Linear Gaussian Regression

Linear Gaussian Models

Sai Sampath Kedari
sampath@umich.edu

Contents

1 Linear Gaussian Regression: Problem Setup

We start with a linear regression model with additive Gaussian noise. At each observation time $t_k \in \mathbb{R}$, the measured output $y_k \in \mathbb{R}$ is modeled as

$$y_k = \theta_1 + \theta_2 t_k + \eta_k, \qquad \eta_k \sim \mathcal{N}(0, \sigma^2),$$

where

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \in \mathbb{R}^2$$

denotes the unknown regression parameters.

Collecting n observations, we introduce the stacked observation vector and design matrix

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, \qquad H = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix} \in \mathbb{R}^{n \times 2}.$$

With this notation, the observation model can be written compactly as

$$Y = H\theta + \eta, \qquad \eta \sim \mathcal{N}(0, \sigma^2 I_n).$$

From a Bayesian perspective, the unknown parameters are treated as random variables. We place a Gaussian prior on θ,

$$\theta \sim \mathcal{N}(\mu_0, C_0),$$

where $\mu_0 \in \mathbb{R}^2$ and $C_0 \in \mathbb{R}^{2\times 2}$ represent the prior mean and covariance.

2 Joint Linear Gaussian Model

We now formulate the linear regression problem within a Bayesian framework and derive the joint Gaussian distribution of the unknown parameters and observations. This joint representation will serve as the foundation for deriving prior predictive distributions, posterior distributions, and recursive updates.

2.1 Bayesian Model Specification

The linear regression model is specified as a hierarchical Gaussian model. First, we place a Gaussian prior on the unknown regression parameters,

$$\theta \sim \mathcal{N}(\mu_0, C_0),$$

where $\theta \in \mathbb{R}^2$ denotes the parameter vector, $\mu_0 \in \mathbb{R}^2$ is the prior mean, and $C_0 \in \mathbb{R}^{2\times 2}$ is the prior covariance.

Given θ, the observations are generated according to the observation model, also referred to as the likelihood,

$$Y \mid \theta \sim \mathcal{N}(H\theta, \sigma^2 I_n),$$

where $H \in \mathbb{R}^{n\times 2}$ is the design matrix and $\sigma^2 I_n$ represents additive Gaussian measurement noise.

This hierarchical specification can be summarized as

$$\theta \sim \mathcal{N}(\mu_0, C_0), \qquad Y \mid \theta \sim \mathcal{N}(H\theta, \sigma^2 I_n).$$

2.2 Why a Joint Distribution

Rather than deriving the posterior distribution $\theta \mid Y$ directly, we first construct the joint distribution of parameters and observations. This approach allows posterior and predictive distributions to be obtained systematically through conditioning and marginalization, and naturally supports recursive estimation.

To this end, we define the stacked random vector

$$Z = \begin{bmatrix} \theta \\ Y \end{bmatrix} \in \mathbb{R}^{2+n}.$$

Since Z is an affine function of Gaussian random variables, it follows that Z is jointly Gaussian. To fully characterize this joint distribution, it suffices to compute its mean and covariance.

2.3 Latent Function, Observations, and Uncertainty Sources

In Bayesian linear regression, the primary object of interest is the latent regression function

$$f = H\theta,$$

which represents the underlying deterministic relationship between input and output.

Accordingly, we distinguish between the following distributions:

- Prior over parameters:
$$\theta \sim \mathcal{N}(\mu_0, C_0),$$
- Prior predictive distribution of the latent function:
$$H\theta \sim \mathcal{N}(H\mu_0,\, HC_0H^\top),$$
- Observation model (likelihood):
$$Y \mid \theta \sim \mathcal{N}(H\theta,\, \sigma^2 I_n),$$
- Posterior distribution over parameters (to be derived later):
$$\theta \mid Y,$$
- Posterior predictive distribution of the latent function:
$$H\theta \mid Y.$$

The observation noise η affects the measurements Y but does not alter the latent function itself. As a result, predictive uncertainty in $H\theta$ reflects parameter uncertainty only, and is reduced as more data are incorporated.

2.4 Joint Mean

From the conditional observation model,

$$Y = H\theta + \eta, \qquad \eta \sim \mathcal{N}(0, \sigma^2 I_n),$$

the conditional expectation of Y given θ is

$$\mathbb{E}[Y \mid \theta] = H\theta.$$

Using the law of total expectation,

$$\mathbb{E}[Y] = \mathbb{E}_\theta[\mathbb{E}[Y \mid \theta]] = \mathbb{E}_\theta[H\theta] = H\,\mathbb{E}[\theta] = H\mu_0.$$

Since $\mathbb{E}[\theta] = \mu_0$, the joint mean of Z is

$$\mathbb{E}[Z] = \begin{bmatrix} \mu_0 \\ H\mu_0 \end{bmatrix}.$$

2.5 Joint Covariance

The covariance matrix of Z admits the block structure

$$\mathrm{Cov}(Z) = \begin{bmatrix} \mathrm{Var}(\theta) & \mathrm{Cov}(\theta, Y) \\ \mathrm{Cov}(Y, \theta) & \mathrm{Var}(Y) \end{bmatrix}.$$

Each block is computed explicitly below.

2.5.1 Parameter Covariance

By definition of the prior,

$$\mathrm{Var}(\theta) = C_0.$$

2.5.2 Cross-Covariance

Using covariance algebra,

$$\mathrm{Cov}(\theta, Y) = \mathrm{Cov}(\theta, H\theta + \eta).$$

Expanding the covariance,

$$\mathrm{Cov}(\theta, Y) = \mathrm{Cov}(\theta, H\theta) + \mathrm{Cov}(\theta, \eta).$$

Since θ and η are independent,

$$\mathrm{Cov}(\theta, \eta) = 0.$$

Using the identity $\mathrm{Cov}(X, AX) = \mathrm{Var}(X)A^\top$, we obtain

$$\mathrm{Cov}(\theta, H\theta) = \mathrm{Var}(\theta)\, H^\top = C_0 H^\top.$$

Therefore,

$$\mathrm{Cov}(\theta, Y) = C_0 H^\top, \qquad \mathrm{Cov}(Y, \theta) = H C_0.$$

2.5.3 Observation Covariance

To compute $\mathrm{Var}(Y)$, we apply the law of total variance,

$$\mathrm{Var}(Y) = \mathbb{E}_\theta[\mathrm{Var}(Y \mid \theta)] + \mathrm{Var}_\theta(\mathbb{E}[Y \mid \theta])\,.$$

From the observation model,

$$\mathrm{Var}(Y \mid \theta) = \sigma^2 I_n,$$

and therefore

$$\mathbb{E}_\theta[\mathrm{Var}(Y \mid \theta)] = \sigma^2 I_n.$$

Moreover,

$$\mathbb{E}[Y \mid \theta] = H\theta,$$

so using the identity $\mathrm{Var}(AX) = A\ \mathrm{Var}(X)\, A^\top$,

$$\mathrm{Var}_\theta(H\theta) = H\ \mathrm{Var}(\theta)\, H^\top = HC_0H^\top.$$

Combining the two terms yields

$$\mathrm{Var}(Y) = HC_0H^\top + \sigma^2 I_n.$$

2.6 Final Joint Distribution

Collecting the mean and covariance results, the joint distribution of parameters and observations is

$$\begin{bmatrix}\theta \\ Y\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_0 \\ H\mu_0\end{bmatrix},\ \begin{bmatrix}C_0 & C_0H^\top \\ HC_0 & HC_0H^\top + \sigma^2 I_n\end{bmatrix}\right).$$

This joint Gaussian formulation provides a unified representation from which prior predictive distributions, posterior distributions over parameters, and recursive estimation algorithms can be derived through marginalization and conditioning.

3 Synthetic Data, Prior, and Prior Predictive Distribution

3.1 Generative Model and Synthetic Data

We consider a linear regression model with additive Gaussian measurement noise. At each observation time $t_k \in \mathbb{R}$, the measurement $y_k \in \mathbb{R}$ is generated according to

$$y_k = \theta_1 + \theta_2 t_k + \eta_k, \qquad \eta_k \sim \mathcal{N}(0, \sigma^2).$$

The unknown regression parameters are collected into the vector

$$\boldsymbol{\theta} = \begin{bmatrix}\theta_1 \\ \theta_2\end{bmatrix} \in \mathbb{R}^2.$$

For the purpose of simulation and visualization, we fix ground-truth parameters

$$\theta_1^{\mathrm{true}} = 1.0, \qquad \theta_2^{\mathrm{true}} = 0.7,$$

and generate a dataset of n noisy observations $\{(t_k, y_k)\}_{k=1}^{n}$. The true regression function

$$y^{\text{true}}(t) = \theta_1^{\text{true}} + \theta_2^{\text{true}} t$$

is evaluated on a dense grid only for visualization and is not used for inference.

Figure 1 shows the generated observations together with the underlying ground-truth regression line.

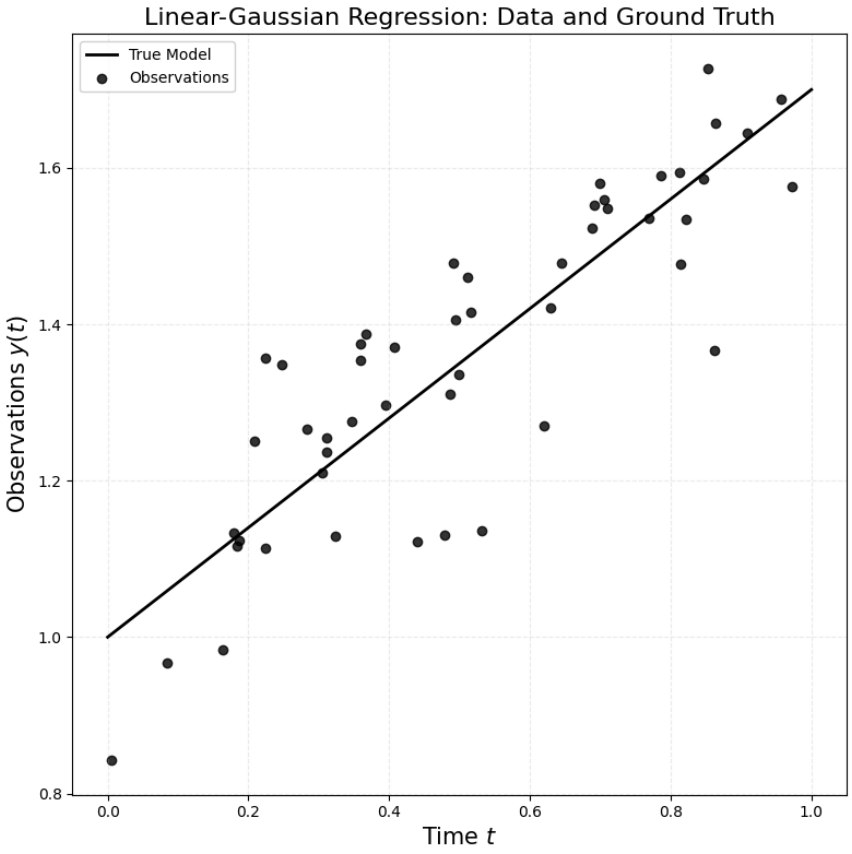

Figure 1: Synthetic linear regression data with additive Gaussian noise and the ground-truth regression function.

3.2 Bayesian Model Specification

In the Bayesian formulation, the regression parameters are treated as random variables. We place independent standard Gaussian priors on the intercept and slope,

$$\theta_1 \sim \mathcal{N}(0, 1), \qquad \theta_2 \sim \mathcal{N}(0, 1).$$

Equivalently, the parameter vector follows a multivariate Gaussian prior

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{C}_0), \qquad \boldsymbol{\mu}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \mathbf{C}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The Bayesian hierarchy for this model can be written compactly as

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{C}_0), \qquad Y \mid \boldsymbol{\theta} \sim \mathcal{N}(H\boldsymbol{\theta}, \sigma^2 I_n),$$

where $Y \in \mathbb{R}^n$ is the stacked observation vector and

$$H = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}$$

is the design matrix.

Figure 2 visualizes the prior distribution through its marginal and joint densities.

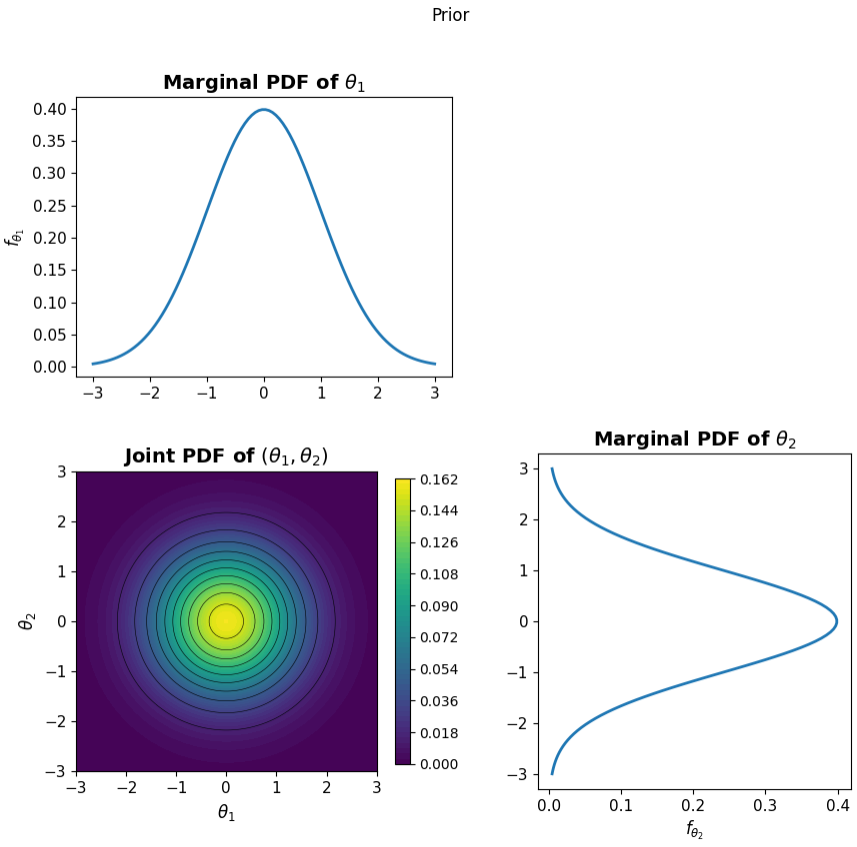

Figure 2: Marginal and joint distributions of the regression parameters under the Gaussian prior.

3.3 Prior Predictive Distribution of the Latent Function

Before observing any data, uncertainty in the parameters induces uncertainty in the regression function itself. The latent function evaluated on a grid of time points t is given by

$$f(t) = \theta_1 + \theta_2 t.$$

Let

$$H_p = \begin{bmatrix} 1 & t_1^{(p)} \\ 1 & t_2^{(p)} \\ \vdots & \vdots \\ 1 & t_m^{(p)} \end{bmatrix}$$

denote the design matrix evaluated on a dense plotting grid. The latent function values can be written as

$$f = H_p \boldsymbol{\theta}.$$

Since $\boldsymbol{\theta}$ is Gaussian, the induced distribution of f is also Gaussian. Using linearity of expectation and covariance,

$$\mathbb{E}[f] = H_p \boldsymbol{\mu}_0, \qquad \mathrm{Var}(f) = H_p \mathbf{C}_0 H_p^\top.$$

This distribution is referred to as the *prior predictive distribution of the latent function*. It captures uncertainty due solely to the unknown parameters, without measurement noise.

Figure 3 shows the prior predictive mean together with a ± 2 standard deviation envelope.

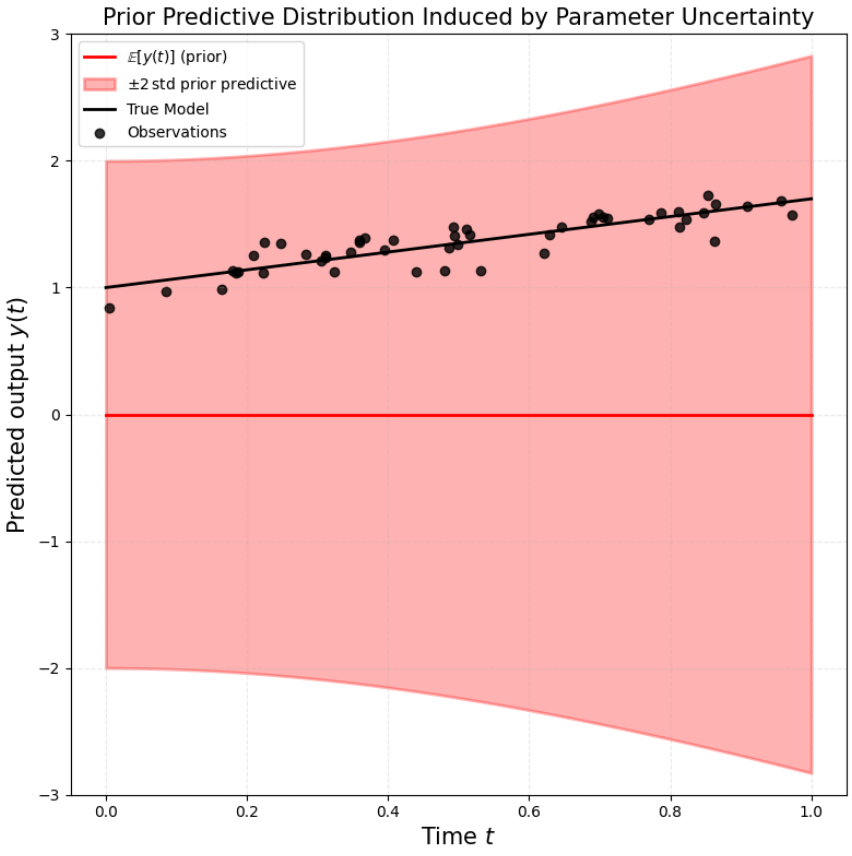

Figure 3: Prior predictive distribution of the latent regression function induced by parameter uncertainty.

This prior predictive distribution provides a baseline against which the effect of data will be assessed. In the next sections, we condition on observations to derive the posterior distribution of

the parameters and the corresponding posterior predictive distribution.

4 Batch Posterior Inference for Linear–Gaussian Regression

In the Bayesian formulation of linear regression, the unknown parameters $\theta = [\theta_1, \theta_2]^\top$ are treated as latent random variables. Given a fixed dataset $Y = y$, our goal is to compute the posterior distribution of θ and the induced posterior uncertainty over the latent regression function.

4.1 Hierarchical Bayesian Model

The linear–Gaussian regression model can be written hierarchically as

$$\begin{aligned} \theta &\sim \mathcal{N}(\mu_0, C_0), \\ Y \mid \theta &\sim \mathcal{N}(H\theta, \sigma^2 I_n), \end{aligned}$$

where θ represents the unknown parameters, and Y denotes the observed data.

This hierarchy makes explicit the Bayesian roles of each component:

$$\text{prior} \;\rightarrow\; \text{likelihood (observation model)} \;\rightarrow\; \text{posterior}.$$

4.2 Joint Gaussian Representation

As derived in Section 2, the stacked random vector

$$Z = \begin{bmatrix} \theta \\ Y \end{bmatrix}$$

is jointly Gaussian with

$$Z \sim \mathcal{N}\left(\begin{bmatrix} \mu_0 \\ H\mu_0 \end{bmatrix}, \begin{bmatrix} C_0 & C_0 H^\top \\ HC_0 & HC_0 H^\top + \sigma^2 I_n \end{bmatrix} \right).$$

This joint representation is central: once (θ, Y) is written as a joint Gaussian, all posterior and predictive quantities follow by conditioning.

4.3 Posterior Distribution of the Parameters

We now condition the joint Gaussian distribution on the observed data $Y = y$.

From the general result on conditional distributions of jointly Gaussian random vectors (derived in the companion report),

$$\mathbf{X} \mid \mathbf{Y} = \mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\mu}_X + \Sigma_{XY}\Sigma_{YY}^{-1}(\mathbf{y} - \boldsymbol{\mu}_Y),\; \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} \right),$$

we identify

$$\mathbf{X} = \theta, \qquad \mathbf{Y} = Y.$$

Substituting the blocks of the joint covariance yields the batch posterior

$$\theta \mid Y = y \sim \mathcal{N}(\mu_{\text{post}}, C_{\text{post}}),$$

with

$$\mu_{\text{post}} = \mu_0 + C_0 H^\top \left(H C_0 H^\top + \sigma^2 I_n \right)^{-1} (y - H\mu_0),$$

$$C_{\text{post}} = C_0 - C_0 H^\top \left(H C_0 H^\top + \sigma^2 I_n \right)^{-1} H C_0.$$

For the dataset considered here, this evaluates numerically to

$$\mu_{\text{post}} \approx \begin{bmatrix} 1.013 \\ 0.682 \end{bmatrix}, \qquad C_{\text{post}} \approx \begin{bmatrix} 0.0010 & -0.0016 \\ -0.0016 & 0.0031 \end{bmatrix}.$$

The posterior mean is close to the true parameters $[\theta_1, \theta_2]^\top = [1.0, 0.7]^\top$, while the posterior covariance shows a dramatic reduction in uncertainty compared to the prior.

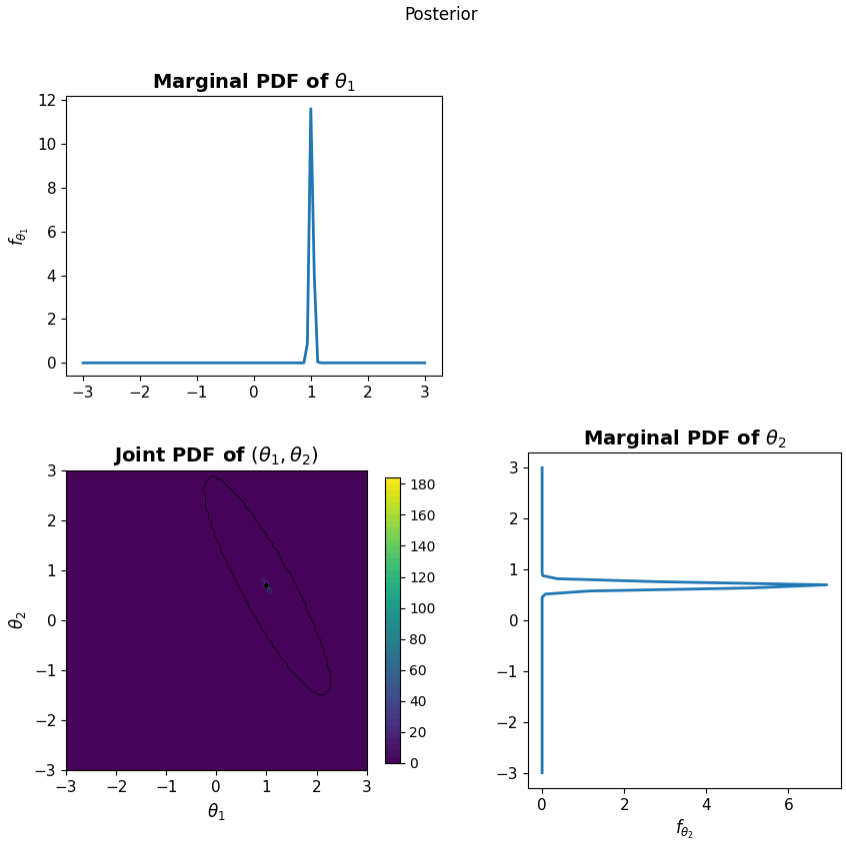

Figure 4: Posterior distribution of (θ_1, θ_2) after observing data. The posterior is sharply concentrated relative to the prior.

4.4 Posterior Predictive Distribution of the Latent Function

Rather than predicting noisy observations, we are often interested in the latent regression function

$$f(t) = H(t)\theta, \qquad H(t) = [1 \;\; t].$$

Conditioned on data, the uncertainty in $f(t)$ arises solely from the posterior uncertainty in θ. The posterior predictive distribution of the latent function is therefore

$$f(t) \mid Y \sim \mathcal{N}\left(H(t)\mu_{\text{post}},\ H(t)C_{\text{post}}H(t)^\top\right).$$

This distribution quantifies uncertainty in the underlying model itself, separate from observation noise. This distinction will be essential when we move to dynamical systems and state estimation.

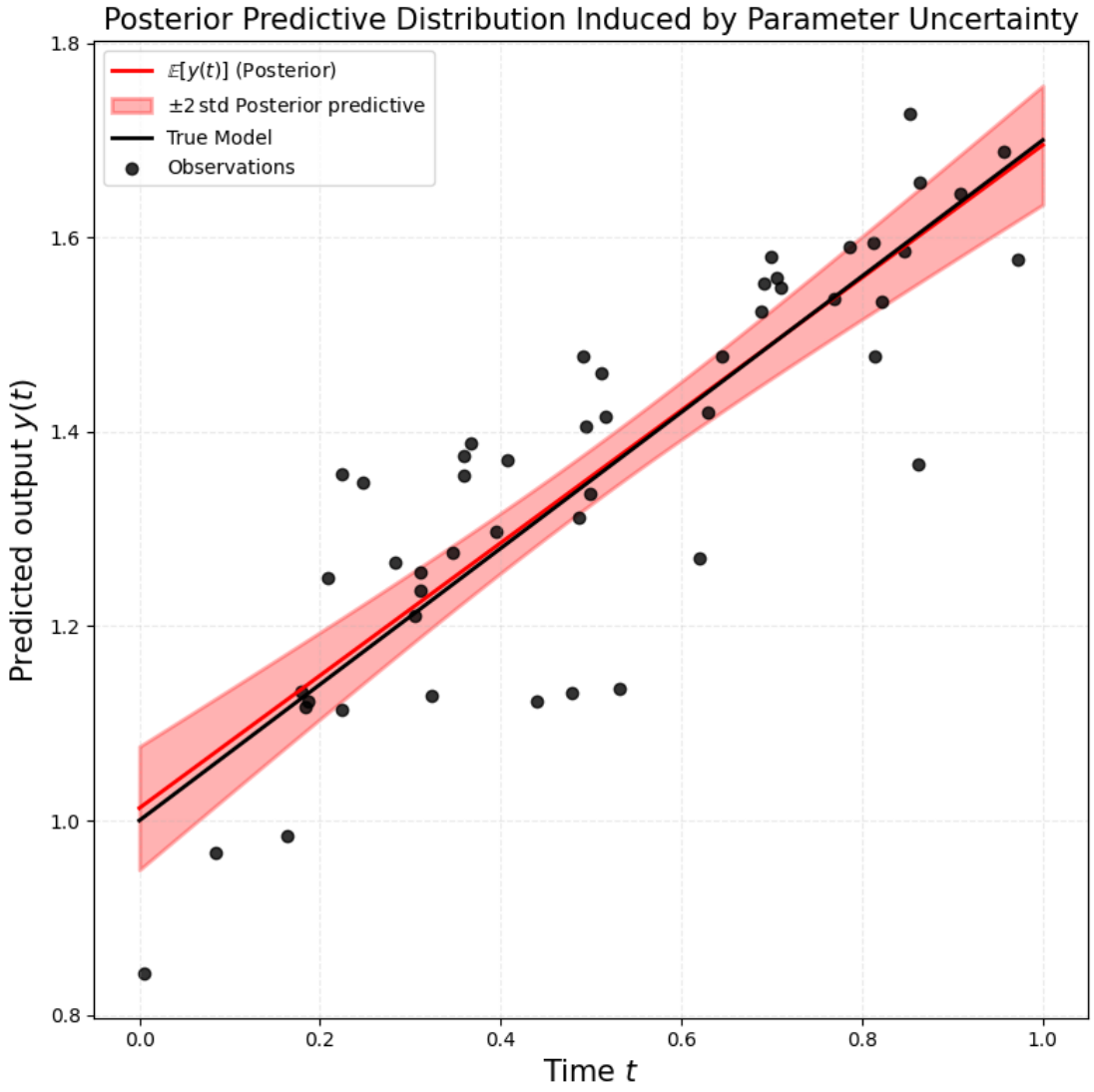

Figure 5: Posterior predictive distribution of the latent regression function. The uncertainty band reflects parameter uncertainty after observing data.

4.5 Interpretation and Outlook

The batch posterior completes the static Bayesian regression problem. All uncertainty updates occur in a single conditioning step using the full dataset.

In the next section, we reformulate this update recursively, showing that the same posterior can be recovered by processing observations sequentially. This transition provides the conceptual bridge from batch regression to Bayesian filtering and, ultimately, to the Kalman filter and its nonlinear extensions.

5 Recursive Linear Regression via Sequential Bayesian Conditioning

In the batch formulation, the posterior distribution $\theta \mid Y_{1:n}$ is obtained by conditioning the prior $\theta \sim \mathcal{N}(\mu_0, C_0)$ on all observations simultaneously. We now derive an equivalent formulation in which the posterior is updated sequentially as new observations arrive.

The central idea is that the posterior after $n-1$ observations becomes the prior for the nth update. This principle underlies recursive Bayesian estimation and filtering.

5.1 Sequential Observation Model

At time step n, we observe a single scalar measurement

$$y_n = H_n\theta + \eta_n, \qquad \eta_n \sim \mathcal{N}(0, \sigma^2),$$

where the observation matrix is the row vector

$$H_n = \begin{bmatrix} 1 & t_n \end{bmatrix} \in \mathbb{R}^{1\times 2}.$$

Conditioned on θ, the likelihood (observation model) is

$$y_n \mid \theta \sim \mathcal{N}(H_n\theta, \sigma^2).$$

5.2 Recursive Prior

Assume that after observing data up to time $n-1$, the posterior over the parameters is

$$\theta \mid Y_{1:n-1} \sim \mathcal{N}(\mu_{n-1}, C_{n-1}).$$

This distribution serves as the prior when incorporating the new observation y_n.

5.3 Posterior Update via Gaussian Conditioning

We now condition the prior $\mathcal{N}(\mu_{n-1}, C_{n-1})$ on the new likelihood $y_n \mid \theta$.

The innovation variance is

$$S_n = H_n C_{n-1} H_n^\top + \sigma^2.$$

The Kalman gain is

$$K_n = C_{n-1} H_n^\top S_n^{-1}.$$

The posterior mean update is

$$\mu_n = \mu_{n-1} + K_n\big(y_n - H_n\mu_{n-1}\big),$$

and the posterior covariance update is

$$C_n = C_{n-1} - K_n H_n C_{n-1}.$$

These equations update the parameter distribution using only the previous posterior and the new data point.

5.4 Equivalence to Batch Linear Regression

After processing n observations, the recursive posterior

$$\theta \mid Y_{1:n} \sim \mathcal{N}(\mu_n, C_n)$$

is algebraically identical to the batch posterior obtained by conditioning on the full dataset $Y_{1:n}$ at once.

Thus, recursive linear regression performs exact Bayesian inference while enabling online, sequential updates.

5.5 Posterior over the Latent Regression Function

The latent regression function is

$$f(t) = H(t)\theta, \qquad H(t) = \begin{bmatrix} 1 & t \end{bmatrix}.$$

Conditioned on data up to time n, the posterior predictive distribution over the latent function is

$$f(t) \mid Y_{1:n} \sim \mathcal{N}\big(H(t)\mu_n,\, H(t)C_n H(t)^\top\big).$$

This distribution captures uncertainty arising solely from the parameters and contracts as more observations are assimilated.

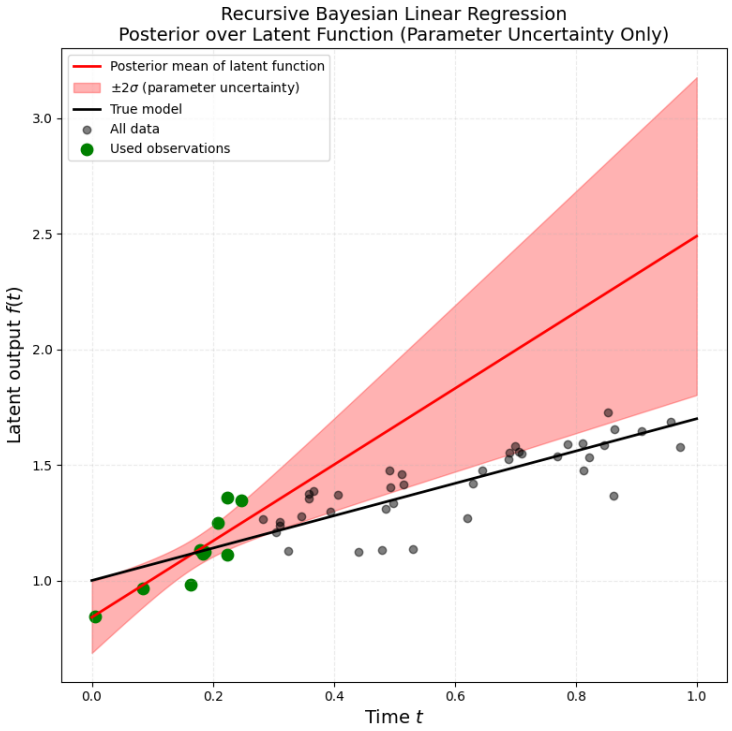

Figure 6: Recursive posterior predictive distribution after 10 observations.

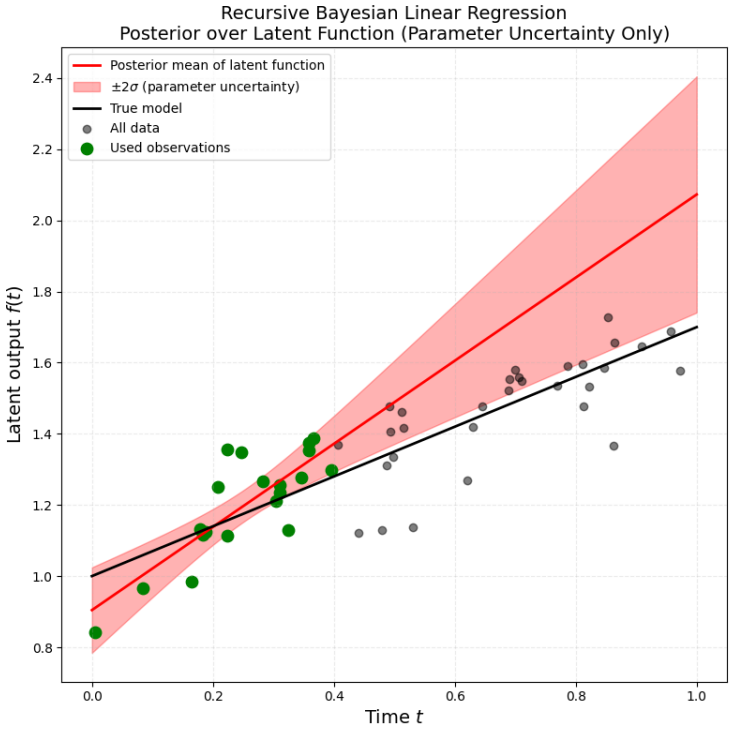

Figure 7: Recursive posterior predictive distribution after 20 observations.

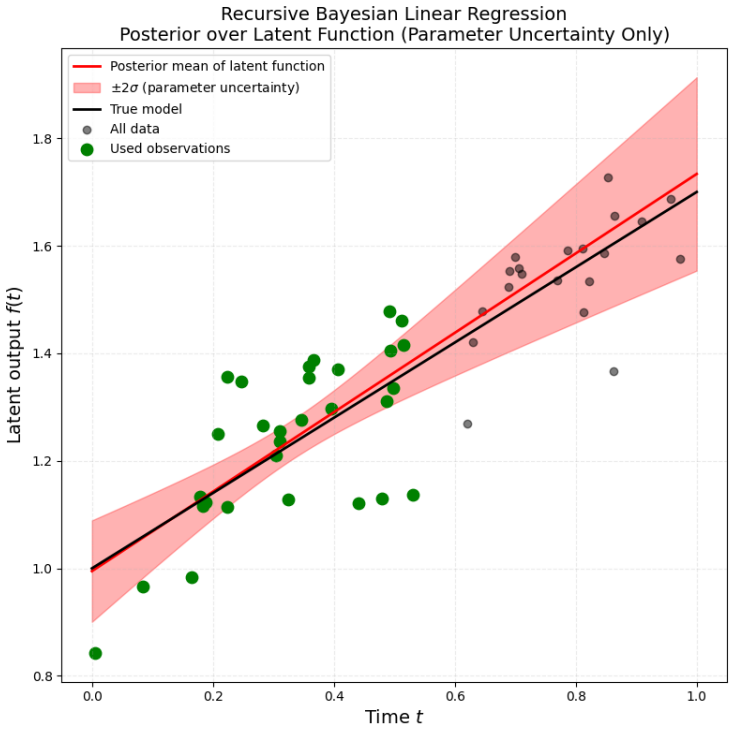

Figure 8: Recursive posterior predictive distribution after 30 observations.

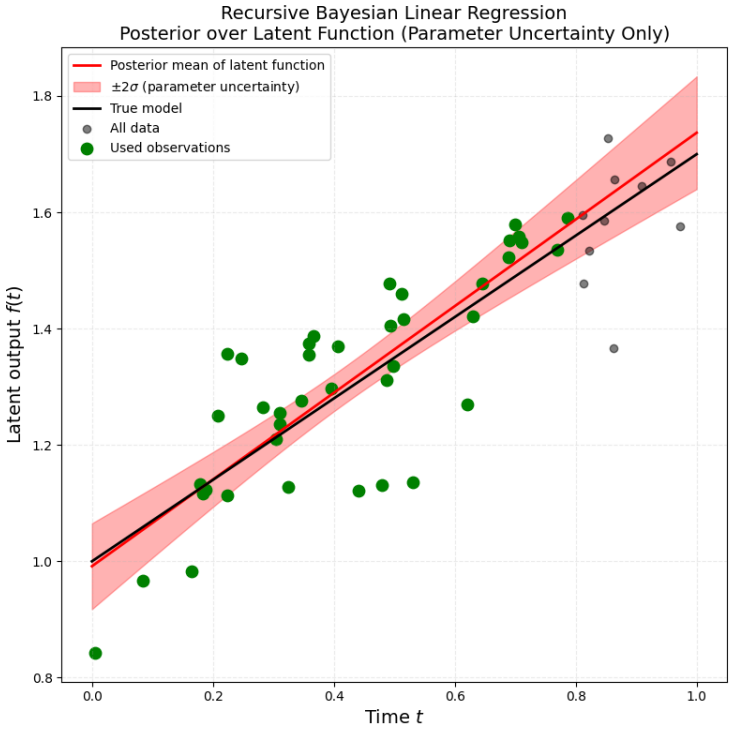

Figure 9: Recursive posterior predictive distribution after 40 observations.

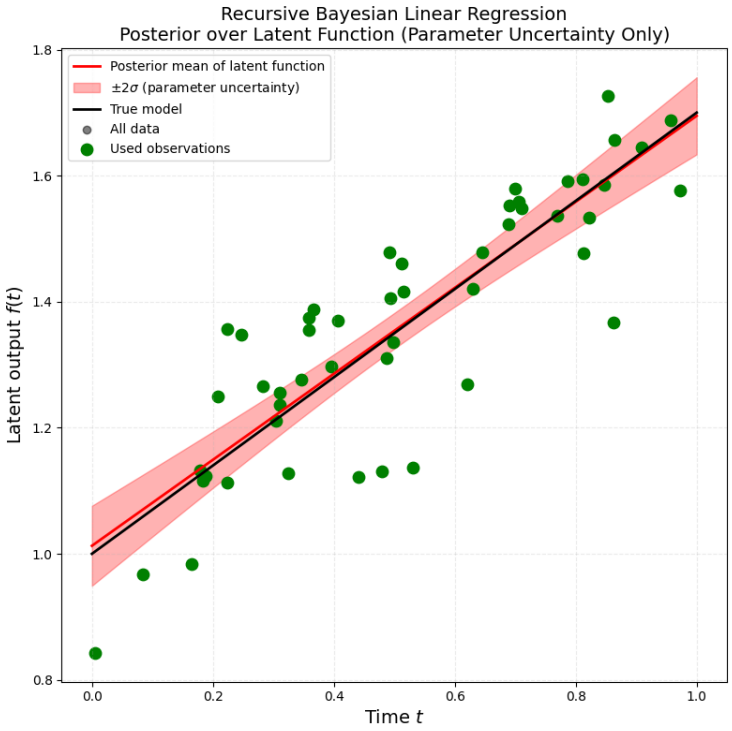

Figure 10: Recursive posterior predictive distribution after 50 observations.

6 Convergence of Recursive and Batch Posterior Means

We conclude this chapter by directly comparing the posterior means obtained from *recursive Bayesian linear regression* and *batch Bayesian linear regression*.

Let

$$\mu_n = \begin{bmatrix} \mu_{\theta_1}^{(n)} \\ \mu_{\theta_2}^{(n)} \end{bmatrix}$$

denote the posterior mean of the parameter vector after conditioning on the first n observations using the recursive update. Let

$$\mu_{\text{batch}} = \begin{bmatrix} \mu_{\theta_1} \\ \mu_{\theta_2} \end{bmatrix}$$

denote the posterior mean obtained by conditioning on the full dataset in a single batch update.

Figure 11 shows the evolution of the recursive posterior means $\mu_{\theta_1}^{(n)}$ and $\mu_{\theta_2}^{(n)}$ as the number of observations increases, together with the corresponding batch posterior means.

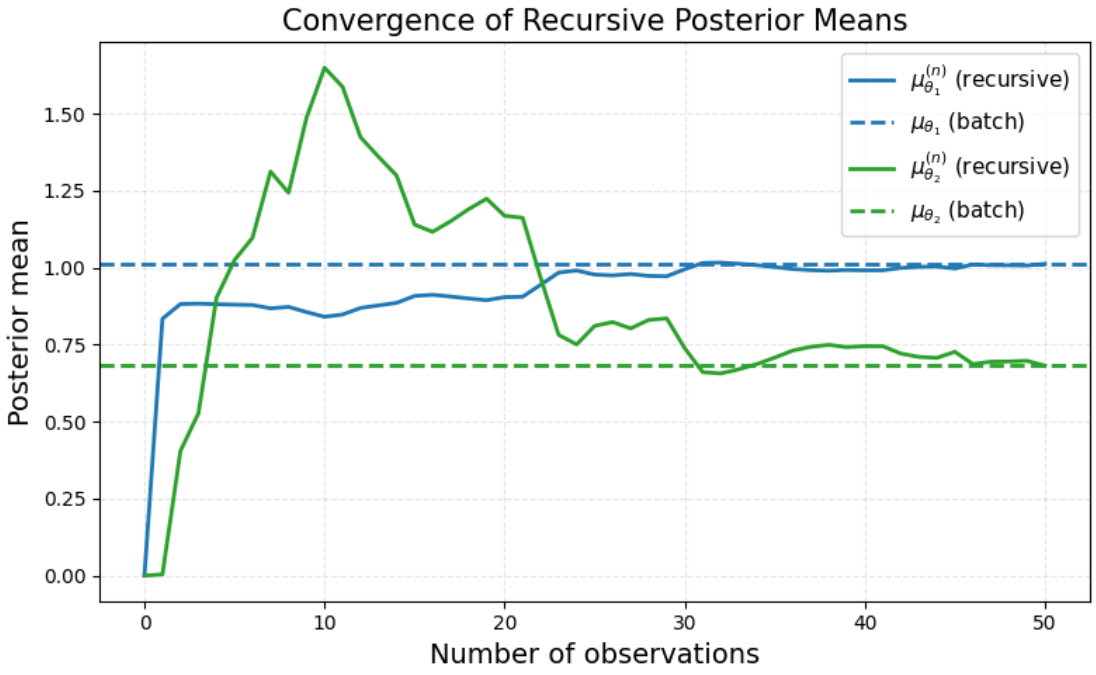

Figure 11: Convergence of recursive posterior means to the batch posterior means. Solid lines show the recursive posterior means $\mu_{\theta_1}^{(n)}$ and $\mu_{\theta_2}^{(n)}$ as a function of the number of observations. Dashed horizontal lines indicate the batch posterior means obtained using all observations simultaneously.

As expected, the recursive posterior means converge to the batch posterior means as more data are assimilated. This confirms that recursive Bayesian linear regression is *exactly equivalent* to batch Bayesian linear regression when the model is linear and Gaussian, differing only in how the data are incorporated.

This equivalence is fundamental and forms the basis for Bayesian filtering and smoothing methods developed in subsequent chapters, where data arrive sequentially and batch inference is computationally impractical.