Mathematical Notes for Bayesian Updates and Posterior Diagnostics in skpro

Arnav K

March 2026

Related PRs

- `https://github.com/sktime/skpro/pull/802`
- `https://github.com/sktime/skpro/pull/771`
- `https://github.com/sktime/skpro/pull/791`
- `https://github.com/sktime/skpro/pull/786`
- `https://github.com/sktime/skpro/pull/785`
- `https://github.com/sktime/skpro/pull/777`

1 How This Document Maps to the PRs

PR	Main contribution	How it maps here
#771	BayesianConjugateLinearRegressor online update in precision form	Sections 2, 3, 4, 5, and 7: conjugate posterior, precision update, diagnostics, and stable computation logic.
#802	BaseBayesianRegressor interface and posterior/predictive workflow	Sections 1 and 6: update contract and predictive distribution representation, including sample-based posterior handling.
#791	Bayesian tutorial notebook improvements	Notation and explanatory consistency for Sections 1 to 3 and 7.
#785, #786, #777	Temporal distribution, baseline estimators, and outlier detection	Useful probabilistic context, but not the direct source of the conjugate linear update derivations.

2 Bayesian Posterior and Sequential Update Contract

Given data

$$D = \{(x_i, y_i)\}_{i=1}^{n}, \tag{1}$$

parameter vector θ, prior $p(\theta)$, and likelihood $p(D \mid \theta)$, Bayes' theorem gives

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}, \qquad p(D) = \int p(D \mid \theta)p(\theta)\, d\theta. \tag{2}$$

For two data blocks D_1 and D_2:

$$p(\theta \mid D_1, D_2) = \frac{p(D_2 \mid \theta, D_1)p(\theta \mid D_1)}{p(D_2 \mid D_1)} \tag{3}$$

$$= \frac{p(D_2 \mid \theta)p(\theta \mid D_1)}{p(D_2 \mid D_1)}, \tag{4}$$

where the second line assumes conditional independence of data given θ.

How I approached this derivation

I started from one practical question: *what property must an online Bayesian update satisfy to be trustworthy?* The answer is posterior consistency across batch and sequential training.

So I used the identity below as the main contract to validate implementation behavior:

$$\text{posterior after } (D_1 \cup D_2) \equiv \text{posterior after update with } D_1 \text{ then } D_2. \tag{5}$$

In code review terms, this is the check I wanted `_update()` to satisfy: sequential updates should not drift away from one-shot fitting on combined data.

3 Gaussian Conjugate Linear Model: Full Derivation

Assume the linear-Gaussian model

$$y = X\beta + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2 I_n), \tag{6}$$

with prior

$$\beta \sim \mathcal{N}(\mu_0, \Sigma_0), \tag{7}$$

where $X \in \mathbb{R}^{n\times d}$, $y \in \mathbb{R}^n$, and $\Sigma_0 \in \mathbb{R}^{d\times d}$ is symmetric positive definite.

The likelihood in exponential form is

$$p(y \mid X, \beta) \propto \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^\top (y - X\beta)\right). \tag{8}$$

The prior is

$$p(\beta) \propto \exp\left(-\frac{1}{2}(\beta - \mu_0)^\top \Sigma_0^{-1}(\beta - \mu_0)\right). \tag{9}$$

Hence

$$p(\beta \mid X, y) \propto p(y \mid X, \beta)p(\beta) \propto \exp\left(-\frac{1}{2}Q(\beta)\right), \tag{10}$$

with

$$Q(\beta) = \frac{1}{\sigma^2}(y - X\beta)^\top (y - X\beta) + (\beta - \mu_0)^\top \Sigma_0^{-1}(\beta - \mu_0) \tag{11}$$

$$= \beta^\top \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}X^\top X\right)\beta - 2\beta^\top \left(\Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2}X^\top y\right) + \text{const.} \tag{12}$$

Define

$$\Sigma_n^{-1} = \Sigma_0^{-1} + \frac{1}{\sigma^2}X^\top X, \qquad \eta_n = \Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2}X^\top y. \tag{13}$$

Completing the square gives

$$Q(\beta) = (\beta - \mu_n)^\top \Sigma_n^{-1}(\beta - \mu_n) + \text{const}, \qquad \mu_n = \Sigma_n \eta_n. \tag{14}$$

Therefore

$$\beta \mid X, y \sim \mathcal{N}(\mu_n, \Sigma_n), \tag{15}$$

with closed forms

$$\Sigma_n = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}X^\top X\right)^{-1}, \qquad \mu_n = \Sigma_n \left(\Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2}X^\top y\right). \tag{16}$$

How I approached this derivation

I treated the conjugate model as a ground-truth reference case.

My workflow was:

1. Write the prior and likelihood in exponential quadratic form.
2. Combine terms and complete the square.
3. Read off posterior mean and covariance in closed form.
4. Use those formulas as a sanity baseline for implementation outputs.

This helped me communicate to reviewers that the implementation is not just numerically plausible, but algebraically aligned with known Bayesian identities.

4 Precision-Form Online Update and Information Accumulation

Let $\Lambda = \Sigma^{-1}$ denote precision and define natural mean parameter $\eta = \Lambda\mu$. Then for a new mini-batch (X_t, y_t) with known noise variance σ^2:

$$\Lambda_t = \Lambda_{t-1} + \frac{1}{\sigma^2} X_t^\top X_t, \qquad \eta_t = \eta_{t-1} + \frac{1}{\sigma^2} X_t^\top y_t. \tag{17}$$

Posterior mean is recovered by

$$\mu_t = \Lambda_t^{-1} \eta_t. \tag{18}$$

This makes the additive structure explicit: each data block contributes positive semidefinite information to precision.

How I approached this derivation

I rewrote the update in precision form because it mirrors how information accumulates.

Instead of re-deriving from scratch at every step, I tracked two additive objects: precision Λ and natural mean parameter $\eta = \Lambda\mu$. Each mini-batch adds a simple term to both. This made the online update logic easy to verify and easy to explain in review, especially for PR #771.

5 Predictive Posterior Derivation

For a new covariate $x_* \in \mathbb{R}^d$:

$$y_* \mid \beta, x_* \sim \mathcal{N}(x_*^\top \beta, \sigma^2), \qquad \beta \mid D \sim \mathcal{N}(\mu_n, \Sigma_n). \tag{19}$$

Using laws of total expectation and variance:

$$\mathbb{E}[y_* \mid x_*, D] = \mathbb{E}_{\beta|D}[\mathbb{E}(y_* \mid x_*, \beta)] = \mathbb{E}_{\beta|D}[x_*^\top \beta] = x_*^\top \mu_n, \tag{20}$$

$$\mathrm{Var}(y_* \mid x_*, D) = \mathbb{E}_{\beta|D}[\mathrm{Var}(y_* \mid x_*, \beta)] + \mathrm{Var}_{\beta|D}[\mathbb{E}(y_* \mid x_*, \beta)] \tag{21}$$

$$= \sigma^2 + x_*^\top \Sigma_n x_*. \tag{22}$$

Hence

$$y_* \mid x_*, D \sim \mathcal{N}\Big(x_*^\top \mu_n,\ x_*^\top \Sigma_n x_* + \sigma^2\Big). \tag{23}$$

How I approached this derivation

I wanted to show maintainers that prediction quality is not only about the mean.

So I used total expectation and total variance to separate:

- aleatoric noise (σ^2), and
- parameter uncertainty ($x_*^\top \Sigma_n x_*$).

That decomposition made it clear why a distribution-valued prediction is the right interface behavior for Bayesian regressors.

6 Posterior Diagnostics Used in Practice

6.1 Trace-based uncertainty contraction

Given prior and posterior covariance Σ_0, Σ_n:

$$R_{\text{trace}} = \frac{\text{tr}(\Sigma_n)}{\text{tr}(\Sigma_0)}. \tag{24}$$

If $R_{\text{trace}} < 1$, total marginal posterior uncertainty has decreased.

6.2 Mean-sigma concentration heuristic

Define component-wise posterior standard deviations

$$\sigma_{n,i} = \sqrt{(\Sigma_n)_{ii}}, \tag{25}$$

and ratios

$$R_i = \frac{|\mu_{n,i}|}{\sigma_{n,i}}, \qquad R_{\text{mean-sigma}} = \frac{1}{d}\sum_{i=1}^{d} R_i. \tag{26}$$

Larger ratios suggest posterior mean components are relatively well-concentrated compared with their uncertainty.

How I approached this derivation

I used diagnostics that are simple, interpretable, and fast to compute.

The goal was not to claim a full theoretical guarantee from a single metric, but to provide practical checks that maintainers can read quickly:

- $R_{\text{trace}} < 1$ as a first signal that uncertainty is shrinking.
- Mean-sigma ratios to see whether posterior location is concentrated.

These became useful review artifacts when discussing whether updates were behaving as expected over multiple online steps.

7 Sample-Based Posterior (Monte Carlo) Form

When posterior objects are represented by samples $\{\theta^{(s)}\}_{s=1}^{S}$ instead of closed-form parameters, empirical moments are:

$$\hat{\mu} = \frac{1}{S}\sum_{s=1}^{S} \theta^{(s)}, \tag{27}$$

$$\hat{\Sigma} = \frac{1}{S-1}\sum_{s=1}^{S}(\theta^{(s)} - \hat{\mu})(\theta^{(s)} - \hat{\mu})^\top. \tag{28}$$

An empirical trace ratio is then

$$\hat{R}_{\text{trace}} = \frac{\text{tr}(\hat{\Sigma})}{\text{tr}(\Sigma_0)}. \tag{29}$$

How I approached this derivation

For sample-based posteriors, I followed the same logic as the conjugate case but with empirical moments.

My approach was: estimate $\hat{\mu}$ and $\hat{\Sigma}$ from samples first, then reuse the same diagnostic language (for example, trace ratio) so comparisons stay consistent across parametric and Monte Carlo implementations.

8 Numerical Stability Notes

Avoid explicit matrix inverses in implementation whenever possible. For instance, rather than directly evaluating

$$\left(\Sigma_0^{-1} + \frac{1}{\sigma^2}X^\top X\right)^{-1}, \tag{30}$$

prefer stable linear algebra routines such as Cholesky factorization and triangular solves.

How I approached this derivation

I treated numerical stability as part of mathematical correctness in practice.

Even when formulas are equivalent on paper, explicit inversion is often fragile. So the implementation target was:

- keep the same Bayesian update equations,
- compute them through stable solves/factorizations,
- verify outputs stay consistent with the analytic derivation.

This framing helped explain why some code-level decisions were about reliability, not changing the underlying model.

Summary of Core Quantities

- Posterior mean: μ_n.
- Posterior covariance: Σ_n.
- Trace ratio: $\text{tr}(\Sigma_n)/\text{tr}(\Sigma_0)$.
- Mean-sigma ratio: $|\mu_i|/\sigma_i$.
- Precision matrix: $\Lambda = \Sigma^{-1}$.
- Predictive variance: $x_*^\top \Sigma_n x_* + \sigma^2$.