wavelengths. If neighboring wavelengths should have similar effects on a response y, fit

$$\hat{\beta}_\lambda = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{\|y - X\beta\|^2}{2n} + \frac{\lambda}{2} \sum_{j=1}^{p-1} (\beta_{j+1} - \beta_j)^2 \right\}.$$

The first-difference matrix D gives $(D\beta)_j = \beta_{j+1} - \beta_j$. Then the sum of squared differences is $\beta^\top P \beta$ with $P = D^\top D$. Larger λ favors smoother coefficients.

Two matrices determine the fit

$A = X^\top X/n$ measures how the loss curves in coefficient space. P measures roughness. Together they give the penalized curvature $H_\lambda = A + \lambda P$ and the equation $H_\lambda \hat{\beta}_\lambda = X^\top y/n$. *Data Curvature* A: directions the data can see. *Penalty* P: directions the prior distrusts. A & P rarely share an eigenbasis.

Why the two influence matrices always commute

Assume A and P are symmetric positive semidefinite ($A, P \succeq 0$, i.e., $v^\top Av, v^\top Pv \geq 0$ for every v). Assume $H_\lambda = A + \lambda P$ is invertible at the candidate λ. Define

$$D_\lambda = H_\lambda^{-1} A, \qquad R_\lambda = H_\lambda^{-1}(\lambda P).$$

D_λ describes how the fit retains coefficient directions, while R_λ gives the complementary shrinkage. Their sum is $H_\lambda^{-1} H_\lambda = I_p$.

A commutator that gives the same answer at every penalty

$$[D_\lambda, R_\lambda] = D_\lambda(I_p - D_\lambda) - (I_p - D_\lambda)D_\lambda = 0.$$

Any comparison based only on this commutator ties every candidate λ.

Parameter Estimation: 2 commutators for tuning

Pair a curvature or penalty matrix with an operation that includes a solve with H_λ.

$$\text{Comm Type I} \qquad C_{\mathrm{I}}(\lambda) = [A, H_\lambda^{-1} P] = A H_\lambda^{-1} P - H_\lambda^{-1} P A,$$

$$\text{Comm Type II} \qquad C_{\mathrm{II}}(\lambda) = [P, H_\lambda^{-1} A] = P H_\lambda^{-1} A - H_\lambda^{-1} A P.$$

For example, apply Type I to a coefficient vector v. The first term applies $H_\lambda^{-1} P$ to v and then A. The second reverses those two operations.

Measuring the difference. The squared Frobenius norm $\|C\|_{\mathrm{F}}^2 = \sum_{i,j} C_{ij}^2$ summarizes the differences over all coordinate directions.

Spherical ridge gives a negative control

When $AP = PA$, the symmetric matrices A and P have a common orthonormal eigenbasis. In direction i, with eigenvalues a_i, p_i, the fitted fraction is

$$\frac{a_i}{a_i + \lambda p_i}.$$

Both compositions act by scalar multiplication along each direction, so $C_{\mathrm{I}}(\lambda) = C_{\mathrm{II}}(\lambda) = 0$ wherever H_λ is invertible.

What happens with $P = I_p$?

Ridge can improve prediction even though both commutator scores are zero for every λ. A flat score gives no preference among the candidates.

How many directions remain in the fit?

The effective degrees of freedom measure how strongly the fitted values respond to the observed responses. For penalized least squares,

$$\hat{y} = \underbrace{X H_\lambda^{-1} X^\top / n}_{\text{smoother matrix}} y, \qquad \mathrm{df}(\lambda) = \mathrm{tr}(H_\lambda^{-1} A).$$

Penalty P	Directions it leaves unpenalized	Nullity r
Ridge I_p	Only the zero vector	0
First differences $D^\top D$	Constant coefficient vectors	1
Graph Laplacian	Constant on each graph component	# components

For fixed $A \succ 0$, $\mathrm{df}(\lambda)$ decreases from p toward the *kernel nullity*, $r = \dim \ker P$. Strong smoothing preserves the unpenalized directions.

The balanced commutator score, $S_C(\lambda)$

For fixed positive definite A, use the effective dimension, $B_r(\lambda)$, to weight the commutator magnitude. For either commutator C,

$$B_r(\lambda) = (\mathrm{df}(\lambda) - r)(p - \mathrm{df}(\lambda)), \qquad S_C(\lambda) = B_r(\lambda)\, \|C(\lambda)\|_{\mathrm{F}}^2.$$

Interpreting the weight, $B_r(\lambda)$

If $r < p$, let $q = (\mathrm{df} - r)/(p - r)$ be the fraction of adjustable degrees of freedom retained. Then $B_r = (p - r)^2 q(1 - q)$. The weight is largest at $q = 1/2$ and vanishes at $q = 0$ and $q = 1$.

Select a maximizer on the candidate grid. The commutator also changes with λ, so its weighted score can peak away from $q = 1/2$. A finite grid can still have a boundary maximum.

The 2 Commutators can select different amounts of smoothing

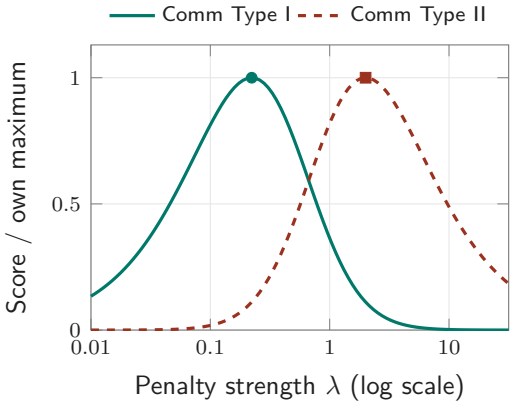
Consider two neighboring coefficients, with

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \qquad P = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

The second coefficient has twice the data curvature of the first. The penalty charges for their difference.

$$S_{\mathrm{I}}(\lambda) = \frac{30\lambda}{(2+3\lambda)^4}, \quad S_{\mathrm{II}}(\lambda) = \frac{24\lambda^3}{(2+3\lambda)^4}.$$

Comm Type I peaks at $\lambda = 2/9$ & Comm Type II peaks at $\lambda = 2$, 9 times larger.



Both curves use only A and P. Choosing between their peaks for prediction requires information about the response.

Computing a candidate penalty strength

1. Choose a structured penalty P, its nullity r, an ordering, and a grid of positive λ values spaced on a log scale. Fix the scaling of P across candidates.
2. For each candidate, form $H_\lambda = A + \lambda P$. Solve linear systems to obtain $H_\lambda^{-1} A$ and, for Type I, $H_\lambda^{-1} P$. Compute df, B_r and S_C.
3. Select a grid value with the largest score. If the score is flat within numerical tolerance, record that the geometry supplies no preference. The paper's simulations used $\lambda = 0$ as the fallback.

For a boundary maximum, inspect the grid range and conditioning. For fixed least-squares X, P, changing y leaves the geometric score unchanged. This is because for fixed design matrix X, penalty matrix P, and penalty parameter λ, both A and $H_\lambda = A + \lambda P$ are completely independent of the response vector y, and hence, the linear operators $H_\lambda^{-1} A$ and $H_\lambda^{-1} P$ are fixed given (X, P, λ).

Logistic regression changes the curvature

In a generalized linear model (GLM), the curvature can change with the fitted coefficients. For binary logistic regression, fit each candidate λ and set

$$\hat{p}_i = \Pr_{\hat{\beta}_\lambda}(Y_i = 1 \mid x_i), \quad W_\lambda = \operatorname{diag}(\hat{p}_i(1-\hat{p}_i)), \quad A_\lambda = X^\top W_\lambda X/n.$$

Why each candidate needs a fit

Changing λ changes the predicted probabilities, then the weights, then A_λ. Evaluate the score with $H_\lambda = A_\lambda + \lambda P$ at a converged fit. The trace and balance now describe local curvature.

A scan uses one fit to the full training set per candidate. Five-fold cross-validation (CV) fits each candidate on five overlapping training subsets and scores the corresponding validation folds.

What Type II sees in a penalty null space

For first differences, a common shift of all coefficients costs zero. More generally, if $v \in \ker P$, then $H_\lambda v = Av$ and $H_\lambda^{-1} Av = v$. Consequently,

$$C_{\mathrm{II}}(\lambda)v = P(H_\lambda^{-1} Av) - H_\lambda^{-1} A(Pv) = 0.$$

A direction can still appear in the output

In the two-coefficient example, Type II sends a contrast to a common shift,

$$C_{\mathrm{II}}(\lambda)\begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{2\lambda}{2+3\lambda}\begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

This asymmetry limits what the Type II score diagnoses. Its effect on prediction depends on the data and the signal.

Experimental evaluation of the tuning rules

Each model used 100 simulated data sets per design, each with 200 observations and 50 features. Each split used 140 observations for training and 60 for testing.

Design	Feature dependence	Penalty
Toeplitz	Correlation decays with distance	First differences
Block	Correlated blocks	Block graph Laplacian
Grouped	Shared factors within groups	Within-group graph Laplacian
Sparse	Independent features	Spherical ridge (negative control)

The candidate grid included 0 and log-spaced values from e^{-6} to e^{4}. Five-fold CV minimized validation MSE or logistic log loss. The *oracle* used the best test MSE or test AUC on that grid, a retrospective benchmark.

Linear regression performance depends on the design

Test mean squared error (MSE) averages squared prediction errors. *Lower is better.* Entries give the median across replications, with the median absolute deviation in parentheses.

Design	Unregularized	Oracle	Five-fold CV	Type I	Type II
Block	1.521 (0.190)	1.431 (0.173)	1.470 (0.176)	1.745 (0.250)	6.593 (0.771)
Grouped	1.568 (0.175)	1.389 (0.153)	1.421 (0.159)	1.425 (0.158)	1.584 (0.223)
Sparse	1.545 (0.214)	1.513 (0.226)	1.532 (0.222)	1.545 (0.214)	1.545 (0.214)
Toeplitz	1.500 (0.206)	1.394 (0.156)	1.429 (0.192)	1.419 (0.173)	2.355 (0.357)

Type I had slightly lower median error than CV in Toeplitz and similar median error in Grouped. In Block, Type II reached a median MSE of 6.593, versus 1.470 for CV.

Dinov, Ho & Ahmed (2026), Table 1.

In *logistic regression*, Type II comes closer to CV

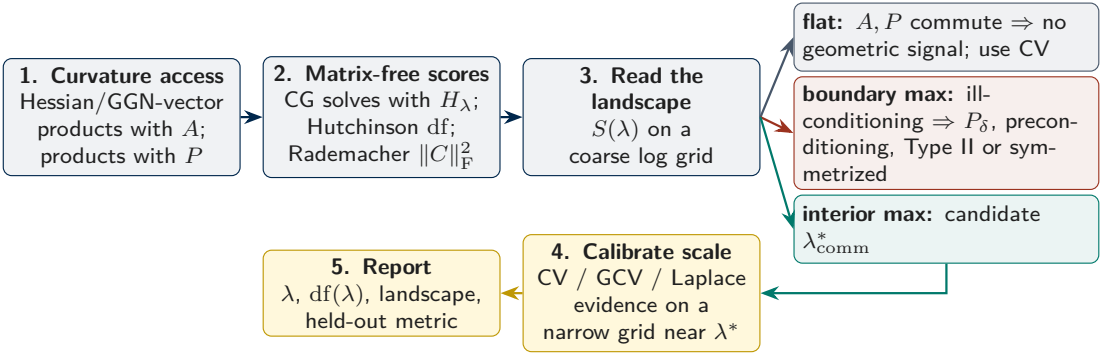
The area under the receiver operating characteristic curve (AUC) measures how often a positive test case ranks above a negative case, counting ties as one-half. *Higher is better*.

Design	Unregularized	Oracle	Five-fold CV	Type I	Type II
Block	0.720 (0.039)	0.888 (0.029)	0.878 (0.030)	0.834 (0.033)	0.875 (0.032)
Grouped	0.710 (0.043)	0.880 (0.030)	0.873 (0.034)	0.834 (0.036)	0.869 (0.032)
Sparse	0.718 (0.055)	0.815 (0.042)	0.806 (0.041)	0.718 (0.055)	0.718 (0.055)
Toeplitz	0.735 (0.045)	0.881 (0.026)	0.870 (0.027)	0.812 (0.040)	0.856 (0.031)

CV had the highest median AUC among feasible selectors in all four designs. In Sparse, the commutators tied at every λ, so both returned the unregularized fit. CV improved AUC in that same design.

Dinov, Ho & Ahmed (2026), Table 3. Entries are median (median absolute deviation).

A scalable workflow for AI pipelines



♦ Cost

Per candidate λ: m probes $\times$ (CG iterations) curvature-vector products, with no fold-specific refits.

Data & Prior Duality

The commutator locates *geometric* tension between data and prior; a data-driven criterion sets the *scale*.

GGN=Generalized Gauss–Newton matrix approximates the Hessian; CG=Conjugate Gradient iterative linear solver; GCV=Generalized Cross-Validation.

Other commutator applications using a structured penalty

Imaging inverse problems. If K maps an image β to measurements, a quadratic reconstruction uses $A = K^\top K/n$ and $P = D^\top D$ for neighboring pixel differences. The commutators compare the directions the scanner measures with those the smoothness penalty suppresses.

A prediction layer on fixed features. With a feature matrix Φ, a squared-error fit gives $A = \Phi^\top \Phi/n$. If the feature coordinates have a known graph, its Laplacian can penalize differences between linked coefficients. Ordinary ridge gives a flat score.

How to test either application. Select λ on training data, then compare reconstruction or prediction error with CV on independent cases. The paper's empirical comparisons cover simulated linear and logistic regression.

What the score tells us

- **Estimation.** $[\Lambda_\theta, \mathsf{M}_T] = \mathsf{M}_{\partial_\theta T}$ detects explicit parameter dependence. For an ordinary statistic, score covariance determines how its mean changes.
- **Efficiency.** In a regular scalar model, the centered sample mean of normal readings lies along the score. This is why it attains the variance bound.
- **Tuning.** A structured penalty (P) and data curvature (A) can favor different directions. The balanced score supplies a candidate λ **without fitting CV folds**. A flat score supplies no candidate preference.
- **Open question.** Which features of the signal and noise predict when either commutator ordering comes close to the best predictive risk? The fixed-design least-squares score uses neither the outcomes nor the noise level.

Reading and reproducible examples

Paper

I. D. Dinov, E. T. W. Ho and S. E. Ahmed (2026). *Statistical commutators with applications to estimation, inference, and regularization tuning*. Statistics and Computing, 36, article 228. doi:10.1007/s11222-026-10991-w

Worked simulations

SOCR DSPA2, Appendix 25. Regularization tuning simulations (R notebook). The notebook includes the selectors, the four simulation designs, and the individual steps used to construct the summary tables.

Background

E. L. Lehmann and G. Casella (1998), *Theory of Point Estimation*, 2nd ed.
A. W. van der Vaart (1998), *Asymptotic Statistics*.

Acknowledgments

Collaborators and colleagues: Eric Ho, S. Ejaz Ahmed, Yueyang Shen, Bojko Bakalov, Milen Velev.

Funding Support NIH UL1 TR002240 and T32 GM141746. NSF 2611649, 1916425 and 1734853.

URL https://socr.umich.edu/ | https://DSPA3.predictive.space/



Appendix

Work locally at an interior θ with common support and reference measure μ. Suppose f_t and $g(\cdot, t)$ are differentiable in t almost everywhere, $\mathbb{E}_t|g| < \infty$ nearby, and, throughout a neighborhood,

$$|\partial_t(g(x,t)f_t(x))| \le G(x), \qquad \int G\, d\mu < \infty.$$

Assume differentiation also holds for $g = 1$. For $g = T$, take finite second moments of T, S_θ and assume $\mathbb{E}_\theta|\partial_\theta T| < \infty$. Then, for $\psi(\theta) = \mathbb{E}_\theta T(X, \theta)$,

$$\mathrm{Var}_\theta(T) I(\theta) \ge \mathrm{Cov}_\theta(T, S_\theta)^2 = \left(\psi'(\theta) - \mathbb{E}_\theta[\partial_\theta T]\right)^2.$$

For $T(X)$ with positive variance, the bound divided by the actual variance is $\psi'(\theta)^2/(I(\theta)\,\mathrm{Var}_\theta\, T) = \cos^2\varphi$.

Exponential families and exact equality

For a regular one-parameter exponential family, write $f_\theta(x) = h(x)\exp\{\eta(\theta)T_s(x) - a(\theta)\}$. Centering the score gives

$$S_\theta = \eta'(\theta)\{T_s - \mathbb{E}_\theta T_s\}.$$

When $I(\theta) > 0$, any statistic affine in T_s satisfies the equality condition for its mean function.

Gamma scale. For independent gamma observations with known shape k and scale θ, $T = \bar{X}/k$ is unbiased for θ and $\mathrm{Var}(T) = \theta^2/(nk) = 1/I(\theta)$.

Fixed truncation. For $N(\theta, 1)$ conditioned on $X \geq 0$, $S_\theta = X - \mathbb{E}_\theta X$. Thus $\partial_\theta \mathbb{E}_\theta X = \mathrm{Var}_\theta\, X = I(\theta)$. At $\theta = 0$ this common value is $1 - 2/\pi$, even though the support stays fixed.

Maximum likelihood and asymptotic score alignment

Let S_i be the score for observation i and $I_1 = \mathbb{E}_\theta S_i^2 > 0$. One sufficient condition for the finite-sample squared cosine to approach one is the mean-square expansion

$$\hat{\theta}_n - \theta = \frac{1}{nI_1} \sum_{i=1}^{n} S_i + R_n, \qquad n\mathbb{E}_\theta[R_n^2] \longrightarrow 0.$$

The leading term lies along the joint score. The remainder has negligible variance on the $1/n$ scale. Convergence in distribution alone gives no such second-moment control.

An exact example

For independent exponential observations with rate θ, the maximum likelihood estimate (MLE) is $1/\bar{X}$. For $n > 2$, its mean is $n\theta/(n-1)$ and its squared score correlation is $(n-2)/n$. This is 0.80 at $n = 10$ and 0.98 at $n = 100$.

Let $S = \nabla_\theta \log f_\theta(X) \in \mathbb{R}^d$, $\mathcal{I} = \mathbb{E}_\theta[SS^\top] \succ 0$, and $T(X) \in \mathbb{R}^m$ be square integrable. Under componentwise regularity, its mean ψ has Jacobian $J_{ij} = \partial_{\theta_j}\psi_i = \mathrm{Cov}(T_i, S_j)$.

$$\mathrm{Cov}(T) \succeq J\mathcal{I}^{-1}J^\top.$$

The difference is the covariance matrix of $T - \psi - J\mathcal{I}^{-1}S$. Every linear combination therefore has nonnegative residual variance.

Normal mean and variance, both unknown

For $n > 1$, let $T = (\bar{X}, s^2)$, where s^2 uses denominator $n-1$. Both coordinates are unbiased, so $J = I_2$. Their covariance is $\mathrm{diag}(\sigma^2/n, 2\sigma^4/(n-1))$, while the bound is $\mathrm{diag}(\sigma^2/n, 2\sigma^4/n)$.

The uniform maximum and its boundary term

For $M_n = \max_i X_i$ from n independent $\mathrm{Uniform}(0, \theta)$ observations, $\Pr(M_n \le m) = (m/\theta)^n$ for $0 \le m \le \theta$. Integrating the resulting density gives

$$\mathbb{E}_\theta M_n = \frac{n\theta}{n+1}, \qquad \mathrm{Var}_\theta(M_n) = \frac{n\theta^2}{(n+1)^2(n+2)}.$$

Inside the sample cube, the derivative of the joint log density is $-n/\theta$. Differentiating the integral for the mean also includes a contribution from each of the n faces moving outward,

$$\frac{d}{d\theta}\mathbb{E}_\theta M_n = \underbrace{-\frac{n^2}{n+1}}_{\text{interior}} + \underbrace{n\,\theta\,\theta^{-n}\theta^{n-1}}_{\text{moving faces}} = \frac{n}{n+1}.$$

On each such face the maximum is θ and the face area is θ^{n-1}.

Classical covariance and quantum uncertainty

For self-adjoint observables A, B with the required products and moments in a common quantum state, write $A_0 = A - \langle A \rangle I$ and $B_0 = B - \langle B \rangle I$. The Robertson–Schrödinger bound is

$$\operatorname{Var}(A) \operatorname{Var}(B) \geq \left| \tfrac{1}{2} \langle A_0 B_0 + B_0 A_0 \rangle \right|^2 + \left| \tfrac{1}{2i} \langle [A, B] \rangle \right|^2 .$$

Real classical observables multiply functions pointwise, hence $[\mathsf{M}_U, \mathsf{M}_V] = 0$. The surviving term gives $\operatorname{Var}(U) \operatorname{Var}(V) \geq \operatorname{Cov}(U, V)^2$. Choosing $U = T$ and $V = S_\theta$ recovers the score-geometric bound in a regular model.

For the Fisher-information connection, see S. L. Braunstein and C. M. Caves (1994), *Physical Review Letters* 72, 3439–3443.

Effective dimension when A may be singular

For fixed $A, P \succeq 0$, assume $G = A + P \succ 0$, equivalently $\ker A \cap \ker P = \{0\}$. Set $K = G^{-1/2} A G^{-1/2}$, so $0 \preceq K \preceq I_p$. If d_i are the eigenvalues of K, cyclic invariance of the trace gives

$$\mathrm{df}(\lambda) = \mathrm{tr}\Big((K + \lambda(I_p - K))^{-1} K\Big) = \sum_i \frac{d_i}{d_i + \lambda(1 - d_i)}, \qquad \lambda > 0.$$

Each term is nonincreasing. At $\lambda \downarrow 0$ it tends to 1 when $d_i > 0$. At $\lambda \to \infty$ only the $d_i = 1$ terms survive. Thus the limiting dimensions are $\mathrm{rank}\, A$ and $\dim \ker P$, respectively.

If A is singular, the paper's $p - \mathrm{df}$ factor can stay positive at the weak-penalty end. The full-rank endpoint argument then needs revision.

Exact structure of the Type II commutator

Let $M = H_\lambda^{-1} A$. In an orthonormal basis for $\mathrm{range}(P) \oplus \ker(P)$, M fixes every null-space vector. Hence

$$P = \begin{pmatrix} P_R & 0 \\ 0 & 0 \end{pmatrix}, \qquad M = \begin{pmatrix} M_{RR} & 0 \\ M_{0R} & I_0 \end{pmatrix}, \qquad [P, M] = \begin{pmatrix} [P_R, M_{RR}] & 0 \\ -M_{0R} P_R & 0 \end{pmatrix}.$$

The right column is zero. The lower-left block carries output from a penalized direction into the null space whenever $M_{0R} \neq 0$.

For Type I and the projector Π_0 onto $\ker P$, $C_{\mathrm{I}} \Pi_0 = -H_\lambda^{-1} P A \Pi_0$ can be nonzero. The two orderings therefore respond differently to unpenalized inputs.

Dinov, Ho & Ahmed (2026), discussion of rank-deficient penalties.

Why the two example peaks are 9-fold apart

For the two-coefficient example, let $d = 2 + 3\lambda$. Direct multiplication gives

$$H_\lambda^{-1}P = \frac{1}{d}\begin{pmatrix} 2 & -2 \\ -1 & 1 \end{pmatrix}, \quad C_{\mathrm{I}} = \frac{1}{d}\begin{pmatrix} 0 & 2 \\ -1 & 0 \end{pmatrix}, \quad C_{\mathrm{II}} = \frac{\lambda}{d}\begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}.$$

Since $H_\lambda^{-1}A = I_2 - \lambda H_\lambda^{-1}P$, $\mathrm{df} = 1 + 2/d$ and $B_1 = 6\lambda/d^2$. Thus

$$S_{\mathrm{I}} = 30\lambda/d^4, \qquad S_{\mathrm{II}} = 24\lambda^3/d^4.$$

For $k = 1, 3$, the logarithmic derivative of λ^k/d^4 is $k/\lambda - 12/d$. Its sign switches from positive to negative at $\lambda = 2k/[3(4-k)]$. This gives $2/9$ and 2, respectively. The large-λ tails in this example decay as λ^{-3} and λ^{-1}.

Vanishing commutators

Commuting symmetric A, P give zero for both orderings. For $\lambda > 0$, there are also converses under simple invertibility conditions.

$$C_{\mathrm{I}} = -\lambda^{-1}[A, H_\lambda^{-1}A] = -\lambda^{-1}[A, H_\lambda^{-1}]A,$$
$$C_{\mathrm{II}} = -\lambda[P, H_\lambda^{-1}P] = -\lambda[P, H_\lambda^{-1}]P.$$

If $A \succ 0$, then $C_{\mathrm{I}} = 0$ implies $[A, H_\lambda^{-1}] = 0$, hence $[A, H_\lambda] = \lambda[A, P] = 0$. If $P \succ 0$, the second line similarly gives $C_{\mathrm{II}} = 0 \Rightarrow [P, A] = 0$.

A nearly flat *sampled* score curve can also reflect the grid range, scaling, or numerical precision. Checking $[A, P]$ directly is possible when the matrices are available.

Computing scores without storing dense matrices

For large p, draw independent vectors z_j with independent entries equal to -1 or 1 with equal probability. Since $\mathbb{E}[zz^\top] = I_p$,

$$\operatorname{tr}(H_\lambda^{-1} A) = \mathbb{E}[z^\top H_\lambda^{-1} A z], \qquad \|C\|_{\mathrm{F}}^2 = \mathbb{E}[\|Cz\|^2].$$

Average each expression over the probes. For example, $C_1 z = A H_\lambda^{-1} P z - H_\lambda^{-1} P A z$ requires matrix–vector products and two linear solves. Conjugate gradients applies when $H_\lambda \succ 0$.

With exact solves these are unbiased estimates of the trace and squared norm separately. Plugging them into the nonlinear balanced score can introduce bias. At a GLM candidate, first obtain the local curvature from its fit.

M. F. Hutchinson (1989), *Communications in Statistics: Simulation and Computation* 18, 1059–1076.

Possible stabilizations for structured penalties

- **Weak ridge component.** $P + \delta I_p$ with $\delta > 0$ gives a positive cost to all directions. This changes the model, including the formerly unpenalized directions.
- **Combining orderings.** Compare both curves, or explore a weighted combination of their squared norms after specifying the relative scaling. The balance factor and the null-space fit can be inspected separately.
- **Calibration.** Compare a geometric choice with validation error. In least squares, the optimal trade-off depends on the signal and noise even when A and P are fixed.

Testing these proposals requires evaluation on independent outcomes. The simulation tables compare the original Type I and Type II rules.

Performance relative to the test-set oracle

Each entry is the median of 100 *within-replication* ratios to the oracle. A ratio of one means the method matched that split's best test performance on the grid.

	Linear MSE / oracle MSE			Logistic AUC / oracle AUC		
Design	CV	Type I	Type II	CV	Type I	Type II
Block	1.008	1.158	4.492	0.993	0.955	0.992
Grouped	1.015	1.024	1.140	0.996	0.952	0.988
Sparse	1.017	1.008	1.008	0.987	0.899	0.899
Toeplitz	1.018	1.012	1.650	0.990	0.934	0.974

For MSE, a larger ratio is worse. For AUC, a smaller ratio is worse. In Block, Type II's linear error was typically about 4.5 times the oracle's, while its logistic discrimination was close to the oracle's.

Dinov, Ho & Ahmed (2026), Tables 2 and 4.

Paired comparisons with cross-validation

Each test uses 100 paired simulation runs. For linear models the difference is $\log \mathrm{MSE}_{\mathrm{comm}} - \log \mathrm{MSE}_{\mathrm{CV}}$ (*positive favors CV*). For logistic models it is $\mathrm{AUC}_{\mathrm{comm}} - \mathrm{AUC}_{\mathrm{CV}}$ (*positive favors the commutator*).

	Linear, paired t		Logistic, paired t	
Design	Type I	Type II	Type I	Type II
Block	10.887	56.552	-10.003	-0.252
Grouped	1.165	10.528	-11.391	-4.878
Sparse	0.436	0.436	-14.618	-14.618
Toeplitz	-4.786	29.711	-12.041	-6.047

For logistic Block, Type II gave $p = 0.8016$. For linear Grouped, Type I gave $p = 0.2468$. These paired tests assess a zero mean difference. Establishing practical equivalence would require a prespecified margin.

Dinov, Ho & Ahmed (2026), Table 5. Nominal two-sided tests, 99 degrees of freedom.