From Weighted Bayesian Regression to State–Space Models

Overview

These notes unify weighted Bayesian regression, its classical special cases (OLS/WLS/Ridge), and their principled extensions via linear–Gaussian state–space models. We present common notation, batch and sequential updates (including information form), show how recursive least squares is a Kalman filter, and collect frequently used extensions (global/per-feature forgetting, heteroscedasticity, multivariate outputs, robustness, learning Q and R, and handling missing/irregular data). An appendix catalogs state–space variants with concise formulas.

Contents

1 Common Notation **1**

2 Weighted Bayesian Regression (WBR) **1**

3 Classical Estimators as Special Cases **2**

4 Per-Feature Decay in WBR: A Direct (Ad-hoc) Approach **2**

5 State–Space (DLM) View; RLS is a Kalman Filter **3**

6 WBR as a Subset of a State–Space Model **3**

7 Using the State–Space Framework for Extensions **4**

Appendix: Variants — Formulas **6**

1 Common Notation

- Data stream: $\{(x_t, y_t)\}_{t=1}^T$ with $x_t \in \mathbb{R}^d$ (column vector), $y_t \in \mathbb{R}$.
- Coefficients: $\theta \in \mathbb{R}^d$ (static) or $\theta_t \in \mathbb{R}^d$ (time-varying).
- Observation noise: $\varepsilon_t \sim \mathcal{N}(0, \sigma_t^2)$; define weight $w_t := \sigma_t^{-2}$. For multi-output $y_t \in \mathbb{R}^m$, use covariance $R_t \succ 0$ (strictly positive definite) instead of a scalar variance.
- Prior (static case): $\theta \sim \mathcal{N}(\mu_0, \Sigma_0)$. *Information form:* $P_0 := \Sigma_0^{-1}$ and $J_0 := P_0\mu_0$.
- Shorthands: $X = [x_1, \ldots, x_T]^\top$, $W = \mathrm{diag}(w_1, \ldots, w_T)$, $y = [y_1, \ldots, y_T]^\top$.

2 Weighted Bayesian Regression (WBR)

Model. For $t = 1, \ldots, T$,

$$y_t \mid \theta \sim \mathcal{N}(x_t^\top \theta,\ \sigma_t^2), \qquad \theta \sim \mathcal{N}(\mu_0, \Sigma_0). \tag{1}$$

Batch posterior (information form).

$$P_T = P_0 + \sum_{t=1}^{T} w_t\, x_t x_t^\top, \tag{2}$$

$$J_T = J_0 + \sum_{t=1}^{T} w_t\, y_t x_t, \tag{3}$$

$$\Sigma_T = P_T^{-1}, \qquad \mu_T = \Sigma_T J_T. \tag{4}$$

Sequential (RLS/KF) update. Starting from $(\mu_{t-1}, \Sigma_{t-1})$:

$$S_t = x_t^\top \Sigma_{t-1} x_t + \sigma_t^2, \tag{5}$$

$$K_t = \Sigma_{t-1} x_t\, S_t^{-1}, \tag{6}$$

$$\mu_t = \mu_{t-1} + K_t\big(y_t - x_t^\top \mu_{t-1}\big), \tag{7}$$

$$\Sigma_t = \Sigma_{t-1} - K_t x_t^\top \Sigma_{t-1}. \tag{8}$$

(Use Sherman–Morrison for an $O(d^2)$ rank-one covariance update.)

Predictive distribution. For a new $x_\star$,

$$y_\star \mid x_\star, D \sim \mathcal{N}\big(x_\star^\top \mu_T,\ x_\star^\top \Sigma_T x_\star + \sigma_\star^2\big). \tag{9}$$

Unknown σ^2 (optional). A Normal–Inverse-Gamma prior yields a Student-t predictive; in practice one may use a plug-in or empirical Bayes estimate for σ^2.

3 Classical Estimators as Special Cases

OLS (homoscedastic MLE) with a vague prior. If $w_t \equiv \sigma^{-2}$ and $P_0 \to 0$,

$$\hat{\theta}_{\text{OLS}} = (X^\top X)^{-1} X^\top y. \tag{10}$$

WLS/GLS (heteroscedastic/correlated). With diagonal weights W (or full R in GLS),

$$\hat{\theta}_{\text{WLS}} = (X^\top W X)^{-1} X^\top W y, \tag{11}$$

which matches the WBR MAP under a flat prior.

Ridge (Gaussian MAP). If $\theta \sim \mathcal{N}(0, \tau^2 I)$, i.e., $P_0 = \lambda I$, $\lambda = \tau^{-2}$,

$$\hat{\theta}_{\text{ridge}} = (X^\top W X + \lambda I)^{-1} X^\top W y = \mu_T \quad \text{(WBR posterior mean)}. \tag{12}$$

4 Per-Feature Decay in WBR: A Direct (Ad-hoc) Approach

To down-weight older information at different rates per coefficient, a heuristic information-form rule is

$$P_t = \Lambda^{1/2} P_{t-1} \Lambda^{1/2} + w_t x_t x_t^\top, \tag{13}$$

$$J_t = \Lambda^{1/2} J_{t-1} + w_t y_t x_t, \tag{14}$$

with $\Lambda = \mathrm{diag}(\delta_1, \ldots, \delta_d)$ and $\delta_j \in (0, 1]$.

Remarks. PSD is preserved via $\Lambda^{1/2}(\cdot)\Lambda^{1/2}$, but cross-covariances are shrunk in a way tied to geometric means of the discounts, which can distort structure and feels ad-hoc. A common variant is to decompose $P_t = P_{\text{prior}} + P_t^{\text{data}}$ and decay only P_t^{data}. This motivates a principled state–space treatment via process noise Q_t.

5 State–Space (DLM) View; RLS is a Kalman Filter

State evolution.

$$\theta_t = F_t\theta_{t-1} + \omega_t, \qquad \omega_t \sim \mathcal{N}(0, Q_t). \tag{15}$$

Observation.

$$y_t = H_t\theta_t + v_t, \qquad v_t \sim \mathcal{N}(0, R_t), \tag{16}$$

with $H_t := x_t^\top$ in regression.

Kalman filter (covariance form). Prediction:

$$\mu_{t|t-1} = F_t\mu_{t-1|t-1}, \tag{17}$$

$$\Sigma_{t|t-1} = F_t\Sigma_{t-1|t-1}F_t^\top + Q_t. \tag{18}$$

Update:

$$S_t = H_t\Sigma_{t|t-1}H_t^\top + R_t, \tag{19}$$

$$K_t = \Sigma_{t|t-1}H_t^\top S_t^{-1}, \tag{20}$$

$$\mu_{t|t} = \mu_{t|t-1} + K_t\big(y_t - H_t\mu_{t|t-1}\big), \tag{21}$$

$$\Sigma_{t|t} = (I - K_tH_t)\Sigma_{t|t-1}. \tag{22}$$

Special case (RLS). Static coefficients: $F_t = I$, $Q_t = 0$. With $H_t = x_t^\top$ and $R_t = \sigma_t^2$, the updates coincide with the WBR sequential recursion in Section 2.

Information form (sketch). Let $Y = \Sigma^{-1}$, $\eta = Y\mu$. The measurement update becomes $Y_{t|t} = Y_{t|t-1} + H_t^\top R_t^{-1}H_t$, $\eta_{t|t} = \eta_{t|t-1} + H_t^\top R_t^{-1}y_t$. (Propagation requires Woodbury; square-root forms are often more stable.)

6 WBR as a Subset of a State–Space Model

Choose $F_t = I$, $Q_t = 0$, $H_t = x_t^\top$, $R_t = \sigma_t^2$. Then

$$\Sigma_T^{-1} = \Sigma_0^{-1} + \sum_{t=1}^{T} w_tx_tx_t^\top, \tag{23}$$

$$\mu_T = \Sigma_T\Big(\Sigma_0^{-1}\mu_0 + \sum_{t=1}^{T} w_ty_tx_t\Big), \tag{24}$$

which is exactly the WBR posterior.

7 Using the State–Space Framework for Extensions

(a) Global exponential forgetting. Let $\delta \in (0, 1]$ be a discount. Inflate the prior covariance:

$$\Sigma_{t|t-1} = \delta^{-1}\Sigma_{t-1|t-1} \quad \iff \quad Q_t = (\delta^{-1} - 1)\,\Sigma_{t-1|t-1}. \tag{25}$$

(b) Per-feature (or block) discounting. Use diagonal or block-diagonal $D \succ 0$:

$$\Sigma_{t|t-1} = D^{-1}\Sigma_{t-1|t-1}D^{-\top}, \qquad Q_t = \Sigma_{t|t-1} - \Sigma_{t-1|t-1}. \tag{26}$$

This gives interpretable feature/group-wise forgetting rates δ_j and preserves cross-structure more cleanly than the ad-hoc rule in Section 4.

(c) Time-varying coefficients. Random-walk coefficients: $F_t = I$, $Q_t = \mathrm{diag}(q_1, \dots, q_d)$. AR(1) coefficients: $F_t = \phi I$, $|\phi| < 1$, with Q set to achieve a desired stationary variance via $S = \phi^2 S + Q$ (scalar case: $Q = (1 - \phi^2)S$).

(d) Heteroscedastic or correlated observation noise (online WLS/GLS). Let R_t vary over t (heteroscedastic) or be full (GLS). The KF recursions are unchanged.

(e) Augmented states (trend/seasonality + regression). Example with level ℓ_t and regression β_t:

$$\begin{bmatrix} \ell_t \\ \beta_t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \ell_{t-1} \\ \beta_{t-1} \end{bmatrix} + \omega_t, \quad \omega_t \sim \mathcal{N}(0, Q_t), \tag{27}$$

$$y_t = \begin{bmatrix} 1 & x_t^\top \end{bmatrix} \begin{bmatrix} \ell_t \\ \beta_t \end{bmatrix} + v_t, \quad v_t \sim \mathcal{N}(0, R_t). \tag{28}$$

(f) Multi-output regression. If $y_t \in \mathbb{R}^m$, use $H_t \in \mathbb{R}^{m\times d}$ and full $R_t \in \mathbb{R}^{m\times m}$; $S_t = H_t\Sigma_{t|t-1}H_t^\top + R_t$.

(g) Robustness. Student-t observation noise down-weights outliers. As a scale-mixture: $v_t \mid \lambda_t \sim \mathcal{N}(0, R_t/\lambda_t)$, $\lambda_t \sim \mathrm{Gamma}(\nu/2, \nu/2)$; an EM-style weight is $\mathbb{E}[\lambda_t \mid \nu_t] = \frac{\nu+1}{\nu+\nu_t^2/R_t}$ for scalar residual ν_t.

(h) Learning Q and R. *Offline:* EM with Rauch–Tung–Striebel (RTS) smoothing to maximize the innovation log-likelihood. *Online:* innovation matching, discounted ML, or Bayesian hierarchies for Q, R.

(i) Missing/irregular data. Missing y_t: skip the update (equivalently $R_t \to \infty$). Irregular sampling: when $\dot{\theta} = A\theta + w$, $F(\Delta t) = e^{A\Delta t}$ and $Q(\Delta t) = \int_0^{\Delta t} e^{A\tau}Q_c e^{A^\top\tau}\,d\tau$.

Cheat sheet of key mappings.

- Ridge $\Leftrightarrow$ Gaussian prior with $P_0 = \lambda I$.
- RLS $\equiv$ KF with $F = I$, $Q = 0$, $H_t = x_t^\top$, $R_t = \sigma_t^2$.
- Global decay: $\Sigma_{t|t-1} = \delta^{-1}\Sigma_{t-1|t-1} \Leftrightarrow Q_t = (\delta^{-1} - 1)\Sigma_{t-1|t-1}$.

- Per-feature decay: choose $D = \mathrm{diag}(\delta_1, \ldots, \delta_d)$ and set $\Sigma_{t|t-1} = D^{-1}\Sigma_{t-1|t-1}D^{-\top}$.
- Online WLS/GLS: encode weights/correlation in R_t (scalar/diagonal/full).

Appendix: State–Space / Dynamic Regression Variants — Formulas

A) Dynamics for the coefficients (state evolution)

$$\text{Static (RLS): } \theta_t = \theta_{t-1}, \quad Q_t = 0. \tag{29}$$

$$\text{Random walk: } \theta_t = \theta_{t-1} + \omega_t,\ \omega_t \sim \mathcal{N}(0, Q_t). \tag{30}$$

$$\text{AR(1): } \theta_t = \Phi\theta_{t-1} + \omega_t,\ \omega_t \sim \mathcal{N}(0, Q), \quad S = \Phi S \Phi^\top + Q. \tag{31}$$

$$\text{Polynomial trend: } \begin{bmatrix} \ell_t \\ b_t \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \ell_{t-1} \\ b_{t-1} \end{bmatrix} + \eta_t. \tag{32}$$

$$\text{Seasonal (trig.): } \begin{bmatrix} u_t \\ v_t \end{bmatrix} = \begin{bmatrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{bmatrix} \begin{bmatrix} u_{t-1} \\ v_{t-1} \end{bmatrix} + \eta_t,\ \omega = \tfrac{2\pi}{s}. \tag{33}$$

$$\text{Switching: } \theta_t = F_{z_t}\theta_{t-1} + \omega_t,\ \omega_t \sim \mathcal{N}(0, Q_{z_t}),\ z_t \sim \text{Markov}(P). \tag{34}$$

B) Observation models (what you're fitting)

$$\text{Gaussian regression: } y_t = x_t^\top \theta_t + v_t,\ v_t \sim \mathcal{N}(0, R_t). \tag{35}$$

$$\text{Poisson: } y_t \mid \theta_t \sim \text{Pois}(\lambda_t),\ \log\lambda_t = H_t\theta_t. \tag{36}$$

$$\text{Binomial/Logistic: } y_t \mid \theta_t \sim \text{Binom}(n_t, p_t),\ \text{logit}(p_t) = H_t\theta_t. \tag{37}$$

$$\text{Student-}t\text{ (robust): } v_t \mid \lambda_t \sim \mathcal{N}(0, R_t/\lambda_t),\ \lambda_t \sim \text{Gamma}(\nu/2, \nu/2). \tag{38}$$

Multivariate output: $y_t \in \mathbb{R}^m$ with $S_t = H_t\Sigma_{t|t-1}H_t^\top + R_t$.

C) Noise structures & weighting (R/WLS/GLS, robustness)

$$\text{Heteroscedastic } R_t : \text{time-varying weights } w_t = R_t^{-1}. \tag{39}$$

$$\text{GLS: } R_t \text{ full; ARMA errors via state augmentation.} \tag{40}$$

$$\text{Adaptive robust: } R_t \leftarrow (1-\gamma)R_{t-1} + \gamma\,\nu_t\nu_t^\top. \tag{41}$$

$$\text{Global discount: } \Sigma_{t|t-1} = \delta^{-1}\Sigma_{t-1|t-1},\ Q_t = (\delta^{-1} - 1)\Sigma_{t-1|t-1}. \tag{42}$$

$$\text{Matrix discount: } \Sigma_{t|t-1} = D^{-1}\Sigma_{t-1|t-1}D^{-\top},\ Q_t = \Sigma_{t|t-1} - \Sigma_{t-1|t-1}. \tag{43}$$

D) Nonlinear and/or non-Gaussian inference

EKF (measurement $y_t = h(\theta_t) + v_t$):

$$H_t = \left.\frac{\partial h}{\partial \theta}\right|_{\mu_{t|t-1}}, \quad \nu_t = y_t - h(\mu_{t|t-1}), \quad S_t = H_t\Sigma_{t|t-1}H_t^\top + R_t, \tag{44}$$

$$K_t = \Sigma_{t|t-1}H_t^\top S_t^{-1}, \quad \mu_{t|t} = \mu_{t|t-1} + K_t\nu_t, \quad \Sigma_{t|t} = (I - K_tH_t)\Sigma_{t|t-1}. \tag{45}$$

UKF/CKF: sigma-point transforms for $(\mu, \Sigma) \mapsto$ nonlinear images.
Particle filter (bootstrap): $\theta_t^{(i)} \sim p(\theta_t \mid \theta_{t-1}^{(i)})$, $w_t^{(i)} \propto w_{t-1}^{(i)}\, p(y_t \mid \theta_t^{(i)})$, $\sum_i w_t^{(i)} = 1$.
Rao–Blackwellized PF: PF on non-Gaussian/discrete, KF on linear–Gaussian blocks.

E) Estimating unknown parameters (Q, R, priors, etc.)

Innovation log-likelihood:

$$\log p(y_{1:T}) = -\tfrac{1}{2}\sum_{t=1}^{T}\big(\log|S_t| + \nu_t^\top S_t^{-1}\nu_t + m\log 2\pi\big). \tag{46}$$

EM for LDS (sketch): from RTS smoothing, with $\hat{s}_t = \mathbb{E}[\theta_t]$, $\hat{S}_t = \mathbb{E}[\theta_t\theta_t^\top]$, $\hat{S}_{t,t-1} = \mathbb{E}[\theta_t\theta_{t-1}^\top]$,

$$\Phi^\star = \Big(\sum_{t=2}^{T}\hat{S}_{t,t-1}\Big)\Big(\sum_{t=1}^{T-1}\hat{S}_t\Big)^{-1}, \tag{47}$$

$$Q^\star = \frac{1}{T-1}\Big(\sum_{t=2}^{T}\hat{S}_t - \Phi^\star\sum_{t=2}^{T}\hat{S}_{t,t-1}^\top\Big), \tag{48}$$

$$R^\star = \frac{1}{T}\sum_{t=1}^{T}\Big(y_t y_t^\top - H_t\hat{s}_t y_t^\top - y_t\hat{s}_t^\top H_t^\top + H_t\hat{S}_t H_t^\top\Big). \tag{49}$$

Online/adaptive: innovation matching $R_t \leftarrow (1-\gamma)R_{t-1} + \gamma\,\nu_t\nu_t^\top$; set Q_t via $\Sigma_{t|t-1} - F_t\Sigma_{t-1|t-1}F_t^\top$.

F) Computational forms & numerics

Information update:

$$Y_{t|t} = Y_{t|t-1} + H_t^\top R_t^{-1}H_t, \qquad \eta_{t|t} = \eta_{t|t-1} + H_t^\top R_t^{-1}y_t, \tag{50}$$

with $Y = \Sigma^{-1}$, $\eta = Y\mu$. **Square-root KF:** maintain $SS^\top = \Sigma$ and use QR/Cholesky updates.
RTS smoother:

$$A_t = \Sigma_{t|t}F_{t+1}^\top\Sigma_{t+1|t}^{-1}, \tag{51}$$

$$\mu_{t|T} = \mu_{t|t} + A_t(\mu_{t+1|T} - \mu_{t+1|t}), \tag{52}$$

$$\Sigma_{t|T} = \Sigma_{t|t} + A_t(\Sigma_{t+1|T} - \Sigma_{t+1|t})A_t^\top. \tag{53}$$

G) Structural time-series as state–space (templates)

Local level: $\ell_t = \ell_{t-1} + \eta_t$, $y_t = \ell_t + \varepsilon_t$.
Local linear trend: state matrix $\left(\begin{smallmatrix}1 & 1\\ 0 & 1\end{smallmatrix}\right)$; $y_t = \ell_t + \varepsilon_t$.
Seasonal (trigonometric or dummy); dynamic harmonic regression via Fourier regressors with drifting coefficients.
ARIMA: embed ARMA error as a state block for exact likelihood with missing data.
Dynamic factor: $y_t = \Lambda f_t + \epsilon_t$, $f_t = Af_{t-1} + u_t$.

H) High-dimensional & large-scale variants

Ensemble KF (EnKF): Monte Carlo covariance via M particles.
Low-rank + diagonal Σ: $\Sigma \approx USU^\top + D$ (Woodbury updates).
Sparse precision: maintain $Y = \Sigma^{-1}$ and exploit sparsity.
Distributed fusion: sum information from sensors i: $Y = \sum_i Y^{(i)}$, $\eta = \sum_i \eta^{(i)}$.

I) Constraints & regularization on states

Equality constraints $A\theta = b$ (projection):

$$\mu^{\text{proj}} = \mu - \Sigma A^\top (A\Sigma A^\top)^{-1}(A\mu - b), \tag{54}$$

$$\Sigma^{\text{proj}} = \Sigma - \Sigma A^\top (A\Sigma A^\top)^{-1} A\Sigma. \tag{55}$$

Positivity: reparametrize $\theta_{t,j} = \text{softplus}(\tilde{\theta}_{t,j})$ and filter $\tilde{\theta}_t$.
Sparsity: Laplace priors (MAP $\to$ dynamic lasso) or spike-and-slab on coefficients.

J) Regime changes, changepoints, mixtures

Bayesian online changepoint detection (hazard $H(r)$; run-length posterior via predictive probs).
SLDS: HMM over z_t with an LDS per state; IMM or forward–backward with mixture pruning.
Mixture innovations: heavy-tail or spike-and-slab on ω_t to allow shocks.

K) Irregularities in the data stream

Missing y_t: skip the update (prediction only).
Irregular sampling (from $\dot{\theta} = A\theta + w$): $F(\Delta t) = e^{A\Delta t}$ and $Q(\Delta t) = \int_0^{\Delta t} e^{A\tau} Q_c e^{A^\top \tau}\, d\tau$.
Delayed/out-of-order: fixed-lag RTS smoothing to revise past states.

L) Control & decision-making (closed loop)

LQG: KF state estimate + LQR; discrete Riccati for the control gain.
Dual control: exploration vs exploitation (Bayes-optimal intractable; use MPC heuristics).
Kalman MPC: receding-horizon optimization with KF predictions.

M) Practical glue for dynamic regression in markets

Per-feature decay: diagonal Q or matrix discount $D = \text{diag}(\delta_j)$.
Heteroscedasticity: R_t driven by volatility proxies.
Cross-zone correlation: multivariate y_t with full R_t or factor-error $R_t = \Psi + \Gamma\Gamma^\top$.
Regime awareness: switching F_t, Q_t, R_t keyed to a latent state z_t.
Robustness: Student-t observation or adaptive inflation of R_t.
Explainability: report smoothed $\theta_{t|T}$ and credible bands $x^\top \Sigma_{t|T} x$.