Linear Regresson Problems

ELG 5218 – Uncertainty Evaluation in Engineering Measurements and Machine Learning

Instructor: Miodrag Bolić, University of Ottawa

Date: January 28, 2026

ELG 5218 – Uncertainty Evaluation in Engineering Measurements and Machine Learning

Instructor: Miodrag Bolic, University of Ottawa
Date: January 28, 2026

PART A: CONCEPTUAL QUESTIONS

A1. Classical vs Bayesian Linear Regression

Question. What is the fundamental difference between classical (frequentist) linear regression and Bayesian linear regression?

Answer.

Classical linear regression typically estimates a single point value of the weight vector w by minimizing the sum of squared errors (or equivalently maximizing a Gaussian likelihood). This yields one "best" estimate $\hat{w}$ with no full uncertainty quantification.

Bayesian linear regression places a prior distribution $p(w)$ on the weights and computes the full posterior distribution $p(w \mid X, y)$. This provides:

- A distribution over plausible weight vectors, not just a single estimate.
- Credible intervals for each component of w.
- A predictive distribution for y_* with uncertainty bands.
- Automatic regularization through the prior (e.g., Gaussian prior corresponds to ridge).

The posterior captures uncertainty due to finite data (epistemic) and, through the Gaussian observation model, also reflects inherent observation noise (aleatoric).

A2. Why is Bayesian Linear Regression Conjugate (with Gaussian Prior)?

Question. With a Gaussian likelihood and a Gaussian prior on w, why is Bayesian linear regression conjugate? What does conjugacy buy us?

Answer.

Conjugacy means the prior and likelihood combine to produce a posterior in the same family as the prior. The likelihood over all data is multivariate Gaussian in y with mean Xw, and the prior is Gaussian in w. Multiplying two exponentials of quadratic forms in w yields another quadratic in w, hence a Gaussian posterior.

Conjugacy allows us to have: Posterior 正比 Likelihood x Prior

- Closed-form posterior $p(w \mid X, y)$.
- Closed-form posterior predictive $p(y_* \mid x_*, X, y)$.
- Simple, interpretable updates: precisions and means "add" in intuitive ways.

A3. Ridge Regression as MAP in Bayesian Linear Regression

Question. Explain the relationship between ridge regression and Bayesian linear regression with a zero-mean Gaussian prior on w.

Answer.
Ridge regression solves

$$\hat{w}_{\text{ridge}} = \arg\min_w \left\{ \frac{1}{2\sigma^2} \|y - Xw\|^2 + \frac{\lambda_0}{2} \|w\|^2 \right\}.$$

Bayesian linear regression with likelihood $y \mid w \sim \mathcal{N}(Xw, \sigma^2 I)$ and prior $w \sim \mathcal{N}(0, \lambda_0^{-1} I)$ has log-posterior

$$\log p(w \mid X, y) = -\frac{1}{2\sigma^2} \|y - Xw\|^2 - \frac{\lambda_0}{2} \|w\|^2 + \text{const.}$$

Maximizing this log-posterior is equivalent to the ridge optimization above. Thus, ridge is the MAP estimate in a conjugate Bayesian linear model with Gaussian prior. The regularization parameter λ_0 is the prior precision.

A4. Aleatoric vs Epistemic Uncertainty in Linear Regression

Question. In Bayesian linear regression, distinguish aleatoric and epistemic uncertainty in the predictive distribution.

Answer.
For a new input x_*, the predictive distribution is

$$p(y_* \mid x_*, X, y) = \mathcal{N}\Big(\mu_N^\top x_*, \; \underbrace{\beta^{-1}}_{\text{aleatoric}} + \underbrace{x_*^\top \Sigma_N x_*}_{\text{epistemic}}\Big).$$

- **Aleatoric uncertainty** ($\sigma^2 = \beta^{-1}$) is irreducible noise in observations. Even with infinite data and known w, outcomes fluctuate due to measurement noise or inherent variability.
- **Epistemic uncertainty** ($x_*^\top \Sigma_N x_*$) comes from uncertainty in w due to limited data. As more data are observed, Σ_N shrinks and epistemic uncertainty decreases.

Total predictive variance is the sum of these two contributions.

A5. Behavior of Predictive Uncertainty Far from Training Data

Question. Qualitatively, how does the predictive variance $\sigma^2 + x_*^\top \Sigma_N x_*$ behave when x_* lies far outside the span of the training inputs?

Answer. 拉大左個uncertainty band

- Predictive variance grows as we move away from training data.
- The model becomes more uncertain (higher epistemic uncertainty) in extrapolation regions.

This is desirable: Bayesian linear regression "knows what it doesn't know" and inflates uncertainty outside the training domain.

A6. Effect of a Strong Prior on Posterior and Predictions

Question. What happens to the posterior over w and the predictive distribution if the prior precision λ_0 becomes very large (strong prior), assuming the prior mean is zero?

Answer.
As $\lambda_0 \to \infty$, the prior becomes extremely concentrated around $w = 0$: *同var反比！

- Posterior covariance Σ_N shrinks toward zero; the posterior collapses toward $w = 0$ regardless of the data (prior dominates).
- Posterior mean μ_N is pulled very close to zero.
- Predictive mean $x_*^\top \mu_N$ becomes close to zero (underfitting).
- Predictive epistemic variance $x_*^\top \Sigma_N x_*$ is small, so total variance is mainly σ^2. Recall: aleatoric noise + epistemic co-variance

The model becomes very confident but biased toward zero, potentially underfitting even strong signals in the data.

A7. Non-IID

Question. You observe a long time series $\{(x_t, y_t)\}_{t=1}^{\infty}$ where the underlying relationship slowly drifts:

$$y_t = w_t^\top x_t + \epsilon_t, \quad w_t = w_{t-1} + \eta_t,$$

with small process noise η_t.

(a) Why is a static Bayesian linear regression model (fixed w) misspecified in this scenario?

Answer.
(a) The assumption $w_t \equiv w$ is violated; parameters drift over time. A static model pools all data equally, leading to outdated estimates that cannot keep up with recent changes.

PART B: MATHEMATICAL DERIVATIONS

Assume the standard model:

$$y \mid w \sim \mathcal{N}(Xw, \sigma^2 I_N), \qquad w \sim \mathcal{N}(m_0, \lambda).$$

B1. Posterior Predictive Distribution

Problem. For a new input $x_* \in \mathbb{R}^D$, derive the posterior predictive distribution

$$p(y_* \mid x_*, X, y).$$

Answer.
Conditionally on w,

$$y_* \mid w, x_* \sim \mathcal{N}(x_*^\top w, \sigma^2).$$

We must integrate over the posterior of w:

$$p(y_* \mid x_*, X, y) = \int p(y_* \mid x_*, w)\, p(w \mid X, y)\, dw.$$

Since $w \mid X, y \sim \mathcal{N}(\mu_N, \Sigma_N)$ and the conditional is linear-Gaussian, the marginal is Gaussian:

$$y_* \mid x_*, X, y \sim \mathcal{N}\left(x_*^\top \mu_N,\ \sigma^2 + x_*^\top \Sigma_N x_*\right).$$

Mean:

$$\mathbb{E}[y_* \mid x_*, X, y] = x_*^\top \mathbb{E}[w \mid X, y] = x_*^\top \mu_N.$$

Variance:

$$\mathrm{Var}(y_* \mid x_*, X, y) = \mathbb{E}[\mathrm{Var}(y_* \mid w, x_*)] + \mathrm{Var}(\mathbb{E}[y_* \mid w, x_*]) = \sigma^2 + x_*^\top \Sigma_N x_*.$$

B2. Gradient of the Log-Posterior (for MAP / Optimization)

Problem. Derive the gradient of the log-posterior $\nabla_w \log p(w \mid X, y)$ under the conjugate Gaussian model (with fixed σ^2, m_0, λ).

Answer.
Ignoring additive constants, the log-posterior is

$$\log p(w \mid X, y) = -\frac{1}{2\sigma^2}\|y - Xw\|^2 - \frac{1}{2}(w - m_0)^\top \lambda (w - m_0).$$

Gradient:

$$\nabla_w \left(-\frac{1}{2\sigma^2}\|y - Xw\|^2\right) = \frac{1}{\sigma^2} X^\top (y - Xw).$$

$$\nabla_w \left(-\frac{1}{2}(w - m_0)^\top \lambda (w - m_0)\right) = -\lambda (w - m_0).$$

Combined:

$$\nabla_w \log p(w \mid X, y) = \frac{1}{\sigma^2} X^\top (y - Xw) - \lambda (w - m_0).$$

Setting this to zero yields the posterior mean formula. Concave --> ddx == 0 yields maxima.

Mean = Mode (MAP)

B3. Hessian and Concavity of the Log-Posterior

Problem. Derive the Hessian $\nabla_w^2 \log p(w \mid X, y)$ and show the log-posterior is strictly concave.

Answer.
From B2, the gradient is

$$g(w) = \frac{1}{\sigma^2} X^\top (y - Xw) - \lambda(w - m_0).$$

Differentiate again:

$$\nabla_w^2 \log p(w \mid X, y) = -\frac{1}{\sigma^2} X^\top X - \lambda.$$

This Hessian is negative definite because:

- $X^\top X$ is positive semidefinite.
- λ is positive definite (prior covariance invertible).
- Their sum $X^\top X/\sigma^2 + \lambda$ is positive definite.

Thus the Hessian is negative definite, implying the log-posterior is strictly concave and has a unique global maximum (the MAP).

PART C: PARAMETRIC ANALYSIS (What if we change parameters?)

C1. Effect of Increasing Prior Precision λ_0

Question. As λ increases (stronger prior), what happens to the posterior covariance Σ_N and predictive variance?

Answer.
Recall:

$$\Sigma_N^{-1} == \lambda I + \frac{1}{\sigma^2} X^\top X$$

As λ increases:

- Σ_N^{-1} increases, so Σ_N decreases: posterior becomes more concentrated.
- Epistemic component of predictive variance, $x_*^\top \Sigma_N x_*$, decreases.
- Predictions become more certain (narrower credible intervals) but more biased toward the prior mean.

C2. Effect of Increasing Noise Variance σ^2

Question. As σ^2 increases (more observation noise), what happens to the posterior and predictive distribution?

Answer.
Recall:

$$\Sigma_N^{-1} = \lambda I + \frac{1}{\sigma^2} X^\top X.$$

If σ^2 increases decreases):

- The data term $\frac{1}{\sigma^2} X^\top X$ is downweighted relative to the prior.
- Posterior covariance Σ_N grows; posterior is more diffuse because each observation is noisy and hence less informative.
- Posterior mean is more influenced by the prior.
- Predictive variance grows both through σ^2 directly and indirectly via larger Σ_N.

aleatoric epistemic

C3. Behavior as $N \to \infty$ (Bernstein–von Mises Intuition)

Question. Intuitively, what happens to the posterior over w and the predictive distribution as $N \to \infty$ while the model is correctly specified?

Answer.
As N grows:

dominate OVER noises

- The data term dominates the prior: $\Sigma_N^{-1} \approx X^\top X / \sigma^2$; the prior becomes negligible.
- Posterior $p(w \mid X, y)$ becomes sharply peaked around the true parameter $w^\star$ (if the model is correct).

tends to 0 (more reciprocols added)

- Epistemic variance $x_*^\top \Sigma_N x_*$ goes to zero; predictive variance converges to σ^2 (irreducible).
- Predictions approach those of classical least squares; Bayesian credible sets asymptotically match frequentist confidence sets (Bernstein–von Mises).

PART D: OTHER PROBLEMS

D1. Engineering Application – Temperature Sensor Calibration

You calibrate a temperature sensor: input is a voltage x, output is temperature y. You model:

$$y_n = w_0 + w_1 x_n + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2),$$

with prior

$$w \sim \mathcal{N}(0, 10^2 I_2), \quad \sigma^2 = 1 \text{ (assumed known)}.$$

You collect $N = 20$ calibration data points spanning the range $x \in [-1, 1]$.

(a) Explain why Bayesian linear regression is preferable to simple least squares for this calibration problem.

(b) Your posterior summary for w_1 is approximately $\mathbb{E}[w_1 \mid \cdot] = 2.0$, $\mathrm{sd}(w_1 \mid \cdot) = 0.2$. Interpret this physically.

(c) For a new measurement at $x_* = 1.5$ (slightly outside the calibration range), your predictive distribution is $y_* \sim \mathcal{N}(3.1, 1.4^2)$. Comment on both the mean and the inflated variance.

Answer (sketch).
(a) Bayesian regression: Why: Bayesian >> Frequentists

- Gives credible intervals for w_0, w_1, which are crucial in metrology/calibration.
- Regularizes estimates with a prior, avoiding overfitting for small N.
- Provides predictive uncertainty for new measurements, essential for uncertainty budgeting.

(b) $\mathbb{E}[w_1] = 2.0$ with sd $= 0.2$ means the slope is around $2°C$ per volt, with a 95% credible interval roughly $[1.6, 2.4]$. Thus, the sensor's sensitivity is well-estimated but not exact; there is residual epistemic uncertainty about its gain.

(c) The mean $3.1°C$ at $x_* = 1.5$ is the extrapolated prediction. The variance 1.4^2 is larger than at in-range points, reflecting that extrapolation beyond the calibration region is less certain. This is a desirable property: the Bayesian model correctly warns that predictions outside the observed range are less reliable.