SELF-CONSISTENT STOCHASTIC INTERPOLANTS FOR COVARIANCE PRIOR AND POSTERIOR LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Given many noisy covariance estimates, each computed from a few Gaussian samples with its own underlying covariance, how do we sample the covariance that produced a given estimate when no clean covariance is ever seen? This is an inverse problem: we also know the sampling mechanism, so each estimate is a draw from an unknown prior, seen through a channel we can simulate. Expectation–maximization over distributions undoes that blur, but every step requires sampling the posterior of the current prior, which no formula provides. We show that a conditional stochastic interpolant trained on its own reconstructions supplies that sampler and turns the iteration into an algorithm. Two versions target the same posterior over the whole matrix: one transports the matrix itself; the other, for a rotationally invariant prior, learns only the eigenvalue distribution and draws the relative orientation from an explicit conditional law, a completion that adds no distributional error. We prove identifiability exactly when each estimate uses at least as many samples as dimensions, and show that progress in what is observable is weaker than recovery of the prior, which needs further conditions. On synthetic data we recover both invariant and structured priors from noisy estimates alone, and on EEG and market records the posterior improves calibration and decisions.

1 INTRODUCTION

A covariance matrix estimated from a few samples is a poor description of the covariance of the Gaussian distribution that generated them. The standard remedy is shrinkage: the empirical spectrum is corrected, either linearly (Ledoit & Wolf, 2004) or by the nonlinear rule that is asymptotically optimal in Frobenius loss among estimators with the empirical eigenvectors (Ledoit & Péché, 2011; Ledoit & Wolf, 2012). The result is a single matrix, and it does not say which other covariances were compatible with the samples. Yet this uncertainty, in the variance the covariance assigns to a given direction, say, is useful information. Quantifying it is possible when the same estimation problem recurs across units: one covariance per subject in an EEG study, per experimental condition, or per trading period in a market, each with its own limited record. Pooling the records describes an average covariance rather than any one unit's, whereas the collection of noisy estimates carries information about how the covariance varies across units and how certain we can be about any one of them.

To make this precise, imagine that each unit has its own covariance C, drawn from a prior $\pi_\star$ common to all units, and we observe only C_e, the empirical covariance of N Gaussian samples with covariance C. We see many C_e's, never a C, and $\pi_\star$ is unknown. *Given these estimates alone, how can we construct the posterior $P_\star(\cdot \mid C_e)$, which weights each covariance by how well it explains a specific C_e and by how plausible it was to begin with?* Knowing $P_\star$ gives the prior $\pi_\star$ by averaging over the observed C_e's, and point estimates under any loss, the finite-N counterpart of shrinkage.

What we know besides the estimates themselves is the corruption channel: *given a C, we can sample the corresponding C_e*. This yields a consistency condition: the posterior used to sample C given C_e must be such that simulating new C_e's from the C's it produces recovers the observed ensemble. Expectation–maximization enforces that consistency, *provided we can sample from the posterior of a candidate prior*. Conditional stochastic interpolants (Chen et al., 2024) provide such a sampler: an SDE whose endpoint at a given C_e is a posterior draw. Training it needs pairs of empirical and clean covariances, which we do not have, so we generate them (Modi et al., 2026): reconstruct covariances from the recorded estimates with the current sampler, simulate new estimates from those

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.

reconstructions, and train the next sampler on the resulting pairs. In the limit of exact training this loop is the EM iteration itself.

We implement this in two ways, both aimed at the posterior over the full matrix: one evolves the matrix itself, in logarithmic coordinates; the other applies when the prior is invariant under rotations, learns the eigenvalue posterior alone, and completes each draw with the conditional law of the orientation, which is known in closed form.

Contributions.

- The prior is identifiable from the observed estimates if and only if $N \geq d$; below it there are distinct smooth priors, rotationally invariant and compactly supported, that produce the same observations (Section 2, Theorem 1).
- The ideal iteration inherits the $O(1/k)$ observation-space rate of Aubin-Frankowski et al. (2022), and under fixed eigenvalue bounds the prior and posterior converge weakly. An explicit identity separates progress in observation space from recovery of the prior, and a rate for the prior requires a further inverse-stability condition (Section 2.2, Theorem 2).
- Under rotational invariance the reduction to eigenvalues is exact: completing a spectral draw with the known eigenvector law costs no Kullback–Leibler divergence, so accuracy on the spectrum transfers to the full matrix (Section 3.2, Theorem 4).
- Two algorithms realize this from empirical covariances alone, in matrix and in spectral coordinates, and coincide with EM when their regressions are exact (Proposition 3, Section 3).

Related work. We build on three lines of work and compare with two others.

EM over distributions. Updating a full distribution rather than a parameterization of it via EM is classical theory (Aubin-Frankowski et al., 2022), but it leaves open what the iteration can identify, and how progress in observation space relates to prior recovery. We take up both for the covariance channel.

Conditional stochastic interpolants. The E-step of EM requires a sampler for the posterior of a candidate prior, and we take it from Chen et al. (2024).

Self-consistency with a black-box simulator. Training such a sampler needs pairs we do not have, and Modi et al. (2026) supply the device: with a simulator of the corruption, a stochastic-interpolant transport is learned by iterating reconstruction and fresh corruption, with no analytic likelihood and with monotone improvement. The same alternation with a conditional diffusion appears in Rozet et al. (2024), Bai et al. (2024) and Hosseintabar et al. (2025). We specialize it to the Gaussian covariance channel, where the closed-form likelihood makes the posterior analysis possible and the equivariance of the channel makes the spectral reduction exact.

Covariance shrinkage. Linear and nonlinear shrinkage return one matrix per observation (Ledoit & Wolf, 2004; Ledoit & Péché, 2011; Ledoit & Wolf, 2012); we produce the distribution that such a matrix summarizes, from which point estimates follow under a chosen loss, with intervals. Their theory is asymptotic and posits no prior, whereas for us rotational invariance is an assumption that yields an exact reduction, see Section 4 and Appendix G.6. Completing the orientation of a sampled matrix connects to the matrix Bingham samplers of Hoff (2009).

Parametric priors over covariances. Fitting a conjugate prior to an ensemble of covariances and using its closed-form posterior is the classical parametric alternative (Alvarez et al., 2014), extended to covariates by covariance regression (Hoff & Niu, 2012) and applied to brain connectivity as population shrinkage (Rahim et al., 2019). Such a family is exact when it contains the truth and rigid when it does not; Section 4 measures both against an inverse-Wishart prior fitted by marginal likelihood.

2 Problem setup and solvability

2.1 What is observed, what is known, and what is the objective

Given an unknown prior $\pi_\star$ over the set $\mathbb{S}^d_{++}$ of positive-definite symmetric $d \times d$ matrices, let

$$C \sim \pi_\star, \qquad x_i \mid C \overset{\text{iid}}{\sim} \mathcal{N}(0, C), \qquad C_e = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^{\mathrm{T}}. \tag{1}$$

Denote by $K_N(dC_e \mid C)$ the forward conditional law of C_e given C, which we call the *Wishart channel*. The recorded observations are draws from

$$\mu_\star(dC_e) \equiv (K_N \pi_\star)(dC_e) = \int K_N(dC_e \mid C)\, \pi_\star(dC). \tag{2}$$

The problem is then specified by:

Observed: independent draws $C_{e,1}, \ldots, C_{e,M} \sim \mu_\star$, and nothing else: no clean C, no pair (C, C_e), and no knowledge of $\pi_\star$.

Known: the sample size N and the channel K_N itself, which we can sample exactly for any supplied C, since

$$C'_e = \frac{1}{N} C^{1/2} Z Z^{\mathrm{T}} C^{1/2}, \qquad Z \in \mathbb{R}^{d \times N}, \quad Z_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \tag{3}$$

has law $K_N(\cdot \mid C)$.

Objective: samplers for the posterior and the prior distributions, given respectively by Bayes' rule and by marginalization,

$$P_\star(dC \mid C_e) = \frac{k_N(C_e \mid C)\, \pi_\star(dC)}{\int k_N(C_e \mid C)\, \pi_\star(dC)}, \qquad \pi_\star(dC) = \int P_\star(dC \mid C_e)\, \mu_\star(dC_e). \tag{4}$$

where $k_N(C_e \mid C)$ is the density of $K_N(\cdot \mid C)$, given explicitly in Appendix A relative to a reference measure that keeps it valid for every N, including $N < d$ where C_e is singular. By the second identity a posterior sampler is a prior sampler too: draw an index a uniformly, then $C \sim P_\star(\cdot \mid C_{e,a})$.

Whether this objective is attainable at all depends on the sample size, and the dependence is sharp.

Theorem 1 (Sharp sample-size threshold)**.** *The map $\pi \mapsto K_N \pi$ is injective if and only if $N \geq d$; for $N < d$ the failure is witnessed by distinct smooth rotationally invariant priors with common compact support inside $\mathbb{S}^d_{++}$ and identical observed laws.*

Injectivity makes the problem well posed: $\mu_\star$ determines $\pi_\star$ (Appendix B). Below the threshold the two indistinguishable priors have different posterior kernels, so no ensemble determines an unrestricted covariance posterior, and since they are rotationally invariant the spectrum is no better determined than the matrix. The obstruction lies in the problem, not in any particular method, and smaller classes stay recoverable (Proposition 5).

Notation. Subscript e denotes empirical quantities, hats denote fitted ones, $p = d(d+1)/2$, and convergence of measures is written $\to$ in total variation and $\Rightarrow$ weakly. Remaining notation is collected in Appendix A.

2.2 The ideal EM iteration and its missing sampler

As a recipe, (4) is circular: the posterior it defines requires the prior we are trying to find. Breaking the circle requires a starting guess, and any candidate prior π has a posterior of its own, obtained by replacing $\pi_\star$ with π in (4),

$$P_\pi(dC \mid C_e) = \frac{k_N(C_e \mid C)\, \pi(dC)}{\int k_N(C_e \mid C)\, \pi(dC)}. \tag{5}$$

Substituting P_π for the unavailable $P_\star$ in the marginalization (4) returns not $\pi_\star$ but a new prior, and iterating the substitution is exactly expectation–maximization: from an initial π_0, sample the posterior of the current prior (the E-step) and average it over the observed population (the M-step),

$$\boxed{\pi_{k+1} = \mathcal{T}(\pi_k), \qquad \mathcal{T}(\pi)(dC) = \int P_\pi(dC \mid C_e)\, \mu_\star(dC_e).} \tag{6}$$

Below we write P_k for P_π at $\pi = \pi_k$. This is *nonparametric* EM: the entire prior is updated, not a fixed list of parameters (Aubin-Frankowski et al., 2022), and (4) makes the truth a fixed point, $\mathcal{T}(\pi_\star) = \pi_\star$, though not the only one.

Theorem 2 (Convergence of the ideal EM iteration). *Write* $\mathrm{KL}(\cdot\|\cdot)$ *for the Kullback–Leibler divergence, equal to* $+\infty$ *when absolute continuity fails, and let* $\mu_k = K_N\pi_k$ *be the observation law of the* k*th iterate of* (6)*. For every prior* π_k *and its posterior* P_k*,*

$$\int \mathrm{KL}(P_\star\|P_k)\, d\mu_\star = D_k - F_k, \qquad D_k = \mathrm{KL}(\pi_\star\|\pi_k), \quad F_k = \mathrm{KL}(\mu_\star\|\mu_k). \tag{7}$$

In addition, along $\pi_{k+1} = \mathcal{T}(\pi_k)$*:*

(i) Descent in observation space. *If* $D_0 < \infty$*, then* $D_{k+1} \le D_k - F_k$*,* $F_{k+1} \le F_k$*, and*

$$F_k \le \frac{D_0}{k+1}, \tag{8}$$

so that $\mu_k \to \mu_\star$ *in total variation.*

(ii) Recovery of the prior and the posterior. *If moreover* $N \ge d$ *and* $\pi_0, \pi_\star$ *are supported on a common compact set* $\{\lambda_- I_d \preceq C \preceq \lambda_+ I_d\}$ *with* $0 < \lambda_- \le \lambda_+ < \infty$*, then*

$$\pi_k \Rightarrow \pi_\star, \qquad P_k(\cdot \mid C_e) \Rightarrow P_\star(\cdot \mid C_e) \quad \text{for } \mu_\star\text{-almost every } C_e. \tag{9}$$

(iii) Geometric recovery. *If* $D_k \le \chi F_k$ *along the iterates for some* $\chi \ge 1$*, then* $D_k \le (1 - \chi^{-1})^k D_0$*, hence* $\pi_k \to \pi_\star$ *in total variation and the left side of* (7) *tends to zero.*

The eigenvalue bounds in part (ii) make the inverse of K_N weakly continuous: Theorem 1 covers exact equality, where $F = 0$ forces $D = 0$, while the step to convergence needs that continuity, which gives $\pi_k \Rightarrow \pi_\star$ on a compact class. Forcing $D_k \to 0$ is stronger. Part (iii) assumes it, and asks more besides: near $\pi_\star$ the smallest admissible χ is set by a local operator whose spectrum is bounded away from zero only on a restricted class (Appendix D). Neither hypothesis can be checked, since both concern $\pi_\star$, but held-out observations allow observation-space discrepancies to be estimated, and the checkpoint selection of Section 4 uses them (Appendix C).

3 Implementing EM with stochastic interpolants

Suppose we could draw clean covariances $C \sim \pi$ together with their observations $C_e \sim K_N(\cdot \mid C)$. Sampling C given C_e would then be an ordinary supervised problem, solvable by the construction of Chen et al. (2024): interpolate along a noisy path from the observation to its paired covariance, learn the drift of that path by squared-error regression, and integrate the resulting Föllmer SDE from C_e to produce a draw of C. We have no such C, but once a drift has been fitted we have the sampler's own output: run it at the recorded observations and it returns matrices. Passing these back through the known channel (3) produces the pairs the regression was missing (Modi et al., 2026). Fit the next drift on those pairs, reconstruct again, and repeat; a sampleable π_0 starts the recursion, and no clean covariance is ever supplied.

The loop is not circular because it uses two different observation distributions. The drift at stage k is *trained* on synthetic observations distributed as $K_N\pi_k$, but it is *applied* at the recorded observations, distributed as $\mu_\star$. Under exact training its endpoint law at a supplied C_e is the posterior $P_k(\cdot \mid C_e)$ of the current prior (Proposition 3), so averaging the reconstructions over the recorded ensemble rather than the synthetic one returns $\mathcal{T}(\pi_k) = \pi_{k+1}$ exactly. The SDE performs the E-step and the ensemble average the M-step (Figure 1). Bringing $K_N\pi_k$ into agreement with $\mu_\star$ is the consistency condition of Section 1.

The only choice left is the state the SDE evolves, and any invertible encoding will do.

3.1 Matrix coordinates

We encode a covariance as $Y = \Phi(C) = \mathrm{svec}(\log C)$, where svec stacks the entries of a symmetric matrix, the off-diagonal ones scaled by $\sqrt{2}$, into $\mathbb{R}^p$. That scaling makes svec a Frobenius isometry, so least squares in $\mathbb{R}^p$ is least squares on matrices, and $\Phi^{-1}(Y) = \exp(\mathrm{svec}^{-1} Y)$ is positive definite for every finite Y, so interpolation noise cannot leave the cone. Appendix E.1 gives the details, including the shift applied to a singular C_e, and shows that coordinates change neither the conditional target nor any KL divergence between candidate priors.

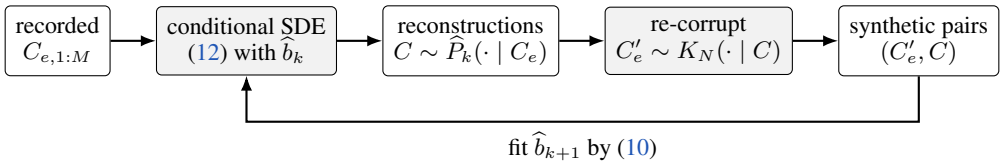

Figure 1: One outer iteration. The learned SDE turns recorded observations into posterior draws under the current prior (the E-step); averaging those draws over the ensemble is the M-step (6); re-corrupting them through the known channel manufactures the paired data that the next regression needs. No clean covariance is ever supplied.

Training pairs come from the current sampleable prior π: draw $C \sim \pi$, form C'_e by (3), and record the encoded pair (Y'_e, Y), held fixed when differentiating the loss. Given such pairs, fix a noise scale $\sigma > 0$ and minimize

$$\begin{aligned} \mathcal{L}_\pi(\widehat{b}) &= \mathbb{E}\big[\|\widehat{b}_t(I_t, Y'_e) - R_t\|_2^2\big], \\ I_t &= (1-t)Y'_e + tY + \sigma(1-t)W_t, \qquad R_t = Y - Y'_e - \sigma W_t, \end{aligned} \tag{10}$$

where the expectation is over $t \sim \mathcal{U}([0,1])$, the pair (Y'_e, Y) drawn from πK_N, and W_t a standard Brownian motion in $\mathbb{R}^p$ independent of that pair. The interpolating path runs from $I_0 = Y'_e$ to $I_1 = Y$, and the minimizer of the loss is its drift,

$$b_t^\pi(y, Y_e) = \mathbb{E}[R_t \mid I_t = y,\ Y'_e = Y_e]. \tag{11}$$

At a supplied observation Y_e, held fixed as the conditioning input, the drift is integrated from that observation,

$$dY_t = \big[(1+t)b_t^\pi(Y_t, Y_e) - Y_t + Y_e\big]dt + \sigma\sqrt{1-t^2}\,dB_t, \qquad Y_0 = Y_e, \tag{12}$$

with a Brownian motion B_t drawn afresh for each posterior sample, and with b^π replaced in practice by a fitted $\widehat{b}$. The noise coefficient is the Föllmer one. Appendix E treats general schedules and the choice of σ, and Appendix F bounds the effect of an imperfect drift.

One pass of the loop then decodes each endpoint to $C = \Phi^{-1}(Y_1)$, draws fresh Gaussian noise, forms C'_e by (3), and records (Y'_e, Y) for the next regression, as Algorithm 1 states.

Proposition 3 (The learned operation supplies the E-step)**.** *Assume the regression moments, interior regularity, uniqueness, and endpoint limits of Appendix E. For synthetic pairs generated from πK_N, the unrestricted minimizer of (10) is b_t^π in (11), and the SDE (12) driven by it has conditional marginals equal to those of the interpolant, so that*

$$\Phi^{-1}(Y_1) \sim P_\pi(\cdot \mid C_e). \tag{13}$$

Exact population reconstruction in Algorithm 1 therefore realizes the EM update (6).

The SDE produces the full conditional law rather than assuming it available, which turns (6) from a formula into a procedure; any failure of the algorithm is then one of approximation and not of the scheme (Appendix E).

3.2 Rotationally invariant priors

A prior is rotationally invariant (RI) if C and QCQ^{T} have the same law for every orthogonal Q. Write

$$C = U\Lambda U^{\mathrm{T}}, \qquad C_e = U_e\Lambda_e U_e^{\mathrm{T}}, \qquad R = U_e^{\mathrm{T}}U, \tag{14}$$

with Λ and Λ_e listing eigenvalues in increasing order. A spectral law ν on these ordered diagonal matrices determines the entire prior, since $\Lambda \sim \nu$ and U Haar are independent, and the posterior factors as

$$\begin{aligned} P_\nu(d\Lambda, dR \mid \Lambda_e) &\propto \nu(d\Lambda)(\det\Lambda)^{-N/2}e^{-\mathcal{V}_{\Lambda,\Lambda_e}(R)}\,dH(R), \\ \mathcal{V}_{\Lambda,\Lambda_e}(R) &= \frac{N}{2}\operatorname{tr}(\Lambda^{-1}R^{\mathrm{T}}\Lambda_e R), \end{aligned} \tag{15}$$

where H is Haar probability measure on orthogonal matrices. The decisive feature is that the conditional orientation law

$$H_N(dR \mid \Lambda, \Lambda_e) \propto e^{-\mathcal{V}_{\Lambda,\Lambda_e}(R)} dH(R) \tag{16}$$

does not depend on ν, so only the spectral posterior $q_\nu(d\Lambda \mid \Lambda_e)$ has to be learned and the rest of the matrix is known. The learned state has dimension d rather than p, while the object sampled is still the full matrix.

The construction of Section 3.1 then runs unchanged in $\mathbb{R}^d$, with $Y = (\log \lambda_i)_{i=1}^d$ and $Y_e = (\log \lambda_{e,i})_{i=1}^d$. The drift acts jointly on all coordinates, conditioned on the entire empirical spectrum, so diagonal storage does not assume independent eigenvalues. Training pairs come from the spectral channel: for a supplied Λ, draw Gaussian Z and set

$$\Lambda'_e = \operatorname{diag}\left[\operatorname{spec}\left(\frac{1}{N}\Lambda^{1/2} Z Z^{\mathrm{T}} \Lambda^{1/2}\right)\right], \qquad K_N^{\mathrm{spec}}(\cdot \mid \Lambda) = \mathrm{Law}(\Lambda'_e \mid \Lambda). \tag{17}$$

Record (Y'_e, Y), fit (10) in $\mathbb{R}^d$, integrate (12) in dimension d, then exponentiate and sort the endpoint coordinates to obtain Λ. Proposition 3 applies verbatim with K_N^{spec} in place of K_N, so the exact loop is again the EM iteration, now over spectral laws; Algorithm 2 in Appendix G states it.

The matrix is completed by sampling (16), a matrix Bingham density for which conditional-update methods are available (Hoff, 2009) and whose normalizing constant cancels in a Metropolis ratio; Appendix G.4 specifies a kernel and proves its invariance and convergence. Writing $\widehat{H}_m$ for the conditional output law after m steps, the returned sample is

$$\Lambda \sim \widehat{q}_{\mathrm{num}}(\cdot \mid \Lambda_e), \qquad R \sim \widehat{H}_m(\cdot \mid \Lambda, \Lambda_e), \qquad C = U_e R \Lambda R^{\mathrm{T}} U_e^{\mathrm{T}}, \tag{18}$$

with Λ held fixed during each orientation run. Note that $U_e \Lambda U_e^{\mathrm{T}}$ is in general *not* a posterior draw, since discarding R discards the relative orientation. What makes the reduction worth its assumption is that this completion is exact.

Theorem 4 (Exact completion preserves posterior error)**.** *Let $\pi_\star$ be rotationally invariant with spectral law $\nu_\star$, and fix $C_e = U_e \Lambda_e U_e^{\mathrm{T}}$. Complete a spectral law $\widehat{q}(\cdot \mid \Lambda_e)$ with the exact H_N and call the resulting matrix law $\widehat{P}^H$. Then*

$$\mathrm{KL}(P_\star(\cdot \mid C_e) \| \widehat{P}^H(\cdot \mid C_e)) = \mathrm{KL}(q_\star(\cdot \mid \Lambda_e) \| \widehat{q}(\cdot \mid \Lambda_e)), \tag{19}$$

where $q_\star = q_{\nu_\star}$. The identity covers repeated eigenvalues when the reconstructed matrix law is used, with no unique eigenbasis required.

The spectrum is recoverable from a reconstructed matrix and the conditional matrix law given it is common to both models, so the KL chain rule gives (19) with no dimension-dependent loss (Appendix G.2); finite MCMC contributes a separate orientation error, bounded in Proposition 12. Accuracy on the spectrum therefore transfers to the full matrix, and the spectral iterates inherit every conclusion of Theorem 2 (Appendix G.3). Only the learned state shrinks: simulating (17) still requires an eigendecomposition, and orientation proposals remain cubic.

4 Numerical Experiments

We present three sets of experiments where the self-consistent learner sees only noisy empirical covariances C_{e}. We first measure recovery of known posteriors on synthetic data (§4.1), then work with real covariances from electroencephalographic (§4.2) and financial (§4.3) time series. We benchmark SC-SI against the sample covariance matrix (SCM), Oracle Approximating Shrinkage (OAS) (Chen et al., 2010), Ledoit–Wolf analytical nonlinear shrinkage (LW-NLS) (Ledoit & Wolf, 2020), and a conjugate inverse-Wishart prior fitted by marginal likelihood (IW-ML). With finite data, self-consistent EM eventually over-sharpens the learned prior, so the outer iteration k is selected on held-out validation pairs $(C_{\mathrm{e}}^{\mathrm{cal}}, C_{\mathrm{e}}^{\mathrm{val}})$ that share a latent covariance prior. Details in Appendix J.

4.1 Toy models with oracles

Setting and aim. We test six d=8 priors with an available oracle posterior $P(C \mid C_{\mathrm{e}})$, from conjugate to badly misspecified. Two are inverse-Wishart: *IW, RI*, with an isotropic scale matrix,

where the fitted family contains the truth and no direction is preferred, and *IW, non-RI*, whose scale matrix has condition number 16 in a fixed random frame, adding a direction to capture. Four leave the family: a two-*factor model* with log-normal idiosyncratic noise, whose non-conjugate posterior is computed by Hamiltonian Monte Carlo (Neal, 2011); the *AR(1)* family $C_{ij} = e^{-\gamma|t_i - t_j|}$, $\log \gamma \sim \mathcal{N}(0, 0.5^2)$, a structured decay with one degree of freedom and an ordering of the coordinates, whose posterior is a one-dimensional quadrature over γ; *iso-diagonal*, a near-flat spectrum in a Haar frame, where a rotation must be learned with no low-rank structure to exploit; and *diagonal*, with no correlation at all. SC-SI sees only $M = 5000$ noisy covariances ($N = 28$) per prior. We report (a) the held-out predictive log-score of $C_{\mathrm{e}}^{\mathrm{val}}$ given $C_{\mathrm{e}}^{\mathrm{cal}}$, relative to the oracle, and (b) the Wasserstein-1 distance between the learned and oracle posteriors for six statistics of C, in units of the oracle's posterior standard deviation, averaged over 128 held-out tasks.

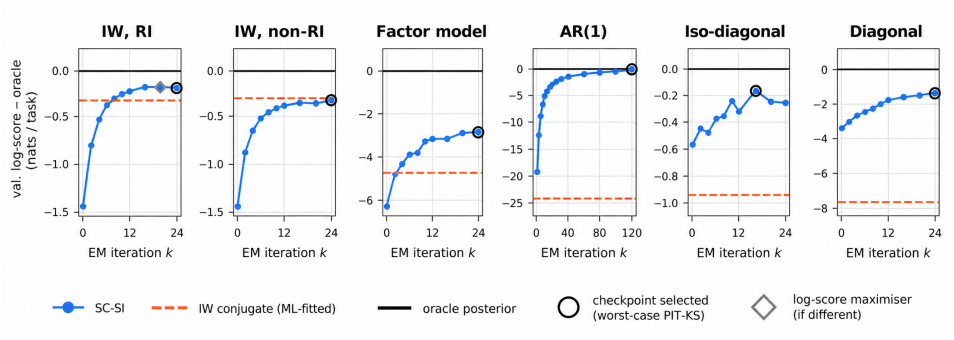

Figure 2: **Self-consistent training approaches the oracle.** Held-out predictive log-score relative to the oracle posterior (300 validation pairs, nats per task) along EM iterations k, on a vertical scale chosen per prior; k=0 is the deconvolved log-Gaussian initialisation. Dashed: IW-ML. Circle: checkpoint selected by the worst-case PIT score; diamond: log-score maximiser, when different. AR(1) is trained for K=120 iterations, the other priors for 24. The sample-covariance plug-in lies 14–26 nats below the oracle (off scale).

Table 1: **Log-score and distance to the oracle posterior in synthetic experiments.** Held-out predictive log-score of $C_{\mathrm{e}}^{\mathrm{val}}$ given $C_{\mathrm{e}}^{\mathrm{cal}}$, as a shortfall from the oracle in nats per task, and Wasserstein-1 distance to the oracle posterior in units of its s.d. κ: condition number; dv: log variance along the top/bottom empirical eigenvector of C_{e}; C_{01}: correlation of the first two coordinates. Bold is the better of SC-SI and IW-ML in each block.

Prior	Method	log-score gap ↓	log det C ↓	log κ ↓	top share ↓	$\mathrm{dv}_{\mathrm{top}}$ ↓	$\mathrm{dv}_{\mathrm{bot}}$ ↓	C_{01} ↓
IW, RI	SC-SI	**0.18**	0.09	0.12	0.17	0.15	0.17	0.08
	IW-ML	0.35	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**
IW, non-RI	SC-SI	0.41	0.08	0.09	0.09	0.09	0.12	0.07
	IW-ML	**0.36**	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**	**0.05**
Factor model	SC-SI	**3.26**	**0.27**	**0.70**	**0.32**	**0.29**	**0.90**	**0.35**
	IW-ML	5.00	**0.27**	1.00	0.65	0.85	1.15	0.45
Iso-diag	SC-SI	**0.19**	**0.11**	**1.61**	**1.79**	**0.62**	**0.48**	**0.23**
	IW-ML	1.08	0.41	3.76	4.05	1.85	1.07	0.60
AR(1)	SC-SI	**0.53**	**0.29**	**0.57**	**0.35**	**3.45**	**0.82**	**0.52**
	IW-ML	23.27	4.59	5.78	8.26	14.88	7.97	11.59
Diagonal	SC-SI	**1.31**	**0.17**	**0.60**	**0.46**	**0.50**	**0.75**	inf
	IW-ML	7.87	0.83	0.86	0.81	0.82	1.45	inf

Results: Self-consistent iterations closely recover the oracle across all six priors (Tab. 1). On the two inverse-Wishart priors SC-SI and IW-ML are both within 0.2–0.4 nats of the oracle, as they should be where the fitted family contains the truth. The ordering flips as soon as it does not: on the factor model SC-SI is 3.3 nats below the oracle against 5.0 for IW-ML, and on AR(1), which no inverse-Wishart can follow, 0.5 against 23.3, a gain of 22.7 nats. The W_1 columns say the same thing more sharply: on AR(1) the log variance along the top empirical eigenvector is off by 3.45 posterior

standard deviations for SC-SI against 14.88 for IW-ML. SC-SI stays within 0.2 nats of the oracle on iso-diagonal and gains 6.6 nats over IW-ML on the diagonal prior. Both report $C_{01} = \infty$ there, since a diagonal prior leaves the oracle posterior of C_{01} with no spread and the normalisation degenerates.

4.2 EEG: CALIBRATED UNCERTAINTY FOR HELD-OUT SUBJECTS

Setting and aim. We use the PhysioNet EEG Motor Movement/Imagery dataset (Goldberger et al., 2000; Schalk et al., 2004), 109 subjects recorded with eyes open and eyes closed on d=8 channels in the alpha band (8–13 Hz), split into non-overlapping 4-s windows. The prior is learned from the 1960 unlabeled window covariances of 70 training subjects; 10 select the EM checkpoint and 29 are held out for testing. We ask whether the posterior is a calibrated predictive law on *unseen subjects* and whether it improves the identification of eyes open versus closed.

Protocols and metrics. For each test window we compute the posterior from a calibration window and score the predictive law of a validation window, under two protocols: *split-half*, where both windows are interleaved 1-s chunks of the same 8-s epoch and thus share one covariance (isolating estimation uncertainty; 812 cases), and *next window*, where the covariance may drift (754 cases). We report the predictive log-score, the coverage of central intervals for directional variances and nonlinear functionals of C, and single-window eyes-open/closed classification with logistic regression on tangent-space features (Barachant et al., 2012).

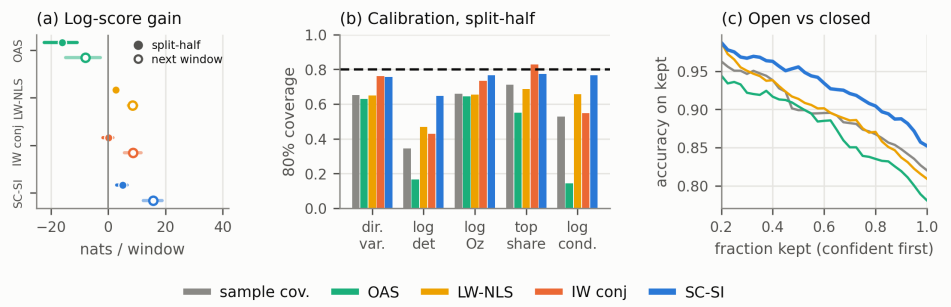

Figure 3: **EEG on held-out subjects.** (a) Predictive log-score gain over the sample covariance with 95% subject-level bootstrap intervals. (b) Coverage of the 80% predictive interval (dashed = nominal) for different functionals of C in *split-half*. (c) Accuracy of eyes-open/closed classification against the fraction of windows kept, most confident first; SC-SI uses the posterior-averaged classifier.

Results. **Log-score.** In the *split-half* test SC-SI gains $+5.0$ nats per window over the sample covariance (95% CI $[3.0, 6.5]$), against $+2.8$ for LW-NLS, $+0.1$ for IW-ML and -16.0 for OAS (Fig. 3a); the paired margin over LW-NLS is $+2.3$ $[0.4, 3.6]$ nats. SC-SI stays first under covariance drift between windows ($+15.7$ against $+8.7$ for IW-ML, $+8.6$ LW-NLS and -7.9 OAS). **Calibration.** For directional variances SC-SI and IW-ML both cover 0.76 at the nominal 80%, whereas the plug-ins reach only 0.63–0.65. The two separate on spectrum-dependent functionals (Fig. 3b): the 80% interval of the log-determinant covers 0.65 for SC-SI against 0.43 for IW-ML and 0.17–0.47 for the plug-ins; for the log condition number, 0.77 against 0.55 and 0.14–0.66. This is consistent with a conjugate prior's single concentration parameter being too rigid for the spectrum. **Decisions.** Averaging predictions over posterior draws raises single-window accuracy to 0.852 from 0.78–0.82 for the plug-ins and lowers the negative log-likelihood to 0.379 from 0.43–0.47; keeping the 50% most confident windows gives 0.956 against 0.90–0.91 (Fig. 3c). Bootstrap intervals exclude zero against the sample-covariance and LW-NLS plug-ins ($+3.2$ $[0.9, 5.5]$ and $+4.3$ $[2.5, 6.2]$ accuracy points).

4.3 FINANCIAL RETURNS: DECISIONS ACROSS MARKET REGIMES

Setting. We use daily returns of the 48 Ken French industry portfolios[1] (1926–2026). A task is a 63-day window and a random basket of d=12 industries observed throughout. The prior is learned from past windows only (non-overlapping, 16 random baskets each, $M \approx$ 4.5–5.7k) and refit on three expanding folds covering 1995–2026, with 8 fixed baskets and monthly forecasts (3008 cases).

[1]Data is accessible at https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Aim and estimators. We test both the posterior calibration through time and a minimum-variance portfolio built from each estimator, scored by its realised variance over the next 21 days relative to OAS. Estimators are the covariance estimates of §4, the SC-SI and IW-ML posterior means, the SC-SI Stein-optimal covariance (inverse of the posterior mean of the inverse), and PCA factor models whose number of factors K is chosen by the Gavish–Donoho threshold (Gavish & Donoho, 2014).

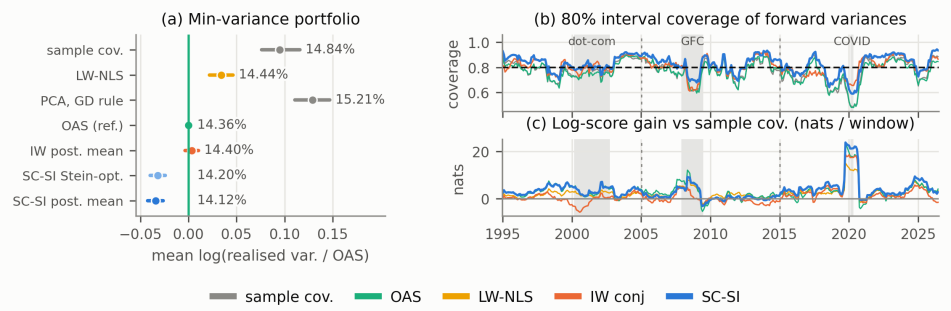

Figure 4: **Financial returns (1995–2026).** (a) Out-of-sample minimum-variance portfolio: mean log ratio of realised 21-day variance to OAS with 95% month-block bootstrap intervals (left is lower risk); the percentages give the annualised realised volatility of the portfolio. (b) 12-month rolling coverage of the central 80% forward interval for the variances of the 12 assets and the equal-weight portfolio. (c) 12-month rolling forward predictive log-score gain over the sample covariance. Shaded: dot-com bust, global financial crisis and COVID; dotted: model refits (2005, 2015).

Results: The posterior improves the decision. Using the SC-SI posterior mean as the covariance input lowers realised variance by 3.4% relative to OAS (log ratio -0.034, 95% CI $[-0.043, -0.026]$; annualised volatility 14.12% vs. 14.36%), and the Stein-optimal covariance is close (-0.032; Fig. 4a). Every other estimator is at or above OAS: the IW-ML posterior mean ($+0.004$, not significant), LW-NLS ($+0.034$) and the sample covariance ($+0.095$). The advantage is steady rather than driven by a few episodes: it holds in each volatility tercile (-0.051, -0.022, -0.030 from low to high; Appendix J.4). Additionally, using the posterior to pick the number of factors does not help: PCA with the SC-SI Bayes rule is $+0.086$ above OAS, better than Gavish–Donoho ($+0.129$) or the IW-ML Bayes rule ($+0.119$) but far from the posterior mean.

Calibration degrades in stressed regimes, less so for posterior-based laws. The 80% coverage of forward variances falls for every method in the global financial crisis and in COVID (Fig. 4b). Troughs are shallower for the posterior methods: in the crisis, 0.68 for SC-SI and 0.61 for IW-ML against 0.60 for the plug-ins; in COVID, 0.59 and 0.66 against 0.48. By volatility tercile (low/mid/high) SC-SI covers $0.84/0.85/0.79$ and IW-ML $0.81/0.81/0.79$, against $0.77/0.80/0.73$ for OAS and $0.77/0.81/0.75$ for the sample covariance: in the top tercile both posterior methods hold 0.79 while the plug-ins fall to 0.73–0.75. In calm markets SC-SI is slightly conservative and IW-ML is closest to nominal.

5 Discussion and Limitations

The ideal EM iteration leaves its E-step unspecified. A conditional stochastic interpolant trained on its own reconstructions supplies that step and turns the iteration into an algorithm that runs on noisy covariance estimates alone. It returns a posterior over the whole matrix rather than a point estimate, so interval estimates and Bayes rules under any loss follow from one object. Under rotational invariance the learned state is the spectrum while the object sampled is still the matrix.

Five limitations are substantive rather than technical, and no result here certifies accuracy from a training loss. Below $N = d$ the unrestricted problem is not identifiable, and no amount of data or training repairs it. Exact EM preserves absolute continuity, so initialization constrains what can ever be reached and point masses are fixed points. Finite ensembles change the object being optimized, and Section 4 shows the outer loop plateauing rather than improving monotonically. Endpoint behavior of the learned drift is assumed, not established. And orientation mixing has only a conservative fixed-spectrum guarantee, weak in high dimension.

AI USE STATEMENT

Generative AI tools (Claude Opus 5 and Claude Sonnet 5) were used to assist with: implementing methods and code; drafting, editing and translating parts of the manuscript; literature search and reference formatting; interpreting experimental results; providing feedback on the presentation of the experiments; and checking, restructuring and drafting parts of the mathematical statements and proofs. All AI-assisted content was reviewed and verified by the authors. The research direction, methodological choices, experimental design and all numerical results were determined and produced by the authors, who take responsibility for the final content of the paper, including all mathematical claims, proofs, code and experimental results.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the implementation of all algorithms and experiments, together with the configurations and hyperparameters used in the experiments in the following anonymized repository https://anonymous.4open.science/r/SCSI_cov-096C. The code include scripts to reproduce the main experimental results and figures reported in the paper. All theoretical results are stated with their assumptions and accompanied by proofs in the appendix.

ETHICS STATEMENT

This work uses only publicly available, de-identified data; the EEG recordings come from the PhysioNet EEG Motor Movement/Imagery dataset, released for research use. No new human-subject data was collected. The authors acknowledge and adhere to the ICLR Code of Ethics.

REFERENCES

Ignacio Alvarez, Jarad Niemi, and Matt Simpson. Bayesian inference for a covariance matrix, 2014. URL https://arxiv.org/abs/1408.4050. arXiv:1408.4050.

Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror descent with relative smoothness in measure spaces, with application to Sinkhorn and EM. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2206.08873.

Seth D. Axen. Efficiently generating inverse-Wishart matrices and their Cholesky factors, 2023. URL https://arxiv.org/abs/2310.15884. arXiv:2310.15884.

Weimin Bai, Yifei Wang, Wenzheng Chen, and He Sun. An expectation-maximization algorithm for training clean diffusion models from corrupted observations, 2024. URL https://arxiv.org/abs/2407.01014. arXiv:2407.01014.

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2012. doi: 10.1109/TBME.2011.2172210.

Yifan Chen, Mark Goldstein, Mengjian Hua, Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Probabilistic forecasting with stochastic interpolants and Föllmer processes, 2024. URL https://arxiv.org/abs/2403.13724. arXiv:2403.13724.

Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, October 2010. ISSN 1941-0476. doi: 10.1109/tsp.2010.2053029. URL http://dx.doi.org/10.1109/TSP.2010.2053029.

Matan Gavish and David L. Donoho. The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014. doi: 10.1109/TIT.2014.2323359.

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23): e215–e220, 2000.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Peter D. Hoff. Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. *Journal of Computational and Graphical Statistics*, 18(2):438–456, 2009. doi: 10.1198/jcgs.2009.07177. URL https://arxiv.org/abs/0712.4166.

Peter D. Hoff and Xiaoyue Niu. A covariance regression model. *Statistica Sinica*, 22(2):729–753, 2012. URL https://arxiv.org/abs/1102.5721.

Danial Hosseintabar, Fan Chen, Giannis Daras, Antonio Torralba, and Constantinos Daskalakis. DiffEM: Learning from corrupted data with diffusion models via expectation maximization, 2025. URL https://arxiv.org/abs/2510.12691. arXiv:2510.12691.

Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1–2):233–264, 2011. doi: 10.1007/s00440-010-0298-3. URL https://arxiv.org/abs/0911.3010.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012. doi: 10.1214/12-AOS989. URL https://arxiv.org/abs/1207.5322.

Olivier Ledoit and Michael Wolf. Analytical nonlinear shrinkage of large-dimensional covariance matrices. *The Annals of Statistics*, 48(5):3043–3065, 2020. doi: 10.1214/19-AOS1921. URL https://doi.org/10.1214/19-AOS1921.

Sho Matsumoto. General moments of the inverse real Wishart distribution and orthogonal Weingarten functions. *Journal of Theoretical Probability*, 25(3):798–822, 2012. doi: 10.1007/s10959-011-0340-0. URL https://arxiv.org/abs/1004.4717.

Chirag Modi, Jiequn Han, Eric Vanden-Eijnden, and Joan Bruna. Generative modeling from black-box corruptions via self-consistent stochastic interpolants. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2512.10857.

Radford M. Neal. MCMC using hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng (eds.), *Handbook of Markov Chain Monte Carlo*, pp. 113–162. Chapman and Hall/CRC, Boca Raton, FL, 2011.

Mehdi Rahim, Bertrand Thirion, and Gaël Varoquaux. Population shrinkage of covariance (PoSCE) for better individual brain functional-connectivity estimation. *Medical Image Analysis*, 54:138–148, 2019. doi: 10.1016/j.media.2019.03.001. URL https://doi.org/10.1016/j.media.2019.03.001.

François Rozet, Gérôme Andry, François Lanusse, and Gilles Louppe. Learning diffusion priors from observations by expectation maximization, 2024. URL https://arxiv.org/abs/2405.13712. arXiv:2405.13712.

Gerwin Schalk, Dennis J. McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R. Wolpaw. BCI2000: A general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004. doi: 10.1109/TBME.2004.827072.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian inference algorithms with simulation-based calibration, 2018. URL https://arxiv.org/abs/1804.06788. arXiv:1804.06788.

A NOTATION AND THE COVARIANCE CHANNEL

A positive-semidefinite (PSD) matrix can have zero eigenvalues; a positive-definite matrix has strictly positive eigenvalues. Throughout, $\mathrm{diag}(v)$ forms a diagonal matrix from v, spec lists eigenvalues in increasing order, $\mathbb{S}^d$ is the set of symmetric matrices, and $\|A\|_F^2 = \mathrm{tr}(A^{\mathrm{T}}A)$. The word *channel* denotes the known conditional distribution of an observation given a supplied clean covariance. To write one likelihood valid even when $N < d$, use the probability measure $\kappa_N = K_N(\cdot \mid I_d)$ as reference. Every covariance-sampling law has a density relative to it:

$$k_N(C_e \mid C) = \frac{dK_N(\cdot \mid C)}{d\kappa_N}(C_e) = (\det C)^{-N/2} \exp\left[-\frac{N}{2}\,\mathrm{tr}\left((C^{-1} - I_d)C_e\right)\right]. \tag{20}$$

Thus $K_N\pi = m_\pi\kappa_N$, where $m_\pi(C_e) = \int k_N(C_e \mid C)\,\pi(dC)$ is the normalizer in (5). This reference measure avoids treating a singular empirical covariance as if it had a density on the full positive-definite cone.

Let $X = (x_1, \ldots, x_N) \in \mathbb{R}^{d\times N}$. Relative to the Gaussian experiment with covariance I_d, its likelihood ratio is

$$\frac{p_C(X)}{p_{I_d}(X)} = (\det C)^{-N/2} \exp\left[-\frac{1}{2}\,\mathrm{tr}\left((C^{-1} - I_d)XX^{\mathrm{T}}\right)\right] = k_N(C_e \mid C). \tag{21}$$

This is a positive function of $C_e = XX^{\mathrm{T}}/N$ alone. Under the reference experiment, take the conditional law of X given C_e, written $L_N(dX \mid C_e)$. Equation (21) shows that the same conditional kernel applies under every C, and hence every mixing prior. Thus

$$\mathrm{Law}_\pi(X) = \int L_N(\cdot \mid C_e)(K_N\pi)(dC_e). \tag{22}$$

In particular, the raw-sample distribution and the empirical-covariance distribution determine one another in this family. This argument does not require $N \geq d$, nor a nonsingular observed covariance. It proves the expression in (20).

For $N \geq d$, C_e is positive definite almost surely and its Lebesgue density on $\mathbb{S}^d_{++}$ is

$$\frac{dK_N(\cdot \mid C)}{dC_e}(C_e) = \frac{N^{Nd/2}}{2^{Nd/2}\Gamma_d(N/2)}(\det C_e)^{(N-d-1)/2}(\det C)^{-N/2} \exp\left[-\frac{N}{2}\,\mathrm{tr}(C^{-1}C_e)\right], \tag{23}$$

where $\Gamma_d(a) = \pi^{d(d-1)/4}\prod_{j=1}^d \Gamma(a + (1-j)/2)$. For $N < d$ this Lebesgue density is unavailable, but (20) remains valid. All posterior formulas in the paper may be read relative to κ_N.

Centering and degrees of freedom. All stated thresholds use a known zero mean and the normalization in (1). If a Gaussian sample mean is subtracted and only the centered covariance is retained, the Wishart degrees of freedom become $N - 1$. After the corresponding known rescaling, the threshold uses those degrees of freedom rather than the raw number of observations. This does not cover a model with an additional unknown distribution of task means.

B PROOFS OF IDENTIFIABILITY AND NON-IDENTIFIABILITY

B.1 IDENTIFICATION FOR $N \geq d$

For a finite signed measure π on $\mathbb{S}^d_{++}$, define its Laplace transform at $B \succeq 0$ by $\mathcal{L}\pi(B) = \int e^{-\mathrm{tr}(BC)}\pi(dC)$. For a probability prior, the joint characteristic function of the N observations is

$$\varphi_\pi(t_1, \ldots, t_N) = \mathbb{E}\exp\left(i\sum_{j=1}^N t_j^{\mathrm{T}}x_j\right) = \mathcal{L}\pi\left(\frac{1}{2}\sum_{j=1}^N t_j t_j^{\mathrm{T}}\right). \tag{24}$$

Every PSD matrix of rank at most N has the displayed representation. If $K_N\pi_1 = K_N\pi_2$, sufficiency (22) implies equality of the raw-sample laws and hence equality of the transforms on this set. For $N \geq d$ the transforms therefore agree for every $B \succ 0$.

For completeness, uniqueness here requires no moment condition on the prior. The transform extends holomorphically to the tube $\{B + iT : B \succ 0,\ T \in \mathbb{S}^d\}$. Differentiation is justified locally by exponential damping, since $C \succ 0$ and $\|C\|_F \le \operatorname{tr} C$. To pass from the real cone to the tube, fix $B_0 \succ 0$ and $T \in \mathbb{S}^d$ and consider the one-variable map $z \mapsto L_{\pi_j}(B_0 + zT)$, holomorphic on a neighbourhood of the strip where $B_0 + \Re(z)T \succ 0$. The two maps agree for real z in an interval around the origin, so by the one-variable identity theorem they agree at $z = i$, giving equality at $B_0 + iT$ for every such T. The values at $B_0 + iT$ are the Fourier transforms of the finite tilted measures $e^{-\operatorname{tr}(B_0 C)}\pi_j(dC)$. Fourier uniqueness identifies these measures; the strictly positive tilting factor then identifies π_1 and π_2. This proves the positive assertion of Theorem 1 for all Borel probability priors.

B.2 Smooth RI counterexamples for $N < d$

Let $r = N + 1 \le d$ and let $e_r(B)$ be the rth elementary symmetric polynomial of the eigenvalues of B. It is a conjugation-invariant homogeneous polynomial, equal to the sum of the principal minors of order r. In particular,

$$e_r(B) = 0 \quad \text{if } \operatorname{rank}(B) \le N, \qquad e_r(tI_d) = \binom{d}{r} t^r > 0 \quad (t > 0). \tag{25}$$

Use the trace inner product on $\mathbb{S}^d$. The matrix gradient has diagonal entries $\partial/\partial C_{ii}$ and off-diagonal entries $\frac{1}{2}\partial/\partial C_{ij}$ in upper-triangular coordinates. This convention ensures $\nabla_C e^{-\operatorname{tr}(BC)} = -Be^{-\operatorname{tr}(BC)}$.

Choose a nonzero, nonnegative, RI function $\vartheta \in C_c^\infty(\mathbb{S}^d_{++})$. For example, a smooth radial bump around aI_d with radius less than a has these properties. Define

$$f(C) = e_r(\nabla_C)\vartheta(C). \tag{26}$$

This is smooth, compactly supported inside $\mathbb{S}^d_{++}$, and RI because replacing the derivatives in this constant-coefficient operator by a matrix B gives the conjugation-invariant polynomial $e_r(B)$. This polynomial is the operator's symbol. Extending ϑ by zero to $\mathbb{S}^d$ introduces no boundary terms. Repeated integration by parts gives, for $B \succeq 0$,

$$\mathcal{L}f(B) = e_r(B)\mathcal{L}\vartheta(B). \tag{27}$$

Indeed, the sign from integration by parts cancels the sign obtained by differentiating the exponential. Thus $\int f(C)\,dC = 0$ and $\mathcal{L}f(B) = 0$ whenever $\operatorname{rank}(B) \le N$. Also $f \not\equiv 0$: at $B = tI_d$, the right-hand side of (27) is strictly positive.

Choose a smooth compactly supported RI probability density p_0 whose support lies inside $\mathbb{S}^d_{++}$ and which is strictly positive on $\operatorname{supp} f$. Such a density is obtained from a slightly larger radial bump. By compactness, $\min_{\operatorname{supp} f} p_0 > 0$ and $\|f\|_\infty < \infty$. Hence for sufficiently small $\varepsilon > 0$,

$$p_+(C) = p_0(C) + \varepsilon f(C), \qquad p_-(C) = p_0(C) - \varepsilon f(C) \tag{28}$$

are distinct nonnegative densities of total mass one. Their supports are compact and bounded away from singular matrices. Equations (24) and (27) show that the two raw-sample characteristic functions agree for every $(t_1, \dots, t_N)$. The raw laws, and therefore their empirical-covariance laws, coincide. This proves the negative assertion, including its restriction to smooth RI priors.

The common observation distribution cannot identify the corresponding posterior either. If both posterior kernels coincided almost everywhere under that distribution, integrating them would give equal priors, contradicting $p_+ \ne p_-$. This is an obstruction to full posterior learning, not merely to a chosen representation of the prior.

B.3 A two-dimensional example

Let $d = 2$, $N = 1$, and $a, b > 0$ with $a \ne b$. Compare the RI priors

$$\begin{aligned} C^{(1)} &= R\,\mathrm{diag}(a, b)R^{\mathrm{T}}, \qquad R \sim H, \\ C^{(2)} &= cI_2, \qquad c = a\cos^2\Theta + b\sin^2\Theta, \quad \Theta \sim \mathrm{Unif}[0, 2\pi]. \end{aligned} \tag{29}$$

They are distinct: one has the fixed unequal eigenvalues (a, b), while the other has two equal random eigenvalues. For any $t \in \mathbb{R}^2$, rotational averaging gives

$$\mathbb{E}\exp\Big(-\tfrac{1}{2}t^{\mathrm{T}}C^{(1)}t\Big) = \mathbb{E}\exp\big(-\tfrac{1}{2}\|t\|_2^2 c\big) = \mathbb{E}\exp\Big(-\tfrac{1}{2}t^{\mathrm{T}}C^{(2)}t\Big)\,. \tag{30}$$

Thus one Gaussian observation has the same law under both models, as does $C_e = xx^{\mathrm{T}}$. This example uses singular prior measures for simplicity; the preceding construction establishes the stronger result for smooth full-dimensional densities.

B.4 Known-basis diagonal priors

Proposition 5 (Identification of a known-basis diagonal prior)**.** *If $C = \mathrm{diag}(c)$ in a fixed known basis, then $c_e = ((C_e)_{11}, \ldots, (C_e)_{dd})$ contains all information about c present in C_e. Its law identifies any joint distribution on $c \in (0, \infty)^d$ for every $N \geq 1$.*

Proof and forward simulator. For $C = \mathrm{diag}(c)$, the part of the raw Gaussian likelihood depending on c is

$$\prod_{j=1}^{d} c_j^{-N/2} \exp\left[-\frac{N(C_e)_{jj}}{2c_j}\right]. \tag{31}$$

It depends only on $\mathrm{diag}\, C_e$, which is therefore sufficient for c even when the full empirical covariance is singular. Write $x_{ij} = \sqrt{c_j}z_{ij}$ with all z_{ij} independent standard normals, independent of c. Then $(C_e)_{jj} = c_j V_j/N$, where $V_j = \sum_i z_{ij}^2$ are independent χ^2_N variables. For $U_j = \log c_j$ and $Z_j = \log(V_j/N)$,

$$\log(C_e)_{jj} = U_j + Z_j, \qquad \mathbb{E}e^{i\omega Z_j} = \left(\frac{2}{N}\right)^{i\omega} \frac{\Gamma(N/2 + i\omega)}{\Gamma(N/2)}. \tag{32}$$

The characteristic function of the observed log-diagonal equals the characteristic function of U multiplied by a product of nonzero factors. Division identifies the characteristic function of U, proving Proposition 5. For fixed N, the Gamma function's large-imaginary-argument asymptotics give

$$\left|\left(\frac{2}{N}\right)^{i\omega} \frac{\Gamma(N/2 + i\omega)}{\Gamma(N/2)}\right| \asymp |\omega|^{N/2-1/2} e^{-\pi|\omega|/2}, \qquad |\omega| \to \infty. \tag{33}$$

Thus high-frequency features of the log-prior distribution have exponentially reduced influence on the observed law. This proof concerns arbitrary joint laws of U, not only independent coordinates.

C Population EM: update, descent, and convergence

C.1 The prior-distribution update

For a current prior π, write $\mu = K_N\pi$ and let $w = d\mu_\star/d\mu$. The channel likelihoods in (20) are strictly positive, so all observation mixtures are equivalent to κ_N; w is therefore defined almost everywhere. Bayes' formula gives

$$\frac{d\mathcal{T}(\pi)}{d\pi}(C) = r_\pi(C), \qquad r_\pi(C) = \int w(C_e)K_N(dC_e \mid C). \tag{34}$$

The right-hand side integrates to one against π.

To identify this with the unrestricted EM step, let a candidate prior have density p relative to a common dominating measure on the permitted latent space. The part of the complete-data EM surrogate depending on p is

$$\int\int \log p(C)\, P_\pi(dC \mid C_e)\mu_\star(dC_e) = \int \log p(C)\, \mathcal{T}(\pi)(dC). \tag{35}$$

Its maximizer is the density of $\mathcal{T}(\pi)$ whenever the cross-entropy expression is finite; equivalently, it minimizes $\mathrm{KL}(\mathcal{T}(\pi)\|p)$. The measure-valued posterior average (6) remains meaningful beyond a global density parametrization. This is the latent-distribution EM update of Aubin-Frankowski et al. (2022). Conditional synthetic training implements the next posterior calculation and is not a different maximizer of this surrogate.

C.2 Proof of Theorem 2

Define joint laws

$$\begin{aligned} J_\star(dC, dC_e) &= \pi_\star(dC)K_N(dC_e \mid C), \\ J_k(dC, dC_e) &= \pi_k(dC)K_N(dC_e \mid C), \\ H_k(dC, dC_e) &= P_k(dC \mid C_e)\mu_\star(dC_e). \end{aligned} \tag{36}$$

Because $J_\star$ and J_k have the same forward conditional, $\mathrm{KL}(J_\star \| J_k) = D_k$. Disintegrating them by C_e instead proves (7). The C-marginals of $J_\star$ and H_k are $\pi_\star$ and π_{k+1}, respectively, and

$$\mathrm{KL}(J_\star \| H_k) = \int \mathrm{KL}(P_\star(\cdot \mid C_e) \| P_k(\cdot \mid C_e))\mu_\star(dC_e) = D_k - F_k. \tag{37}$$

Data processing yields $D_{k+1} \le D_k - F_k$. Starting with finite D_0 ensures that all these quantities are finite at every iteration.

For observation-space monotonicity, put $r_k = d\pi_{k+1}/d\pi_k$. Bayes' formula gives

$$\frac{m_{\pi_{k+1}}(C_e)}{m_{\pi_k}(C_e)} = \int r_k(C) P_k(dC \mid C_e). \tag{38}$$

Jensen's inequality then implies

$$\begin{aligned} F_k - F_{k+1} &= \int \log \frac{m_{\pi_{k+1}}(C_e)}{m_{\pi_k}(C_e)} \mu_\star(dC_e) \\ &\ge \int\int \log r_k(C) P_k(dC \mid C_e)\mu_\star(dC_e) = \mathrm{KL}(\pi_{k+1} \| \pi_k). \end{aligned} \tag{39}$$

The latter KL is finite because $\mathrm{KL}(H_k \| J_k) = F_k$ and marginalization decreases KL. Standard truncation justifies Jensen if individual log terms are not initially bounded. Summing $D_{k+1} \le D_k - F_k$ gives $\sum_{j=0}^{k} F_j \le D_0$, and monotonicity gives $(k+1)F_k \le D_0$.

C.3 Proof of the recovery statements in Theorem 2

For compact-support recovery, let $\mathcal{K} \subset \mathbb{S}^d_{++}$ be the common compact set supporting π_0 and $\pi_\star$. Support preservation keeps every π_k in $\mathcal{K}$. Probability measures on $\mathcal{K}$ are weakly sequentially compact. The channel is weakly continuous there: for a bounded continuous function f of the observation,

$$C \longmapsto \mathbb{E}_Z f\Big(C^{1/2} Z Z^{\mathrm{T}} C^{1/2}/N\Big) \tag{40}$$

is continuous by continuity of the positive square root and dominated convergence. Any subsequential weak limit π of π_k therefore satisfies $K_N\pi_k \Rightarrow K_N\pi$ along the subsequence. But $F_k \to 0$ implies $K_N\pi_k \to \mu_\star$ in total variation by Pinsker's inequality. Thus $K_N\pi = \mu_\star$. For $N \ge d$, Theorem 1 identifies $\pi = \pi_\star$, so every subsequential limit agrees. The same proof applies to any common compact class on which the channel is injective. This proves $\pi_k \Rightarrow \pi_\star$ without a latent-KL conclusion.

To obtain the conditional statement, fix an admissible observed matrix C_e. The function $\ell_{C_e}(C) = k_N(C_e \mid C)$ in (20) is continuous, strictly positive, and bounded above and away from zero on $\mathcal{K}$. For every continuous f on $\mathcal{K}$, weak prior convergence therefore gives

$$\int f(C) P_k(dC \mid C_e) = \frac{\int f(C)\ell_{C_e}(C)\pi_k(dC)}{\int \ell_{C_e}(C)\pi_k(dC)} \longrightarrow \frac{\int f(C)\ell_{C_e}(C)\pi_\star(dC)}{\int \ell_{C_e}(C)\pi_\star(dC)}. \tag{41}$$

The denominator has a strictly positive limit. The likelihood-defined versions of the conditional laws thus converge weakly for every such C_e, and in particular for $\mu_\star$-almost every observation. This argument extends the compact-prior convergence proof to the posterior itself; it does not show convergence for all measurable events or in KL. For any fixed bounded continuous test function, dominated convergence also yields convergence of its conditional-expectation discrepancy averaged over $\mu_\star$. Since covariance and precision entries are bounded continuous on $\mathcal{K}$, their posterior means converge under these same compactness assumptions.

For the stronger statement, the inverse-stability premise gives $D_{k+1} \le D_k - F_k \le (1 - \chi^{-1})D_k$, and induction gives the geometric bound. The posterior identity in Theorem 2 gives the second part

because $0 \le D_k - F_k \le D_k$. For $\chi = 1$ the first update is already exact; the displayed geometric convention at $k = 0$ is understood as the initial bound D_0. This inverse-stability condition is not proved for unrestricted Wishart mixtures and is not inferred from injectivity. It is a separate sufficient hypothesis on the iterates and model class.

C.4 Support restriction and non-optimal fixed points

Equation (34) proves $\pi_{k+1} \ll \pi_k$. If $\pi_k = \delta_{C_0}$, then $P_k(\cdot \mid C_e) = \delta_{C_0}$ for every admissible observation, so $\mathcal{T}(\pi_k) = \pi_k$ even when it does not explain $\mu_\star$. These fixed points do not contradict Theorem 2: finite $\mathrm{KL}(\pi_\star \| \delta_{C_0})$ would force the truth itself to be that point mass. Conversely, a continuous initial density can have full topological support and still fail to dominate an atomic truth. These distinctions are important when interpreting particle implementations and nominally expressive generative models.

D Local information and tractable modes

D.1 Linearization in a specified norm

Under the true joint law define

$$A_N h(C_e) = \mathbb{E}[h(C) \mid C_e], \qquad A_N^* g(C) = \mathbb{E}[g(C_e) \mid C]. \tag{42}$$

For a bounded perturbation $\pi_h = (1+h)\pi_\star$ with $\mathbb{E}_{\pi_\star} h = 0$ and small $\|h\|_\infty$, the following calculation proves

$$\frac{d\mathcal{T}(\pi_h)}{d\pi_\star} - 1 = (\mathrm{Id} - A_N^* A_N)h + O(\|h\|_\infty^2). \tag{43}$$

Write $A = A_N$. For bounded centered h with $\delta = \|h\|_\infty < 1$, the observation law of $\pi_h = (1+h)\pi_\star$ has density $1 + Ah$ relative to $\mu_\star$. Substituting this into (34) gives

$$\frac{d\mathcal{T}(\pi_h)}{d\pi_\star} = (1+h) A_N^* \left[\frac{1}{1 + A_N h} \right]. \tag{44}$$

Using $(1+u)^{-1} = 1 - u + u^2/(1+u)$ gives

$$\frac{d\mathcal{T}(\pi_h)}{d\pi_\star} - 1 = (\mathrm{Id} - A^*A)h - hA^*Ah + (1+h)A^* \left[\frac{(Ah)^2}{1+Ah} \right]. \tag{45}$$

Conditional expectation contracts L^∞, so the last two terms have norm at most $2\delta^2/(1-\delta)$. This proves the L^∞ expansion in (43). The derivative $\mathrm{Id} - A^*A$ extends continuously to $L_0^2(\pi_\star)$, where A^* is the Hilbert adjoint because

$$\mathbb{E}[(Ah)(C_e) g(C_e)] = \mathbb{E}[h(C) g(C_e)] = \mathbb{E}[h(C)(A^* g)(C)]. \tag{46}$$

Conditional Jensen yields $\|Ah\|_2 \le \|h\|_2$, hence $0 \le A^*A \le \mathrm{Id}$. The bounded-perturbation proof is not a claim of Fréchet differentiability in the unrestricted L^2 topology. If bounded nonzero h satisfies $Ah = 0$, the positive priors $(1 \pm \varepsilon h)\pi_\star$ have exactly the same observation law for sufficiently small ε.

D.2 Inverse stability and the local spectrum

The inverse-stability hypothesis $D_k \le \chi F_k$ of Theorem 2 can be read off the same operator. For $\pi = (1+h)\pi_\star$ with $\mathbb{E}_{\pi_\star} h = 0$ and small $\|h\|_\infty$, expanding $-\mathbb{E}_{\pi_\star} \log(1+h)$ and the corresponding expression for the observation law, whose density relative to $\mu_\star$ is $1 + A_N h$, gives

$$\mathrm{KL}(\pi_\star \| \pi) = \tfrac{1}{2}\|h\|_2^2 + O(\|h\|_\infty^3), \qquad \mathrm{KL}(\mu_\star \| K_N \pi) = \tfrac{1}{2}\langle h, A_N^* A_N h \rangle + O(\|h\|_\infty^3). \tag{47}$$

Writing $\eta = \inf\{\langle h, A_N^* A_N h \rangle : \|h\|_2 = 1\}$ over the admissible perturbations, the smallest χ compatible with (47) near $\pi_\star$ is η^{-1}, and $\chi \ge 1$ because $A_N^* A_N \le \mathrm{Id}$. The infimum need not be attained, and it need not be positive: $\eta = 0$ with $A_N h = 0$ for some bounded $h \ne 0$ is non-identifiability, in agreement with the counterexample of Theorem 1, while directions carrying arbitrarily little information give η arbitrarily small and recovery arbitrarily slow. A uniform positive lower bound therefore need not exist on an unrestricted nonparametric class, and inverse stability is a restriction on the class of admissible priors rather than a consequence of $N \ge d$. The Gaussian experiment has finite Fisher information at each fixed C, but a pointwise information bound does not give a uniform inverse-stability bound for a distribution of covariances, which is an infinite-dimensional inverse problem.

D.3 Monotonicity with the number of observations

Couple all sample sizes through one covariance $C \sim \pi_\star$ and an infinite conditionally iid sequence $x_1, x_2, \ldots$. For $\mathcal{F}_N = \sigma(x_1, \ldots, x_N)$ and centered $h \in L^2(\pi_\star)$, sufficiency implies

$$\mathbb{E}[h(C) \mid \mathcal{F}_N] = \mathbb{E}[h(C) \mid C_{e,N}] = A_N h(C_{e,N}). \tag{48}$$

Orthogonal projection onto nested sigma-fields yields

$$\|A_N h\|_2^2 \le \|A_{N+1} h\|_2^2 \le \|h\|_2^2. \tag{49}$$

Conditional strong laws imply $C_{e,N} \to C$ almost surely, so $h(C)$ is measurable with respect to $\mathcal{F}_\infty = \sigma(\bigcup_N \mathcal{F}_N)$, up to completion. The L^2 martingale convergence theorem therefore gives $A_N h(C_{e,N}) \to h(C)$ in L^2. There is no rate uniform in h in this argument.

D.4 An explicit scalar mode

Proposition 6 (Inverse-gamma information mode)**.** *Let $d = 1$ and $c \sim \mathrm{InvGamma}(a, b)$ with density proportional to $c^{-a-1}e^{-b/c}$, $a > 2$, $b > 0$. Put $n = N/2$ and $h(c) = c - b/(a-1)$. Then*

$$A_N^* A_N h = \frac{n}{a+n-1} h, \qquad (\mathrm{Id} - A_N^* A_N) h = \frac{a-1}{a+n-1} h. \tag{50}$$

This is one exact information mode, not a worst-case convergence rate. Since h is unbounded, it is an identity for the square-integrable extension of the linear operator, not an application of the bounded-perturbation remainder to h itself.

Proof. Let $n = N/2$. The scalar channel has likelihood proportional to $c^{-n}e^{-nC_e/c}$ as a function of c. Multiplying by $c^{-a-1}e^{-b/c}$ yields

$$c \mid C_e \sim \mathrm{InvGamma}(a+n, b+nC_e), \qquad \mathbb{E}[c \mid C_e] = \frac{b+nC_e}{a+n-1}. \tag{51}$$

With $\bar{c} = b/(a-1)$ this gives

$$A_N h(C_e) = \frac{n}{a+n-1}(C_e - \bar{c}), \qquad A_N^* A_N h(c) = \frac{n}{a+n-1}\big(\mathbb{E}[C_e \mid c] - \bar{c}\big) = \frac{n}{a+n-1} h(c). \tag{52}$$

The condition $a > 2$ guarantees $h \in L^2(\pi_\star)$; in fact $\mathrm{Var}(c) = b^2/[(a-1)^2(a-2)]$. For a direct numerical test of the derivative, bounded centered truncations of h avoid using the L^∞ remainder outside its domain.

D.5 Tensor products for a factorized diagonal prior

If the diagonal prior factorizes, both the joint experiment and its posterior factorize over coordinates. On product functions $h(c) = \prod_{j=1}^d h_j(c_j)$, conditional expectation consequently satisfies $Ah = \prod_j A_j h_j$. Thus $A^*A = \bigotimes_j (A_j^* A_j)$ on the Hilbert tensor product. If $A_j^* A_j h_j = \eta_j h_j$, the product is an eigenfunction with eigenvalue $\prod_j \eta_j$. Constant factors have eigenvalue one; modes involving several nonconstant factors may have much smaller information eigenvalues. This tensorization assertion requires a factorized prior and is not assumed in Proposition 5.

E Supervised conditional sampling and the Föllmer process

This appendix supplies the proofs and general-schedule details for the conditional sampler specified in Section 3. Both main algorithms repeatedly use this supervised construction on synthetic pairs, in their respective dimensions. The Brownian-interpolant construction and its tunable diffusion follow Chen et al. (2024); we give the algebra in covariance-compatible coordinates and state the qualifications needed here. The distribution-level EM results do not require this particular sampler.

Algorithm 1: Self-consistent covariance reconstruction in matrix-logarithm coordinates

Input: Recorded observations $C_{e,1:M}$; sample size N; sampleable initialization π_0; noise scale $\sigma > 0$; outer budget K.

Output: A conditional sampler for the covariance posterior at any supplied observation.

Encode the recorded observations, $Y_{e,a} = \Phi(C_{e,a})$ for $a = 1, \dots, M$

Draw $C \sim \pi_0$; form C'_e from C by (3); record the pair $(Y'_e, \Phi(C))$

Fit $\widehat{b}_0$ by minimizing (10) on these pairs

for $k = 0, \dots, K-1$ **do**

`/* E-step: reconstruct at recorded observations */`

Sample $Y_{e,a}$ from $Y_{e,1:M}$; integrate (12) with $\widehat{b}_k$; decode $C = \Phi^{-1}(Y_1)$

`/* Re-corrupt: manufacture the next pairs */`

Form C'_e from each reconstructed C by (3); record $(Y'_e, \Phi(C))$

Fit $\widehat{b}_{k+1}$ by minimizing (10) on these pairs, their values held fixed under differentiation

return $\widehat{b}_K$*; one posterior draw at* C_e*: integrate* (12)*, decode* $\Phi^{-1}(Y_1)$

E.1 Coordinates, decoding, and scope

Sections 3 and 3.2 use Y_e for the retained observation expressed in numerical coordinates. We use that same conditioning argument throughout this appendix. Since the input encoding is injective, conditioning on Y_e is equivalent to conditioning on the retained observation. In the spectral and known-basis diagonal versions, that retained observation is Λ_e and c_e, respectively. A different deterministic initialization, such as zero, is also possible provided the observation is supplied separately to the drift.

For full matrices, svec is the linear isometry from $\mathbb{S}^d$ onto $\mathbb{R}^p$ obtained from the Frobenius-orthonormal symmetric basis. In particular $\|\operatorname{svec}(A)\|_2 = \|A\|_F$ for every symmetric A, whether or not A is positive definite. We set $Y = \Phi(C) = \operatorname{svec}(\log C)$ and $C = \Phi^{-1}(Y) = \exp(\operatorname{svec}^{-1} Y)$. This is a bijection from $\mathbb{S}^d_{++}$ to $\mathbb{R}^{d(d+1)/2}$. If $\rho_j = \operatorname{Law}(\Phi(C))$ for $C \sim \pi_j$, then $d\rho_1/d\rho_2(y) = d\pi_1/d\pi_2(\Phi^{-1}(y))$ whenever these derivatives exist. Integrating their logarithms gives equality of KL divergences; applying the inverse map also shows that posterior averaging commutes with the coordinate transformation. When absolute continuity fails, the corresponding KL values on both sides are infinite. This proves the coordinate-invariance assertion of Section 3.1. The same argument applies to the componentwise logarithm in the known-basis diagonal model.

For $N \geq d$, one can take $Y_e = \operatorname{svec}(\log C_e)$. If $N < d$, $Y_e = \operatorname{svec}(\log(C_e + \epsilon_0 I_d))$ with fixed $\epsilon_0 > 0$ is an injective encoding: $C_e = \exp(\operatorname{svec}^{-1} Y_e) - \epsilon_0 I_d$. It changes neither the Wishart likelihood nor the information in C_e. In particular, the shift cannot resolve the non-identifiability of Theorem 1.

Frobenius Brownian motion on symmetric matrices has independent standard Brownian coefficients in the basis E_{ii} and $(E_{ij} + E_{ji})/\sqrt{2}$, $i < j$. Its off-diagonal entries have variance $t/2$, not t. This is equivalent to standard Brownian motion in the svec coordinates. Additive noise does not preserve positive definiteness in the original matrix space. In log coordinates, every finite decoded numerical output is positive definite and can be passed through the next forward simulation. This is a numerical advantage; exact posterior inference is unchanged by the bijection.

For RI, the primary target encoding in Section 3.2 is $Y_i = \log \lambda_i$ with increasing eigenvalue order. It is one-to-one on the ordered spectral state space, but that space is a subset of $\mathbb{R}^d$. The exact endpoint stays in it; a numerical endpoint may be unordered. Decoding by sorting $(e^{Y_i})_i$ always gives a positive ordered spectrum, and this same order is used to form the next training target. Sorting is applied after sampling, not by modifying the interior SDE. It has no effect on the exact endpoint. For an approximate endpoint, any deterministic decoding can only decrease its KL discrepancy from the exact decoded law by data processing. The spectral version of the one-update bound therefore remains an inequality; it need not use equality of KL under the extended sorting map.

An alternative for *strictly* ordered spectra uses log-positive-gap coordinates:

$$Y_1 = \log \lambda_1, \qquad Y_j = \log(\lambda_j - \lambda_{j-1}) \quad (j \geq 2), \qquad \lambda_j = \sum_{i=1}^{j} e^{Y_i}. \tag{53}$$

This gap parametrization is a bijection onto $\mathbb{R}^d$ and enforces strict ordering for every output. It cannot represent repeated eigenvalues exactly. Plain ordered log eigenvalues can represent repetitions, but their law may then have lower-dimensional support. Neither representation alone supplies the smooth-density and endpoint-relative-entropy hypotheses below. For observed spectra with zeros, the injective input $(\log(\lambda_{e,i} + \epsilon_0))_{i=1}^d$ remains valid regardless of which clean spectral coordinates are chosen. The drift is joint across coordinates in both versions.

For the following identities, fix Y_e and write $q(\cdot \mid Y_e)$ for the conditional density of Y. Assume the stated conditional expectations exist, the interior marginal densities and drifts are sufficiently regular for the Fokker–Planck calculation, and the relevant SDE and Fokker–Planck solutions are unique in the considered class. Endpoint assertions require convergence to the intended endpoint law; finite trajectory-KL assertions additionally require the Girsanov absolute-continuity and integrability conditions. Girsanov's formula expresses the likelihood ratio between SDE trajectory distributions with the same noise coefficient and different drifts. These are assumptions on the sampler, not consequences of arbitrary $\pi_\star$. Atomic or lower-dimensional target priors can satisfy the EM theory while falling outside these density and finite-entropy conditions.

E.2 Brownian interpolant and drift regression

With W_t standard Brownian motion in $\mathbb{R}^p$, independent of (Y_e, Y), define

$$I_t = \alpha_t Y_e + \beta_t Y + \sigma_t W_t, \qquad R_t = \dot{\alpha}_t Y_e + \dot{\beta}_t Y + \dot{\sigma}_t W_t, \tag{54}$$

where $\alpha_0 = \beta_1 = 1$ and $\alpha_1 = \beta_0 = \sigma_1 = 0$. Take $\beta_t, \sigma_t > 0$ for $0 < t < 1$ and schedules differentiable there with suitable endpoint limits. Since $W_0 = 0$, σ_0 need not vanish. The minimizer of the pointwise conditional square loss is

$$b_t(y, Y_e) = \mathbb{E}[R_t \mid I_t = y, Y_e]. \tag{55}$$

Equivalently, it minimizes (10) whenever that loss is finite, which holds for regular schedules with finite integrated second moments, and the minimizer is unchanged under any positive reweighting of t.

Itô's formula gives $dI_t = R_t\, dt + \sigma_t\, dW_t$. Thus, for a smooth compactly supported test function f,

$$\frac{d}{dt}\mathbb{E}[f(I_t) \mid Y_e] = \mathbb{E}[\nabla f(I_t)^{\mathrm{T}} b_t(I_t, Y_e) + \tfrac{1}{2}\sigma_t^2 \Delta f(I_t) \mid Y_e]. \tag{56}$$

Writing $\rho_t(\cdot \mid Y_e) = \mathrm{Law}(I_t \mid Y_e)$, the conditional density solves

$$\partial_t \rho_t = -\nabla \cdot (b_t \rho_t) + \tfrac{1}{2}\sigma_t^2 \Delta \rho_t. \tag{57}$$

Uniqueness then identifies these marginals with those of $dY_t = b_t(Y_t, Y_e)dt + \sigma_t dB_t$, $Y_0 = Y_e$, where B_t is a fresh Brownian motion. Its endpoint is $q(\cdot \mid Y_e)$ under the stated endpoint assumptions. No derivative of Brownian motion appears in the regression target. This proves the base-SDE part of Proposition 3 after applying the decoding map (and Φ^{-1} in the full-matrix case). The drift correction below proves the displayed Föllmer version. The same projection argument holds for synthetic spectral or diagonal pairs because it only uses their joint law, not a matrix-specific likelihood. Because $k_N(C_e \mid C) > 0$ for every SPD C and κ_N-almost every observation, every probability mixture $K_N\pi$ is equivalent to κ_N. Thus the conditional identities established for synthetic observations hold $\mu_\star$-almost everywhere as well. Equivalence alone does not bound their approximation error after this change of observation distribution.

For any admissible fitted drift, conditional orthogonality also gives

$$\mathbb{E}[\|\widehat{b}_t(I_t, Y_e') - R_t\|_2^2] = \mathbb{E}[\|b_t^\pi(I_t, Y_e') - R_t\|_2^2] + e_t(\pi, \widehat{b}). \tag{58}$$

Integrating with a positive time weight proves the excess-risk identity used in Appendix F. The weights are deterministic, so whenever the objectives are finite they do not change the unrestricted pointwise minimizer. A finite architecture or finite optimization budget can of course react differently to different weights.

E.3 Tunable diffusion and recovery of the score

The volatility is the coefficient multiplying Brownian noise. For a suitable deterministic scalar volatility g_t, set

$$b_t^g(y, Y_e) = b_t(y, Y_e) + \tfrac{1}{2}(g_t^2 - \sigma_t^2)\nabla_y \log \rho_t(y \mid Y_e). \tag{59}$$

Substitution into the Fokker–Planck equation shows that $dY_t^g = b_t^g(Y_t^g, Y_e)dt + g_t dB_t$ has the same marginals whenever this equation and its initial condition are well posed. The drift correction is essential. In particular, setting $g = 0$ does not provide a regular deterministic ODE that spreads a point mass; singular behavior would have to be treated separately.

The score is the gradient of the log conditional density, $\nabla_y \log \rho_t$. It can be recovered from b_t when the following denominator is nonzero:

$$\zeta_t = \big[t\sigma_t(\dot\beta_t\sigma_t - \beta_t\dot\sigma_t)\big]^{-1}, \qquad j_t(y, Y_e) = \dot\beta_t y + (\beta_t\dot\alpha_t - \dot\beta_t\alpha_t)Y_e. \tag{60}$$

Indeed, Gaussian differentiation and (54) give

$$\nabla_y \log \rho_t(y \mid Y_e) = -\frac{\mathbb{E}[W_t \mid I_t = y, Y_e]}{t\sigma_t}, \tag{61}$$

$$\beta_t b_t(y, Y_e) - j_t(y, Y_e) = (\beta_t\dot\sigma_t - \dot\beta_t\sigma_t)\mathbb{E}[W_t \mid I_t = y, Y_e], \tag{62}$$

so

$$\nabla_y \log \rho_t(y \mid Y_e) = \zeta_t[\beta_t b_t(y, Y_e) - j_t(y, Y_e)]. \tag{63}$$

This algebra permits changing g after learning b. It does not remove the need to handle apparent endpoint singularities in ζ_t. The endpoint regularity conditions in Chen et al. (2024), or directly verified conditions for a chosen schedule, are needed in applying these identities on the closed time interval.

E.4 Path-space error and the general Föllmer choice

Define the approximate $\widehat{b}^g$ by substituting $\widehat{b}$ into both (59) and (63). Then

$$\widehat{b}_t^g - b_t^g = \big[1 + \tfrac{1}{2}\beta_t\zeta_t(g_t^2 - \sigma_t^2)\big](\widehat{b}_t - b_t). \tag{64}$$

Let P^g and $\widehat{\mathsf{P}}^g$ denote the conditional path laws, meaning probability distributions over entire trajectories, of the exact and approximate SDEs with the same initial point and volatility. Under Girsanov's hypotheses,

$$\mathrm{KL}(\mathsf{P}^g \| \widehat{\mathsf{P}}^g) = \frac{1}{2}\int_0^1 \frac{[1 + \frac{1}{2}\beta_t\zeta_t(g_t^2 - \sigma_t^2)]^2}{g_t^2} L_t\, dt, \qquad L_t = \mathbb{E}[\|\widehat{b}_t(I_t, Y_e) - b_t(I_t, Y_e)\|_2^2 \mid Y_e]. \tag{65}$$

The same identity holds on a truncated interval under the corresponding restricted hypotheses. Marginalization bounds endpoint KL by this path KL. Since L_t does not depend on g_t, minimize the prefactor pointwise: writing $a = \beta_t\zeta_t/2$ and $c = 1 - a\sigma_t^2$, the relevant function is $(c + ag_t^2)^2/g_t^2$. For nondegenerate a, c, its minimum occurs at $g_t^2 = |c/a|$. Hence

$$(g_t^{\mathrm{F}})^2 = \left|2t\sigma_t\left(\frac{\dot\beta_t}{\beta_t}\sigma_t - \dot\sigma_t\right) - \sigma_t^2\right|. \tag{66}$$

Zero or singular minimizing values need a limiting or separate well-posedness argument. When $\beta_t/(\sqrt{t}\sigma_t)$ is increasing, the quantity inside the absolute value is nonnegative. For the implemented choice $\alpha_t = 1 - t$, $\beta_t = t$, and $\sigma_t = \sigma(1 - t)$, this gives $(g_t^{\mathrm{F}})^2 = \sigma^2(1 - t^2)$.

The Föllmer process minimizes relative entropy of the trajectory distribution, subject to its prescribed endpoint law and relative to a given reference process. For general schedules that reference is *specified*, not automatically standard Brownian motion. With $v_t = \beta_t^2 + t\sigma_t^2$, the reference in Chen et al. (2024) is

$$dZ_t = \big[\dot\alpha_t Y_e + a_t(Z_t - \alpha_t Y_e)\big]dt + g_t^{\mathrm{F}} dB_t, \qquad a_t = \frac{d}{dt}\log\frac{v_t}{\beta_t}, \qquad Z_0 = Y_e. \tag{67}$$

Under the required regularity, monotonicity, and finite terminal relative entropy, the exact process using (66) is the path-relative-entropy minimizer with terminal density $q(\cdot \mid Y_e)$ for this reference. This is distinct from optimizing the outer EM iteration. It does not establish that g^{F} is optimal for numerical discretization, covariance risk, or finite-data posterior calibration.

E.5 Closed-form Föllmer drift for the main implementation

For the positive scale retained throughout the main implementation, $\alpha_t = 1 - t$, $\beta_t = t$, and $\sigma_t = \sigma(1 - t)$ with $\sigma > 0$, the score-recovery coefficients reduce to $\zeta_t = [\sigma^2 t(1 - t)]^{-1}$ and $j_t(y, Y_e) = y - Y_e$. Formula (66) gives $(g_t^{\mathrm{F}})^2 = \sigma^2(1 - t^2)$, so

$$\tfrac{1}{2}\big((g_t^{\mathrm{F}})^2 - \sigma_t^2\big) = \sigma^2 t(1 - t). \tag{68}$$

Substituting the recovered score into (59) therefore yields

$$b_t^{\pi,\mathrm{F}} = b_t^{\pi} + \sigma^2 t(1 - t)\zeta_t[t b_t^{\pi} - y + Y_e] = (1 + t)b_t^{\pi} - y + Y_e. \tag{69}$$

The same algebra gives the fitted drift in (12) for every $\sigma > 0$. The scale cancels from this explicit drift correction, but changes the training distribution and hence the learned function itself. Unlike the general score formula, this expression has no explicit division by t or $1 - t$, and requires one evaluation of the fitted drift per integration step. The process remains subject to the regularity and integrability hypotheses stated above. Comparing its time grid and error with the base sampler is an empirical and numerical-analysis question, not resolved by the continuous-time path-KL optimum.

E.6 Choosing the scale of the interpolation noise

The scale σ is an algorithmic parameter in the chosen coordinates, not an additional variance in the observation model. For every fixed $\sigma > 0$ satisfying the sampler assumptions, exact conditional training gives the same posterior endpoint and hence the same population EM map. A dependence on N is nevertheless natural for the numerical problem.

In the scalar model, $C_e/C = V/N$ with $V \sim \chi^2_N$. Put $Z_N = \log(V/N)$. Its moment-generating function near zero is

$$\mathbb{E}e^{uZ_N} = \left(\frac{2}{N}\right)^u \frac{\Gamma(N/2 + u)}{\Gamma(N/2)}. \tag{70}$$

Twice differentiating its logarithm at $u = 0$ yields

$$\mathrm{Var}(Z_N) = \psi_1(N/2) = \frac{2}{N} + O(N^{-2}), \tag{71}$$

where ψ_1 is the second derivative of $\log\Gamma$, called the trigamma function. For example its series $\psi_1(a) = \sum_{j=0}^{\infty}(a + j)^{-2}$ and an integral comparison give $\psi_1(a) = a^{-1} + O(a^{-2})$. This motivates a *standard-deviation* scale $\sigma = a\sqrt{2/N}$, not a/N. The constant a includes the time profile of the interpolation variance, $\sigma^2 t(1 - t)^2$, and must be chosen numerically. This is a scale comparison, not an optimality theorem for posterior learning.

The same per-coordinate scale arises near an isotropic covariance at fixed d. Set $C = cI_d$ and $E_N = C_e/c - I_d$. Gaussian sampling gives $\sqrt{N}\ \mathrm{svec}(E_N) \Rightarrow \mathcal{N}(0, 2I_p)$, $p = d(d + 1)/2$, using the Frobenius normalization of svec. Since $\log C_e - \log C = E_N + O_P(N^{-1})$, it follows that

$$\sqrt{N}\ \mathrm{svec}(\log C_e - \log C) \Rightarrow \mathcal{N}(0, 2I_p). \tag{72}$$

This fixed-dimensional local limit does not make matrix-log errors isotropic for arbitrary C, or remain uniform as d/N stays positive. Its leading total squared coordinate scale is proportional to $2p/N$, not $2/N$.

For an RI prior, the observed log eigenvalues do not obey independent scalar chi-square channels. Gaps, repetitions, and d/N affect their noise, so the exact spectral simulator in (17) is the appropriate calibration tool. Re-corruption of draws from the current prior can assess this scale without clean draws from $\pi_\star$. Smaller σ is not automatically better: the conditional approximation bounds contain $1/\sigma^2$, and both regression and numerical integration change with the scale. The ideal EM theorems are independent of this numerical tuning.

E.7 Numerical sampling

The main implementation integrates the Föllmer SDE (12). More generally, once a compatible fitted drift $\widehat{b}^g$ and noise coefficient g have been chosen, Euler–Maruyama uses a grid $0 = t_0 < \cdots < t_J = 1$ and independent $z_j \sim \mathcal{N}(0, I_p)$:

$$Y_{j+1} = Y_j + \widehat{b}^g_{t_j}(Y_j, Y_e)\Delta t_j + g_{t_j}\sqrt{\Delta t_j}\, z_j, \qquad Y_0 = Y_e. \tag{73}$$

The observation Y_e stays fixed throughout a path; independent paths use independent Gaussian draws. Decode only the endpoint, using Φ^{-1} for full matrices and sorted exponentiation for spectra. The base coefficients are $(\widehat{b}, \sigma(1-t))$; the Föllmer coefficients are $((1+t)\widehat{b} - y + Y_e, \sigma\sqrt{1-t^2})$. The numerical error term in the supplementary reconstruction bounds concerns the difference between this discrete endpoint law and the continuous-time law. Choosing a step size does not itself certify that error.

F Drift approximation, outer-update error, and finite ensembles

A learned conditional kernel need not equal P_ρ for any prior ρ under the fixed channel. Nevertheless, its reconstructions define an actual output law. The purpose of this section is to compare that output with the exact EM update of the *prior supplying its synthetic training pairs*, using the regression error of the conditional interpolant. All population-risk quantities below are evaluated at the fitted network; no assumption of exact finite-data optimization is made. The ideal identifiability and EM recovery results do not rely on the bounds in this appendix.

F.1 Supplementary one-update bound from drift regression

For a current prior π, define the excess drift error and its Föllmer-weighted integral by

$$e_t = \mathbb{E}_{\pi K_N, W} \|\widehat{b}_t(I_t, Y_e') - b_t^\pi(I_t, Y_e')\|_2^2, \tag{74}$$

$$\mathcal{E}_\pi^{\mathrm{F}} = \frac{1}{2\sigma^2} \int_0^1 \frac{1+t}{1-t} e_t \, dt. \tag{75}$$

Conditional-expectation orthogonality gives $\mathcal{L}_\pi(\widehat{b}) - \mathcal{L}_\pi(b^\pi) = \int_0^1 e_t \, dt$. Thus e_t excludes the unavoidable random variation of the regression target. The singular endpoint weights are not automatically controlled by a small unweighted training loss.

Theorem 7 (From drift regression to a reconstruction update)**.** *Assume the endpoint and change-of-drift hypotheses in Appendix E, finite $\mathcal{E}_\pi^{\mathrm{F}}$, and $d\mu_\star/d\mu_\pi \le B_\pi < \infty$. Let $\widehat{P}^{\mathrm{F}}$ be the learned continuous-time covariance kernel and $\widehat{\pi}^+ = \int \widehat{P}^{\mathrm{F}}(\cdot \mid C_e)\mu_\star(dC_e)$. Then*

$$\int \mathrm{KL}(P_\pi(\cdot \mid C_e) \| \widehat{P}^{\mathrm{F}}(\cdot \mid C_e)) \mu_\pi(dC_e) \le \mathcal{E}_\pi^{\mathrm{F}}, \tag{76}$$

$$\mathrm{KL}(\mathcal{T}(\pi) \| \widehat{\pi}^+) \le B_\pi \mathcal{E}_\pi^{\mathrm{F}}. \tag{77}$$

The corresponding result for other admissible noise coefficients is given in Appendix E.

The overlap bound B_π transfers synthetic-training error to real observations; a positive likelihood does not imply such a bound. The proof combines the trajectory calculation of Chen et al. (2024) with posterior averaging (Appendix F).

To include error in the current prior and numerical integration, write $D(\pi) = \mathrm{KL}(\pi_\star \| \pi)$, $F(\pi) = \mathrm{KL}(\mu_\star \| \mu_\pi)$, and let $\widehat{P}^{\mathrm{F}}_{\mathrm{num}}$ be the numerical endpoint law. Total variation, $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$, measures the largest discrepancy in event probabilities. If $D(\pi) < \infty$ and $\mathbb{E}_{\mu_\star} \mathrm{TV}(\widehat{P}^{\mathrm{F}}, \widehat{P}^{\mathrm{F}}_{\mathrm{num}}) \le \delta_{\mathrm{num}}$, the KL identity and Pinsker's inequality give

$$\boxed{\mathbb{E}_{\mu_\star} \mathrm{TV}(P_\star, \widehat{P}^{\mathrm{F}}_{\mathrm{num}}) \le \sqrt{\frac{D(\pi) - F(\pi)}{2}} + \sqrt{\frac{B_\pi \mathcal{E}_\pi^{\mathrm{F}}}{2}} + \delta_{\mathrm{num}}.} \tag{78}$$

Kernels in this display are evaluated at the same C_e. The terms separate prior learning, conditional regression, and integration error. They are theoretical quantities, not certificates computable from raw training loss. The bound concerns a current prior; it is not unrestricted finite-data neural convergence. The remaining subsections give finite-iteration tracking and discuss empirical sampling and endpoint control. These results are not needed for the ideal population theorems.

F.2 Changing the observation distribution

Proposition 8 (A conditional one-step approximation bound)**.** *Fix a prior π with observation law $\mu = K_N\pi$, exact posterior P_π, and approximate kernel $\widehat{P}$. Suppose $d\mu_\star/d\mu \le R$ almost everywhere*

and

$$\varepsilon = \int \mathrm{KL}(P_\pi(\cdot \mid C_e) \| \widehat{P}(\cdot \mid C_e)) \mu(dC_e) < \infty. \tag{79}$$

For $\widehat{\pi}^+ = \int \widehat{P}(\cdot \mid C_e) \mu_\star(dC_e)$,

$$\mathrm{KL}(\mathcal{T}(\pi) \| \widehat{\pi}^+) \le R\varepsilon, \qquad \mathrm{TV}(\mathcal{T}(\pi), \widehat{\pi}^+) \le \sqrt{R\varepsilon/2}, \tag{80}$$

where total variation is $\sup_B |\rho(B) - \rho'(B)|$.

Proof. The two joint laws with common observation marginal $\mu_\star$ and conditional kernels P_π and $\widehat{P}$ have KL divergence equal to the $\mu_\star$-average of the conditional KL. The density-ratio assumption bounds it by $R\varepsilon$. Marginalization and Pinsker's inequality give the two assertions. □

The density-ratio condition transfers an average conditional error under the synthetic observation law to the real observation law. Common support alone is insufficient. In particular, this lemma does not use a triangle inequality for KL, and cannot be applied by adding its right-hand side directly to the latent KL recursion of Theorem 2.

F.3 Proof of Theorem 7

For the base and Föllmer coefficients with scale σ, define $\mathcal{E}_\pi^{\mathrm{base}} = (2\sigma^2)^{-1} \int_0^1 e_t/(1-t)^2 \, dt$ and $\mathcal{E}_\pi^{\mathrm{F}} = (2\sigma^2)^{-1} \int_0^1 (1+t) e_t/(1-t) \, dt$. Write $\mathcal{E}_\pi^g$ for either value when its corresponding noise coefficient is chosen. The Föllmer value is exactly (75) for the scale $\sigma > 0$ retained in the main implementation. Fix π, an observation C_e, and its encoding Y_e. We prove the statement for general linear-schedule noise scale $\sigma > 0$ and both the base and Föllmer samplers; Theorem 7 is the Föllmer case with this same σ. The coordinate transform Φ is invertible, so endpoint KL in covariance coordinates equals endpoint KL in Y coordinates. The exact and learned base SDEs have common initial state and volatility $\sigma(1-t)$. Under the stated Girsanov conditions, their conditional path KL is

$$\frac{1}{2\sigma^2} \int_0^1 \frac{\mathbb{E}[\|\widehat{b}_t(I_t, Y_e) - b_t^\pi(I_t, Y_e)\|_2^2 \mid C_e]}{(1-t)^2} \, dt. \tag{81}$$

For the Föllmer variant, (12) gives $\widehat{b}_t^{\mathrm{F}} - b_t^{\pi,\mathrm{F}} = (1+t)(\widehat{b}_t - b_t^\pi)$ and $g_t^2 = \sigma^2(1-t^2)$. The conditional path KL therefore has weight

$$\frac{(1+t)^2}{2\sigma^2(1-t^2)} = \frac{1+t}{2\sigma^2(1-t)}. \tag{82}$$

The exact base and Föllmer SDEs share the conditional interpolant marginals, so both expectations are evaluated using the same synthetic (I_t, Y_e'). Marginalizing to time one and averaging against μ_π gives (76). Proposition 8, with $R = B_\pi$, then yields (77) and the bound $\mathrm{TV}(\mathcal{T}(\pi), \widehat{\pi}^+) \le \sqrt{B_\pi \mathcal{E}_\pi^g/2}$. For spectral sampling, replace matrix variables by their spectral counterparts and the forward channel by K_N^{spec}; the same argument applies, using data processing at the final sort/decoding step. The known-basis diagonal version is identical with the simulator stated in Appendix B.4. All assertions require finite energy on the full interval and the corresponding endpoint absolute continuity; a calculation only on $[0, 1-\tau]$ controls that intermediate marginal, not automatically the intended covariance posterior.

F.4 Proof of the full posterior-reconstruction bound

The following argument applies to either noise choice covered by the preceding proof; take $g = g^{\mathrm{F}}$ for the bound stated in this appendix. For a current prior π with $D(\pi) < \infty$, the joint-law calculation from Theorem 2 does not require π to belong to an exact EM trajectory. It gives

$$\int \mathrm{KL}(P_\star(\cdot \mid C_e) \| P_\pi(\cdot \mid C_e)) \mu_\star(dC_e) = D(\pi) - F(\pi). \tag{83}$$

Conditional Pinsker's inequality and Jensen's inequality for the square root imply

$$\int \mathrm{TV}(P_\star(\cdot \mid C_e), P_\pi(\cdot \mid C_e)) \mu_\star(dC_e) \le \sqrt{\frac{D(\pi) - F(\pi)}{2}}. \tag{84}$$

Theorem 7 and the observation-overlap bound give, before marginalizing the latent variable,

$$\int \mathrm{KL}(P_\pi(\cdot \mid C_e) \| \widehat{P}^g(\cdot \mid C_e)) \mu_\star(dC_e) \le B_\pi \mathcal{E}_\pi^g. \tag{85}$$

Apply Pinsker and Jensen again, and insert P_π and $\widehat{P}^g$ between $P_\star$ and $\widehat{P}^g_{\mathrm{num}}$ in the total-variation triangle inequality. The assumed numerical error then proves (78). This is an average over real observations, not a bound for every individual observation. It controls probabilities of posterior events and bounded test functions, not unbounded moments such as arbitrary covariance or precision entries without additional assumptions.

The ratio bound and weighted regression error are assumptions about population laws. They are not estimated by the finite training objective alone. In particular, neither a low empirical regression loss nor a plausible output covariance supplies the first term, which concerns the unknown true prior. The bound describes which errors must be controlled; it is not a directly computable stopping certificate.

F.5 Finite-iteration tracking of the exact map

The one-step error can be propagated in a genuine metric. The following elementary stability lemma makes explicit one sufficient route, without assuming that the EM map is contractive.

Lemma 9 (Overlap-dependent stability of posterior averaging)**.** *For priors ρ, π, if $d\mu_\star / d(K_N \rho) \le R$, then*

$$\mathrm{TV}(\mathcal{T}(\rho), \mathcal{T}(\pi)) \le 2R \ \mathrm{TV}(\rho, \pi). \tag{86}$$

Proof. Set $\mu_\rho = K_N \rho$ and $\mu_\pi = K_N \pi$. Inserting the joint law $\mu_\pi(dC_e) P_\pi(dC \mid C_e)$ and using the TV triangle inequality gives

$$\begin{aligned} \int \mathrm{TV}(P_\rho(\cdot \mid C_e), P_\pi(\cdot \mid C_e))\, \mu_\rho(dC_e) &\le \mathrm{TV}(\rho K_N, \pi K_N) + \mathrm{TV}(\mu_\rho, \mu_\pi) \\ &\le 2\, \mathrm{TV}(\rho, \pi). \end{aligned} \tag{87}$$

The first joint TV equals $\mathrm{TV}(\rho, \pi)$ because the latent coordinate is retained and the forward kernel is common; the second decreases under the forward channel. Change measure from μ_ρ to $\mu_\star$ and marginalize the latent output to obtain (86). □

Corollary 10 (Tracking for finitely many learned updates)**.** *Let ρ_k be the priors supplying the synthetic training pairs of successive learned drifts. Suppose Theorem 7 applies at each ρ_k with overlap bound R_k and energy $\mathcal{E}_k$. Let $H_k = \int \widehat{P}_k(\cdot \mid C_e) \mu_\star(dC_e)$ be the learned continuous-time population output, and suppose $\mathrm{TV}(\rho_{k+1}, H_k) \le \delta_k$. For the exact reference trajectory $\pi_{k+1} = \mathcal{T}(\pi_k)$ with $\pi_0 = \rho_0$, write $d_k = \mathrm{TV}(\rho_k, \pi_k)$. Then*

$$d_{k+1} \le 2R_k d_k + \sqrt{R_k \mathcal{E}_k / 2} + \delta_k, \tag{88}$$

and, for each finite K,

$$d_K \le \sum_{j=0}^{K-1} \left(\sqrt{R_j \mathcal{E}_j / 2} + \delta_j \right) \prod_{\ell=j+1}^{K-1} 2R_\ell. \tag{89}$$

Proof. Insert H_k, $\mathcal{T}(\rho_k)$, and $\mathcal{T}(\pi_k)$ between ρ_{k+1} and π_{k+1} in the TV triangle inequality. Use the numerical/ensemble premise, Theorem 7, and Lemma 9, respectively. Iterate the resulting recurrence from $d_0 = 0$. □

The overlap bounds can make this estimate conservative: $R_k \ge 1$, so (89) is not a uniform-in-iteration convergence bound. It does establish approximation of any fixed finite exact trajectory as the stated errors vanish with controlled overlap. If an additional stability estimate gives a factor $r < 1$ in the *same metric* on an invariant model class, the factor $2R_k$ can be replaced by r, yielding a geometric accumulated-error bound. The L^2 linearization in Appendix D does not by itself establish such TV contraction.

F.6 Endpoint-compatible parametrization and excess risk

For the linear schedules, the exact endpoint regression is $b_1^\pi(y, Y_e) = y - Y_e$ because Y and W_1 are conditionally independent at $t = 1$. One possible architecture therefore writes

$$\widehat{b}_\theta(t, y, Y_e) = y - Y_e + (1 - t)u_\theta(t, y, Y_e). \tag{90}$$

For $t < 1$ let $u_t^\pi = (b_t^\pi - y + Y_e)/(1 - t)$. Whenever this residual is well defined and the relevant expectations are finite, put $e_t^u = \mathbb{E}_{\pi K_N, W}\|u_\theta(t, I_t, Y_e') - u_t^\pi(I_t, Y_e')\|_2^2$. Then

$$e_t = (1 - t)^2 e_t^u, \qquad \mathcal{E}_\pi^{\mathrm{base}} = \frac{1}{2\sigma^2}\int_0^1 e_t^u\, dt, \qquad \mathcal{E}_\pi^{\mathrm{F}} = \frac{1}{2\sigma^2}\int_0^1 (1 - t^2)e_t^u\, dt. \tag{91}$$

Thus a square-integrable residual error gives finite endpoint energies for both samplers. This is a conditional design implication, not a regularity guarantee for every prior. The raw regression loss should be evaluated as in (10), without dividing the random target by $1 - t$. Multiplying the *raw* square loss by the singular weights in (75) can give an infinite objective because its irreducible conditional variance need not vanish. The theory controls weighted excess risk, which must not be confused with that raw objective. Endpoint cutoff, finite-variance time weighting, and the residual parametrization are distinct implementation choices to evaluate.

F.7 Finite observation and synthetic ensembles

For a finite observed ensemble, let $\widehat{\mu}_M = M^{-1}\sum_{a=1}^M \delta_{C_{e,a}}$. The empirical objective is

$$\widehat{\ell}_M(\pi) = \frac{1}{M}\sum_{a=1}^M \log m_\pi(C_{e,a}), \tag{92}$$

where the reference κ_N is fixed. In this Gaussian experiment the observation law is nonatomic, so $\mathrm{KL}(\widehat{\mu}_M\|K_N\pi) = +\infty$. It is therefore incorrect to replace $\mu_\star$ by $\widehat{\mu}_M$ directly inside Theorem 2. Exact finite-data EM instead increases (92), with the same Jensen argument as for population EM after replacing the observation integral by a finite average (when the logarithms are finite). This remains an empirical likelihood statement even if drift regression and conditional sampling are exact. Recovery as M grows, or recovery for a fixed M, requires separate statistical assumptions; neither follows from the population theorem alone. Likewise, the density-ratio premise of Proposition 8 does not hold with an atomic empirical marginal against a continuous synthetic one. Regularized model classes, held-out validation, and early stopping are plausible design choices, not established statistical guarantees in this draft.

In Corollary 10, δ_k can include the discrepancy caused by replacing the population observation average by the finite ensemble and by numerical integration, but no distribution-free rate for this quantity is asserted. Training on finitely many synthetic pairs contributes to the population regression error $\mathcal{E}_k$. These errors may depend on all previous iterates and on the same reused dataset. A finite cached particle population may also create atomic target laws outside the endpoint-KL assumptions. Fresh draws from a continuous implicit prior and a generalization analysis are different from an exact finite-particle EM update; they should not be conflated.

F.8 Supplementary RI reconstruction bound with incomplete learning

For a current spectral prior ν, put $D_\nu = \mathrm{KL}(\nu_\star\|\nu)$ and $F_\nu = \mathrm{KL}(\mu_\star^{\mathrm{spec}}\|K_N^{\mathrm{spec}}\nu)$. Assume the spectral version of Theorem 7, with overlap bound B_ν, weighted error $\mathcal{E}_\nu^{\mathrm{F}}$, and average numerical spectral error δ_{spec}. With $\widehat{q}_{\mathrm{num}}$ in (18), define

$$\overline{\varepsilon}_m = \int\int \varepsilon_m(\Lambda, \Lambda_e)\widehat{q}_{\mathrm{num}}(d\Lambda \mid \Lambda_e)\mu_\star^{\mathrm{spec}}(d\Lambda_e). \tag{93}$$

For $D_\nu < \infty$, the resulting full reconstruction satisfies

$$\boxed{\mathbb{E}_{\mu_\star}\,\mathrm{TV}(P_\star, \widehat{P}_m) \le \sqrt{\frac{D_\nu - F_\nu}{2}} + \sqrt{\frac{B_\nu \mathcal{E}_\nu^{\mathrm{F}}}{2}} + \delta_{\mathrm{spec}} + \overline{\varepsilon}_m.} \tag{94}$$

The first three terms are the supplementary general-case bound applied to spectra. The fourth is solely the cost of finite orientation sampling. The fixed-spectrum MCMC guarantee in Appendix G.4 need not be uniform over dimensions or spectral draws; efficient mixing must be measured rather than inferred from the known density.

Proof. Apply the proof of (78) to the spectral model $(\nu, K_N^{\mathrm{spec}})$. The required exact conditional-sampling statement follows from the same Brownian-interpolant argument in dimension d. Decoding by exponentiation and sorting cannot increase the discrepancy of numerical states; at exact endpoints sorting changes nothing. Under the preceding spectral assumptions,

$$\int \mathrm{TV}(q_\star(\cdot \mid \Lambda_e), \widehat{q}_{\mathrm{num}}(\cdot \mid \Lambda_e))\mu_\star^{\mathrm{spec}}(d\Lambda_e) \le \sqrt{\frac{D_\nu - F_\nu}{2}} + \sqrt{\frac{B_\nu \mathcal{E}_\nu^{\mathrm{F}}}{2}} + \delta_{\mathrm{spec}}. \tag{95}$$

Now apply (100) with $\widehat{q} = \widehat{q}_{\mathrm{num}}$ at each observation. The bound depends on C_e only through Λ_e, and its average is therefore over $\mu_\star^{\mathrm{spec}}$. Adding the generated-spectrum average of ε_m proves (94). In particular, exact orientation sampling contributes no additional error to the spectral learning theorem. Spectral simulation costs and finite-time MCMC errors do not disappear from the implementation merely because the statistical reduction is exact.

G ROTATIONAL INVARIANCE AND FULL POSTERIOR RECONSTRUCTION

An *orthogonal orbit* is the set of matrices $Q\Lambda Q^{\mathrm{T}}$ obtained by rotating a fixed diagonal matrix. Splitting an RI matrix law into its eigenvalue distribution and its conditional distribution on each orbit is called disintegration. The latter conditional distribution is uniform on that orbit. This explains why the spectral reduction differs from an invertible matrix-logarithm transformation.

Theorem 11 (Exact RI reduction)**.** *For an RI prior, $P(\Lambda \in \cdot \mid C_e) = q_\nu(\cdot \mid \Lambda_e)$. If π_1, π_2 are RI with spectral laws ν_1, ν_2, then*

$$\mathrm{KL}(\pi_1 \| \pi_2) = \mathrm{KL}(\nu_1 \| \nu_2), \qquad \mathrm{KL}(K_N \pi_1 \| K_N \pi_2) = \mathrm{KL}(K_N^{\mathrm{spec}} \nu_1 \| K_N^{\mathrm{spec}} \nu_2). \tag{96}$$

For RI truth and initialization, exact population EM preserves RI and equals spectral posterior averaging. The prior-identification threshold remains $N \geq d$ within RI priors.

G.1 PROOF OF THEOREM 11

For an ordered positive spectrum Λ, let $\mathcal{H}_\Lambda$ be the pushforward of normalized Haar measure under $U \mapsto U\Lambda U^{\mathrm{T}}$. Every RI prior disintegrates as $\pi = \int \mathcal{H}_\Lambda \, \nu(d\Lambda)$; this can be seen by averaging any conditional law over the orthogonal group. The kernel $\mathcal{H}_\Lambda$ is common to every RI prior, including at repeated eigenvalues. Because spectra of distinct orbits differ, the KL chain rule gives

$$\mathrm{KL}(\pi_1 \| \pi_2) = \mathrm{KL}(\nu_1 \| \nu_2) + \int \mathrm{KL}(\mathcal{H}_\Lambda \| \mathcal{H}_\Lambda) \nu_1(d\Lambda) = \mathrm{KL}(\nu_1 \| \nu_2). \tag{97}$$

This identity also holds with value $+\infty$ by absolute-continuity considerations. Orthogonal equivariance of the Gaussian channel makes every observation mixture RI. Disintegration on its PSD orbits, including singular ones, gives the second equality in (96).

Conditional on Λ, the spectrum of $C^{1/2} Z Z^{\mathrm{T}} C^{1/2}/N$ is independent of the externally chosen eigenbasis, by orthogonal invariance of Z. This proves (17). To compute the posterior, represent the prior with latent independent $(\Lambda, U) \sim \nu \otimes H$. For a fixed observation choose any orthogonal diagonalization $C_e = U_e \Lambda_e U_e^{\mathrm{T}}$ and set $R = U_e^{\mathrm{T}} U$. Substitution into (20) yields (15), after dropping the factor $\exp(N \operatorname{tr} C_e/2)$ that is constant in (Λ, R). The formula is valid for every $N \geq 1$. It depends only on Λ_e, proving spectral sufficiency.

Bayesian equivariance implies $P_\pi(Q\, dC\, Q^{\mathrm{T}} \mid Q C_e Q^{\mathrm{T}}) = P_\pi(dC \mid C_e)$ in pushforward notation. Averaging against the RI law $\mu_\star$ shows that $\mathcal{T}(\pi)$ is RI. Taking the spectrum of this posterior average gives exactly the spectral EM update in Theorem 11. The reduction concerns the full joint law of the ordered diagonal spectrum, not merely a pooled one-dimensional spectral histogram. The $N \geq d$ identification result follows by lifting spectral laws to RI matrix laws and using Theorem 1; the negative construction in Appendix B is already RI. This proves the threshold assertion in Theorem 11.

G.2 FULL POSTERIOR COMPLETION: PROOF OF THEOREM 4

Fix an observation $C_e = U_e \Lambda_e U_e^{\mathrm{T}}$, and write $\mathcal{O}_{\Lambda, C_e}$ for the conditional law of $C = U_e R \Lambda R^{\mathrm{T}} U_e^{\mathrm{T}}$ when $R \sim H_N(\cdot \mid \Lambda, \Lambda_e)$. This is a probability kernel supported on matrices with spectrum Λ. It is independent of the spectral prior. Thus

$$P_\star(dC \mid C_e) = \int \mathcal{O}_{\Lambda, C_e}(dC) q_\star(d\Lambda \mid \Lambda_e), \qquad \widehat{P}^H(dC \mid C_e) = \int \mathcal{O}_{\Lambda, C_e}(dC) \widehat{q}(d\Lambda \mid \Lambda_e). \tag{98}$$

The mapping $C \mapsto (\operatorname{spec} C, C)$ retains all information about C. In these joint coordinates the two laws have different spectral marginals but identical conditional matrix laws. The chain rule for relative entropy therefore proves (19), including infinite values. No unique eigenvector frame is needed. Repeated population eigenvalues merely make the representation by R nonunique; the conditional matrix law remains well defined. A different empirical eigenbasis within a tied eigenspace changes the representation of R but not the exact matrix law.

Exact completion also preserves total variation:

$$\mathrm{TV}(P_\star, \widehat{P}^H) = \mathrm{TV}(q_\star, \widehat{q}). \tag{99}$$

One inequality follows because applying a common probability kernel cannot increase total variation. The other follows by extracting the ordered spectrum from the matrix. Equivalently, the conditional matrix laws for distinct spectra have disjoint spectral supports.

Proposition 12 (Finite orientation-sampling error). *Fix $C_e = U_e \Lambda_e U_e^{\mathrm{T}}$ and a spectral law $\widehat{q}(\cdot \mid \Lambda_e)$. Let $\widehat{P}^H$ and $\widehat{P}_m$ be its completions using H_N and $\widehat{H}_m$, respectively. Write $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$ for total variation, the largest discrepancy in event probabilities, and set $\varepsilon_m(\Lambda, \Lambda_e) = \mathrm{TV}(H_N(\cdot \mid \Lambda, \Lambda_e), \widehat{H}_m(\cdot \mid \Lambda, \Lambda_e))$. Then*

$$\mathrm{TV}(P_\star, \widehat{P}_m) \le \mathrm{TV}(q_\star, \widehat{q}) + \int \varepsilon_m(\Lambda, \Lambda_e) \widehat{q}(d\Lambda \mid \Lambda_e). \tag{100}$$

No uniqueness of an eigenvector frame is required, so this includes repeated eigenvalues.

Proof. Let $\widehat{\mathcal{O}}^m_{\Lambda, C_e}$ be the matrix law obtained from $\widehat{H}_m$. Deterministic reconstruction gives

$$\mathrm{TV}(\mathcal{O}_{\Lambda, C_e}, \widehat{\mathcal{O}}^m_{\Lambda, C_e}) \le \varepsilon_m(\Lambda, \Lambda_e). \tag{101}$$

Insert the exact completion $\widehat{P}^H$ between $P_\star$ and $\widehat{P}_m$, apply (99), and average the preceding conditional bound under $\widehat{q}$. This proves (100). The orientation error is averaged under the generated spectral law, not the unknown true spectral posterior. On matrix space the orientation bound may be conservative because sign flips and rotations within repeated population eigenspaces produce the same matrix.

If instead conditional orientation KL is available, the same chain rule at the augmented (Λ, R) level and deterministic reconstruction give the alternative inequality

$$\begin{aligned} \mathrm{KL}(P_\star \| \widehat{P}_m) \le{}& \mathrm{KL}(q_\star \| \widehat{q}) \\ &+ \int \mathrm{KL}(H_N(\cdot \mid \Lambda, \Lambda_e) \| \widehat{H}_m(\cdot \mid \Lambda, \Lambda_e)) q_\star(d\Lambda \mid \Lambda_e). \end{aligned} \tag{102}$$

For finite MCMC runs without a full-dimensional transition component, initialization at a point can give a singular output and an infinite conditional KL. An atomic component by itself does not force this KL to be infinite if the output also has a sufficiently positive density. The Haar-mixture kernel below supplies such a density after a step, but the TV formulation also applies to other samplers. The total-variation formulation does not require an everywhere positive approximation density.

G.3 Proof of Corollary 13

Recall the statement.

Corollary 13 (Convergence of the ideal RI reconstruction). *Assume exact spectral regression and exact SDE sampling at each population iteration, with the sampler hypotheses of Proposition 3, and $D_0^\nu < \infty$. Then*

$$F_k^\nu \le \frac{D_0^\nu}{k+1}, \qquad \int \mathrm{KL}(P_\star \| P_k^H)\, d\mu_\star = D_k^\nu - F_k^\nu. \tag{103}$$

If $N \ge d$ and $\nu_0, \nu_\star$ have support in a common compact set of positive spectra, then $\nu_k \Rightarrow \nu_\star$ and $P_k^H(\cdot \mid C_e) \Rightarrow P_\star(\cdot \mid C_e)$ for $\mu_\star$-almost every observation. If $D_k^\nu \le \chi F_k^\nu$ holds along the iterates instead, the averaged full-posterior KL in (103) is at most $(1 - \chi^{-1})^k D_0^\nu$.

Lift ν_k to the RI prior $\pi_k = \int \mathcal{H}_\Lambda \nu_k(d\Lambda)$. Theorem 11 shows that these are precisely the full-matrix population EM iterates from the RI initialization, and that their D_k, F_k equal D_k^ν, F_k^ν. Exact orientation completion of q_{ν_k} is the posterior of π_k, so $P_k^H = P_{\pi_k}$. Theorem 2 proves (103).

If the initial and true spectra belong to a common compact set of positive ordered diagonal matrices, their eigenvalues have common positive lower and finite upper bounds. Its orthogonal rotations form a compact subset of $\mathbb{S}^d_{++}$. For $N \ge d$, Theorem 2 yields weak convergence of the lifted priors and their matrix posteriors. Taking spectra gives $\nu_k \Rightarrow \nu_\star$. This proof covers repeated eigenvalues at the distribution level; application through an SI still requires the stated sampler hypotheses. Under $D_k^\nu \le \chi F_k^\nu$, the stronger averaged-KL rate follows from the geometric bound of Theorem 2. Exact orientation completion does not add an approximation term. None of these conclusions claims that a finite run of orientation MCMC has already reached its target.

G.4 A concrete orientation MCMC kernel and its guarantee

The following is a specified baseline sampler for Algorithm 2. It is not claimed to be optimal among orthogonal-matrix samplers. The target in (16) belongs to the matrix Bingham family studied by Hoff (2009); the simple mixture below is chosen so its invariance and convergence have a short proof.

Algorithm 2: Self-consistent spectral reconstruction with orientation completion (RI priors). Both algorithms target the full covariance posterior; they differ in what is learned and what is known.

Input: Empirical spectra $\Lambda_{e,1:M}$; sample size N; sampleable spectral prior ν_0; noise scale $\sigma > 0$; outer budget K; orientation budget m.

Output: A conditional sampler for the full covariance posterior at any supplied observation.

Draw $\Lambda \sim \nu_0$; form Λ'_e by (17); record (Y'_e, Y); fit $\widehat{b}_0$ by (10)

for $k = 0, \dots, K-1$ **do**

  Integrate (12) in $\mathbb{R}^d$ at recorded $\Lambda_{e,a}$ with $\widehat{b}_k$; decode population spectra Λ

  Form fresh Λ'_e from each Λ by (17); record (Y'_e, Y)

  Fit $\widehat{b}_{k+1}$ by (10) in $\mathbb{R}^d$, targets held fixed

`/* Inference at` $C_e = U_e \Lambda_e U_e^{\mathrm{T}}$ `(none of this occurs in training) */`

Draw Λ from the final spectral sampler conditioned on Λ_e

Holding Λ fixed, run m steps of the orientation chain (105) targeting (16) to obtain R

return $C = U_e R \Lambda R^{\mathrm{T}} U_e^{\mathrm{T}}$

Fix $\lambda_i > 0$ and $\lambda_{e,i} \geq 0$, and set $\mathcal{V}(R) = \mathcal{V}_{\Lambda,\Lambda_e}(R)$. Choose $\omega \in (0, 1]$ and a symmetric angle distribution, such as a centered normal with fixed variance. At each step, independently choose one of two proposals. With probability $1 - \omega$, pick $i < j$ uniformly and propose $R' = RG_{ij}(\theta)$, where the (i, j) block of $G_{ij}(\theta)$ is

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}. \tag{104}$$

With probability ω, propose a fresh independent Haar matrix R'. For either proposal, accept with probability

$$\min\{1, \exp[\mathcal{V}_{\Lambda,\Lambda_e}(R) - \mathcal{V}_{\Lambda,\Lambda_e}(R')]\}, \tag{105}$$

and otherwise retain R; the unknown normalizing constant cancels. For $d = 1$ the local proposal can be the identity. A Haar draw on $O(d)$ can be obtained by QR decomposition of an independent Gaussian matrix, with diagonal-sign correction of the triangular factor. This samples both determinant components. Plane rotations alone stay in the initial determinant component, although that sign itself does not affect the reconstructed covariance.

Proposition 14 (Validity and a conservative fixed-spectrum mixing bound)**.** *Let P_{Λ,Λ_e} be this mixture kernel and let H_N denote its target at the fixed pair (Λ, Λ_e). Define*

$$\Delta_{\Lambda,\Lambda_e} = \max_{R \in O(d)} \mathcal{V}(R) - \min_{R \in O(d)} \mathcal{V}(R) < \infty. \tag{106}$$

Then P_{Λ,Λ_e} is reversible with invariant law H_N, and for any starting distribution η and integer $m \geq 1$,

$$\mathrm{TV}(\eta P^m_{\Lambda,\Lambda_e}, H_N) \leq \left(1 - \omega e^{-\Delta_{\Lambda,\Lambda_e}}\right)^m. \tag{107}$$

Proof. Haar measure is invariant under right multiplication. Since θ and $-\theta$ have the same law and $G_{ij}(\theta)^{-1} = G_{ij}(-\theta)$, the local proposal is symmetric with respect to Haar measure. The independent-Haar proposal is symmetric as well. The Metropolis acceptance ratio consequently gives detailed balance for each component with density $f(R) = e^{-\mathcal{V}(R)}/Z_{\Lambda,\Lambda_e}$ relative to Haar measure, and hence for their mixture.

The compactness of $O(d)$ and continuity of $\mathcal{V}$ give finite extrema. Put $M_{\Lambda,\Lambda_e} = \sup_R f(R)$. Since $Z_{\Lambda,\Lambda_e} \geq e^{-\max \mathcal{V}}$, we have $M_{\Lambda,\Lambda_e} \leq e^{\Delta_{\Lambda,\Lambda_e}}$. For the independent-Haar Metropolis transition and every state R, its accepted part satisfies

$$\min\{1, f(R')/f(R)\}\, dH(R') \geq M^{-1}_{\Lambda,\Lambda_e} f(R')\, dH(R') = M^{-1}_{\Lambda,\Lambda_e} H_N(dR'). \tag{108}$$

Thus the full mixture transition has a common component of weight at least $\epsilon = \omega e^{-\Delta_{\Lambda,\Lambda_e}}$ equal to a fresh draw from H_N. Write it as $P_{\Lambda,\Lambda_e} = \epsilon H_N + (1 - \epsilon)\widetilde{P}$, where $\widetilde{P}$ is a probability kernel preserving H_N when $\epsilon < 1$. Iteration gives (107); if $\epsilon = 1$ the first step is exact. This also proves uniqueness of the invariant law. □

For example,

$$\Delta_{\Lambda,\Lambda_e} \le \frac{N}{2}\,\mathrm{tr}(\Lambda_e)\big(\lambda_{\min}^{-1} - \lambda_{\max}^{-1}\big), \tag{109}$$

since $\lambda_{\max}^{-1} I_d \preceq R\Lambda^{-1}R^{\mathrm{T}} \preceq \lambda_{\min}^{-1} I_d$. This bound can be extremely large. Proposition 14 is a fixed-pair convergence guarantee, not a practical dimension-uniform rate or a recommendation to use many global proposals. It implies an upper bound on the ε_m term in Proposition 12; integrating it under the generated spectra bounds $\bar{\varepsilon}_m$. For fixed spectral output law, dominated convergence also gives $\bar{\varepsilon}_m \to 0$ as $m \to \infty$, without a uniform useful rate. Other valid Bingham samplers can replace this kernel; their finite-run error must then be analyzed or assessed separately.

Cost of local updates and factored output. Let $A = R^{\mathrm{T}}\Lambda_e R$ and $a_i = \lambda_i^{-1}$. For the rotation convention above,

$$\mathcal{V}(RG_{ij}(\theta)) - \mathcal{V}(R) = \frac{N}{2}(a_i - a_j)\left[(A_{jj} - A_{ii})\sin^2\theta + 2A_{ij}\sin\theta\cos\theta\right]. \tag{110}$$

The acceptance decision uses this two-by-two block. Upon acceptance, update two columns of R and two rows and columns of A; this costs $O(d)$ with cached matrices. A sweep through all pairs is still $O(d^3)$, and additional sweeps may be needed. Initializing A or proposing an independent Haar matrix has dense cubic cost. These are operation counts, not mixing guarantees.

Retaining (U_e, R, Λ) is a full sample representation: for any vector v,

$$Cv = U_e R[\Lambda(R^{\mathrm{T}}(U_e^{\mathrm{T}}v))], \qquad C^{-1}v = U_e R[\Lambda^{-1}(R^{\mathrm{T}}(U_e^{\mathrm{T}}v))], \qquad \log\det C = \sum_i \log\lambda_i. \tag{111}$$

The two vector operations cost $O(d^2)$ with dense factors; forming the full product is unnecessary. Generic full samples still require quadratic storage. The learned spectral state is smaller, but the output orientation generally has $d(d-1)/2$ continuous degrees of freedom. No linear-cost full-matrix sampling claim follows from RI alone.

G.5 Posterior mean and repeated eigenvalues

Let $M(C_e) = \mathbb{E}[C \mid C_e]$, which exists almost surely under the moment assumption. For any Q with $QC_eQ^{\mathrm{T}} = C_e$, equivariance gives $QM(C_e)Q^{\mathrm{T}} = M(C_e)$. In an empirical eigenbasis, arbitrary coordinate sign flips force all off-diagonal entries of $U_e^{\mathrm{T}}M(C_e)U_e$ to vanish. If C_e has a repeated eigenvalue, arbitrary orthogonal rotations in its eigenspace force the corresponding diagonal block of the mean to be a scalar multiple of the identity. The diagonal entries depend only on Λ_e by equivariance and are precisely $\mathbb{E}[u_{e,i}^{\mathrm{T}}Cu_{e,i} \mid \Lambda_e]$. Using the latent orientation representation gives the expression for ξ_i in (114).

For any coefficient vector z, Frobenius orthogonality gives the finite-sample identity

$$\|C - U_e\,\mathrm{diag}(z)U_e^{\mathrm{T}}\|_F^2 = \|U_e^{\mathrm{T}}CU_e - \mathrm{diag}(\xi)\|_F^2 + \|\xi - z\|_2^2. \tag{112}$$

Thus $z = \xi$ is the unconstrained diagonal oracle in a fixed empirical eigenbasis. In repeated-eigenvalue blocks, this oracle can depend on the chosen basis and need not itself be an equivariant estimator; the posterior mean remains block-scalar and well defined. For $N \ge d$, the Gaussian sample spectrum is simple almost surely, so this distinction disappears with probability one.

G.6 From posteriors to covariance estimates

Both algorithms learn a distribution before any decision loss is chosen, and the loss then selects the estimate. Under squared Frobenius loss the posterior mean is optimal; under the Stein loss $L_S(A,C) = \mathrm{tr}(C^{-1}A) - \log\det(C^{-1}A) - d$ the optimal estimate averages the precision and inverts afterwards:

$$\widehat{C}_F = \mathbb{E}[C \mid C_e], \qquad \widehat{C}_S = (\mathbb{E}[C^{-1} \mid C_e])^{-1}. \tag{113}$$

Under RI both share the empirical eigenvectors, and

$$\widehat{C}_F = U_e\,\mathrm{diag}(m(\Lambda_e))U_e^{\mathrm{T}}, \qquad m_i(\Lambda_e) = \mathbb{E}[\xi_i \mid \Lambda_e], \quad \xi_i = u_{e,i}^{\mathrm{T}}Cu_{e,i} = \sum_j \lambda_j R_{ij}^2. \tag{114}$$

The distinction this exposes is worth stating plainly: $\mathbb{E}[\lambda_i \mid \Lambda_e]$ estimates a population eigenvalue, whereas $m_i(\Lambda_e)$ estimates the variance along an empirical direction. Ledoit–Wolf and Ledoit–Péché shrinkage approximate the latter directional oracle, not the mean population spectrum (Ledoit & Péché, 2011; Ledoit & Wolf, 2012), and Proposition 15 makes the averaged connection precise. Neither shrinkage coefficients nor posterior means substitute for population-spectrum draws in Algorithm 2. The rest of this appendix gives the moment assumptions, the Stein coefficients, and two ways of computing the estimates after learning.

G.7 Posterior learning and loss-specific Bayes actions

The spectral SI at iteration k targets $q_{\nu_k}(d\Lambda \mid \Lambda_e)$. This distribution describes uncertainty in the ordered population eigenvalues under the *current* prior. The superscript (r) in $\Lambda^{(r)}$ indexes posterior draws; a hat in $\widehat{\Lambda}_L$ denotes decision coefficients instead. Conditional orientation sampling with (15) converts a spectral draw into a matrix posterior draw. Aligning a sampled Λ with U_e omits this orientation and does not perform that conversion.

The choice between Frobenius and Stein loss is a decision-theoretic operation on this posterior, not a choice of the posterior itself. In particular, for $Q = P_\pi(\cdot \mid C_e)$ define

$$\mathcal{R}_L(A \mid C_e) = \int L(A, C)\, Q(dC), \qquad \widehat{C}_L(C_e) = \operatorname*{arg\,min}_{A \succ 0} \mathcal{R}_L(A \mid C_e). \tag{115}$$

The underlying prior π may be true or learned. All conditional expectations below are with respect to this same Q.

Squared Frobenius loss. Assume $\mathbb{E}[\|C\|_F^2 \mid C_e] < \infty$. With $M_e = \mathbb{E}[C \mid C_e]$, expansion gives

$$\mathcal{R}_F(A \mid C_e) = \mathbb{E}[\|C - M_e\|_F^2 \mid C_e] + \|A - M_e\|_F^2. \tag{116}$$

Since $M_e \succ 0$, it is the unique minimizer. Under RI, the symmetry argument above gives $M_e = U_e \operatorname{diag}(m(\Lambda_e)) U_e^{\mathrm{T}}$. Thus $\widehat{\Lambda}_F = \operatorname{diag}(m(\Lambda_e))$, where $m_i(\Lambda_e) = \mathbb{E}[\sum_j \lambda_j R_{ij}^2 \mid \Lambda_e]$. This averages covariance entries in the empirical frame; it is not generally the average of the ordered population eigenvalues.

Stein loss and its convention. For

$$L_S(A, C) = \operatorname{tr}(C^{-1}A) - \log\det(C^{-1}A) - d, \tag{117}$$

assume $\mathbb{E}[\|C^{-1}\|_F + |\log\det C| \mid C_e] < \infty$ and set $P_e = \mathbb{E}[C^{-1} \mid C_e] \succ 0$. Then

$$\mathcal{R}_S(A \mid C_e) = \operatorname{tr}(P_e A) - \log\det A + \mathbb{E}[\log\det C \mid C_e] - d. \tag{118}$$

Its derivative is $P_e - A^{-1}$ and $-\log\det A$ is strictly convex on the SPD cone, so the unique minimizer is $\widehat{C}_S = P_e^{-1}$. Equivalently, the excess risk is

$$\mathcal{R}_S(A \mid C_e) - \mathcal{R}_S(P_e^{-1} \mid C_e) = \operatorname{tr}(P_e A) - \log\det(P_e A) - d \geq 0, \tag{119}$$

with equality only at $A = P_e^{-1}$. The inequality follows by diagonalizing $P_e^{1/2} A P_e^{1/2}$ and using $z - \log z - 1 \geq 0$. Under RI, $P_e = U_e \operatorname{diag}(r(\Lambda_e)) U_e^{\mathrm{T}}$, where

$$\zeta_i = u_{e,i}^{\mathrm{T}} C^{-1} u_{e,i} = \sum_j \lambda_j^{-1} R_{ij}^2, \qquad r_i(\Lambda_e) = \mathbb{E}[\zeta_i \mid \Lambda_e], \qquad (\widehat{\Lambda}_S)_{ii} = r_i(\Lambda_e)^{-1}. \tag{120}$$

Inversion occurs *after* averaging the precision. Neither averaging the reciprocals ζ_i^{-1} nor using $\mathbb{E}[\lambda_i \mid \Lambda_e]$ gives these coefficients in general. Invariance under the stabilizer of C_e forces both P_e and P_e^{-1} to be scalar on repeated empirical eigenspaces, just as for M_e.

The reverse convention $L_S^{\mathrm{rev}}(A, C) = \operatorname{tr}(A^{-1}C) - \log\det(A^{-1}C) - d$ instead has Bayes action $\mathbb{E}[C \mid C_e]$ when the corresponding covariance and log-determinant moments are finite. Indeed, its A-dependent risk is $\operatorname{tr}(A^{-1}M_e) + \log\det A$, uniquely minimized at M_e. Naming the convention is therefore necessary: "Stein-optimal posterior mean" is not correct for (117).

Why square-loss drift regression does not select a decision. The minimizer of (10) is $\mathbb{E}[R_t \mid I_t, C'_e]$, a conditional mean of a *drift target*. Proposition 3 shows that integrating the resulting stochastic dynamics produces the entire conditional endpoint law, not its mean. Positive time weights preserve this exact regression minimizer, although weights and the coordinate map can affect finite-capacity approximation. In particular, training in log-covariance coordinates does not make the terminal output $\exp(\mathbb{E}[\log C \mid C_e])$ or select a log-Euclidean point estimate. The same separation holds for supervised conditional interpolation; self-consistency addresses missing clean training pairs, not a missing decision loss.

G.8 Computing estimates after learning the spectral prior

There are two ways to extract a loss-specific estimate without modifying spectral EM. In either case, the target is the Bayes action for the *learned* RI prior. Matching the true-prior Bayes action additionally requires prior/posterior accuracy and suitable moment control; the posterior KL bounds in this paper alone do not control unbounded covariance or precision moments.

Posterior moment calculation. Draw $(\Lambda^{(r)}, R^{(r)})$ from (15), $r = 1, \dots, J$, and compute

$$\widehat{m}_i = \frac{1}{J}\sum_{r=1}^{J}\sum_{j}\lambda_j^{(r)}(R_{ij}^{(r)})^2, \qquad \widehat{r}_i = \frac{1}{J}\sum_{r=1}^{J}\sum_{j}(\lambda_j^{(r)})^{-1}(R_{ij}^{(r)})^2. \tag{121}$$

Then $U_e \operatorname{diag}(\widehat{m})U_e^{\mathrm{T}}$ and $U_e \operatorname{diag}((\widehat{r}_i^{-1})_i)U_e^{\mathrm{T}}$ are consistent Monte Carlo approximations of the two actions under independent exact draws and the stated moments. The reciprocal estimate is generally biased at finite J. Conditional integration over R can replace orientation sampling whenever tractable; means over a repeated empirical eigenspace can be tied to preserve equivariance at finite J. Spectral draws alone directly estimate spectral functionals such as trace and log-determinant, but do not by themselves yield the empirical-frame coefficients in (114) or (120).

A separate supervised shrinkage predictor. After fitting a spectral prior ν, independently generate

$$\Lambda \sim \nu, \qquad U \sim H, \qquad C = U\Lambda U^{\mathrm{T}}, \qquad C'_e \sim K_N(\cdot \mid C), \quad C'_e = U'_e\Lambda'_e U_e'^{\mathrm{T}}. \tag{122}$$

For a positive vector-valued predictor $f_\theta(\Lambda'_e)$, fit

$$\mathcal{J}_L(\theta) = \mathbb{E}_{\pi_\nu K_N}\left[L\big(U'_e \operatorname{diag}(f_\theta(\Lambda'_e))U_e'^{\mathrm{T}}, C\big)\right]. \tag{123}$$

Here π_ν is the RI matrix prior induced by ν. Assume the relevant population risks are finite and the predictor class can represent the conditional minimizers. The entire empirical spectrum is supplied to f_θ, and equal empirical eigenvalues have tied outputs when rotation equivariance is required. Define the known synthetic targets $\xi'_i = (u'_{e,i})^{\mathrm{T}} C u'_{e,i}$ and $\zeta'_i = (u'_{e,i})^{\mathrm{T}} C^{-1} u'_{e,i}$. By (112), the Frobenius objective differs from $\mathbb{E}\sum_i (f_i(\Lambda'_e) - \xi'_i)^2$ by a term independent of f. The Stein objective differs from

$$\mathbb{E}\sum_i \big(f_i(\Lambda'_e)\zeta'_i - \log f_i(\Lambda'_e)\big) \tag{124}$$

by a term independent of f. Conditioning on Λ'_e proves that the unrestricted population minimizers satisfy

$$f_i^F(\Lambda_e) = \mathbb{E}_{\pi_\nu}[\xi_i \mid \Lambda_e], \qquad f_i^S(\Lambda_e) = \big(\mathbb{E}_{\pi_\nu}[\zeta_i \mid \Lambda_e]\big)^{-1}. \tag{125}$$

Thus one learned prior supports both estimators, and the chosen downstream loss changes only the decision predictor. Haar sampling of U in (122) is prior sampling, not sampling from the nonuniform posterior orientation law. This route avoids that posterior group-sampling problem entirely when only a point estimate is needed. For simple empirical spectra, one can further simplify by taking $C = \Lambda$ when computing the invariant synthetic features $(\Lambda'_e, \xi', \zeta')$; a common orthogonal rotation changes none of them. For tied eigenspaces, use block-averaged targets with tied predictor outputs.

Neither predictor is fed back as the reconstructed target. Spectral EM re-corrupts draws from $q_{\nu_k}(d\Lambda \mid \Lambda_e)$; replacing them by $f^F(\Lambda_e)$, $f^S(\Lambda_e)$, or ξ defines a different latent distribution and loses the EM interpretation. These coefficients optimize decisions in the observed eigenbasis; they are not population-spectrum samples. This distinction is independent of whether the posterior was learned from supervised pairs or through self-consistency.

G.9 Nonlinear shrinkage and an averaged Bayes comparison

For $d/N \to q \in (0,1)$, the Ledoit–Péché expression for asymptotic directional-variance coefficients has the form

$$\delta_i^{\mathrm{LP}}(\Lambda_e) = \frac{\lambda_{e,i}}{|1-q-q\lambda_{e,i}\breve{m}_F(\lambda_{e,i})|^2}, \qquad m_F(z) = \int_{[0,\infty)} \frac{F(du)}{u-z}, \qquad \breve{m}_F(x) = \lim_{\epsilon \downarrow 0} m_F(x+i\epsilon), \quad x>0. \tag{126}$$

Here F is the limiting sample spectral distribution under the assumptions of the random-matrix results, and the displayed integral defines its Stieltjes transform. It is not the raw finite-sample eigenvalue measure. Ledoit–Wolf nonlinear shrinkage estimates the required spectral quantities; see Ledoit & Péché (2011); Ledoit & Wolf (2012). Neither method is defined as the posterior mean of the ordered population eigenvalues under our unknown ensemble prior.

Proposition 15 (An averaged connection to the RI Bayes estimate)**.** *For a square-integrable vector $\delta(\Lambda_e)$ measurable in the sample spectrum, and $m(\Lambda_e) = \mathbb{E}[\xi \mid \Lambda_e]$,*

$$\frac{1}{d}\mathbb{E}\|m(\Lambda_e) - \delta(\Lambda_e)\|_2^2 \le \frac{1}{d}\mathbb{E}\|\xi - \delta(\Lambda_e)\|_2^2. \tag{127}$$

Thus any averaged squared-error approximation of the directional oracle ξ by LP-type coefficients gives the corresponding approximation of the RI Frobenius Bayes coefficients.

Proof and scope. Conditional expectation is an orthogonal projection in L^2. Since $m(\Lambda_e) = \mathbb{E}[\xi \mid \Lambda_e]$ and $\delta(\Lambda_e)$ is measurable in Λ_e,

$$\mathbb{E}\|\xi - \delta(\Lambda_e)\|_2^2 = \mathbb{E}\|\xi - m(\Lambda_e)\|_2^2 + \mathbb{E}\|m(\Lambda_e) - \delta(\Lambda_e)\|_2^2. \tag{128}$$

This proves (127). For a sequence of RI models indexed by d, N, vanishing of the first displayed right-hand risk $d^{-1}\mathbb{E}\|\xi - \delta(\Lambda_e)\|_2^2$ implies vanishing of $d^{-1}\mathbb{E}\|m(\Lambda_e) - \delta(\Lambda_e)\|_2^2$ without further argument. However, that premise must itself be established in the considered random-matrix regime. Weak convergence of a weighted spectral measure alone does not imply it. One must also distinguish a deterministic limiting F, for which (126) defines a spectrum-measurable oracle, from an unobserved task-dependent random limiting law. In the latter situation a coefficient using that hidden law is not necessarily measurable in the observed spectrum, and the proposition cannot be applied without additional conditioning or estimation.

Under RI, the excess Bayes Frobenius risk of the estimator $U_e \operatorname{diag}(\delta(\Lambda_e))U_e^{\mathrm{T}}$ is exactly $\mathbb{E}\|m(\Lambda_e) - \delta(\Lambda_e)\|_2^2$. This connects coefficient approximation to point-estimation risk, but says nothing about the quality of posterior draws. Sampling $\operatorname{diag}(\xi)$ in the empirical frame discards off-diagonal posterior fluctuations; sampling Λ in that frame discards the relative orientation. Neither generally reproduces the full matrix posterior.

H Conjugate benchmarks and exact reference quantities

The experiments of Section 4 score reconstructions against a hidden *prior*. Scoring against exact *conditional posteriors* requires a family in which the posterior is available in closed form, and this appendix collects the relevant quantities for the conjugate inverse-Wishart family: the posterior itself, an independent reference sampler, analytic moments and conditional distributions, and the diagnostics they support, including one that distinguishes recovery of a posterior distribution from recovery of a single covariance estimate. In all of them the generator knows the prior and can compute its posterior, while the learner receives only the empirical-covariance ensemble, N, and the forward simulator.

H.1 Inverse-Wishart examples: one family for both implementations

We use the following explicit convention for the inverse-Wishart distribution:

$$C \sim \mathrm{IW}_d(\nu_0, \Psi_0), \qquad p(C) = \frac{|\Psi_0|^{\nu_0/2}}{2^{\nu_0 d/2}\Gamma_d(\nu_0/2)} (\det C)^{-(\nu_0+d+1)/2} \exp\left[-\frac{1}{2}\operatorname{tr}(\Psi_0 C^{-1})\right],$$
$$\Gamma_d(a) = \pi^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\left(a + \frac{1-i}{2}\right). \tag{129}$$

The density is relative to the independent entries of a symmetric matrix. Here ν_0 is a scalar number of degrees of freedom and $\Psi_0 \succ 0$ is a scale matrix. The law is proper for $\nu_0 > d-1$. Its usefulness here is conjugacy: multiplying its density by the Gaussian covariance likelihood gives another inverse-Wishart density (Alvarez et al., 2014). In the known-zero-mean model of (1), collecting the powers of $\det C$ and the terms in C^{-1} gives

$$\boxed{C \mid C_e \sim \mathrm{IW}_d(n, B), \qquad n = \nu_0 + N, \quad B = \Psi_0 + NC_e.} \tag{130}$$

This formula uses $NC_e = \sum_i x_i x_i^{\mathrm{T}}$, without centering by an estimated mean. It also holds for singular C_e because it follows from the raw Gaussian likelihood and $B \succ 0$. For reconstruction experiments without a restrictive prior assumption, the principal tests nevertheless use $N \geq d$, in accordance with Theorem 1. A known-prior reference posterior below this threshold does not make the unknown-prior problem identifiable.

RI example. Set $\nu_0 = d + 1 + \kappa_0$ and $\Psi_0 = \kappa_0 c_0 I_d$, where $c_0 > 0$ and $\kappa_0 > 2$. The prior is RI and $\mathbb{E}C = c_0 I_d$. Both Algorithms 1 and 2 can be tested against exactly the same full posterior. In particular,

$$\mathbb{E}[C \mid C_e] = \frac{N}{N+\kappa_0} C_e + \frac{\kappa_0}{N+\kappa_0} c_0 I_d. \tag{131}$$

Thus the exact posterior mean is a linear shrinkage rule in this benchmark. Recovering it alone is not evidence that a full posterior has been learned, nor does this example establish a benefit over nonlinear shrinkage for arbitrary population spectra. The posterior still has uncertainty in both eigenvalues and orientations.

Non-RI example. Replace $c_0 I_d$ by a fixed nonspherical $C_0 \succ 0$ and take $\Psi_0 = \kappa_0 C_0$. Keep the directions of C_0 fixed across tasks rather than rotating it independently for every task. This introduces shared directional information that the general method can learn. Formula (131) then holds with C_0 in place of $c_0 I_d$. A controlled choice is $C_0 = Q_0 \operatorname{diag}(c_1, \ldots, c_d) Q_0^{\mathrm{T}}$, with Q_0 drawn once and fixed, and c_i geometrically spaced before normalizing their mean to one. Normalize the chosen scale matrix, not the sampled covariances: fixing the trace of each draw would destroy (130).

Moment and convergence assumptions. The second-moment formulas below require $n > d + 3$. Choosing $\kappa_0 > 6$ also supplies finite fourth moments before conditioning, which is useful when estimating the sampling error of second-moment estimates (Matsumoto, 2012). Higher moments must be checked individually. These priors have unbounded support, so the compact-support recovery corollary in the main text does not apply automatically. Neither an explicit reference posterior nor good finite-data performance proves the unrestricted population convergence claim without its additional assumptions.

H.2 Independent reference samples and a full-matrix check

At a fixed C_e, compute a lower-triangular Cholesky factor L_B with $B = L_B L_B^{\mathrm{T}}$. Independently generate a lower-triangular matrix T whose entries satisfy

$$T_{ii} = \sqrt{\chi^2_{n-i+1}}, \qquad T_{ij} \sim \mathcal{N}(0,1)\ (i > j), \qquad T_{ij} = 0\ (i < j), \tag{132}$$

with all displayed random variables mutually independent. The Bartlett construction gives $TT^{\mathrm{T}} \sim W_d(n, I_d)$. Consequently,

$$C^{\mathrm{ref}} = L_B (TT^{\mathrm{T}})^{-1} L_B^{\mathrm{T}} \sim \mathrm{IW}_d(n, B). \tag{133}$$

In code, solve $TD = L_B^{\mathrm{T}}$ and return $D^{\mathrm{T}} D$ instead of explicitly forming the inverse. This is a reference construction, not a claim of an optimal implementation; direct inverse-Wishart factor samplers are also available (Axen, 2023). Independent reference draws require no MCMC or equilibration period. Diagonalizing independent C^{ref} gives independent reference spectra; under RI these have exactly the desired law $P_\star(d\Lambda \mid \Lambda_e)$.

The same construction supplies a strong diagnostic for *full* posterior samples. Given a generated C, form

$$W = L_B^{\mathrm{T}} C^{-1} L_B, \qquad T = \mathrm{chol}(W). \tag{134}$$

For a correct posterior draw, $W \sim W_d(n, I_d)$ and the Cholesky entries have the independent distributions in (132). Test the diagonal squares against their chi-squared distributions and the lower off-diagonal entries against standard normals. Check dependence between entries as well: correct one-dimensional histograms alone do not establish the joint distribution. This transformation is applied separately at each observation, with its own B; it can detect orientation errors even when the spectrum of C has been reproduced.

H.3 Analytic posterior quantities for testing accuracy

Fix C_e and put $r = n - d - 1 > 2$. The following formulas provide checks without a Monte Carlo approximation to the reference moments. First and second inverse-Wishart moments can be obtained from the inverse-Wishart moment identities; see, for example, Matsumoto (2012). With the convention (129), they give

$$M_\star = \mathbb{E}[C \mid C_e] = \frac{B}{r}, \qquad \Omega_\star = \mathbb{E}[C^{-1} \mid C_e] = nB^{-1}, \tag{135}$$

$$\mathrm{Cov}(C_{ij}, C_{k\ell} \mid C_e) = \frac{2B_{ij}B_{k\ell} + r(B_{ik}B_{j\ell} + B_{i\ell}B_{jk})}{(r+1)r^2(r-2)}. \tag{136}$$

The precision mean in (135) also follows directly from $C^{-1} \mid C_e \sim W_d(n, B^{-1})$. The corresponding Stein-optimal covariance estimate, for the loss convention in Appendix G.7, is $\Omega_\star^{-1} = B/n$, not $M_\star$.

Covariance and precision means. For J generated draws, estimate these quantities by $\widehat{M} = J^{-1}\sum_j C^{(j)}$ and $\widehat{\Omega} = J^{-1}\sum_j (C^{(j)})^{-1}$. Report

$$e_M = \frac{\|\widehat{M} - M_\star\|_F}{\|M_\star\|_F}, \qquad e_\Omega = \frac{\|\widehat{\Omega} - \Omega_\star\|_F}{\|\Omega_\star\|_F}. \tag{137}$$

Also compare entrywise variances and a specified subset of covariances using (136). The full array has $O(d^4)$ entries, so large-dimensional tests should fix a reproducible subset rather than promise to estimate every entry. Mean errors and variance errors answer different questions: a sampler returning only $M_\star$ would pass a first-moment check and fail uncertainty checks.

Directional-variance distributions. For a nonzero direction v fixed conditional on C_e, the scalar $V_v = v^{\mathrm{T}}Cv$ has distribution

$$\begin{aligned} V_v \mid C_e &\sim \mathrm{InvGamma}\left(\frac{r+2}{2}, \frac{v^{\mathrm{T}}Bv}{2}\right), \\ \mathbb{E}[V_v \mid C_e] = \frac{v^{\mathrm{T}}Bv}{r}, &\qquad \mathrm{Var}(V_v \mid C_e) = \frac{2(v^{\mathrm{T}}Bv)^2}{r^2(r-2)}. \end{aligned} \tag{138}$$

Here the inverse-gamma density is proportional to $x^{-a-1}e^{-b/x}$ for shape a and scale b. To see the distributional statement, rotate and rescale v to a coordinate direction and use the one-dimensional inverse-Wishart marginal (Alvarez et al., 2014). Choose coordinate vectors, empirical eigenvectors, and selected sums $(u_{e,i} + u_{e,j})/\sqrt{2}$. The sums involve off-diagonal covariance entries and therefore test more than variances along empirical axes. Directions may depend on the observation but not on the generated posterior sample being tested.

Let F_{v,C_e} be the exact cumulative distribution function in (138). For each held-out observation, check the empirical distribution of

$$U_v^{(j)} = F_{v,C_e}\left(v^{\mathrm{T}}C^{(j)}v\right). \tag{139}$$

Its correct marginal distribution is uniform on $[0, 1]$. Quantile plots, the maximum absolute difference from the uniform cumulative distribution, and errors in fixed tail probabilities provide conditional diagnostics. Serially correlated MCMC outputs are not independent uniform observations, so independence-based significance thresholds must not be used without a justified treatment of that correlation.

An orientation-sensitive RI check. For the RI example, write $C_e = U_e \Lambda_e U_e^{\mathrm{T}}$, set $\widetilde{C} = U_e^{\mathrm{T}} C U_e$, and let $b_i = \kappa_0 c_0 + N\lambda_{e,i}$. Then $U_e^{\mathrm{T}} B U_e = \mathrm{diag}(b_i)$, and (136) implies, for $i \neq j$,

$$\mathbb{E}[\widetilde{C}_{ij} \mid C_e] = 0, \qquad \boxed{\mathbb{E}[\widetilde{C}_{ij}^2 \mid C_e] = \frac{b_i b_j}{(r+1)r(r-2)}}, \qquad \mathrm{Var}(\widetilde{C}_{ii} \mid C_e) = \frac{2b_i^2}{r^2(r-2)}. \tag{140}$$

The ratio of each sampled off-diagonal second moment to this exact value has target one, and returning $U_e \Lambda U_e^{\mathrm{T}}$ without orientation completion makes every such ratio zero regardless of spectral accuracy. This is the sharpest available separation between matching a shrinkage estimate and recovering the full posterior.

Spectral summaries. Compare joint ordered log spectra and the conditional distributions of $\mathrm{tr}\, C$, $\log\det C$, the largest eigenvalue, and the condition number $\lambda_{\max}/\lambda_{\min}$. The latter distributions can be compared with reference draws rather than approximated by functions of posterior mean eigenvalues. The log determinant also has analytic moments:

$$\begin{aligned} \mathbb{E}[\log\det C \mid C_e] &= \log\det B - d\log 2 - \sum_{i=1}^{d} \psi\left(\frac{n+1-i}{2}\right), \\ \mathrm{Var}(\log\det C \mid C_e) &= \sum_{i=1}^{d} \psi_1\left(\frac{n+1-i}{2}\right). \end{aligned} \tag{141}$$

Here ψ and ψ_1 are the first and second derivatives of $\log\Gamma$. These formulas follow from $\log\det C = \log\det B - \sum_i \log T_{ii}^2$ in (133) and the independent chi-squared diagonal variables. Spectral checks do not validate orientation; both kinds of diagnostics are necessary. Accuracy of arbitrary unbounded moments is not implied by a distributional error bound without suitable tail control.

H.4 Mixtures with explicit posteriors and an exact small-scale EM control

A finite mixture allows more varied posterior shapes while keeping the reference computable. Let p_h be the density of $\mathrm{IW}_d(\nu_h, \Psi_h)$ and take

$$\pi_\star(dC) = \sum_{h=1}^{H} w_h p_h(C)\, dC, \qquad w_h > 0, \quad \sum_h w_h = 1. \tag{142}$$

Applying (130) component by component gives

$$\begin{aligned} P_\star(dC \mid C_e) &= \sum_{h=1}^{H} \widetilde{w}_h(C_e)\ \mathrm{IW}_d(\nu_h + N, \Psi_h + NC_e)(dC), \\ \widetilde{w}_h(C_e) &= \frac{w_h z_h(C_e)}{\sum_\ell w_\ell z_\ell(C_e)}, \qquad z_h(C_e) = \frac{|\Psi_h|^{\nu_h/2}\Gamma_d((\nu_h+N)/2)}{\Gamma_d(\nu_h/2)|\Psi_h + NC_e|^{(\nu_h+N)/2}}. \end{aligned} \tag{143}$$

The expression for z_h is the ratio of inverse-Wishart normalizing constants after updating; powers of two and observation factors common to all components cancel. Calculate weights from log determinants and log gamma functions, subtracting the largest log weight before exponentiation. Reference sampling first draws h with probability $\widetilde{w}_h$ and then uses (133) for that component.

For an RI mixture, take $\Psi_h = \kappa_h c_h I_d$ and $\nu_h = d + 1 + \kappa_h$. For a non-RI mixture, use fixed nonspherical $C_{0,h}$ in place of $c_h I_d$. All relevant component moments must exist. Means and one-dimensional cumulative distributions are posterior-weighted averages of the component quantities. Variances include both within-component and between-component terms: if a_h and s_h^2 are the component mean and variance of $f(C)$, then

$$\mathbb{E}[f(C) \mid C_e] = \sum_h \widetilde{w}_h a_h, \qquad \mathrm{Var}(f(C) \mid C_e) = \sum_h \widetilde{w}_h(s_h^2 + a_h^2) - \left(\sum_h \widetilde{w}_h a_h\right)^2. \tag{144}$$

Mixture uncertainty need not remain appreciable at every observation, so accuracy separates naturally into observations with uncertain component membership and those dominated by one component,

under a grouping rule fixed in advance. Generated matrices should not be labelled by a hard nearest-component rule; their exact component probabilities are available,

$$a_h(C, C_e) = \frac{\widetilde{w}_h(C_e) p_h^{\mathrm{post}}(C \mid C_e)}{\sum_\ell \widetilde{w}_\ell(C_e) p_\ell^{\mathrm{post}}(C \mid C_e)}, \qquad \mathbb{E}[a_h(C, C_e) \mid C_e] = \widetilde{w}_h(C_e), \tag{145}$$

and compare the sampled average with the right-hand side. This avoids arbitrary labels for overlapping components. A mixture of isotropic-scale inverse-Wisharts still has a posterior mean of the form $a(C_e)C_e + b(C_e)I_d$. It tests distributional learning and data-dependent shrinkage, not general spectral shrinkage functions.

An exact empirical EM reference. There is a useful additional control for the SI–EM implementation. Start from any specified finite inverse-Wishart mixture and use a very small recorded ensemble. Its posterior at each observation is explicit by (143), so

$$\pi_{k+1} = \frac{1}{M} \sum_{a=1}^{M} P_{\pi_k}(\cdot \mid C_{e,a}) \tag{146}$$

is again an exactly specified finite mixture. Its component count is at most M times the preceding count; no approximate component merging, pruning, or projection is permitted when calling this the exact reference. At small M and for a few iterations it therefore gives an exactly known reference trajectory, against which a learned reconstruction and its conditional posteriors can be compared at the same iteration; this isolates approximation of the EM step from the finite-data difference between EM and the true prior. The empirical log likelihood is available for a monotonicity check from (143), up to terms fixed by the observations. The exponential growth in component count makes this a diagnostic rather than a scalable method, and it concerns empirical EM rather than the population theorem.

H.5 Conditional discrepancies and decision accuracy

Two properties make a discrepancy informative here. It must be computed conditionally on each C_e before being averaged across tasks, since pooling generated matrices across observations can conceal a sampler that largely ignores its conditioning input; and it must compare distributions rather than moments, since a sampler returning the posterior mean passes every first-moment check.

A joint distribution discrepancy. Supplement the analytic checks by maximum mean discrepancy (MMD), a difference between kernel averages from two distributions (Gretton et al., 2012). For two independent draws X, X' from the generated law and Z, Z' from the reference, define

$$\begin{aligned} \mathrm{MMD}_h^2 &= \mathbb{E}k_h(X, X') + \mathbb{E}k_h(Z, Z') - 2\mathbb{E}k_h(X, Z), \\ k_h(A, B) &= \exp\left[-\frac{\| \log A - \log B \|_F^2}{2h^2} \right]. \end{aligned} \tag{147}$$

The Gaussian kernel in invertible log coordinates distinguishes probability distributions, so the entire matrix law is compared without estimating a density in dimension $d(d+1)/2$; finite-sample power, however, depends on dimension and bandwidth, and the sample estimator has a nonzero fluctuation that an exact-reference against exact-reference comparison at matching sample sizes quantifies. Independent-sample test thresholds are not valid for correlated MCMC output.

Coverage and rank checks. Central posterior intervals for selected $f(C)$ can be compared with the exact conditional intervals, and the frequency with which the hidden $f(C)$ falls inside them measures calibration across tasks. Simulation-based rank checks (Talts et al., 2018) are weaker than they appear in this setting: averaged over observations drawn from the benchmark model, some of them are satisfied even by a sampler that ignores C_e, so the fixed-observation tests (139) and (134) are the primary ones.

Connection to LW/LP. Covariance-estimation risk and posterior accuracy are different quantities. Under squared Frobenius loss, the exact benchmark action is $M_\star$, and for an estimate $A = A(C_e)$,

$$\mathbb{E}[\|A - C\|_F^2 \mid C_e] - \mathbb{E}[\|M_\star - C\|_F^2 \mid C_e] = \|A - M_\star\|_F^2. \tag{148}$$

The same identity holds after averaging independent Monte Carlo randomness used to construct A. Under the stated Stein convention, compare with $\Omega_\star^{-1}$; for mixtures, invert the posterior mixture mean of C^{-1}, not the individual component means before averaging. Include the empirical covariance, linear Ledoit–Wolf shrinkage, an implemented nonlinear shrinkage estimator associated with the LP oracle, and an inverse-Wishart prior fitted using only the same noisy ensemble. The last receives the same noisy data but assumes a restricted family; fit its parameters by maximizing the observed marginal likelihood obtained from z_1 in (143), not by using hidden clean covariances. The known-prior posterior and the supervised SI have additional information and are diagnostic references, not equally informed competitors. Neither LW nor nonlinear shrinkage supplies posterior intervals by itself. For empirical-PCA variance and covariance risk, compare the coefficients $\mathbb{E}[\xi_i \mid \Lambda_e]$, not $\mathbb{E}[\lambda_i \mid \Lambda_e]$.

I TRANSPORTING SAMPLES INSTEAD OF MATRICES

The state of the interpolant need not be a matrix. This appendix develops the variant whose state is a single Gaussian sample $x \in \mathbb{R}^d$, which makes every training example $O(d)$ and never forms a matrix logarithm or an eigendecomposition. It is a construction with its cost stated, not a method we evaluate.

I.1 ONE CONDITIONAL SUFFICES

Given C, the N samples behind C_e and any further samples are exchangeable, and the likelihood depends on C only through their pooled second moment. Conditioning on C_e together with samples $x_1, \dots, x_{j-1}$ already generated is therefore the same as conditioning on a single empirical covariance with a larger count,

$$S_m = \frac{1}{m}\Big(NC_e + \sum_{i<j} x_i x_i^{\mathrm{T}}\Big), \qquad m = N + j - 1. \tag{149}$$

A single drift $b_t(\cdot\,, S, m)$ thus covers every step of the generation. It is trained exactly as in Section 3.1: draw $C \sim \pi_k$, draw m samples from $\mathcal{N}(0, C)$ to form S with m drawn over the intended range, draw one further sample as the regression target, and interpolate between a draw from $\mathcal{N}(0, S)$ and that target.

I.2 A POSTERIOR DRAW IS THE LIMIT OF THE SEQUENCE

Proposition 16 (Sample-space posterior draws)**.** *Fix C_e, set $S_N = C_e$, and generate $x_1, x_2, \dots$ sequentially, each from the predictive law associated with the current (S_m, m) and updated through (149). Then the joint law of $x_{1:n}$ is $\int \mathcal{N}(0, C)^{\otimes n} P_\pi(dC \mid C_e)$ for every n; conditionally on C the samples are independent; and $S_m \to C$ almost surely, with $C \sim P_\pi(\cdot \mid C_e)$.*

Proof. The sequential predictive of an exchangeable model reproduces its joint law, which gives the first assertion, and the second is the conditional independence in that representation. Conditionally on C the x_i are i.i.d. with covariance C, so the strong law gives $S_m \to C$; the limit is the de Finetti measure of the sequence, distributed as $P_\pi(\cdot \mid C_e)$. □

Three consequences are worth stating, since they are what makes the variant usable. The covariance is never assembled as a separate step: S_m is an empirical approximation to the latent posterior draw C, with relative Frobenius root-mean-square error $O(\sqrt{d/m})$ under Gaussian sampling. Further samples from that same covariance do not require the network, since $\mathcal{N}(0, S_m)$ can be sampled directly once S_m has stabilized. And independent posterior draws come from independent runs at the same C_e, which diverge because their early samples differ: the posterior spread is resolved at the start of the sequence and is exhausted thereafter.

I.3 WHAT IT COSTS

Pinning C to relative Frobenius accuracy ε needs $m \sim d/\varepsilon^2$, so one posterior draw costs that many network evaluations, each $O(d)$, with an $O(d^2)$ rank-one update of S. Against a single integration of a drift in dimension $p = d(d+1)/2$, the trade is a cheaper training example for a dearer draw, and it favours this variant only when d is large and the accuracy demanded of an individual draw is modest. Time-discretization error is also incurred once per sample rather than once per covariance.

I.4 A CHECK AVAILABLE BEFORE ANY OUTER LOOP

At $m = N$ a single step is the posterior predictive of a fresh sample given C_e. Under the inverse-Wishart prior of Appendix H this law is a multivariate t in closed form, so the learned conditional can be validated directly, at a fixed observation, before it is used inside the self-consistency iteration.

I.5 ALTERNATIVES

The batch can instead be transported at once, with state $(x_1, \ldots, x_n) \in \mathbb{R}^{nd}$ and a permutation-equivariant drift, which costs one integration rather than m and imposes exchangeability through the architecture. A shared latent drawn once per batch, with the samples conditionally independent given it, is cheaper still but exact only if that latent encodes C, which is an assumption on the fitted model rather than a property of the construction.

J EXPERIMENTAL DETAILS

This appendix describes the three experiments of Section 4 and their evaluation protocol (§J.1). The toy models with oracle posteriors is discussed in (§J.2), the EEG experiment in (§J.3) and the financial-returns experiment in (§J.4). We also discuss the cross-validation rule used to select the EM iteration across experiments (§J.5).

J.1 IMPLEMENTATION AND PROTOCOL COMMON TO ALL EXPERIMENTS

Algorithm, network and training. All experiments run Algorithm 1 (matrix-logarithm version). A covariance is encoded as $Y = \mathrm{svec}(\log C)$ (upper triangle, off-diagonals scaled by $\sqrt{2}$, dimension $p = d(d{+}1)/2$) and the noise level is $\sigma = \kappa\sqrt{2/N}$ with $\kappa = 1$. The drift $b_\theta(t, I, Y_\mathrm{e})$ is an MLP on the standardised interpolant state and observation with Fourier features of t (frequencies $2^j\pi$, $j{=}0, \dots, 7$), four hidden layers with SiLU activations, a zero-initialised output layer and a skip term $t\,(I - Y_\mathrm{e})$ that imposes the exact endpoint limit $b_1(y, Y_\mathrm{e}) = y - Y_\mathrm{e}$. The width is 256 for the toys (1024 at $d{=}20$) and 1024 for EEG and finance. Each fit minimises $\mathbb{E}\,\|b_\theta(t, I_t, Y'_\mathrm{e}) - R_t\|^2$ with $I_t = (1{-}t)Y'_\mathrm{e} + tY + \sigma(1{-}t)W_t$ and $R_t = Y - Y'_\mathrm{e} - \sigma W_t$, drawing t, the Brownian state and the Wishart corruption afresh at every step (AdamW, batch 1024, cosine learning-rate decay, gradient-norm clipping at 5, weight moving average 0.999, which is the network used for sampling). The initial fit (4000 steps, learning rate $10^{-3} \to 10^{-4}$) uses a sampleable prior π_0: a log-Gaussian fitted to the encoded ensemble by moment-matching deconvolution (channel-noise covariance subtracted, result inflated by 1.5). Each outer iteration $k = 1, \dots, K$ reconstructs every recorded covariance with n_rec posterior draws (8; 6 in finance) by the Föllmer SDE (Euler–Maruyama, 64 steps), then refits the warm-started network for 800 steps (learning rate $5 \cdot 10^{-4} \to 10^{-4}$) on the $M\,n_\mathrm{rec}$ reconstructions, re-corrupted by the channel. Matrix functions and the SDE run in double precision; test-time draws use the same 64-step sampler.

Baselines. *SCM*: C_e itself. *OAS*: linear shrinkage of C_e toward $(\mathrm{tr}\,C_\mathrm{e}/d)\,I$. *LW-NLS*: analytical nonlinear shrinkage (case $N > d$), which keeps the eigenvectors of C_e and shrinks the eigenvalues using a kernel estimate of the sample spectral density. Both use only (C_e, N), with N the effective sample size of the experiment. *IW-ML*: a conjugate inverse-Wishart prior $\mathrm{IW}(\nu_0, \Psi_0)$, with full scale matrix and $\nu_0 > d{+}1$, fitted by maximising the closed-form marginal likelihood of the training ensemble (600-800 Adam iterations); its posterior $\mathrm{IW}(\nu_0{+}N, \Psi_0{+}NC_\mathrm{e})$ is sampled exactly. A point estimate $\hat{C}$ becomes a predictive law by adding channel noise around it, $C_\mathrm{e}^\mathrm{val} \mid \hat{C} \sim \mathrm{Wishart}_N(\hat{C}/N)$, and carries no estimation uncertainty of its own.

Scores. The *predictive log-score* of a held-out $C_\mathrm{e}^\mathrm{val}$ given a companion $C_\mathrm{e}^\mathrm{cal}$ is

$$\log \frac{1}{J} \sum_{j=1}^{J} p_N(C_\mathrm{e}^\mathrm{val} \mid C_j),$$

with C_j posterior draws given $C_\mathrm{e}^\mathrm{cal}$ ($J{=}64$ for validation, 256 for test) and p_N the Wishart density ($J{=}1$ for a plug-in). Terms free of C are dropped, so log-scores are comparable across methods within an experiment; on real data we report differences to the sample covariance. The *coverage* of central intervals of level $\ell \in \{0.5, 0.8, 0.95\}$ for a functional $f(C)$ uses the probability integral transform (PIT): each posterior draw is pushed through the channel, the PIT is the fraction of the resulting draws of f below $f(C_\mathrm{e}^\mathrm{val})$, and the coverage is the fraction of cases whose PIT lies between $(1{-}\ell)/2$ and $(1{+}\ell)/2$. Confidence intervals are 95% percentile bootstrap intervals over independent units (1000 resamples of subjects for EEG, of months for finance).

J.2 TOY MODELS WITH ORACLES

Priors and oracles. We consider six covariance priors, designed to cover both conjugate and non-conjugate settings, as well as isotropic, structured, and diagonal geometries.

IW, RI. We use an inverse-Wishart prior $C \sim \mathrm{IW}(\nu_0, \Psi_0)$ with $\nu_0 = d + 1 + \kappa_0 = 25$ and $\Psi_0 = \kappa_0 I_8 = 16 I_8$, with $\kappa_0 = 16$. The prior is isotropic, with $\mathbb{E}[C] = I_8$, and admits a closed-form conjugate posterior.

IW, non-RI. We retain the same inverse-Wishart family, but set $\Psi_0 = 16Q\,\mathrm{diag}(\lambda)Q^\top$, where Q is a fixed random rotation (seed 123) and λ is a geometrically spaced spectrum from 1 to 16, normalized to unit mean. This gives a condition number of 16 with a fixed but arbitrary orientation.

Factor model. We use $C = BB^\top + \mathrm{diag}(\exp(\ell))$, with $B \in \mathbb{R}^{8\times 2}$. The first factor has positive loadings, $b_1 \sim \mathcal{N}(1.0, 0.3^2)$, while the second has sign-free loadings, $b_2 \sim \mathcal{N}(0, 0.6^2)$. The idiosyncratic log-variances follow $\ell \sim \mathcal{N}(\log 0.4, 0.4^2)$, giving a median idiosyncratic variance of 0.4.

AR(1). We use the Toeplitz covariance $C_{ij} = \rho^{|i-j|}$ with $\rho \sim \mathrm{Uniform}(0.3, 1.0)$, excluding the singular endpoint $\rho = 1$.

Iso-diag. We set $C = U\,\mathrm{diag}(\mathrm{spikes})U^\top$, where U is a Haar-random orthogonal matrix and the eigenvalues are independently drawn from a truncated normal $\mathcal{N}(3.0, 1.0^2)$ restricted to $[2, 4]$. This produces a relatively concentrated spectrum with random orientation.

Diagonal. Finally, we use $C = \mathrm{diag}(\exp(u))$, with independent $u_i \sim \mathcal{N}(0, 0.5^2)$. The marginal variance has median 1, with approximately $[0.61, 1.65]$ spanning one standard deviation on the log scale, and all off-diagonal correlations are zero.

The posterior is exact for both inverse wishart law (RI and non-RI). For Factor, Iso-diagonal and diagonal models, we run a batched Hamiltonian Monte Carlo Neal (2011) on $(B, \log\psi)$: 8 chains with 256 retained draws each per test task, thinning 4, 1500 warm-up iterations with step size adapted to a 75% acceptance rate, 20 leapfrog steps with 10% step-size jitter. For AR(1) the posterior is one-dimensional, $p(\log\gamma \mid C_\mathrm{e}) \propto \mathcal{N}(\log\gamma; 0, 0.5^2)\exp\{-\frac{N}{2}[\mathrm{tr}(C(\gamma)^{-1}C_\mathrm{e}) + \log\det C(\gamma)]\}$, evaluated on 3001 grid points over ± 6 prior standard deviations.

Training and evaluation. We use a four-layer MLP, with 256 neurons per layer and SiLU activations. We run 4,000 initial steps followed by 800 EM steps, with $K = 24$ EM iterations for all models except AR(1), for which we use $K = 120$. For $d = 8$, a 24-iteration run takes approximately 80,s on a single A100. Evaluation is performed on 128 held-out tasks, each consisting of one observation from a fresh latent covariance matrix. We use 1,024 posterior draws per task for SC-SI and IW-ML, and 2,048 draws for the oracle. For each of six statistics of C—the log-determinant, log condition number, top-eigenvalue share, log variance along the top and bottom eigenvectors of C_e, and the correlation $C_{01}/\sqrt{C_{00}C_{11}}$—we measure the Wasserstein-1 distance W_1 between the empirical distributions produced by the learner and the oracle. The distance is normalized by the oracle posterior standard deviation and averaged across tasks.

J.3 EEG

Data and effective sample size. We use runs R01 (eyes open) and R02 (eyes closed) of the PhysioNet EEG Motor Movement/Imagery dataset: 109 subjects, 160 Hz. Eight channels (Fz, C3, Cz, C4, Pz, O1, Oz, O2) are re-referenced to the average of all 64 channels, band-pass filtered to 8–13 Hz (FIR, 2 Hz transition bands), trimmed by 2 s at both ends and cut into 14 non-overlapping 4-s windows ($n = 640$ samples) per run. The observation of a task (subject, condition, window) is $C_\mathrm{e} = \frac{1}{n}XX^\top$; the condition is not given to the prior. Successive samples of a band-limited signal are correlated, so N is an effective sample size. The autocorrelation formula $N_\mathrm{eff}/n = (1 + 2\sum_{k\geq 1}\rho_k^2)^{-1}$ has median 0.067 over channels and runs ($N_\mathrm{eff} \approx 43$, as the time–bandwidth product $2BT = 40$). The test–retest variability of interleaved half-windows on the *training* subjects is larger (alpha bursts make window power more variable) and corresponds to $N \approx 24$–36; we use $N = 28$ ($N/d = 3.5$). An earlier run with a 256-unit network at $N = 43$ gives the same ordering of SC-SI, IW-ML and OAS (split-half gain $+11.4$ $[8.7, 13.6]$ nats for SC-SI, $+5.7$ $[3.3, 7.7]$ for IW-ML, -13.6 for OAS).

Splits and training. Subjects are split at random into 70 training, 10 validation and 29 test subjects. The prior is learned from the $70 \times 2 \times 14 = 1960$ unlabeled training covariances (width 1024, K=30, 800 steps per iteration, $n_\mathrm{rec} = 8$). The checkpoint is selected on 260 validation pairs (window j as calibration, window j+1 of the same run as validation) with the worst-case PIT score over all six functionals; the log-score selects the same iteration ($k = 8$). For classification every subject must be tested once, so the whole procedure (prior, classifier, checkpoint) is repeated in 5 subject-disjoint folds with $22, 22, 22, 22, 21$ test subjects, 10 validation subjects and $\approx$ 77 training subjects; the selected iterations are $5, 8, 12, 12, 16$.

Table 2: **EEG: log-score and calibration** (29 held-out subjects). (a) Predictive log-score minus that of the sample covariance (nats per window, 95% subject-level bootstrap intervals; bold: best) and coverage of the central 50/80/95% intervals of eleven directional variances. (b) 80% coverage of nonlinear functionals in the split-half protocol (nominal 0.80; bold: closest); Riem. dist.: affine-invariant distance between the eyes-open and eyes-closed covariances.

	(a) Split-half (812 cases)		Next window (754 cases)	
Method	gain ↑	coverage	gain ↑	coverage
Sample cov.	0	0.38/0.65/0.84	0	0.31/0.54/0.72
OAS	-16.0 [-22.4, -10.8]	0.37/0.63/0.82	-7.9 [-14.8, -2.7]	0.29/0.53/0.70
LW-NLS	$+2.8$ [2.2, 3.4]	0.38/0.65/0.84	$+8.6$ [7.1, 10.0]	0.31/0.54/0.72
IW-ML	$+0.1$ [-1.8, 1.7]	0.46/0.76/0.92	$+8.7$ [5.6, 11.6]	0.39/0.66/0.85
SC-SI	$+\mathbf{5.0}$ [**3.0**, **6.5**]	0.47/0.76/0.91	$+\mathbf{15.7}$ [**12.3**, **18.8**]	0.38/0.65/0.82
Long-run proxy (non-causal)	–	–	$+23.5$ [20.1, 27.2]	–

(b) Method	$\log\det C$	log Oz	top share	$\log\kappa$	Riem. dist.
Sample cov.	0.34	0.66	0.71	0.53	0.26
OAS	0.17	0.64	0.55	0.14	0.39
LW-NLS	0.47	0.66	0.69	0.66	0.38
IW-ML	0.43	0.73	**0.83**	0.55	0.36
SC-SI	**0.65**	**0.77**	0.77	**0.77**	**0.41**

Protocols and metrics. The posterior is computed from a calibration window $C_{\mathrm{e}}^{\mathrm{cal}}$ and predicts a validation window $C_{\mathrm{e}}^{\mathrm{val}}$ through the channel. In the *split-half* protocol the two windows are the even and odd 1-s chunks of the same 8-s epoch (in both directions, 7 epochs per run), so they share one covariance and the test isolates estimation uncertainty (812 cases). In the *next-window* protocol the validation window follows the calibration window in the same run, so the covariance may drift (754 cases). Coverage is measured for eleven directional variances $v^\top C v$ (the eight channel axes and the top three principal components of the pooled training covariance) and, in the split-half protocol, for four functionals (log-determinant, log power at Oz, top-eigenvalue share, log condition number) and for the affine-invariant Riemannian distance between the eyes-open and eyes-closed covariances of the same epoch and half. The long-run reference of Table 2 is the plug-in built from the run average of all other windows; it is not causal and only indicates the attainable score.

Log-score and calibration. Table 2 gives the results behind Figure 3a,b. The paired margins of SC-SI in the split-half protocol are $+2.3$ $[0.4, 3.6]$ nats over LW-NLS and $+4.9$ $[3.1, 6.6]$ over IW-ML. SC-SI and IW-ML are close to nominal on directional variances while the plug-ins under-cover; coverage falls below nominal for every method in the next-window protocol, where the covariance moves and the common-C channel has no term for that.

Classification. The task is eyes open vs. eyes closed from a single window. Features are the tangent-space coordinates $\mathrm{svec}\log(C_{\mathrm{ref}}^{-1/2} C C_{\mathrm{ref}}^{-1/2})$ at the arithmetic mean of the training covariances (Barachant et al., 2012), standardised, with an ℓ_2-regularised logistic regression (inverse strength 1). The plug-in classifiers apply it to the sample, OAS or LW-NLS covariance of the test window. SC-SI averages the probability over 64 posterior draws of the test window, with the regression trained either on the sample covariances or, as in the main text, on four posterior draws per training window. The confidence is $|p - \frac{1}{2}|$ and "acc@q" is the accuracy on the fraction q of windows with the highest confidence. Table 3 pools the five folds (3052 windows). The gain of SC-SI is small: $+1.4$ points over the sample covariance and $+1.5$ over LW-NLS, with lower interval ends at 0.1 and 0.0; SC-SI is better for 47, 51 and 55 subjects against 26, 25 and 25 for the sample covariance, LW-NLS and OAS. On the confident half of the windows the gain is significant against OAS only. Its calibration error (0.030) is below those of the plug-ins (0.036–0.044). On a single 29-subject split the gain was larger ($+3$ to $+4$ points) and does not carry over to the whole cohort.

Table 3: **EEG: eyes-open vs. eyes-closed classification**, 5-fold subject-disjoint cross-validation over all 109 subjects. Top: pooled metrics (acc@q: accuracy on the q most confident windows; NLL: negative log-likelihood; ECE: expected calibration error; bold: best). Bottom: paired differences SC-SI−baseline in percentage points (accuracy) and nats (NLL) with 95% subject-level bootstrap intervals, number of subjects with higher/lower accuracy, and Wilcoxon signed-rank p.

Method	acc ↑	acc@80% ↑	acc@50% ↑	NLL ↓	ECE ↓
Sample cov. (plug-in)	0.792	0.849	**0.915**	0.464	0.042
OAS (plug-in)	0.781	0.825	0.871	0.492	0.036
LW-NLS (plug-in)	0.791	0.844	0.906	0.466	0.044
SC-SI, LR trained on C_e	0.805	**0.860**	0.910	**0.443**	**0.023**
SC-SI, LR trained on posterior draws	**0.806**	0.859	0.911	**0.443**	0.030

	Δacc	Δacc@50%	ΔNLL	better/worse	p
vs Sample cov.	+1.4 [0.1, 2.8]	−0.5 [−2.4, 1.3]	−0.021 [−0.043, 0.000]	47/26	0.026
vs OAS	+2.6 [0.4, 4.6]	+3.7 [1.2, 6.2]	−0.048 [−0.082, −0.014]	55/25	0.002
vs LW-NLS	+1.5 [0.0, 3.0]	+0.4 [−1.9, 2.6]	−0.023 [−0.045, −0.002]	51/25	0.015

J.4 Financial returns

Data, tasks and folds. We use the daily value-weighted returns (in percent) of the 48 Ken French industry portfolios, July 1926 to July 2026, with a zero-mean model; industries not yet listed are never used. A task is a window of 63 trading days and a random basket of $d = 12$ industries observed throughout it; its observation is the second-moment matrix of the window, and the forecast target is that of the following 21 trading days. The prior is learned from past windows only and refit on three expanding folds (training before 1995, 2005 and 2015; test months 1995–2004, 2005–2014, 2015–2026), so no test month contributes to the prior that scores it. A training ensemble is made of the non-overlapping windows preceding the fold with 16 random baskets each ($M \approx 4.5$–5.7k); the last three years before each cutoff are held out for checkpoint selection (calibration window and next 21 days, 8 random baskets, monthly). The network has width 1024 and is trained for $K = 30$ iterations of 800 steps with $n_{\mathrm{rec}} = 6$; the iteration is selected per fold (§J.5). The test set has 8 fixed baskets per fold, drawn among industries observed without gaps from two years before the fold to its end, and a forecast every 21 trading days: 376 months, 3008 cases. Cases of the same month are strongly dependent, so intervals are month-block bootstrap intervals and rolling or regime summaries first average over the 8 baskets of a month.

Effective sample size. The Wishart channel needs degrees of freedom for the estimation window (N) and the forecast window (N_f). We set both to the same fraction s of the window length, $N = \mathrm{round}(63s)$ and $N_f = \mathrm{round}(21s)$, and choose s on a validation fold that precedes all test folds (training before 1985, validation 1985–1994) by minimising the total absolute calibration error of the 50/80/95% intervals of the forward asset variances under IW-ML (Table 4). The error decreases monotonically as s decreases and is smallest at the smallest scale tried, $s = 0.3$: $N = 19$, $N_f = 6$ ($N/d = 1.6$, $N_f < d$). A kurtosis-based estimate from the same data ($N_{\mathrm{eff}} = 2N/(2 + \kappa_{\mathrm{ex}})$) gives 37–55, so the small value inflates forecast uncertainty and absorbs the drift between estimation and forecast windows. OAS, LW-NLS and IW-ML use the same N.

Estimators and decision. The decision is the unconstrained global minimum-variance portfolio built from an estimate $\hat{C}$ of the window covariance (weights $w \propto \hat{C}^{-1}\mathbf{1}$ summing to one), scored by its realised variance over the next 21 days. We report the mean of the log ratio of realised variance to that of OAS (negative is better) and the annualised volatility of the realised portfolio return. The estimators are the sample covariance, OAS, LW-NLS, the posterior means of IW-ML and SC-SI, the SC-SI Stein-optimal covariance (inverse of the posterior mean of the inverse, optimal for the Stein loss $L_{\mathrm{S}}(\hat{C}, C) = \mathrm{tr}(\hat{C}^{-1}C) - \log\det(\hat{C}^{-1}C) - d$), and PCA factor models $C_K = \sum_{k \le K} \ell_k u_k u_k^\top + s_K^2(I - P_K)$ with s_K^2 the mean of the remaining eigenvalues. The number of factors is fixed (K=1, 3), set by the Gavish–Donoho threshold (unknown noise level: $\omega(\beta)^2$ times the median eigenvalue, $\beta = d/N$), or by the Bayes rule minimising the posterior expected Stein loss $\mathbb{E}[L_{\mathrm{S}}(C_K(C_e), C) \mid C_e]$ under the SC-SI or the IW-ML posterior. The SC-SI rule picks K

Table 4: **Finance: calibration of the effective sample size** on the validation fold (train < 1985, validation 1985–1994), IW-ML posterior. s: fraction of the window length; N, N_f: degrees of freedom for the 63-day and 21-day windows; coverage of the 50/80/95% forward intervals of the twelve asset variances; calibration error: sum of absolute deviations from the nominal levels (bold: smallest).

s	N	N_f	coverage 50/80/95	calibration error
0.30	19	6	0.48/0.81/0.95	**0.036**
0.45	28	9	0.43/0.74/0.91	0.180
0.60	38	13	0.37/0.66/0.85	0.372
0.80	50	17	0.34/0.60/0.79	0.523
1.00	63	21	0.31/0.55/0.74	0.641

Table 5: **Finance: out-of-sample minimum-variance portfolios** (1995–2026, 3008 cases). Ann. vol.: annualised volatility of the realised portfolio return (%); log ratio: mean log of realised 21-day variance over that of OAS with 95% month-block bootstrap intervals (negative is better; bold: best). PCA: factor model with K fixed, chosen by the Gavish–Donoho threshold, or by the Bayes rule under the IW-ML or the SC-SI posterior.

Estimator	ann. vol. (%)	log ratio to OAS ↓
OAS	14.36	0
Sample cov.	14.84	$+0.095$ $[0.076, 0.116]$
LW-NLS	14.44	$+0.034$ $[0.022, 0.046]$
IW-ML posterior mean	14.40	$+0.003$ $[-0.004, 0.010]$
SC-SI posterior mean	14.12	$\mathbf{-0.034}$ $[\mathbf{-0.043}, \mathbf{-0.026}]$
SC-SI Stein-optimal	14.20	-0.032 $[-0.040, -0.024]$
PCA, K=1	16.08	$+0.234$ $[0.213, 0.254]$
PCA, K=3	14.97	$+0.093$ $[0.075, 0.110]$
PCA, Gavish–Donoho K	15.21	$+0.129$ $[0.111, 0.147]$
PCA, Bayes K (IW-ML)	15.04	$+0.119$ $[0.098, 0.136]$
PCA, Bayes K (SC-SI)	14.89	$+0.086$ $[0.068, 0.103]$

with mean 2.54 (IW-ML: 3.00; Gavish–Donoho: 1.72) and changes it by 0.53 per month (0.23 for Gavish–Donoho): the choice is genuinely uncertain, and a hard choice discards that uncertainty (Table 5).

Calibration, log-score and regimes. Forward predictive laws are scored on the second-moment matrix of the next 21 days ($N_f = 6$), and calibration is measured on the twelve asset variances and the equal-weight direction. Regimes are terciles of the trailing 63-day realised volatility of the equal-weight portfolio of all listed industries at the forecast origin (no look-ahead): thresholds 12.2% and 16.5% annualised, 126, 125 and 125 months. The advantage of the SC-SI posterior mean is steady rather than driven by a few episodes: it holds in each tercile (Table 6b) while the sample covariance and LW-NLS lose ground as volatility rises. In calm markets SC-SI is slightly conservative and IW-ML is the closest to nominal; in the top tercile the plug-ins fall to 0.72–0.75 while both posterior methods stay at 0.79 (Table 6a). The log-score gain of SC-SI over the sample covariance ($+3.4$ $[2.3, 4.8]$) is not separated from that of OAS ($+3.0$ $[1.8, 4.7]$): intervals overlap in every regime and the gains are concentrated in a few episodes, chiefly a spike shared by all methods in early 2020.

J.5 Checkpoint selection

The worst-case PIT score. Self-consistent EM on a finite ensemble eventually over-sharpens the prior, so the iteration k must be chosen without clean covariances. The aggregate predictive log-score is a poor guide: it can keep improving while individual statistics deteriorate, because it trades them against others that dominate the density fit (for instance the variance along the bottom eigenvector of C_{e}). We use validation pairs $(C_{\mathrm{e}}^{\mathrm{cal}}, C_{\mathrm{e}}^{\mathrm{val}})$ sharing a latent covariance and six functionals f: log-determinant, log condition number, top-eigenvalue share, the correlation of the first two coordinates,

Table 6: **Finance: calibration, log-score and regimes.** (a) Coverage of the central 50/80/95% forward intervals over all 3008 cases; 80% coverage by tercile of trailing realised volatility (low/mid/high; bold: closest to 0.80); forward log-score gain over the sample covariance in nats per window with 95% month-block bootstrap intervals (bold: best). (b) Mean log ratio of realised variance to OAS in each tercile (negative is better; bold: best per tercile).

	overall	80% by volatility tercile			log-score
(a) Method	50/80/95	low	mid	high	gain ↑
Sample cov.	0.48/0.77/0.91	0.77	0.81	0.75	0
OAS	0.47/0.77/0.91	0.77	**0.80**	0.73	+3.0 [1.8, 4.7]
LW-NLS	0.47/0.76/0.91	0.77	0.80	0.72	+2.3 [1.6, 3.3]
IW-ML	0.50/0.81/0.94	**0.81**	0.81	**0.79**	+1.0 [0.1, 2.1]
SC-SI	0.53/0.83/0.94	0.84	0.85	0.79	**+3.4 [2.3, 4.8]**

(b) Estimator	low vol.	mid vol.	high vol.
SC-SI posterior mean	**−0.051 [−0.065, −0.037]**	**−0.022 [−0.037, −0.006]**	−0.030 [−0.044, −0.016]
SC-SI Stein-optimal	−0.046 [−0.059, −0.032]	−0.021 [−0.035, −0.006]	**−0.031 [−0.045, −0.017]**
IW-ML posterior mean	+0.002 [−0.008, 0.011]	−0.001 [−0.013, 0.011]	+0.010 [−0.005, 0.027]
LW-NLS	+0.008 [−0.009, 0.026]	+0.036 [0.014, 0.058]	+0.059 [0.033, 0.086]
Sample cov.	+0.028 [−0.002, 0.055]	+0.104 [0.069, 0.139]	+0.154 [0.118, 0.191]

and the variances $v^\top C v$ along the top and the bottom eigenvector v of $C_{\mathrm{e}}^{\mathrm{cal}}$. For each f we push 64 posterior draws given $C_{\mathrm{e}}^{\mathrm{cal}}$ through the channel, compute the PIT of $f(C_{\mathrm{e}}^{\mathrm{val}})$ under the resulting law, and measure the KS distance of the PITs to uniformity over the validation pairs. The score of an iteration is the *largest* of these six distances, so it cannot trade one functional against another, and we keep the iteration with the smallest score. It uses only $(C_{\mathrm{e}}^{\mathrm{cal}}, C_{\mathrm{e}}^{\mathrm{val}}, N)$ and therefore applies unchanged to real data. When the forecast window has fewer effective observations than assets ($N_f < d$, finance), every simulated Wishart draw is rank deficient and the PIT of the log-determinant and condition number is constant; both are then dropped. Validation pairs are 300 pairs of independent observations of fresh latent covariances (toys), consecutive windows of 10 validation subjects (EEG), and calibration/forecast pairs from the last three years before each cutoff (finance).

K ADDITIONAL EXPERIMENTAL FIGURES

This appendix collects figures from the experiments of Section 4 that do not fit in the main text.

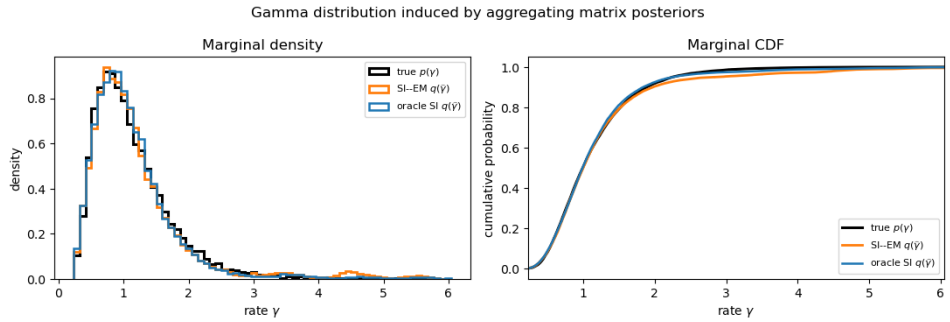

Figure 5: AR(1) experiment. The rate γ is not an input, a latent coordinate, or a training target; applying the deterministic AR(1) projection to matrix posterior draws after the fact, and aggregating over observations, induces the displayed law for it. Self-consistent training gives $W_1 = 0.092$ on the γ scale against the true prior, with the supervised oracle at 0.046, so most of the residual is the full-ma

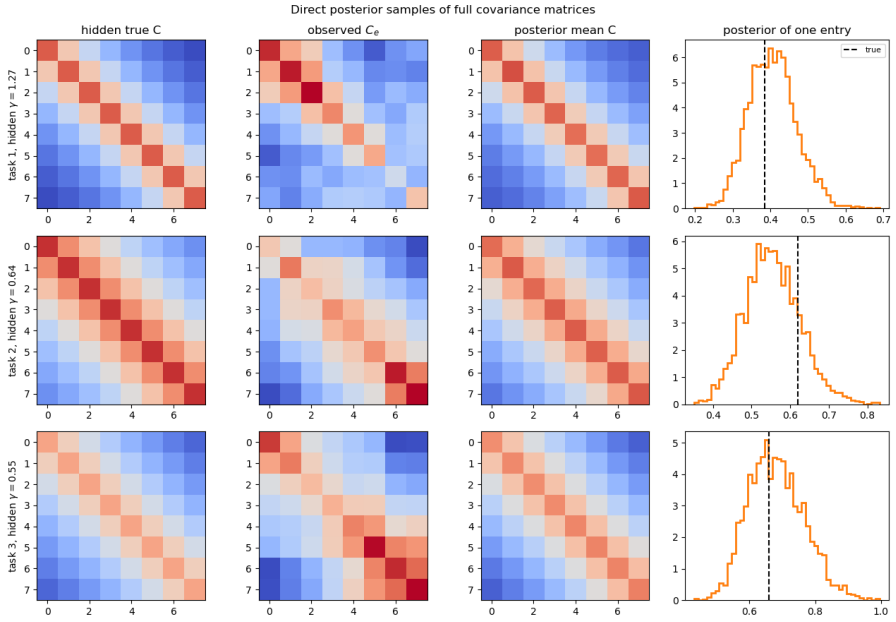

Figure 6: AR(1) experiment, fixed-observation posteriors. For several held-out empirical matrices, the hidden covariance, the conditioning empirical matrix, the posterior mean, and posterior uncertainty in selected entries. No parameter is inferred inside the sampler.

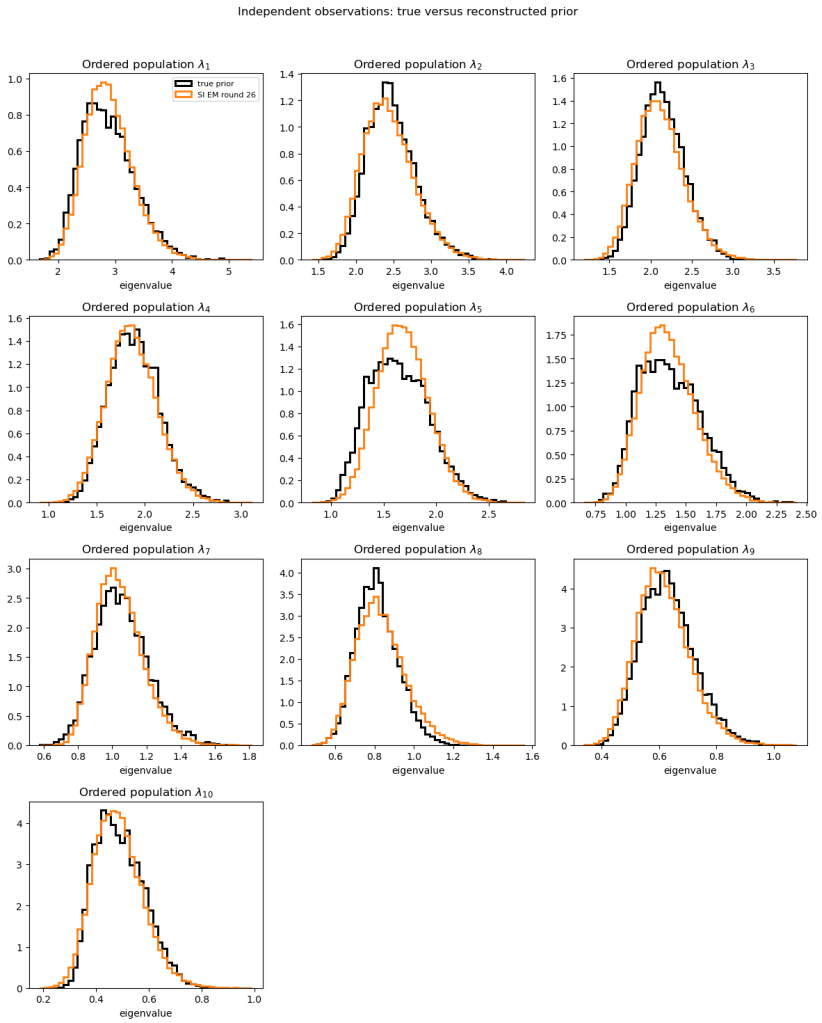

Figure 7: Rotationally invariant experiment, fresh-observation reconstruction. The selected drift is applied to a new ensemble of empirical spectra, four draws per spectrum, and the aggregate law of each ordered eigenvalue is compared with an independent draw from the hidden prior. The hidden spectrum behind a given observation need not match that observation's posterior draw; the comparison is between distributions.