Spectral Calibration Without a Split

Learning Covariance Eigenvalue Corrections One Observation at a Time

Peter Cotton

Note — `microprediction/precise`

Abstract

Cross-validated covariance cleaning learns the variance belonging to an empirical eigenvector by projecting held-out data onto it, which requires partitioning a sample. A stream needs no partition. The eigenvector basis carried forward from past data is already out of sample with respect to the observation about to arrive, so every observation labels every eigenvalue at once. We give the conditional identity this rests on and the one-pass estimator that follows, which costs $O(p^2)$ per observation and keeps bounded state. In five regimes the learned calibration beats an uncleaned exponentially weighted covariance everywhere, and a simpler estimator beats it wherever dependence drifts: the analytical map applied to a rolling window, where equal weights make the asymptotics exact with $n = W$.

1 The gap

Nonlinear shrinkage of a covariance spectrum, which moves each sample eigenvalue by its own amount while keeping the sample eigenvectors, has an analytical solution under equally weighted, identically distributed observations [Ledoit and Péché, 2011, Ledoit and Wolf, 2020], and a random-matrix counterpart in the rotationally invariant estimators [Bun et al., 2017]. Both derivations assume a fixed population covariance observed through sampling noise. Financial dependence is not fixed. It drifts, and common modes whose strength varies dominate it. Manolakis et al. [2026] make this the explicit motivation for replacing the analytical singular-value map with a learned one, fixing the empirical basis from theory so that only the part where the derivation is known to fail is learned. Their map is trained offline, on a corpus.

The weighted case has since been worked out analytically: Oriol [2024] give non-linear shrinkage formulas for weighted sample covariances, the exponentially weighted case explicitly. That is the principled route for a recency-weighted analytical cleaner, and we would take it. We ask instead whether the correction can be *learned from the stream itself*, with bounded state, no corpus and no pretraining.

2 The identity

Let $\mathcal{F}_{t-1}$ denote everything observable before time t, and let $U_{t-1} = [u_1 \cdots u_p]$ be an orthonormal basis computed entirely from $\mathcal{F}_{t-1}$. Suppose $\mathbb{E}[x_t \mid \mathcal{F}_{t-1}] = 0$ and $\mathrm{Cov}(x_t \mid \mathcal{F}_{t-1}) = \Sigma_t$. Project the new observation,

$$z_t = U_{t-1}^\top x_t .$$

Because u_i is $\mathcal{F}_{t-1}$-measurable it passes through the conditional expectation, giving

$$\mathbb{E}\left[z_{i,t}^2 \mid \mathcal{F}_{t-1}\right] \;=\; u_i^\top \Sigma_t \, u_i, \tag{1}$$

which is exactly the variance a cleaned spectrum must assign to that empirical eigenvector. One observation supplies a noisy but conditionally unbiased label for *every* eigenvalue simultaneously. Equation (1) requires neither Gaussianity nor stationarity; it is a statement about conditioning, and the only structural demand is that U_{t-1} not be built using x_t.

That the squared projections are the right thing to predict follows from an exact decomposition. For any orthogonal U, any $d \in \mathbb{R}^p$, and $z = U^\top x$,

$$\left\| U \operatorname{diag}(d) U^\top - xx^\top \right\|_F^2 \;=\; \sum_i \left(d_i - z_i^2\right)^2 \;+\; \underbrace{\sum_{i \neq j} z_i^2 z_j^2}_{\text{free of } d}. \tag{2}$$

The Frobenius norm is invariant under orthogonal conjugation, and $U^\top (U \operatorname{diag}(d) U^\top - xx^\top) U = \operatorname{diag}(d) - zz^\top$, whose diagonal entries are $d_i - z_i^2$ and whose off-diagonal entries are $-z_i z_j$. So learning to predict squared projections minimizes, term for term, the same instantaneous objective as learning the covariance matrix within the chosen basis. The residual term is the price of holding the eigenvectors fixed, and no choice of d can touch it.

Neither statement is new. Equation (1) is what makes cross-validated cleaning work: Bartz [2016] estimates cleaned eigenvalues by projecting held-out folds onto eigenvectors fitted on the remainder, and Lamrani et al. [2025] analyze how such a sample should optimally be split, finding the train–test ratio should scale as the square root of the dimension. The observation we have not found recorded is that in a streaming setting the split need not be constructed at all. Every arriving observation is already held out with respect to the basis carried forward, so the data cost of cross-validation, the reason one must decide how much sample to sacrifice, simply does not arise. What was a design choice becomes a property of the arrow of time.

3 The estimator

State: a running mean m, an exponentially weighted second moment S, a basis U, a spectrum λ, and the accumulators of a calibration map f. On observing x_t, with weight w:

$$\begin{aligned}
&\delta = x_t - m, \qquad z = U^\top \delta, \\
&\quad \text{train } f \text{ on the pairs } (\text{features}_i,\ z_i^2), \\
&\lambda \leftarrow (1-w)\lambda + w z^2, \\
&m \leftarrow m + w\delta, \qquad S \leftarrow (1-w)S + w\,\delta\delta^\top,
\end{aligned}$$

refreshing (λ, U) from an eigendecomposition of S every K observations, and reporting $\hat{\Sigma}_t = U \operatorname{diag}\big(f(\lambda)\big) U^\top$. The ordering matters: the labels are formed against the basis and the mean that predate x_t, which is what makes (1) apply.

The λ recursion is not an approximation. For a basis held fixed between refreshes,

$$u_i^\top S_t u_i = (1-w)\, u_i^\top S_{t-1} u_i + w\left(u_i^\top \delta\right)^2,$$

so the exponentially weighted Rayleigh quotient obeys the same recursion as the covariance and is available at $O(p)$ from projections already computed. Only the basis needs an eigendecomposition. The per-observation cost is therefore $O(p^2)$ for the projection and the outer-product update, plus $O(p^3/K)$ amortized, and nothing grows with elapsed stream length. Stale eigenvectors are the price. They bound how quickly the estimator can follow a rotation of the true eigenbasis.

4 What has to be learned

The map f must be positive, must treat equal eigenvalues equally, and must forget. We use kernel-weighted running averages over a fixed grid of feature centers, a Nadaraya–Watson estimator with exponential forgetting,

so that the prediction is a ratio of two nonnegative accumulators. It is positive because the labels are squares, it pools directions with identical features by construction, and it forgets at a controlled rate. A few dozen bins suffice; nothing here needs a neural framework.

The design question is what to index the map by. Three choices separate cleanly: *rank*, the position of an eigenvalue in the sorted spectrum; *eigenvalue*, its observed size relative to the mean, on a log scale; and *shared*, the observed size together with a global summary of the spectrum's shape (we use the dispersion of the log spectrum). Rank calibration can memorize a useful spectral profile; the concern is that it will keep imposing that profile after the structure generating it is gone, since the position of an eigenvalue says nothing about whether it still refers to anything.

5 Experiments

Gaussian streams in $p = 32$ dimensions, 6000 observations, eight seeds, exponential weight $r = 0.02$, basis refreshed every 10 observations, calibration forgetting $\rho = 0.02$, scoring from observation 1000 onward. The metric is mean squared Frobenius error divided by the squared norm of the truth; lower is better. Five regimes: an isotropic covariance; a broad, static population spectrum; an abrupt rotation of the factor directions at the midpoint; a factor structure that disappears at the midpoint; and one in which the factor structure and its absence alternate repeatedly, the only regime in which a spectral shape *recurs* and can therefore be recalled.

We compare against an uncleaned exponentially weighted covariance, the linear shrinkers `LedoitWolf` and `OAS` at the same decay, and the analytical nonlinear shrinker of Ledoit and Wolf [2020] over an expanding sample.

	uncleaned	linear, weighted		analytical nonlinear		learned calibration		
regime	raw EW	L–W	OAS	expanding	window	rank	eigenvalue	shared
isotropic	0.340	**0.0007**	**0.0007**	0.0007	0.0049	0.0026	0.0018	0.0018
broad spectrum	0.206	0.204	0.199	**0.0072**	0.066	0.135	0.147	0.147
factor rotation	0.243	0.238	0.231	0.206	**0.067**	0.191	0.186	0.186
factors vanish	0.299	0.127	0.124	0.239	**0.059**	0.105	0.103	0.103
regimes alternate	0.330	0.131	0.128	0.217	**0.104**	0.114	0.109	0.109

Table 1: Mean squared Frobenius error relative to the squared norm of the truth; lower is better, best in each row in bold. Eight seeds, $p = 32$, 6000 observations each. The windowed column uses $W = 250$; the learned columns use the same exponential weight $r = 0.02$ as the linear shrinkers.

The learned correction is dominated by a simpler analytical one. On every regime where the dependence changes, applying the analytical map to a rolling window beats every learned variant: 0.067 against 0.186 under a rotation, 0.059 against 0.103 when the factor structure dies, 0.104 against 0.109 under alternation.

The learned calibration does what we set out to show it could. It beats an uncleaned exponentially weighted covariance everywhere, and beats exponentially weighted linear shrinkage by fifteen to twenty percent on exactly the non-stationary regimes that motivated it. But linear shrinkage was the wrong bar. A rolling window gives the analytical nonlinear map a sample it can forget without approximating anything, since equal weights inside a window of length W are exactly the sample the asymptotics describe, and that cheap route captures the adaptivity the learned route was introduced to supply.

The gap narrows to five percent on the fastest-drifting regime. That is the one direction in which a case for learning could be rebuilt: the window's advantage should shrink as the drift rate approaches the window length, and a learned map need not commit to a timescale in advance.

Rank calibration remembers, and that cuts both ways. Indexing by spectral position is the best learned variant on the static broad spectrum (0.135 against 0.147), where a stable profile pays to memorize, and the

worst on every regime where the structure changes. Conditioning on the observed eigenvalue instead removes that failure at a modest cost elsewhere, That is what one would hope. An eigenvalue's size still refers to something after the structure that produced it is gone, whereas its rank does not.

Global shape conditioning does not help. We expected the third variant, indexing the map by the observed eigenvalue *and* a summary of the spectrum's overall shape, to dominate the second, on the reasoning that it could recall a calibration when a shape recurred. It does not. The two agree to three decimal places in every regime, including the alternating one built specifically to reward recall. The reason is visible in the diagnostics. The shape statistic ranges over roughly 0.3–1.1 *across* these regimes but moves only about 0.4 *within* any single stream, so the discriminating variation is between problems rather than along a trajectory, and a map that one estimator carries through one stream never sees enough of the axis to use it. We report this as a negative result rather than tuning until it disappears.

6 Relation to prior work

The analytical line [Ledoit and Péché, 2011, Ledoit and Wolf, 2020, Bun et al., 2017] derives the map rather than learning it, and is the right default whenever its assumptions hold; the expanding-sample column above shows how much it wins by when they do, and how badly it fails when they do not. The cross-validation line [Bartz, 2016, Lamrani et al., 2025] uses the same identity as we do but must construct its split. Oriol [2024] extends the analytical treatment to weighted samples and is the principled way to build a recency-weighted analytical cleaner; substituting a moment-matched effective sample size into an equally weighted formula, as is sometimes done, is a scalar approximation that does not encode the weight distribution. Manolakis et al. [2026] learn the singular-value map, as we do, but offline and for the rectangular cross-covariance problem. For that problem the corresponding sequential label is immediate, $(u_i^\top x_t)(v_i^\top y_t)$, though assembling a valid joint covariance additionally requires compatibility with the marginal blocks: an arbitrary cross-block correction need not preserve positive semidefiniteness.

Incremental eigenspace tracking is a separate and well-developed literature [Weng et al., 2003]; we take the basis by periodic refresh, which is crude, and a better tracker would directly relax the one cost we identified.

7 Limitations

The labels are severely noisy. Under Gaussianity z_i^2 has variance $2(u_i^\top \Sigma u_i)^2$, so a single observation carries roughly one effective bit per direction, and the map's forgetting rate trades adaptation against that noise. The eigenvector basis is refreshed rather than tracked, which bounds adaptation to abrupt rotation.

All evidence here is synthetic and Gaussian, and we constructed the regimes to separate the failure modes we suspected. There is no real equity study. On those regimes the construction does not win: a rolling window carrying the analytical map beats it wherever the dependence drifts, which is the case the learned map was meant to answer.

What survives is the identity, the fact that a stream needs no split to exploit it, and an estimator cheap enough that the question can be revisited on real data at drift rates faster than a window can follow. The open question is not whether spectral cleaning can be made online. It plainly can. It is how calibration information should be shared across directions and forgotten as the spectrum changes, and the negative result on shape conditioning says the obvious first answer is not the right one.

References

Daniel Bartz. Cross-validation based nonlinear shrinkage. *arXiv preprint arXiv:1611.00798*, 2016.

Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: Tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.

Lamia Lamrani, Christian Bongiorno, and Marc Potters. Optimal data splitting for holdout cross-validation in large covariance matrix estimation. *arXiv preprint arXiv:2503.15186*, 2025.

Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151:233–264, 2011.

Olivier Ledoit and Michael Wolf. Analytical nonlinear shrinkage of large-dimensional covariance matrices. *The Annals of Statistics*, 48(5):3043–3065, 2020.

Efstratios Manolakis, Christian Bongiorno, and Rosario Nunzio Mantegna. Physics-informed singular-value learning for cross-covariances forecasting in financial markets. *Finance Research Letters*, 110:110602, 2026.

Benoit Oriol. Asymptotic non-linear shrinkage and eigenvector overlap for weighted sample covariance. *arXiv preprint arXiv:2410.14420*, 2024.

Juyang Weng, Yilu Zhang, and Wey-Shiuan Hwang. Candid covariance-free incremental principal component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):1034–1040, 2003.