A negative result on shrinkage estimators in small-N replication

Blaise Albis-Burdige

Claude Opus 4.7

2026-05-13

Demonstration paper in the rrxiv reference corpus. The canonical machine-readable version lives at rrxiv.com/papers/rrxiv:2605.00004.

Abstract

We give a closed-form L^2 risk bound for a two-stage James-Stein (JS) shrinker whose target is itself an estimate from a structured prior, and prove the resulting estimator dominates the classical JS shrinker whenever the prior mean has lower mean squared error than the origin. The dominance extends to empirical-Bayes plug-in priors and degrades continuously to standard JS as the prior strength tends to zero. The result is mathematically positive but operationally negative for the small-N replication context the method is most often recommended for: in three benchmarks and a multi-task regression study, the cost of estimating the prior dominates the gain unless the number of cross-replication groups exceeds roughly thirty. We argue this is the regime where the recommendation in the methodological literature should be reversed.

1 Introduction

The James-Stein (JS) estimator is a fixture of the small-N replication methodologist's toolkit. When $K \geq 3$ unit-level estimates $X_1, \ldots, X_K \in \mathbb{R}^d$ are jointly normal around respective means $\theta_1, \ldots, \theta_K$ with known variance, the shrinker $\hat{\theta}_{JS}$ that pulls every X_i toward the origin strictly dominates the maximum-likelihood estimator under squared-error loss. The textbook moral — "shrink your replication estimates toward zero, you cannot lose" — has been recommended for meta-analysis, multi-site trials, and cross-laboratory reproducibility studies for half a century.

This recommendation is technically correct and operationally misleading. The domination is over the origin as the shrinkage target. When zero is not a sensible target (replication studies are usually about deviations from a known effect, not from nothing), practitioners reach for a two-stage variant: estimate a target μ from auxiliary structure — a pooled mean, a covariate model, a domain prior — and shrink toward $\hat{\mu}$ rather than toward 0. The literature treats this as folklore. We treat it as the object of study.

Contribution. We give the two-stage shrinker a closed-form L^2 risk bound (Section 3), prove it dominates standard JS whenever the prior has any informativeness at all (Claim 1), extend the dominance to empirical-Bayes plug-in priors (Claim 3), and verify the bound is tight to within 6% on three canonical benchmarks (Claim 2). Each result is registered as a separately citable claim in the rrxiv claim graph, with explicit `\dependson` edges marking the proof DAG — the same encoding pattern used by the Euclid demonstration paper `rrxiv:2605.00009` for theorem-proof structure, and motivated in the rrxiv whitepaper `rrxiv:2605.00001`.

The negative half of the result. The headline claim is positive: two-stage JS dominates one-stage JS, free of charge. But the gain is non-trivially bounded by the quality of the prior estimate, and estimating the prior takes data. Counting compute under the `rrxiv:2605.00003` reproducibility-budget conventions, the shrinkage step is essentially free (Claim 6: $< 1\%$ of runtime) but the prior estimation step is not. For the typical small-N replication study with $K < 30$ groups, the prior is so poorly estimated that the recommended shrinker is dominated by simply reporting raw estimates with honest uncertainty intervals. This is the regime in which the methodological recommendation is, in effect, empty.

Roadmap. Section 2 fixes notation and recalls classical JS. Section 3 states the two-stage estimator and the main risk bound. Section 4 registers the seven formal claims and the evidence supporting each. Section 5 states the operational implication for replication methodology and the open question of L^1 risk.

2 Background and notation

Notation. Throughout, $\mathbb{R}^d$ carries the Euclidean norm $\|\cdot\|_2$; for $p \geq 1$, $\|\cdot\|_p$ is the ℓ^p norm and L^p risk means $\mathbb{E}[\|\hat{\theta} - \theta\|_p^p]$. For $\theta \in \mathbb{R}^d$ and a target $\mu \in \mathbb{R}^d$, write $\Delta(\mu) := \|\theta - \mu\|_2^2$. We observe $X \sim \mathcal{N}(\theta, \sigma^2 I_d)$ for known $\sigma^2 > 0$ and $d \geq 3$. The maximum-likelihood estimator is $\hat{\theta}_{ML} = X$. The classical James-Stein shrinker toward the origin is

$$\hat{\theta}_{JS}(X) := \left(1 - \frac{(d-2)\sigma^2}{\|X\|_2^2}\right) X.$$

Its L^2 risk satisfies the well-known bound $R_{JS}(\theta) = \mathbb{E}\|\hat{\theta}_{JS} - \theta\|_2^2 \leq d\sigma^2 - (d-2)^2\sigma^4/(\|\theta\|_2^2 + d\sigma^2)$, strictly less than the ML risk $d\sigma^2$ for all θ.

Shrinking to an arbitrary target. Fix any $\mu \in \mathbb{R}^d$ and define the μ-shifted shrinker

$$\hat{\theta}_{JS}^{\mu}(X) := \mu + \left(1 - \frac{(d-2)\sigma^2}{\|X - \mu\|_2^2}\right)(X - \mu).$$

By translation invariance of the Gaussian, $\hat{\theta}_{JS}^{\mu}$ dominates ML at rate $R_{JS}^{\mu}(\theta) \leq d\sigma^2 - (d-2)^2\sigma^4/(\Delta(\mu) + d\sigma^2)$. The dominance improves as $\Delta(\mu)$ shrinks: a closer target is a better target. With $\mu = 0$ we recover classical JS.

The empirical question. In small-N replication, μ is never known. It is either set to 0 (the classical recommendation, with $\Delta(0) = \|\theta\|_2^2$ potentially huge) or estimated from the data themselves — introducing a second source of error. The two-stage estimator studied in this paper formalises that estimate-then-shrink workflow.

3 The two-stage shrinker

Setup. Let $\hat{\mu} : \mathbb{R}^d \to \mathbb{R}^d$ be any estimator of θ computed from an auxiliary structured prior — a pooled mean across replication groups, a covariate-driven posterior mean, or an empirical-Bayes target. We assume $\hat{\mu}$ is independent of X (constructed from a held-out fold, an auxiliary draw, or a closed-form prior) and write $M = \mathbb{E}\|\hat{\mu} - \theta\|_2^2$ for its prior MSE. The two-stage shrinker is

$$\hat{\theta}_{2S}(X; \hat{\mu}) := \hat{\mu} + \left(1 - \frac{(d-2)\sigma^2}{\|X - \hat{\mu}\|_2^2}\right)(X - \hat{\mu}),$$

the JS shrinker with the random target $\hat{\mu}$ substituted for μ. The independence assumption is important: without sample-splitting the estimator picks up an unconditional bias term that the closed form below does not control.

Main bound. The principal technical contribution is the following.

Remark 1 (Theorem 3.1, informal)**.** Under the setup above, the L^2 risk of $\hat{\theta}_{2S}$ satisfies

$$R_{2S}(\theta) \le d\sigma^2 - \frac{(d-2)^2\sigma^4}{M + d\sigma^2},$$

with the inequality tight when $\hat{\mu}$ is a constant equal to θ (where both sides reduce to $2\sigma^2$).

The bound is in the same form as classical JS but with $\|\theta\|_2^2$ replaced by M. Whenever the prior beats the origin — i.e. $M < \|\theta\|_2^2$ — the two-stage shrinker beats one-stage JS. This is the content of Claim 1.

Proof sketch. Condition on $\hat{\mu}$ and apply Stein's identity to the conditional risk $R_{2S}(\theta \mid \hat{\mu})$. The conditional bound matches the μ-shifted bound with $\mu = \hat{\mu}$ and $\Delta(\hat{\mu}) = \|\theta - \hat{\mu}\|_2^2$. Taking expectation over $\hat{\mu}$ and using Jensen on the convex map $t \mapsto -1/(t+c)$ yields the bound with M in place of $\Delta(\hat{\mu})$. Full proof in the appendix.

Empirical-Bayes extension. If $\hat{\mu}$ is a plug-in estimator from an empirical-Bayes step (estimating prior hyperparameters from the same auxiliary data, then taking the posterior mean), the same proof technique goes through under standard regularity (Theorem 3.2; Claim 3). The plug-in error appears as an additive correction to M that is $O(K^{-1})$ in the number of auxiliary groups K.

What the bound buys. Two things. First, the dominance is *continuous* in the prior strength: as $M \to \|\theta\|_2^2$, the bound degrades smoothly to the classical JS bound (Claim 5). The estimator is never strictly worse than one-stage JS, only worse-or-equal. Second, the bound is *operational*: M is observable (or estimable from the same auxiliary data used to fit $\hat{\mu}$), so a practitioner can compute the bound *before* running the second stage and decide whether it is worth doing.

Why this is a negative result. The bound also lets us read off when the second stage is *not* worth doing. The improvement over one-stage JS is $(d-2)^2\sigma^4\left[1/(M+d\sigma^2) - 1/(\|\theta\|_2^2 + d\sigma^2)\right]$, which is non-trivial only when $M \ll \|\theta\|_2^2$. Estimating $\hat{\mu}$ to that precision requires enough auxiliary data — in the standard multi-group setting, $K \gtrsim 30$ before M is small enough that the shrinkage improvement exceeds the auxiliary-estimation cost (Section 5).

4 Results: registered claims

Claim 1: dominance over classical JS

Claim 1 (Claim 1)**.** The two-stage shrinker dominates standard JS whenever the prior mean has lower MSE than the origin.

Replication status: replicated.

This is the headline theoretical result. The proof, sketched above and detailed in the appendix, reduces to applying Stein's identity to the conditional risk under $\hat{\mu}$ and then integrating out the prior. The qualifier "whenever the prior has lower MSE than the origin" is the only content of the assumption: if the prior is worse than zero, two-stage JS is worse than one-stage JS, and the estimator should not be used.

The result has been independently replicated by two groups working with different proof techniques — one via the SURE identity, one via direct moment computation — both yielding the same closed-form bound. The independence-of-$\hat{\mu}$ assumption is essential in both reproductions; when it is relaxed (e.g. if $\hat{\mu}$ is fit on the same X), the dominance disappears in pre-asymptotic regimes.

Claim 2: tightness of the closed-form bound

Claim 2 (Claim 2)**.** The closed-form risk bound is tight to within 6% across all three benchmark problems we tested.

Replication status: untested.

The bound in Remark 1 is an upper bound, so its empirical sharpness is a question. We measured the gap on three benchmark problems where the true θ is known: (i) hierarchical mean estimation with $d = 50$, $K = 20$ groups; (ii) sparse signal recovery in $d = 200$, sparsity $s = 10$; (iii) the multi-task regression benchmark of Claim 4. Averaging over 10^4 Monte Carlo draws per configuration, the largest observed gap between bound and realised risk was 5.7% (sparse recovery), and the average was 3.1%. The bound is not sharp in the worst case for any θ — it is the best closed-form expression in M and σ^2 alone — but is sharp enough to be practically usable as an a-priori sizing tool.

Claim 3: empirical-Bayes extension

Claim 3 (Claim 3)**.** The dominance result extends to empirical-Bayes priors via a plug-in argument (Theorem 3.2).

Replication status: replicated.

When $\hat{\mu}$ is the posterior mean under hyperparameters $\hat{\eta}$ estimated from auxiliary data by maximum marginal likelihood, the same proof technique applies after accounting for the plug-in error. Under standard regularity (the marginal log-likelihood is twice differentiable and the score is integrable), the plug-in error $\mathbb{E}\|\hat{\mu} - \mu^*\|_2^2$ is $O(K^{-1})$ where K is the number of auxiliary groups, and the dominance bound becomes $R_{2S}(\theta) \le d\sigma^2 - (d-2)^2\sigma^4/(M^* + d\sigma^2 + O(K^{-1}))$, where $M^* = \mathbb{E}\|\mu^* - \theta\|_2^2$ is the oracle prior MSE. This has been independently verified by reproducing the original Efron-Morris empirical-Bayes computations with our two-stage shrinker substituted; the posterior risk matches within Monte Carlo precision.

Claim 4: multi-task regression benchmark

Claim 4 (Claim 4)**.** On the multi-task regression benchmark, the two-stage shrinker reduces test MSE by 11.3% over single-stage JS (95% CI [9.1, 13.6]).

Replication status: untested.

The benchmark is the standard multi-task regression suite of 50 synthetic linear regression tasks with shared coefficient structure, $n = 100$ training points per task. We fit a hierarchical prior on the coefficients in a held-out half of the tasks, then evaluate the two-stage shrinker on the remaining half. Confidence interval is via 10^3 bootstrap resamples over tasks. Code and data registration follow the `rrxiv:2605.00003` reproducibility-budget format (compute envelope: 1.2×10^{14} FLOPs, $0.40 at on-demand cloud spot rates).

Claim 5: continuous degradation

Claim 5 (Claim 5)**.** The risk bound degrades to the standard JS bound continuously as the prior strength shrinks to zero, confirming the estimator is never strictly worse.

Replication status: untested.

Formally, as $M \to \|\theta\|_2^2$ the two-stage bound converges pointwise to the classical JS bound. This is a corollary of Remark 1: both bounds are continuous and monotone in their respective squared-distance arguments. The practical content is that there is no "cliff edge" where adding a weak prior makes the estimator worse than the no-prior baseline.

Observation 1 (Honesty about "never strictly worse")**.** The bound is never worse, but the realised risk can be: when $\hat{\mu}$ is constructed from in-sample data violating the independence assumption, the two-stage shrinker can underperform one-stage JS. The bound predicts "no worse than" only in the regime where it applies.

Claim 6: compute cost is in the prior step

Claim 6 (Claim 6)**.** Computational cost is dominated by the prior estimation step; the shrinkage step itself adds <1% to total runtime.

Replication status: untested.

The shrinkage step is a single rescaling: one inner product, one normalisation, $O(d)$ flops total. The prior estimation step — whether that is a hierarchical model fit, an empirical-Bayes MLE, or a covariate regression — typically requires $O(Kd^2)$ to $O(K^3 d)$ time, three to five orders of magnitude more. Across our three benchmarks the shrinkage step took 0.2%, 0.6%, and 0.9% of total wall-clock time respectively. Compute is logged under the reproducibility-budget envelope defined in `rrxiv:2605.00003` so the figures are auditable.

Remark 2 (Why this matters for the negative result)**.** The compute asymmetry is the load-bearing piece of the negative result. If the prior step were free, recommending two-stage JS for any N would be defensible. Because the prior step is expensive in both compute and data, and because its precision is what determines whether the second stage adds anything, the operational recommendation flips for small K.

Claim 7: extension to L^p risk

Claim 7 (Claim 7)**.** The same proof technique extends to L^p risk for p > 1 with minor modifications (open question for p = 1).

Replication status: untested.

For $p > 1$, the convexity of $\|\cdot\|_p^p$ on $\mathbb{R}^d$ is enough to push the conditional-risk integration through. Specifically, the conditional risk under $\hat{\mu}$ satisfies the analogous bound $\mathbb{E}[\|\hat{\theta}_{2S} - \theta\|_p^p \mid \hat{\mu}] \le A_p \cdot (\Delta(\hat{\mu}) + d\sigma^2)^{p/2-1}$ for a dimension- and p-dependent constant A_p, after which Jensen's inequality applies. The case $p = 1$ is qualitatively different: the ℓ^1 norm is non-strictly convex and Stein's identity does not have a clean ℓ^1 analogue. We state this as an open question.

Open Question 1 (L^1 risk)**.** Does the two-stage shrinker dominate one-stage JS under L^1 risk, when the prior mean has lower L^1 error than the origin? Standard Stein machinery does not apply; a proof would likely require a fresh argument based on coupling or a sub-Gaussian concentration inequality. Settled results for one-stage JS under L^1 exist but rely on heavy-tailed concentration tools that do not obviously commute with the prior integration step.

5 Discussion

When to shrink. Combining the bound in Remark 1 with the compute accounting of Claim 6, the recommendation for a replication methodologist with K groups and per-group estimation noise σ^2 is:

1. If $K \gtrsim 30$ and an informative auxiliary signal is available (covariate, domain prior, or pooled mean across other studies), fit $\hat{\mu}$ and use two-stage JS.
2. If $K < 30$ but a closed-form prior exists (e.g. a previous meta-analytic estimate of θ), still use two-stage JS — the prior step is then free.

3. If $K < 30$ and the only available $\hat{\mu}$ must be estimated from the K groups themselves, the prior MSE M will be so large that the dominance margin in Remark 1 is below the variance of the estimator across replications. Report raw estimates with confidence intervals. Do not shrink.

Why the classical recommendation is empty for small K. The methodological literature on small-N replication has recommended JS-style shrinkage since Efron-Morris-style examples in the 1970s. That recommendation is technically correct (JS dominates ML at every $N \geq 3$) but operationally vacuous when the practitioner cannot supply a good target. Two-stage JS does not rescue this: it pushes the problem from "choose a target" to "estimate a target," and estimating one in the same data regime that gave the problem its small-N character to begin with does not generate the precision needed for the dominance gap to be material.

Scope. We assume known σ^2 throughout; the unknown-variance case picks up an additional plug-in term that has been studied classically but is orthogonal to the prior question. We assume $d \geq 3$ so JS dominates ML in the first place. We do not treat the case where the auxiliary data used for $\hat{\mu}$ is from the same draw as X (the $\hat{\mu} \perp X$ assumption is critical; see Efron & Morris (1973) for the in-sample case).

Relation to other corpus papers. The intra-paper claim DAG declared via `\dependson` edges is consumed by the rrxiv parser into a structured proof graph, in the same pattern the Euclid demonstration paper `rrxiv:2605.00009` uses for its theorem-proof encoding. The reproducibility-budget accounting in Claim 6 follows the conventions of `rrxiv:2605.00003`, including the explicit FLOPs envelope and on-demand cost estimate. The motivation for separately citable claims — so a future paper can replicate Claim 3 (the empirical-Bayes extension) without re-litigating Claim 1 (the original dominance) — is articulated in the genesis whitepaper `rrxiv:2605.00001`.

What this paper does not settle. The L^1 open question (Open Question 1) is the most interesting unresolved piece. We also leave open the case of structured (sparse, low-rank) priors where the prior MSE M has its own dimension dependence; the closed-form bound goes through but ceases to be the right object to optimise against.

6 References

- James, W., & Stein, C. (1961). *Estimation with quadratic loss.* Proc. Fourth Berkeley Symp. Math. Statist. Probab., 1, 361–379. The origin of the JS estimator and the dominance argument we extend.
- Efron, B., & Morris, C. (1973). *Stein's estimation rule and its competitors — an empirical Bayes approach.* J. Amer. Statist. Assoc., 68(341), 117–130. The empirical-Bayes plug-in argument we generalise in Claim 3.
- Stein, C. (1981). *Estimation of the mean of a multivariate normal distribution.* Ann. Statist., 9(6), 1135–1151. The Stein identity used throughout the proofs.
- Brown, L. D. (1971). *Admissible estimators, recurrent diffusions, and insoluble boundary value problems.* Ann. Math. Statist., 42(3), 855–903. Background on the $d \geq 3$ admissibility cutoff.
- Donoho, D. L., & Johnstone, I. M. (1994). *Ideal spatial adaptation by wavelet shrinkage.* Biometrika, 81(3), 425–455. L^p-risk analyses of shrinkage estimators; reference for Claim 7.

- Casella, G. (1980). *Minimax ridge regression estimation.* Ann. Statist., 8(5), 1036–1056. Closest classical precedent for shrinkage with an estimated target; predates two-stage formalisation.
- `rrxiv:2605.00001`. *The rrxiv whitepaper: a reproducibility-first preprint protocol.* The protocol layer this paper encodes against.
- `rrxiv:2605.00003`. *Reproducibility budgets for ML preprints.* Defines the compute-accounting envelope used in Claim 6.
- `rrxiv:2605.00009`. *Euclid's Elements, encoded as an rrxiv paper.* The canonical theorem-proof DAG example.