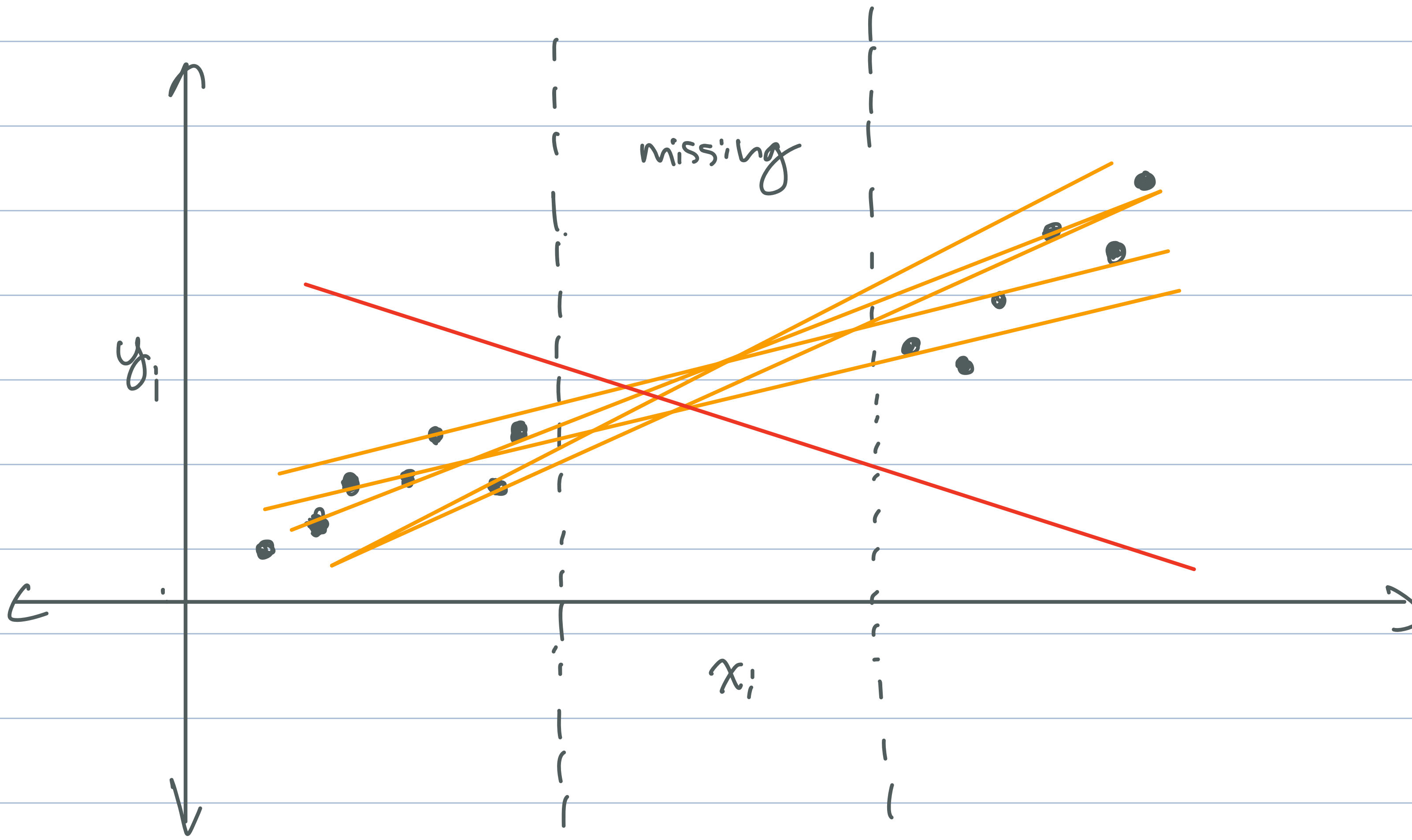
Motivating example

We have a data set of (x, y) pairs with **censoring** for a certain range of x. (In other words, we never get to observed any (x,y) pair for x in the censored region.)



This might happen if x = time and y = temperature and our thermometer stops recording for some period due to power outage.

This might also happen if x = biomarker, and y = health status, and the only patients that come to the hospital are those with extreme values of the biomarker.

We want to **predict** the relationship between x and y in the censored region.
We cannot do this without **making assumptions = modeling.**
There is a clear upward trend in the plot, and let's also say that we have expert background knowledge about (x,y) which supports a monotonic upward trend.

To start we could fit a regression line. However, there are multiple linear trends that all seem plausible and consistent with the data we observe. The orange lines are all consistent with our data, while the red line is not.

In making predictions about the censored region, we would ideally **characterize our uncertainty** about which the true trend. A natural way to accomplish this is by averaging our prediction according to the **posterior probability of each trend:**

Posterior predictive distribution:

$$p(y_{n+1} \mid x_{n+1}, X, Y, -) = \mathbb{E}_{\beta \mid X, Y}\left[p(y_{n+1} \mid x_{n+1}, \beta, -) \right]$$

$$= \int_{-\infty}^{\infty} \underbrace{p(\beta \mid X, Y)}_{\text{Posterior}} \underbrace{p(y_{n+1} \mid x_{n+1}, \beta)}_{\text{Likelihood of new data point}} \, d\beta$$

This is the motivation for a Bayesian approach to regression. We want to specify our assumptions via a **model** which then gives us a principled way to encode our uncertainty via the **posterior** and **posterior predictive** distributions.

Bayesian linear regression

$$y_i = x_i^T \beta + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$Y \equiv \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, \quad X \equiv \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}$$

$$\varepsilon \equiv \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \in \mathbb{R}^n, \quad \varepsilon \sim \mathcal{N}(0, I\sigma^2)$$

multivariate Gaussian

Vectorized form: $\underset{n\times 1}{Y} = \underset{n\times p}{X} \underset{p \times 1}{\beta} + \underset{n\times 1}{\varepsilon}$

Equivalently: $Y \sim \mathcal{N}(X\beta, \Sigma)$

"$I\sigma^2$

Prior:

$$\underset{p\times 1}{\beta} \sim \mathcal{N}(\underset{p\times 1}{m_0}, \underset{p\times p}{L_0^{-1}})$$

Prior mean

inverse of prior precision matrix

Hyperparameters (fixed / known): σ^2, m_0, L_0

Model: $P(Y, \beta \mid X, \sigma^2, m_0, L_0)$

Multivariate Gaussian PDF

$$\mathcal{N}(y ; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)\right)$$

$$\underset{y}{\propto} \exp\left(-\frac{1}{2} y^T \Sigma^{-1} y + y^T \Sigma^{-1} \mu\right)$$

Kernel of the multivariate Gaussian

Posterior calculation:

"and everything else"

$$p(\beta \mid y, x, -)$$

$$\propto_\beta \mathcal{N}(\beta; m_0, L_0^{-1}) \prod_{i=1}^{n} \mathcal{N}(y_i; x_i^T\beta, \sigma^2)$$

$$\propto_\beta \exp\left(-\tfrac{1}{2}\beta^T L_0 \beta + \beta^T L_0 m_0\right) \prod_{i=1}^{n} \exp\left(-\tfrac{1}{2}(y_i - x_i^T\beta)^T \Sigma^{-1} (y_i - x_i^T\beta)\right)$$

$$\propto_\beta \exp\left(-\tfrac{1}{2}\beta^T L_0 \beta + \beta^T L_0 m_0\right) \prod_{i=1}^{n} \exp\left(-\tfrac{1}{2}\beta^T x_i \Sigma^{-1} x_i^T \beta + \beta^T x_i \Sigma^{-1} y_i\right)$$

$$\propto_\beta \exp\Bigg(-\tfrac{1}{2}\beta^T \underbrace{\left[L_0 + \sum_{i=1}^{n} x_i \Sigma^{-1} x_i^T\right]}_{= L_n} \beta + \beta^T \underbrace{\left[L_0 m_0 + \sum_{i=1}^{n} x_i \Sigma^{-1} y_i\right]}_{= L_n m_n}\Bigg)$$

Kernel of a multivariate Gaussian!

posterior precision matrix

$$m_n = L_n^{-1} L_n m_n = \left(L_0 + \sum_{i=1}^{n} x_i \Sigma^{-1} x_i^T\right)^{-1}\left(L_0 m_0 + \sum_{i=1}^{n} x_i \Sigma^{-1} y_i\right)$$

posterior mean

$$p(\beta \mid X, Y, -) = \mathcal{N}(\beta; m_n, L_n^{-1})$$

This is an example of **Gaussian-Gaussian conjugacy.**

Since $\Sigma = I\sigma^2$ in this case:

$$m_n = \left(L_0 + \tfrac{1}{\sigma^2} X^T X\right)^{-1}\left(L_0 m_0 + \tfrac{1}{\sigma^2} X^T y\right)$$

where $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$

The MAP (maximum a posteriori) solution is:

$$\hat{\beta}^{MAP} = m_n = \left(L_0 + \frac{1}{\sigma^2} X^T X\right)^{-1} \left(L_0 m_0 + \frac{1}{\sigma^2} X^T y\right)$$

As the variance of the prior goes to infinity, the MAP becomes the

$$\lim_{L_0 \to 0} \hat{\beta}^{MAP} = (X^T X)^{-1} X^T y = \hat{\beta}^{MLE}$$

A Gaussian prior with (near) infinite variance is a flat or "non-

The MAP solution generalizes the ridge regression

$$\text{if } m_0 = 0 \text{ and } L_0 = \frac{\alpha}{\sigma^2}:$$

$$\hat{\beta}^{MAP} = \left(\frac{\alpha}{\sigma^2} + \frac{1}{\sigma^2} X^T X\right)^{-1} \left(\frac{1}{\sigma^2} X^T y\right) = (\alpha + X^T X)^{-1} X^T y = \hat{\beta}_{ridge}$$

Why did this happen? Priors as regularizers:

$$\hat{\beta}^{MAP} = \underset{\beta}{\operatorname{argmax}}\ P(\beta \mid y, X, -)$$

$$= \underset{\beta}{\operatorname{argmin}}\ -\log P(y \mid \beta, X) - \log P(\beta).$$

$m_0 = 0$:

$$= \underset{\beta}{\operatorname{argmin}}\ \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \frac{L_0}{2} \sum_j \beta_j^2$$

$L_0 = \frac{\alpha}{\sigma^2}$:

$$= \underset{\beta}{\operatorname{argmin}}\ \underbrace{\sum_{i=1}^{n} (y_i - x_i^T \beta)^2}_{\text{squared loss}} + \underbrace{\alpha \sum_{j=1}^{p} \beta_j^2}_{\ell_2\text{-regularization w/ strength } \alpha} = \hat{\beta}_{ridge}$$

(this is the loss function for ridge

To summarize: the negative log posterior under a Gaussian-Gaussian model has the same minimizer as the ridge regression loss (i.e. L2-regularized least squares).

This is not an "accident". Priors act as regularizers. The L2 regularization term comes from the Gaussian prior, and a different prior would lead to a different regularizers (e.g., L1 regularization corresponds to a Laplace prior).

Posterior predictive calculation:

$$p(y_{n+1} \mid x_{n+1}, X, Y, -) = \mathbb{E}_{\beta \mid X, Y}\left[p(y_{n+1} \mid x_{n+1}, \beta, -)\right]$$

$$= \int \underbrace{p(\beta \mid X, Y)}_{\text{Posterior}} \; \underbrace{p(y_{n+1} \mid x_{n+1}, \beta)}_{\text{Likelihood of new data point}} \, d\beta$$

$$= \int \mathcal{N}(\beta; m_n, L_n^{-1}) \, \mathcal{N}(y_{n+1}; x_{n+1}^T \beta, \sigma^2) \, d\beta$$

A fact about Gaussians (which we won't prove) is the following:

If $y = Az + b$ and $z \sim \mathcal{N}(\cdots)$, then $y \sim \mathcal{N}(\cdots)$, for A and b fixed and known.

Applying this to our setting, we have that:

$$y_{n+1} = x_{n+1}^T \beta + \varepsilon_{n+1} \quad \text{and} \quad \beta \sim \mathcal{N}(m_n, L_n^{-1})$$

Therefore: $y_{n+1} \sim \mathcal{N}(\mu_n, S_n)$

We know that P(y | ...) is Gaussian, we can then solve for its mean and covariance:

$$\mu_n = \mathbb{E}_{\beta, \varepsilon}\left[x_{n+1}^T \beta + \varepsilon_{n+1}\right] = x_{n+1}^T \mathbb{E}_{\beta}[\beta] + \mathbb{E}_{\varepsilon_{n+1}}[\varepsilon_{n+1}] = x_{n+1}^T m_n$$

$$S_n = \mathrm{Cov}_{\beta, \varepsilon}\left[x_{n+1}^T \beta + \varepsilon_{n+1}\right] = x_{n+1} \mathrm{Cov}(\beta) x_{n+1}^T + \mathrm{Cov}(\varepsilon_{n+1})$$

$$= x_{n+1} L_n^{-1} x_{n+1}^T + \sigma^2 I$$

Marginal likelihood and prior predictive distribution:

What if there are no training data points? $(n=0)$

$$n=0 \rightarrow \begin{array}{l} X^T X = \sum_{i=1}^{n} x_i x_i^T = 0 \\ X^T y = \sum_{i=1}^{n} x_i y_i = 0 \end{array} \rightarrow \begin{array}{l} m_n = L_0^{-1}(L_0 m_0) = m_0 \\ L_n = L_0 \end{array}$$

The "posterior" parameters are just the **prior parameters** when n=0

The posterior predictive becomes the "**prior predictive**" when n=0:

$$P(y_{n+1} \mid x_{n+1}, y_{1:n}, x_{1:n})$$

$$= P(y_1 \mid x_1)$$

...which is just the **marginal likelihood** of a single data point

...which is Gaussian:

$$= \mathcal{N}(y_1 ; x_1^T m_0 , x_1 L_0 x_1^T)$$

The marginal likelihood as a sequence of posterior predictives:

$$P(y_{1:n} \mid x_{1:n}) = \prod_{i=1}^{n} \underbrace{P(y_i \mid x_i, y_{<i}, x_{<i})}_{\text{posterior predictive at } i}$$

$$= \prod_{i=1}^{n} \mathcal{N}(y_i ; x_i^T m_i , \sigma^2 + x_i L_i x_i^T)$$

$$L_i = L_0 + \frac{1}{\sigma^2} \sum_{k<i} x_k x_k^T$$

$$m_i = L_i^{-1}(L_0 m_0 + \frac{1}{\sigma^2} \sum_{k<i} x_i y_i)$$

Bayesian updating: for each data point, the new "prior" is the posterior given all the data up to that point.

An alternative way to derive the marginal likelihood uses the vectorized form of the likelihood for all n data points:

$$Y \sim X\beta + \varepsilon \;, \quad \beta \sim \mathcal{N}(m_0, L_0^{-1})$$

$$y \sim \mathcal{N}(\mu, S)$$

$$\mu = \mathbb{E}[X\beta + \varepsilon] = X m_0$$

$$S = \mathrm{Cov}(X\beta + \varepsilon) = X L_0^{-1} X^T + I\sigma^2$$

$$p(y \mid X, m_0, L_0^{-1}) = \mathcal{N}(y \,;\, X m_0, \, X L_0^{-1} X^T + I\sigma^2)$$

So the **marginal likelihood** can be expressed either as a product of univariate Gaussian distributions (each of which is a kind of posterior predictive) or equivalently as a single multivariate Gaussian.

Model evaluation via the prior/posterior predictive distribution

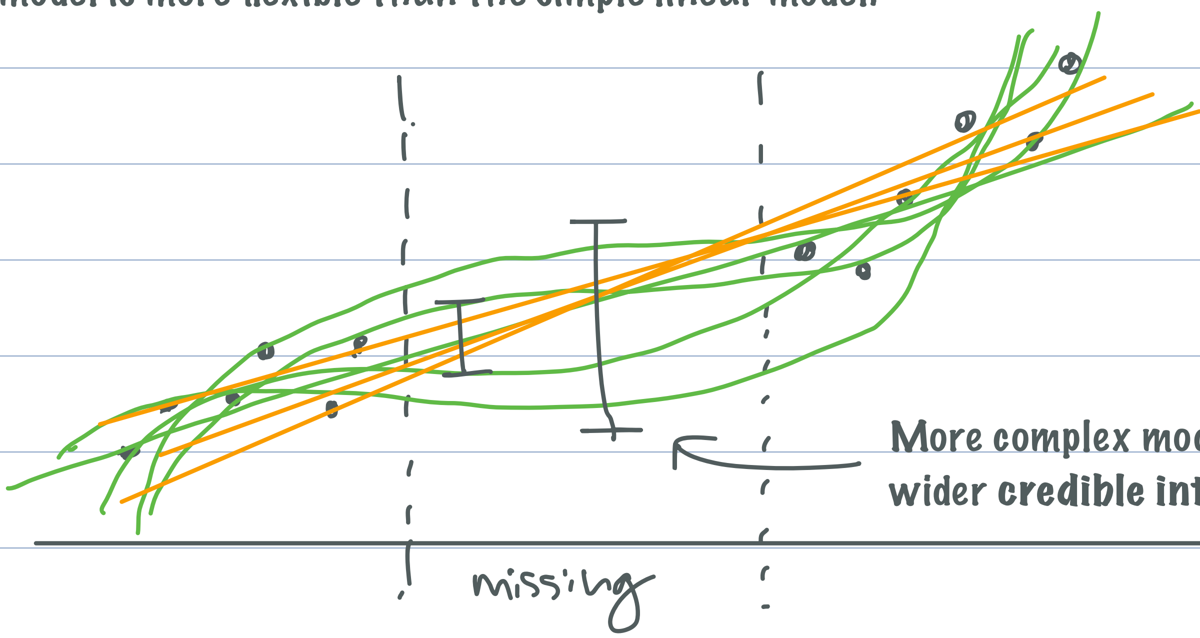
Consider a more complex model class for Bayesian linear regression:

$$P(y \mid x, \beta) = \mathcal{N}(y;\ \text{poly}_3(x)^T \beta,\ \sigma^2)$$

$$\text{e.g. if } x \in \mathbb{R} \rightarrow \text{poly}_3(x) = [x, x^2, x^3]$$

$$\hookrightarrow \text{poly}_3(x)^T \beta = \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

This model is more flexible than the simple linear model:

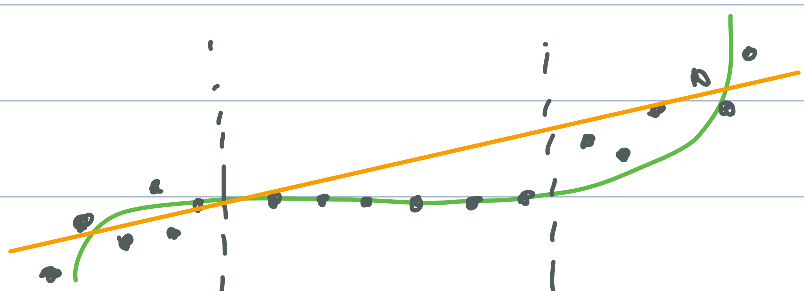


More complex model has wider credible interval

All of the green curves correspond to a possible values of beta in the cubic (more complex) model which give high-ish likelihood to the observed data points.

All of the orange curves correspond to possible values of beta in the linear (simple) model which give high-ish likelihood to the observed data points.

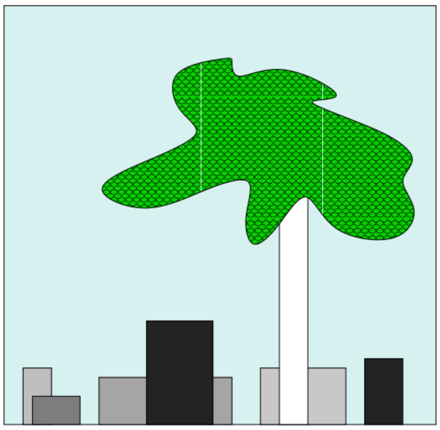
Certain models (i.e., curves) in the complex model class could fit the heldout data points much better than any plausible model in the linear class...



...however there are **many more** models in the complex model class that are consistent with the training data.

The principle of "Occam's razor" says we should take the **simplest** hypothesis that is consistent with the data. In this case a **model class is a hypothesis** and the **model evidence** (aka the marginal likelihood) accounts for simplicity.

See Chapter 28 of Mackay:



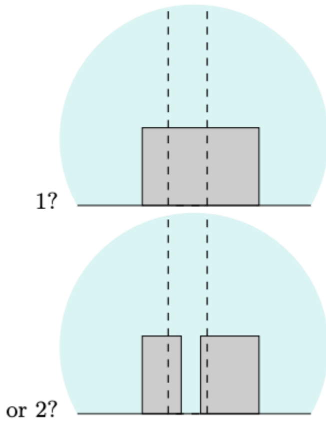
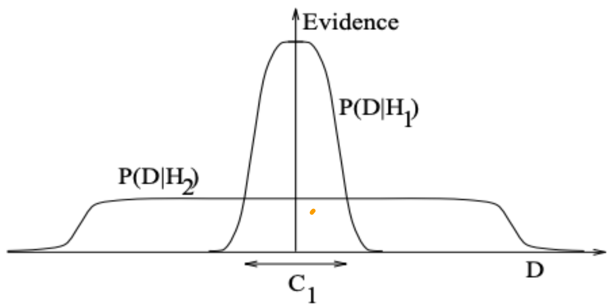


Figure 28.2. How many boxes are behind the tree?



Probabilistic models, in general:

modeling assumptions (including hyperparameters)

$$P(D, Z \mid H)$$

data latent

likelihood prior

Posterior

$$P(Z \mid D, H) = \frac{P(D \mid Z, H)\, P(Z \mid H)}{P(D \mid H)}$$

marginal likelihood
model evidence
normalizing constant
prior predictive distribution
...

When comparing two model classes (H1 vs H2), the **Bayes factor**, rewards simplicity:

$$BF = \frac{P(D \mid H_1)}{P(D \mid H_2)}$$

Similarly, we can compare **posterior** (rather than prior) **predictive** probabilities, if we create a train-test split of the data:

$$P(D^{test} \mid D^{train}, H_1) \quad vs. \quad P(D^{test} \mid D^{train}, H_2)$$

$$P(D^{test} \mid D^{train}, H)$$

$$= \int P(D^{test} \mid D^{train}, z, H)\, P(z \mid D^{train}, H)\, dz$$

Model selection via the prior/posterior predictive distribution:

Model evaluation and selection are naturally related. Just as it makes sense to use the marginal likelihood to evaluate, it also makes sense to use it to select.

A common example of this are **empirical** or **objective Bayesian** procedures for choosing the prior in a data-driven manner.

Marginal likelihood for the regression model above:

hyperparameter of the likelihood
$y_i \sim N(x_i^T\beta, \sigma^2)$

$$P(Y_{1:n} \mid X_{1:n}, \sigma^2, \underbrace{m_0, L_0})$$

hyperparameters of the prior
$\beta \sim N(m_0, L_0^{-1})$

For example, for fixed values of the other hyperparameters we could do **type-II maximum likelihood** to set the prior mean:

$$\hat{m}_0 = \underset{m_0}{\operatorname{argmax}}\; P(Y \mid X, \sigma^2, m_0, L_0)$$