Weighted Bayesian Linear Regression (WIP)

10/07/2025

1 Introduction

Bayesian linear regression models a dataset of paired observations

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

with the generative rule

$$y_i = x_i^\top \theta + \varepsilon_i, \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \tag{1}$$

and places a multivariate Gaussian prior on the unknown weight vector

$$\theta \sim \mathcal{N}(\mu_0, \Sigma_0). \tag{2}$$

Here, the inputs $x_i \in \mathbb{R}^d$ are treated as fixed and known, the outputs $y_i \in \mathbb{R}$ are random through the noise term ε_i, and σ^2 is assumed known and homoscedastic. The inference task is to compute the posterior distribution

$$p(\theta \mid X, y)$$

and, from it, the predictive distribution for a new point

$$p(y^* \mid x^*, X, y).$$

2 Notation and Quantities

Moment parameters (Σ, μ) and natural parameters (P, J) are inter-convertible via

$$P = \Sigma^{-1}, \quad J = P\,\mu.$$

Symbol	Dim.	Category	Description
n	–	fixed	number of observations
d	–	fixed	number of features
x_i	$d \times 1$	data	input feature vector of sample i
X	$n \times d$	data	design matrix; rows are $x_i^\top$
y_i	1×1	data	scalar response of sample i
y	$n \times 1$	data	column vector $[y_1, \ldots, y_n]^\top$
θ	$d \times 1$	random	weight vector to be inferred
θ_j	1×1	random	j-th component of θ
ε_i	1×1	random	noise term, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
σ^2	1×1	hyper-param	known noise variance (homoscedastic)
μ_0	$d \times 1$	hyper-param	prior mean of θ
Σ_0	$d \times d$	hyper-param	prior covariance of θ
$P_0 = \Sigma_0^{-1}$	$d \times d$	hyper-param	prior precision matrix
$J_0 = P_0 \, \mu_0$	$d \times 1$	hyper-param	prior information vector
Σ_n	$d \times d$	derived	posterior covariance after all n points
μ_n	$d \times 1$	derived	posterior mean
$P_n = \Sigma_n^{-1}$	$d \times d$	derived	posterior precision
$J_n = P_n \, \mu_n$	$d \times 1$	derived	posterior information vector
x^*	$d \times 1$	data	new (test) input
y^*	1×1	random	predictive output corresponding to x^*
σ^{*2}	1×1	derived	predictive variance $x^{*\top} \Sigma_n x^* + \sigma^2$

Table 1: Notation and Quantities

3 Derivation of the Posterior and Update

(The equivalent compact matrix statements are relegated to the appendix.)

3.1 Prior assumption (moment form)

We place a multivariate normal prior on the weight vector $\theta \in \mathbb{R}^d$:

$$P(\theta) = \frac{1}{Z} \exp\Big\{ -\tfrac{1}{2} (\theta_j - \mu_{0,j}) \, \Sigma_{0,jk}^{-1} \, (\theta_k - \mu_{0,k}) \Big\}, \tag{3.1}$$

where repeated indices $j, k = 1, \ldots, d$ are summed over. Here

$$\mu_{0,j} = \mathbb{E}[\theta_j], \qquad \Sigma_{0,jk} = \mathbb{E}\big[(\theta_j - \mu_{0,j})(\theta_k - \mu_{0,k})\big],$$

and the normalizing constant is

$$Z = (2\pi)^{d/2} \, \det(\Sigma_0)^{1/2}.$$

3.2 Likelihood for a single datum

For an observation (x_i, y_i) with feature components $x_{i,j}$ and scalar response y_i:

$$P(y_i \mid x_i, \theta) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\Bigl\{-\tfrac{1}{2\sigma^2}\,(y_i - \theta_j\, x_{i,j})^2\Bigr\}. \tag{3.2}$$

3.3 Natural parametrisation

Writing a Gaussian in natural form

$$P(x) = \frac{1}{Z} \exp\bigl\{J_\ell\, x_\ell - \tfrac{1}{2}\, P_{\ell m}\, x_\ell\, x_m\bigr\}, \tag{3.3}$$

one shows (by expanding $(x-\mu)^\top \Sigma^{-1}(x-\mu)$ and dropping constants) that

$$J_j = \Sigma^{-1}_{jk}\, \mu_k, \qquad P_{jk} = \Sigma^{-1}_{jk}. \tag{3.4}$$

We therefore define the prior information vector $J_{0,j}$ and prior precision matrix $P_{0,jk}$ from μ_0, Σ_0 via (3.4).

3.4 Posterior from one datum

Multiplying (3.1) and (3.2) and retaining only terms in θ:

$$\begin{aligned} P(y_i \mid x_i, \theta)\, P(\theta) &\propto \exp\Bigl[-\frac{1}{2\sigma^2}(y_i - \theta_j x_{i,j})^2 + J_{0,\ell}\,\theta_\ell - \tfrac{1}{2}\, P_{0,\ell m}\,\theta_\ell\,\theta_m\Bigr] \quad (3.5) \\ &= \exp\Bigl[\tfrac{y_i}{\sigma^2}\, x_{i,\ell}\,\theta_\ell - \tfrac{1}{2\sigma^2}\, x_{i,\ell}\, x_{i,m}\,\theta_\ell\,\theta_m + J_{0,\ell}\,\theta_\ell - \tfrac{1}{2}\, P_{0,\ell m}\,\theta_\ell\,\theta_m\Bigr]. \quad (3) \end{aligned}$$

Grouping linear and quadratic terms identifies an updated Gaussian with

$$J_{1,\ell} = J_{0,\ell} + \frac{y_i}{\sigma^2}\, x_{i,\ell}, \qquad P_{1,\ell m} = P_{0,\ell m} + \frac{1}{\sigma^2}\, x_{i,\ell}\, x_{i,m}. \tag{3.6}$$

3.5 Posterior after t data points

By induction over $D = \{(x_s, y_s)\}_{s=1}^t$, one obtains

$$J_{t,\ell} = J_{0,\ell} + \frac{1}{\sigma^2} \sum_{s=1}^{t} y_s\, x_{s,\ell}, \qquad P_{t,\ell m} = P_{0,\ell m} + \frac{1}{\sigma^2} \sum_{s=1}^{t} x_{s,\ell}\, x_{s,m}. \tag{3.7}$$

Since each outer-product $x_{s,\ell} x_{s,m}/\sigma^2$ is positive semi-definite, the diagonal entries $P_{t,jj}$ grow (and variances shrink) with t.

To return to moment form, invert P_t to get

$$\Sigma_t = P_t^{-1},$$

and multiply to recover the posterior mean

$$\mu_{t,j} = (\Sigma_t)_{jk}\, J_{t,k}. \tag{3.8}$$

Hence the full posterior is

$$\theta \mid D \;\sim\; \mathcal{N}(\mu_t, \Sigma_t). \tag{3.9}$$

Matrix form	Index form
$J_{\text{final}} = J_0 + \dfrac{1}{\sigma^2}\sum_{i=1}^{n} y_i\, x_i$	$J_{\text{final},j} = J_{0,j} + \dfrac{1}{\sigma^2}\sum_{i=1}^{n} y_i\, x_{i,j}$
$P_{\text{final}} = P_0 + \dfrac{1}{\sigma^2}\sum_{i=1}^{n} x_i\, x_i^\top$	$P_{\text{final},jk} = P_{0,jk} + \dfrac{1}{\sigma^2}\sum_{i=1}^{n} x_{i,j}\, x_{i,k}$

Table 2: Batch updates in natural (information) form

Matrix form	Index form
$\Sigma_{\text{final}} = \big(\Sigma_0^{-1} + \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i\, x_i^\top\big)^{-1}$	$\Sigma_{\text{final},jk} = \big(\Sigma_0^{-1} + \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i\, x_i^\top\big)^{-1}_{jk}$
$\mu_{\text{final}} = \Sigma_{\text{final}}\big(\frac{1}{\sigma^2}\sum_{i=1}^{n} y_i\, x_i\big)$	$\mu_{\text{final},j} = (\Sigma_{\text{final}})_{jk}\, \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i\, x_{i,k}$

Table 3: Posterior mean and covariance in moment form

3.6 Implication: variance shrinks with sample size

Equation (3.7) shows each new datum adds the positive-semidefinite outer product $\sigma^{-2}x_{s,\ell}x_{s,m}$ to P_t. Since $\Sigma_t = P_t^{-1}$, larger P_t implies smaller diagonal entries of Σ_t, i.e. monotonically decreasing posterior variances.

3.7 Predictive distribution (index form)

For a new feature vector x_j^*, the predictive distribution is

$$y^* \mid x^*, D \;\sim\; \mathcal{N}\Big(x_j^*\, \mu_{t,j},\; x_\ell^*\, \Sigma_{t,\ell m}\, x_m^* + \sigma^2\Big). \tag{3.10}$$

4 Assumptions & Immediate Effects

#	Modelling assumption	Why it is imposed	Direct mathematical effect
A1	Gaussian prior: $\theta \sim \mathcal{N}(\mu_0, \Sigma_0)$	Encodes prior beliefs in closed-form; conjugate to Gaussian likelihood.	Posterior remains Gaussian (no numerical integration).
A2	Linear law: $y_i = x_i^\top \theta + \varepsilon_i$	Defines the regression relationship.	Likelihood is quadratic in θ.
A3	i.i.d. noise: $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$	Simplest continuous noise; matches A1 for conjugacy.	Each datum contributes outer-product $\sigma^{-2} x_i\, x_i^\top$ to the precision update.
A4	Known homoscedastic variance σ^2	Removes one layer of inference; keeps algebra linear in sufficient statistics.	σ^2 appears only as the common scaling $1/\sigma^2$.
A5	Deterministic inputs X (no uncertainty in x_i)	Standard supervised-learning setup; simplifies independence statements.	Conditional independence: $p(y \mid \theta, X) = \prod_i p(y_i \mid x_i, \theta)$.
A6	Observations independent given θ	Follows from A2 & deterministic X.	Likelihood factorizes; updates are sums over data.
A7	Prior independent of inputs: $p(\theta \mid X) = p(\theta)$	Keeps model symmetric in x; typical unless using functional priors.	Posterior precision = prior precision + data precision.

Table 4: Assumptions and their direct mathematical effects

Key consequences

- **Closed-form posterior & predictive.** Conjugacy of A1 + A3 makes Eqs. (3.7)–(3.10) analytic; no MCMC or variational inference needed.
- **Monotone variance decrease with sample size.** Each $\sigma^{-2} x_i x_i^\top$ is positive semidefinite $\Rightarrow$ precision P_t grows $\Rightarrow$ each posterior variance component in $\Sigma_t = P_t^{-1}$ shrinks.
- **Regularisation effect.** With diagonal prior $\Sigma_0 = \tau^2 I$, the posterior mean equals the ridge/MAP estimator; smaller τ^2 gives stronger

shrinkage.

- **Rank-one online updates.** Since data enter via $x_i x_i^\top$ and $y_i x_i$, each new sample modifies P and J by rank-1—enabling $O(d^2)$ sequential algorithms (cf. Sherman–Morrison in App. A.5).
- **Potential mismatch in heavy-tailed or heteroscedastic settings.** If noise is not well-modelled by a single σ^2, A3 and the ever-shrinking Σ_t can produce over-confident predictions.

5 Always-Zero Features and Their Revival

We treat one coordinate—call it k—whose value is identically zero in every sample collected so far:

$$x_{i,k} = 0, \quad \forall\, i \leq t.$$

5.1 Index-notation derivation

The online natural-parameter updates (Eq. (3.6)) are

$$J_{t+1,\ell} = J_{t,\ell} + \frac{y_{t+1}}{\sigma^2}\, x_{t+1,\ell}, \tag{4}$$

$$P_{t+1,\ell m} = P_{t,\ell m} + \frac{1}{\sigma^2}\, x_{t+1,\ell}\, x_{t+1,m}. \tag{5}$$

Putting $\ell = k$ (the zero feature) gives

Information vector

$$J_{t+1,k} = J_{t,k} + \sigma^{-2}\, y_{t+1}\, x_{t+1,k} = J_{t,k} + 0 = J_{t,k}.$$

Precision matrix

$$P_{t+1,km} = P_{t,km} + \sigma^{-2}\, x_{t+1,k}\, x_{t+1,m} = P_{t,km}, \quad P_{t+1,mk} = P_{t,mk}.$$

By induction,

$$J_{t,k} = J_{0,k}, \qquad P_{t,km} = P_{0,km} \quad \forall\, m,\ t.$$

Hence the posterior marginal of θ_k remains the prior:

$$\theta_k \mid D_t \;\sim\; \mathcal{N}(\mu_{0,k},\, \Sigma_{0,kk}).$$

All other coordinates ($\ell \neq k$) update as usual, never involving $x_{i,k}$; their means and variances evolve as if the k-th feature were absent. Since predictions multiply θ_k by $x_k^* = 0$, the predictive distribution for y^* is unchanged from the reduced model without feature k.

5.2 Matrix view (same result, compact)

Partition

$$x = \begin{bmatrix} 0 \\ \bar{x} \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_k \\ \bar{\theta} \end{bmatrix}, \quad P_t = \begin{bmatrix} P_{kk}(t) & P_{k\bar{k}}(t) \\ P_{\bar{k}k}(t) & P_{\bar{k}\bar{k}}(t) \end{bmatrix}, \quad J_t = \begin{bmatrix} J_k(t) \\ \bar{J}(t) \end{bmatrix}.$$

The rank-one increment

$$\sigma^{-2}\, x\, x^\top = \sigma^{-2} \begin{bmatrix} 0 \\ \bar{x} \end{bmatrix} \begin{bmatrix} 0 & \bar{x}^\top \end{bmatrix} = \sigma^{-2} \begin{bmatrix} 0 & 0 \\ 0 & \bar{x}\, \bar{x}^\top \end{bmatrix}$$

has zeros in row and column k, so

$$P_{t+1} = P_t + \sigma^{-2}\, x\, x^\top = \begin{bmatrix} P_{kk}(t) & P_{k\bar{k}}(t) \\ P_{\bar{k}k}(t) & P_{\bar{k}\bar{k}}(t) + \sigma^{-2}\, \bar{x}\, \bar{x}^\top \end{bmatrix}, \quad J_{t+1} = \begin{bmatrix} J_k(t) \\ \bar{J}(t) + \sigma^{-2}\, y\, \bar{x} \end{bmatrix},$$

leaving row/column k unchanged.

5.3 What if the feature becomes non-zero later?

Suppose at step $r > t$ we observe $x_{r,k} \neq 0$.

Update mechanics Apply the same formulas (4)–(5); now $\sigma^{-2} y_r x_{r,k}$ and $\sigma^{-2} x_{r,k} x_{r,m}$ are nonzero, so row/column k begin accumulating information.

Posterior mean No retroactive correction is needed. The prior-only mean $\mu_{0,k}$ shifts toward the data-driven value once evidence arrives.

Posterior variance Σ_{kk} remains at its prior level until the first informative sample, then shrinks normally. Early ignorance thus reflects honest uncertainty, not bias.

Other coefficients Their existing estimates are unaffected except via any covariance with θ_k. Future updates couple them through the new $x_{r,k} x_{r,m}$ terms exactly as in the standard update.

No special re-initialisation is required; the standard Bayesian update handles the transition seamlessly.

6 Forgetting / Exponential-Decay Schemes

Below are two exponential–decay ("forgetting") schemes. In both cases we start from the standard natural–parameter recursion

$$P_{t+1} = P_t + \frac{1}{\sigma^2}\, x\, x^\top, \tag{6}$$

$$J_{t+1} = J_t + \frac{1}{\sigma^2}\, y\, x, \tag{7}$$

and insert a decay step before folding in the new sample.

6.1 Single global decay factor λ (same weight for every feature and every past sample)

Choose a fixed scalar $0 < \lambda < 1$. The update becomes

$$\begin{aligned} P_{t+1} &= \lambda\, P_t \;+\; \frac{1}{\sigma^2}\, x\, x^\top, \\ J_{t+1} &= \lambda\, J_t \;+\; \frac{1}{\sigma^2}\, y\, x. \end{aligned} \tag{5.1}$$

Interpretation. Each older datum is down-weighted by λ^{age}. The pair (P, J) still represents the sufficient statistics of a weighted Gaussian model, so $\Sigma_t = P_t^{-1}$ and $\mu_t = \Sigma_t\, J_t$ remain valid. The effective sample size is approximately $(1-\lambda)^{-1}$; smaller λ forgets faster.

6.2 Feature-specific decay factors $\lambda_1, \ldots, \lambda_d$

Let

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d), \qquad 0 < \lambda_j \le 1.$$

The update is

$$\begin{aligned} P_{t+1} &= \Lambda\, P_t\, \Lambda \;+\; \frac{1}{\sigma^2}\, x\, x^\top, \\ J_{t+1} &= \Lambda\, J_t \;+\; \frac{1}{\sigma^2}\, y\, x. \end{aligned} \tag{5.2}$$

If features are grouped, set the same λ within each group so that Λ has repeated diagonal entries.

Interpretation. Row j and column k of the old precision are multiplied by $\lambda_j \lambda_k$, so cross-terms involving rapidly-forgotten features fade quickly. The information-vector component J_j decays by λ_j. Heterogeneous λ_j lets you "freeze" slow-changing channels ($\lambda_j \approx 1$) while aggressively forgetting noisy ones ($\lambda_j \ll 1$).

7 Exponential Decay Derivations

We derive each forgetting rule twice:

- *Weighted-sum view* (shows where the decay factors come from).
- *Recursive natural-parameter update* (what you implement).

Throughout, the un-decayed natural parameters after t samples are

$$P_t = P_0 + \frac{1}{\sigma^2} \sum_{i=1}^{t} x_i \, x_i^\top, \qquad J_t = J_0 + \frac{1}{\sigma^2} \sum_{i=1}^{t} y_i \, x_i. \tag{6.0}$$

7.1 Single global decay factor λ

7.1.1 Weighted-sum derivation

Give sample i the weight $w_i = \lambda^{t-i}$ with $0 < \lambda < 1$. Define

$$\widetilde{P}_t = P_0 + \frac{1}{\sigma^2} \sum_{i=1}^{t} w_i \, x_i \, x_i^\top, \quad \widetilde{J}_t = J_0 + \frac{1}{\sigma^2} \sum_{i=1}^{t} w_i \, y_i \, x_i. \tag{6.1}$$

At time $t+1$:

$$\begin{aligned}
\widetilde{P}_{t+1} &= P_0 + \frac{1}{\sigma^2} \sum_{i=1}^{t+1} \lambda^{t+1-i} x_i \, x_i^\top = \lambda \Big[P_0 + \tfrac{1}{\sigma^2} \sum_{i=1}^{t} \lambda^{t-i} x_i \, x_i^\top \Big] + \frac{1}{\sigma^2} \, x_{t+1} \, x_{t+1}^\top \\
&= \lambda \, \widetilde{P}_t + \frac{1}{\sigma^2} \, x_{t+1} \, x_{t+1}^\top, & (6.2) \\
\widetilde{J}_{t+1} &= \lambda \, \widetilde{J}_t + \frac{1}{\sigma^2} \, y_{t+1} \, x_{t+1}. & (6.3)
\end{aligned}$$

7.1.2 Recursive implementation

Start from (P_t, J_t), apply decay, then add the new sample:

$$P_t \;\leftarrow\; \lambda \, P_t, \quad J_t \;\leftarrow\; \lambda \, J_t,$$

$$P_{t+1} \;\leftarrow\; P_t + \frac{1}{\sigma^2}\, x_{t+1} x_{t+1}^{\top}, \quad J_{t+1} \;\leftarrow\; J_t + \frac{1}{\sigma^2}\, y_{t+1} x_{t+1}.$$

This reproduces Eq. (5.1).

7.2 Feature-specific decay factors $\lambda_1, \ldots, \lambda_d$

Let

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d), \quad 0 < \lambda_j \le 1.$$

Now the weight for sample i on coordinate j is λ_j^{t-i}.

7.2.1 Weighted-sum derivation

Define

$$\widetilde{P}_t = P_0 + \frac{1}{\sigma^2} \sum_{i=1}^{t} (\Lambda^{\,t-i} x_i)(\Lambda^{\,t-i} x_i)^{\top}, \quad \widetilde{J}_t = J_0 + \frac{1}{\sigma^2} \sum_{i=1}^{t} \Lambda^{\,t-i} y_i\, x_i. \tag{6.4}$$

At time $t+1$:

$$\widetilde{P}_{t+1} = P_0 + \frac{1}{\sigma^2} \sum_{i=1}^{t+1} (\Lambda^{\,t+1-i} x_i)(\Lambda^{\,t+1-i} x_i)^{\top} = \Lambda\, \widetilde{P}_t\, \Lambda + \frac{1}{\sigma^2}\, x_{t+1}\, x_{t+1}^{\top}, \tag{6.5}$$

$$\widetilde{J}_{t+1} = \Lambda\, \widetilde{J}_t + \frac{1}{\sigma^2}\, y_{t+1}\, x_{t+1}. \tag{6.6}$$

7.2.2 Recursive implementation

Decay first, per feature:

$$P_t \;\leftarrow\; \Lambda\, P_t\, \Lambda, \quad J_t \;\leftarrow\; \Lambda\, J_t,$$

then add the usual outer-product and cross-term. This reproduces Eq. (5.2).

7.3 Numerical update of the covariance

After decay, let the interim covariance be $\Sigma_{t+1/2} = P_{t+1/2}^{-1}$. Then apply the Sherman–Morrison rank-one update:

$$\Sigma_{t+1} = \Sigma_{t+1/2} - \frac{\Sigma_{t+1/2}\, x\, x^{\top}\, \Sigma_{t+1/2}}{\sigma^2 + x^{\top}\, \Sigma_{t+1/2}\, x}. \tag{6.7}$$

For the global-λ case, $P_{t+1/2} = \lambda P_t$ so $\Sigma_{t+1/2} = \lambda^{-1} \Sigma_t$; for the feature-specific case, $P_{t+1/2} = \Lambda P_t \Lambda$ so $\Sigma_{t+1/2} = \Lambda^{-1} \Sigma_t \Lambda^{-1}$.

Practical checkpoint. If every $\lambda_j = 1$ we recover the original Bayesian update. Setting $\lambda_j < 1$ is equivalent to adding process noise on θ_j at each step, hence "forgets" old evidence. Heterogeneous λ_j lets you down-weight stale information selectively while preserving long-term memory on stable channels.

Appendix A, Compact Matrix-Notation Equations

(All indices suppressed; every symbol matches the indexed derivation in Section 3.)

A.1 Prior

$$p(\theta) = \mathcal{N}(\theta \mid \mu_0, \Sigma_0) \;\propto\; \exp\!\big[-\tfrac{1}{2}\,(\theta-\mu_0)^{\top}\Sigma_0^{-1}(\theta-\mu_0)\big]. \tag{A.1}$$

Natural parameters:

$$J_0 = \Sigma_0^{-1}\,\mu_0, \qquad P_0 = \Sigma_0^{-1}.$$

A.2 Likelihood for all data

$$p(y \mid X, \theta) \;\propto\; \exp\!\big[-\tfrac{1}{2\sigma^2}\,(y - X\theta)^{\top}(y - X\theta)\big]. \tag{A.2}$$

A.3 Batch posterior after n observations

$$P_n = P_0 + \frac{1}{\sigma^2}\,X^{\top}X, \qquad J_n = J_0 + \frac{1}{\sigma^2}\,X^{\top}y. \tag{A.3}$$

Recover moment parameters:

$$\Sigma_n = P_n^{-1}, \qquad \mu_n = \Sigma_n\,J_n.$$

Hence

$$\theta \mid X, y \;\sim\; \mathcal{N}(\mu_n, \Sigma_n). \tag{A.4}$$

A.4 Online rank–one update (new pair (x, y))

$$P_{t+1} = P_t + \frac{1}{\sigma^2}\,x\,x^{\top}, \qquad J_{t+1} = J_t + \frac{1}{\sigma^2}\,y\,x. \tag{A.5}$$

A.5 Sherman–Morrison covariance update

(avoid full inversion after each sample)

$$\Sigma_{t+1} = \Sigma_t - \frac{\Sigma_t\,x\,x^{\top}\,\Sigma_t}{\sigma^2 + x^{\top}\,\Sigma_t\,x}. \tag{A.6}$$

A.6 Predictive distribution for a new input x^*

$$y^* \mid x^*, X, y \;\sim\; \mathcal{N}\big(x^{*\top} \mu_n,\; x^{*\top} \Sigma_n \, x^* + \sigma^2\big). \tag{A.7}$$

A Extra notes

B Estimating or Relaxing σ^2

Approach	How to get σ^2	Pros / cons
Plug-in MLE	$\hat{\sigma}^2 = \frac{1}{n-d} \sum_{i=1}^{n} (y_i - x_i^\top \mu_n)^2$	Cheap; works if noise truly is i.i.d. Gaussian.
Empirical Bayes	Maximise the marginal likelihood $p(y \mid X, \sigma^2)$. In closed form for BLR: $-\frac{1}{2}\big[n \log \sigma^2 + \log\det(\Sigma_0) - \log\det(\Sigma_n)\big]$.	Uses data to tune σ^2; remains conjugate.
Hierarchical model	Prior $\sigma^2 \sim \mathrm{IG}(\alpha_0, \beta_0)$. Posterior is IG; integrate out σ^2 for a Student-t predictive.	Yields heavier tails and robustness to outliers.
Heteroscedastic noise	Model $\varepsilon_i \sim N(0, \sigma_i^2)$ with $\sigma_i^2 = g(z_i; \phi)$ (e.g. log-linear on extra features z_i).	Captures input-dependent uncertainty; essential for sensor-noise.
Robust alternatives	Keep Gaussian prior on θ but use a heavy-tailed prior or directly use a Student-t likelihood.	Prevents the "variance always shrinks" pathology; better with outliers.

Table 5: Approaches to estimating or relaxing σ^2.

If you still want a running estimate within an exponential-decay filter, update

$$\beta_{t+1} \;=\; \lambda\, \beta_t \;+\; \tfrac{1}{2}\,(y_t - x_t^\top \mu_t)^2,$$

keeping α fixed. Then the posterior mean variance is

$$\hat{\sigma}^2_{t+1} \;=\; \frac{\beta_{t+1}}{\alpha - 1}.$$

B.1 Decaying the Prior vs. Data

Yes—with the update rule written in the "decay" note you are shrinking the entire precision matrix, and that includes the original prior. Concretely, after each new sample (x_t, y_t) the note does

$$P_{t+1} \;=\; \lambda\, P_t \;+\; \frac{1}{\sigma^2}\, x_t\, x_t^\top, \qquad J_{t+1} \;=\; \lambda\, J_t \;+\; \frac{1}{\sigma^2}\, y_t\, x_t.$$

Because the very first P_0, J_0 (your prior) are inside P_t, J_t, they get multiplied by $\lambda < 1$ every step. Their weight therefore dwindles at exactly the same exponential rate as the accumulated data.

How to not decay the prior while still forgetting data Maintain two separate precision/information blocks:

$$P_t = P_{\text{prior}} + P_{\text{data},\, t}, \qquad J_t = J_{\text{prior}} + J_{\text{data},\, t}.$$

Update only the data part:

$$P_{\text{data},\, t+1} = \lambda\, P_{\text{data},\, t} + \frac{1}{\sigma^2}\, x_t\, x_t^\top, \qquad J_{\text{data},\, t+1} = \lambda\, J_{\text{data},\, t} + \frac{1}{\sigma^2}\, y_t\, x_t.$$

Then the posterior mean and covariance remain

$$\Sigma_{t+1} = P_{t+1}^{-1}, \qquad \mu_{t+1} = \Sigma_{t+1}\, J_{t+1},$$

but P_{prior} (often chosen as Σ_0^{-1}) keeps its full weight forever.

Scenario	Influence of μ_0, Σ_0	Intuition
No decay (classical BLR)	Posterior precision grows as $$P_n = P_0 + \frac{1}{\sigma^2} X^\top X.$$ After $n \gg d$, the data dominate (eigenvalues of Σ_0 become negligible) and μ_n approaches the ordinary-least-squares/ridge solution.	Prior only protects against small-sample over-fitting.
Global decay $\lambda < 1$	Each step multiplies *all* information—including the frozen prior—by λ. In steady state the prior acts like an *extra* sample worth $\approx 1/(1-\lambda)$ observations—small but never disappearing.	Good if you want beliefs to drift slowly with the incoming data stream.
Feature-specific decay Λ	Component j retains a separate effective sample size $\approx 1/(1-\lambda_j)$. Setting $\lambda_j \approx 1$ for a "stable" coefficient makes its prior weight persistent, while a noisy coefficient with $\lambda_j \ll 1$ forgets quickly.	Lets you mix rigid and plastic parts in the same model.
"Zero-feature" corner case	If a feature is always zero, its posterior stays exactly $\mathcal{N}(\mu_{0,k}, \Sigma_{0,kk})$ until the first non-zero value arrives. The prior therefore *completely* determines early predictions along that axis.	Shows the prior is the only information channel when data are silent.

Table 6: Comparison of forgetting/decay scenarios and the influence of the prior

Decay the prior?	When it makes sense	When it is risky
Yes	[nosep,left=0pt]Your initial prior was just a rough starting guess you intend to phase out. You want the whole system to "forget" ancient assumptions as the environment changes.	[nosep,left=0pt]Prior encodes physical constants or hard constraints (e.g. wheel-base length 2 m). Prior reflects expensive expert knowledge you do not want to dilute.
No (keep prior fixed)	[nosep,left=0pt]Prior expresses structural information you always trust. Prior serves as a regulariser protecting against rank-deficiency when features become rare.	[nosep,left=0pt]You truly believe initial prior should count less as time goes on. (E.g. you began with a very vague or even mis-specified belief.)

Table 7: When (not) to decay the prior