Gaussian Models Problems – Solutions

ELG 5218 – Uncertainty Evaluation in Engineering Measurements and Machine Learning

Instructor: Miodrag Bolić, University of Ottawa

Date: January 28, 2026

PART A: GAUSSIAN INTUITION

A1. Why precisions (not variances) add

(a) **Solution:** The likelihood for n observations is:

Subs数代入

$$p(x_1, \ldots, x_n \mid \mu) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

The prior is:

$$p(\mu) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right).$$

The posterior (unnormalized) is:

$$p(\mu \mid x) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right).$$

Rewriting the exponent by completing the square: the quadratic form in μ is:

$$-\frac{1}{2}\left[\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right](\mu - \text{const})^2,$$

so the posterior precision is:

$$\lambda_n = \frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}.$$

Thus precisions add.

A2. Precision-weighted averaging: limiting cases

(a) **Solution:** As $n \to \infty$:

Limit

$$w = \frac{n/\sigma^2}{n/\sigma^2 + 1/\sigma_0^2} \to \frac{n/\sigma^2}{n/\sigma^2} = 1.$$

Thus $\mu_n \to 1 \cdot \bar{x} + 0 \cdot \mu_0 = \bar{x}$. The posterior mean converges to the sample mean.

(b) **Solution:** As $\sigma_0^2 \to 0$:

Limit

$$w = \frac{n/\sigma^2}{n/\sigma^2 + 1/\sigma_0^2} \to \frac{n/\sigma^2}{\infty} = 0.$$

Thus $\mu_n \to 0 \cdot \bar{x} + 1 \cdot \mu_0 = \mu_0$. The posterior mean approaches the prior mean.

(c) **Solution:** With infinite data, the posterior is driven entirely by the data (MLE). With an infinitely strong prior, the posterior is entirely determined by prior belief and ignores the data. These represent the limits of data dominance and prior dominance, respectively.

PART B : SEQUENTIAL GAUSSIAN UPDATES

B1. Learning rate dynamics in online Gaussian learning

(a) **Solution:** After $n-1$ observations, $\lambda_{n-1} = 1/\sigma_0^2 + (n-1)/\sigma^2$. Thus:

代式 $$w_n = \frac{\lambda}{\lambda_{n-1} + \lambda} = \frac{1/\sigma^2}{1/\sigma_0^2 + (n-1)/\sigma^2 + 1/\sigma^2} = \frac{1/\sigma^2}{1/\sigma_0^2 + n/\sigma^2}.$$

(b) **Solution:** To show w_n is decreasing in n, note that:

Show Trend: d Gradient Change! $$\frac{dw_n}{dn} = \frac{d}{dn}\left(\frac{1/\sigma^2}{1/\sigma_0^2 + n/\sigma^2}\right) = -\frac{(1/\sigma^2)^2}{(1/\sigma_0^2 + n/\sigma^2)^2} < 0.$$

The numerator is constant, but the denominator grows with n, so w_n decreases monotonically.

(c) **Solution:** As more data accumulate, the posterior precision increases, making it already highly concentrated. Each new observation contributes diminishing information relative to the accumulated certainty, so its weight (step size) shrinks accordingly.

PART C : UNKNOWN VARIANCE – NORMAL–GAMMA

C1. Interpreting Normal–Gamma hyperparameters

(a) **Solution:** In the Normal–Gamma prior, κ_0 determines the precision of the prior distribution on μ given λ: $\mathrm{Var}(\mu \mid \lambda) = 1/(\kappa_0\lambda)$. It is interpreted as an "effective prior sample size" because the posterior update $\kappa_n = \kappa_0 + n$ shows that κ_0 combines additively with the number of observations, just like prior sample size in a Bayesian learning model. A larger κ_0 corresponds to more confident prior beliefs about μ.

(b) **Solution:** The first term $\sum(x_i - \bar{x})^2$ is the within-sample sum of squared deviations from the sample mean (data scatter). The second term $\frac{\kappa_0 n}{\kappa_0+n}(\bar{x} - \mu_0)^2$ represents the disagreement between the sample mean and prior mean, weighted by an effective sample size $\frac{\kappa_0 n}{\kappa_0+n}$ (harmonic mean of κ_0 and n). Together, they accumulate uncertainty in variance estimation.

(c) **Solution:** A large discrepancy between $\bar{x}$ and μ_0 increases β_n, which increases the posterior shape parameter and scale of the Gamma distribution for λ. This represents greater posterior uncertainty about the variance (heavier-tailed Gamma), reflecting conflict between data and prior on the mean. Intuitively, when data and prior disagree on μ, the extra variation is attributed to either poor prior specification or true variance larger than expected.

PART D : STUDENT-t MARGINAL AND ROBUSTNESS

D1. Tail behavior: Gaussian vs Student-t

(a) **Solution:** For large $|\mu - \mu_n|$, the Student-t distribution assigns more mass to extreme values. The Student-t density decays as $(\mu - \mu_n)^{-(\nu+1)}$ whereas the Gaussian decays as $\exp(-c(\mu - \mu_n)^2)$, and polynomial decay is slower than exponential decay.

PART E : DATA-DRIVEN GAUSSIAN ANALYSIS

E1. Simulated Gaussian data with unknown variance

(a) **Solution:** Yes, $\hat{\mu} \approx 4.9$ is very close to the true value $\mu^\star = 5$. The estimate lies within one posterior standard deviation (0.5) of the truth, which is consistent with the Bayesian posterior being calibrated to the true parameter. The posterior mean is a reasonable point estimate.

(b) **Solution:** The diagnostics show good convergence and mixing: (i) The trace plot is stationary with no trend, (ii) The ACF decays rapidly, reaching near-zero by lag 10, indicating relatively low autocorrelation, (iii) ESS $\approx$ 1600 out of 4000 iterations gives an effective sample size ratio of 40%, which is acceptable, (iv) These combined suggest the chain has converged and is mixing reasonably well.

(c) **Solution:** Even though σ^2 is unknown (and Gamma-distributed in the posterior), the posterior variance for μ is relatively small (0.5^2) because of the relatively large sample size ($n = 20$). The posterior precision $\lambda_n = \kappa_0 + n = 1 + 20 = 21$, which provides strong information about μ even as σ^2 is integrated over. The data itself are quite informative about the mean, independent of variance uncertainty.

OTHER QUESTIONS

Problem 1 – Robustness: Known vs Unknown Variance

(a) **Solution (Case A):** For known variance, the posterior for μ is:

$$p(\mu \mid x) \propto \exp\left(-\frac{n}{2\sigma^2}(\mu - \bar{x})^2\right) \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right).$$

Completing the square, the posterior is $\mu \mid x \sim \mathcal{N}(\mu_n, \sigma_n^2)$ where:

$$\sigma_n^2 = \frac{1}{n/\sigma^2 + 1/\sigma_0^2} = \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2},$$

$$\mu_n = \sigma_n^2\left(\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right).$$

Plugging in: $\mu_0 = 20$, $\sigma_0^2 = 25$, $\sigma^2 = 25$, $n = 5$, $\bar{x} = 16.0$:

$$\sigma_n^2 = \frac{25 \cdot 25}{5 \cdot 25 + 25} = \frac{625}{150} = \frac{25}{6} \approx 4.17,$$

$$\mu_n = \frac{25}{6}\left(\frac{5 \cdot 16}{25} + \frac{20}{25}\right) = \frac{25}{6}(3.2 + 0.8) = \frac{25}{6} \cdot 4 = \frac{100}{6} \approx 16.67.$$

(b) **Solution (Case B):** For Normal–Gamma prior with $\mu_0 = 20, \kappa_0 = 1, \alpha_0 = 2.5, \beta_0 = 12.5$, and data $\bar{x} \approx 16.0$, $\sum(x_i - \bar{x})^2 = 67.6$:

$$\kappa_n = \kappa_0 + n = 1 + 5 = 6,$$

$$\mu_n = \frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_n} = \frac{1 \cdot 20 + 5 \cdot 16}{6} = \frac{100}{6} \approx 16.67,$$

$$\alpha_n = \alpha_0 + \frac{n}{2} = 2.5 + 2.5 = 5,$$

$$\beta_n = \beta_0 + \frac{1}{2}\left[67.6 + \frac{1 \cdot 5}{6}(16 - 20)^2\right] = 12.5 + \frac{1}{2}\left[67.6 + \frac{5}{6} \cdot 16\right].$$

$$\beta_n = 12.5 + \frac{1}{2}(67.6 + 13.33) = 12.5 + 40.47 \approx 52.97.$$

The marginal posterior for μ is Student-t with:

$$\nu_n = 2\alpha_n = 10, \quad s_n^2 = \frac{\beta_n}{\alpha_n \kappa_n} = \frac{52.97}{5 \cdot 6} = \frac{52.97}{30} \approx 1.77.$$

(c) **Solution:** Case B's marginal posterior has heavier tails than Case A's Gaussian because: (i) Uncertainty in σ^2 is explicitly captured by the Gamma posterior on λ, (ii) The marginal posterior $p(\mu \mid x) = \int p(\mu \mid \lambda, x)p(\lambda \mid x)d\lambda$ is a mixture of Gaussians with different variances. Large-variance components place probability mass in the tails, (iii) This tail probability makes the model robust to outliers: extreme observations are explained by large variance rather than shifting μ dramatically. The Student-t form with finite degrees of freedom (10 here) reflects genuine uncertainty about whether outliers are extreme events or signals.

(d) **Solution:** Adding outlier $x_6 = 60$: The new $\bar{x}$ becomes $(16 \cdot 5 + 60)/6 \approx 26.67$.

- **Case A:** The posterior mean is forced upward to around 22.5 (weighted average of 26.67 and 20), and the credible interval widens modestly because variance is fixed. The fixed variance of 25 does not increase despite the extreme observation.
- **Case B:** The posterior mean moves less dramatically (to around 24) because the large deviation $(26.67 - 20)$ increases β_n substantially, inflating the posterior variance estimate. The Student-t tails accommodate the outlier without pulling μ as far. Case B is more robust; outliers increase variance estimates rather than forcing changes in the mean.

Problem 2 – Sequential Bayesian Filtering vs Exponential Moving Average

(a) **Solution:** Starting with $\mu_0, \lambda_0 = 1/\sigma_0^2$, upon observing x_t:

$$p(\mu \mid x_t, \mu_{t-1}, \lambda_{t-1}) \propto \exp\left(-\frac{\lambda}{2}(x_t - \mu)^2 - \frac{\lambda_{t-1}}{2}(\mu - \mu_{t-1})^2\right).$$

Completing the square, the posterior precision is $\lambda_t = \lambda_{t-1} + \lambda$ and the posterior mean is:

$$\mu_t = \frac{\lambda_{t-1}\mu_{t-1} + \lambda x_t}{\lambda_t} = \mu_{t-1} + \frac{\lambda}{\lambda_{t-1} + \lambda}(x_t - \mu_{t-1}) = \mu_{t-1} + w_t(x_t - \mu_{t-1}),$$

where $w_t = \lambda/(\lambda_{t-1} + \lambda)$.

(b) **Solution:** The EMA uses fixed α independent of time. The Bayesian update uses $w_t = \lambda/(\lambda_{t-1} + \lambda)$, which decreases over time. For large t, $\lambda_{t-1} \approx t\lambda/\sigma_0^2 + \lambda_0$, so:

$$w_t \approx \frac{\lambda}{t\lambda + \text{const}} \to 0 \text{ as } t \to \infty.$$

Bayesian updating is data-adaptive: early observations receive high weight (more uncertainty), later observations receive lower weight (more confidence). The EMA does not adapt, applying constant weight α regardless of information accumulation.

(c) **Solution:** For the Bayesian filter, you track (μ_t, λ_t) sequentially. The posterior distribution at time t is fully specified; you can compute credible intervals, tail probabilities, etc., even with 10^6 observations. For the EMA, you only store a single number m_t. You lose all uncertainty quantification: you cannot compute $\mathrm{Var}(\mu \mid m_t)$ without storing additional historical information. The EMA is computationally lighter but cannot quantify posterior uncertainty with only (m_t) in hand.

Problem 3 – Multivariate Gaussian Fusion and Geometry

1. 代式

(a) **Solution:** Likelihood: $p(y \mid \mu) \propto \exp(-\frac{1}{2}(y_1-\mu)^\top \Sigma_1^{-1}(y_1-\mu) - \frac{1}{2}(y_2-\mu)^\top \Sigma_2^{-1}(y_2-\mu))$ Prior: $p(\mu) \propto \exp(-\frac{1}{2}(\mu-\mu_0)^\top \Sigma_0^{-1}(\mu-\mu_0))$. Posterior (proportional to prior $\times$ likelihood):

2. Posterior: Conjugagcy ratios! $\Sigma_n^{-1} = \Sigma_0^{-1} + n_1\Sigma_1^{-1} + n_2\Sigma_2^{-1}$, (or) 2. show lambda w.r.t. covariance (and) 3. show relationship from reciprocals

$$\mu_n = \Sigma_n \left(\Sigma_0^{-1}\mu_0 + n_1\Sigma_1^{-1}\bar{y}_1 + n_2\Sigma_2^{-1}\bar{y}_2\right).$$

(b) **Solution:**

- As n_1 increases with n_2 fixed: The term $n_1\Sigma_1^{-1}$ dominates, pulling the posterior toward sensor 1's data and information. If Σ_1 is ellipsoidal, the posterior ellipse C_n becomes more elongated along the directions of high precision in sensor 1, and contracts in those directions.
- If Σ_1 has strong correlation aligned to one axis (elongated ellipse) and Σ_2 is spherical: The posterior combines information from orthogonal sources. The posterior ellipse will have intermediate shape, less elongated than sensor 1 alone, reflecting regularization from sensor 2's more uniform covariance structure.

(c) **Solution:**

- *Numerical:* If Σ_1 is nearly singular (one eigenvalue ≈ 0), naive matrix inversion is ill-conditioned. Small errors in Σ_1 produce large errors in Σ_1^{-1}, leading to numerical instability.
- *Bayesian:* Including Σ_0^{-1} and Σ_2^{-1} adds positive-definite contributions to the posterior precision matrix, improving conditioning. Even if $n_1\Sigma_1^{-1}$ is rank-deficient, the sum Σ_n^{-1} becomes full-rank and well-conditioned. This is implicit regularization via Bayesian fusion.

(d) **Solution:** Diagnostic checklist:

(a) Verify that the posterior mean μ_n is reasonable relative to the true μ (within $2\sigma_n$ distances).

(b) Check that the posterior ellipse C_n is smaller than the prior and likelihood ellipses (information fusion reduces uncertainty).

(c) Confirm alignment of principal axes: posterior axes should align with directions of strong agreement between sensors.

(d) Sensitivity check: perturb Σ_1, Σ_2 slightly and recompute; posterior should be stable (not highly sensitive to small input changes).

(e) Compare 95% credible region volume: should be smaller than individual sensor uncertainties.

(e) **Solution:** For high dimensions:

- *Low-rank structure:* Assume $\Sigma_i = D_i + U_i V_i^\top$ (diagonal plus low-rank), reducingparameter count and inversion cost to $O(dk^2)$ where $k \ll d$.
- *Sparse precision matrices:* If posterior precision $\Lambda_n = \Sigma_n^{-1}$ is sparse, use sparse matrix algorithms ($O(d^\alpha)$ with $\alpha < 3$) and variational inference with sparse graphical model structure.
- *Pros/cons:* Low-rank is interpretable and fast when applicable; sparse structure exploits conditional independence but requires problem structure to exist. Both reduce dimensions but may lose information.

Problem 4 – Normal–Gamma Posterior Geometry and Student-t Marginals

(a) **Solution:** The joint posterior for (μ, λ) given x factors as:

$$p(\mu, \lambda \mid x) \propto p(x \mid \mu, \lambda) p(\mu \mid \lambda) p(\lambda).$$

Likelihood: $\prod_i \mathcal{N}(x_i \mid \mu, 1/\lambda)$. Prior: $\mu \mid \lambda \sim \mathcal{N}(\mu_0, 1/(\kappa_0 \lambda))$, $\lambda \sim \text{Gamma}(\alpha_0, \beta_0)$. The posterior is Normal–Gamma with:

$$\kappa_n = \kappa_0 + n, \quad \mu_n = \frac{\kappa_0 \mu_0 + n\bar{x}}{\kappa_n},$$

$$\alpha_n = \alpha_0 + \frac{n}{2}, \quad \beta_n = \beta_0 + \frac{1}{2}\left[\sum (x_i - \bar{x})^2 + \frac{\kappa_0 n}{\kappa_n}(\bar{x} - \mu_0)^2\right].$$

(b) **Solution:** Marginal $p(\mu \mid x)$:

$$p(\mu \mid x) = \int p(\mu \mid \lambda, x) p(\lambda \mid x) d\lambda.$$

Given λ, the conditional posterior is $\mathcal{N}(\mu_n, 1/(\kappa_n \lambda))$. Marginalizing:

$$p(\mu \mid x) = \int \frac{\sqrt{\kappa_n \lambda}}{\sqrt{2\pi}} \exp\left(-\frac{\kappa_n \lambda}{2}(\mu - \mu_n)^2\right) \frac{\beta_n^{\alpha_n}}{\Gamma(\alpha_n)} \lambda^{\alpha_n - 1} e^{-\beta_n \lambda} d\lambda.$$

Substituting $z = \lambda$ and completing the integral (recognizing a Gamma integral):

$$p(\mu \mid x) \propto \left[1 + \frac{\kappa_n (\mu - \mu_n)^2}{\beta_n / \alpha_n}\right]^{-(\alpha_n + 1/2)}.$$

This is a Student-t with $\nu_n = 2\alpha_n$ degrees of freedom, location μ_n, and scale squared $s_n^2 = \beta_n/(\alpha_n \kappa_n)$.

(c) **Solution:** Visual agreement of histogram with analytical Student-t provides evidence that: (i) The MCMC sampler correctly samples (
mu, λ) pairs with the right joint distribution, (ii) The marginal for μ is correctly computed. However, this does not rule out: (i) Slow mixing in λ (ESS might be low), (ii) Autocorrelation in μ due to dependence on λ, (iii) Multi-modality in the joint space that the marginal histogram masks.

(d) **Solution:** For the Gamma posterior $\lambda \mid x \sim \text{Gamma}(\alpha_n, \beta_n)$:

$$\mathbb{E}[\lambda \mid x] = \frac{\alpha_n}{\beta_n}, \quad \text{Var}[\lambda \mid x] = \frac{\alpha_n}{\beta_n^2}.$$

Interpretation: (i) The effective sample size on precision is $\kappa_n = \kappa_0 + n$, analogous to a sample size. (ii) As $n \to \infty$ with fixed prior, $\alpha_n \approx n/2$ and $\beta_n \approx ns^2/2$, so $\mathbb{E}[\lambda] \approx 1/s^2$ (MLE) and $\text{Var}[\lambda] \approx 1/(ns^4)$, showing posterior concentration at rate $1/n$.

(e) **Solution:**

- *Varying α_0:* Large α_0 (informative prior on precision) produces a Gamma posterior with larger α_n, concentrating the posterior variance estimate. This reduces posterior tail probability on extreme λ and makes $p(\mu \mid x)$ closer to Gaussian (fewer tails). Small α_0 allows the posterior to spread and accommodate outliers (heavier tails on μ).
- *Varying β_0:* Large β_0 places prior belief on small λ (large σ^2), making the posterior variance estimate larger and $p(\mu \mid x)$ heavier-tailed (more robust to outliers). Small β_0 produces tighter posterior on σ^2 and less robust behavior.

Heavy-tailed μ posteriors (small α_0, β_0) accommodate outliers better because uncertainty in σ^2 is explicitly preserved.

Problem 5 – Monte Carlo Estimation and Diagnostics in Gaussian Models

(a) **Solution:** Convergence and mixing assessment:

- *Trace plots:* "Hairy" and stationary traces indicate the chain is actively exploring the target distribution and not stuck in one region. No trends or drift suggest burn-in was sufficient.
- *ACF:* Rapid decay (near zero by lag 10) indicates low autocorrelation and good mixing. The correlation $\rho(1) \approx 0.30$ is moderate but acceptable.
- *ESS:* ESS ≈ 1500 out of 4000 iterations gives ESS per iteration $= 1500/4000 = 37.5\%$, indicating the chains are effectively sampling independent draws at $\approx 1/3$ the rate of iid samples. This is adequate.
- $\hat{R}$*:* $\hat{R} \approx 1.01$ is excellent (target < 1.05), indicating the 4 chains have mixed and converged to the same distribution.
- **Conclusion:** The chain has converged and mixed well for μ. Inference on μ is reliable.

(b) **Solution:** Slower mixing in λ ($\hat{R} \approx 1.05$, ESS ≈ 300) may still produce a reasonable marginal histogram for μ if: (i) μ and λ are moderately correlated (but not perfectly), so non-convergence in λ marginally averages out, (ii) the burn-in and thinning were designed to remove dependent samples. However, this is concerning: (a) Non-convergence in λ means the posterior variance estimate is unreliable, (b) If μ and λ are strongly dependent, poor mixing in λ implies poor exploration of the μ conditional distribution as well. Partial non-convergence is acceptable only if the primary inferential interest is in μ and λ is a nuisance parameter; unacceptable if σ^2 (or functions of it) are of interest.

(c) **Solution:** The Monte Carlo standard error (MCSE) of $\hat{\mu}$ is:

$$\text{MCSE}(\hat{\mu}) = \frac{\text{sd}(\mu)}{\sqrt{\text{ESS}}}.$$

Using the given values:

$$\text{MCSE}(\hat{\mu}) = \frac{1.3}{\sqrt{1500}} = \frac{1.3}{38.73} \approx 0.0336.$$

This is negligible relative to the posterior sd (1.3), representing only $\approx 2.6\%$ of the posterior standard deviation. For practical purposes, MCMC Monte Carlo error is small and point estimates are reliable.

(d) **Solution:** For a nonlinear functional $g(\mu, \sigma^2)$ (e.g., a quantile or cost function), the MCSE of $\hat{g}$ depends on both marginal and joint properties. Slow mixing in λ (and thus σ^2) means: (i) Estimates of $\mathbb{E}[g(\mu, \sigma^2)]$ are noisy even if μ marginal is well-estimated, (ii) The posterior covariance between μ and σ^2 may be poorly estimated. To empirically check: (a) Compute $g(\mu^{(s)}, \sigma_2^{(s)})$ for each MCMC draw and compare univariate ESS of this functional to ESS of μ alone; (b) If functional ESS $\ll$ univariate ESS of μ, then joint mixing is poor; (c) Estimate MCSE of $\hat{g}$ separately and compare to the marginal MCSE of μ.

(e) **Solution:** Reparameterizations and alternative MCMC schemes:

- *Gibbs sampling:* Use conditional conjugacy of Normal–Gamma to alternate sampling $\mu \mid \lambda, x$ and $\lambda \mid \mu, x$. Both conditionals are exactly samplable (Gaussian and Gamma). Gibbs can improve mixing over random-walk Metropolis by updating in the latent-data space.
- *Reparameterization:* Use $\log \lambda$ or centered parameterization $(\mu - \bar{x})$ to improve geometry and reduce correlation between μ and λ.
- *HMC:* If exact sampling is infeasible, HMC with automatic differentiation can be faster than Gibbs for high-dimensional posteriors, but is overkill here.
- **Justification:** The posterior geometry exhibits strong coupling between μ and λ (heavier λ allows larger deviations in μ). Gibbs exploits conditional conjugacy to decouple sampling in a natural way, improving mixing. HMC would be computationally expensive for this relatively simple problem. Reparameterization helps precondition the posterior to reduce correlation, improving any sampler's efficiency.