Bayesian Linear Regression

June 10, 2021

Table of contents

Why Bayesian Linear Regression?

Why Bayes? (Linear Regression Version)

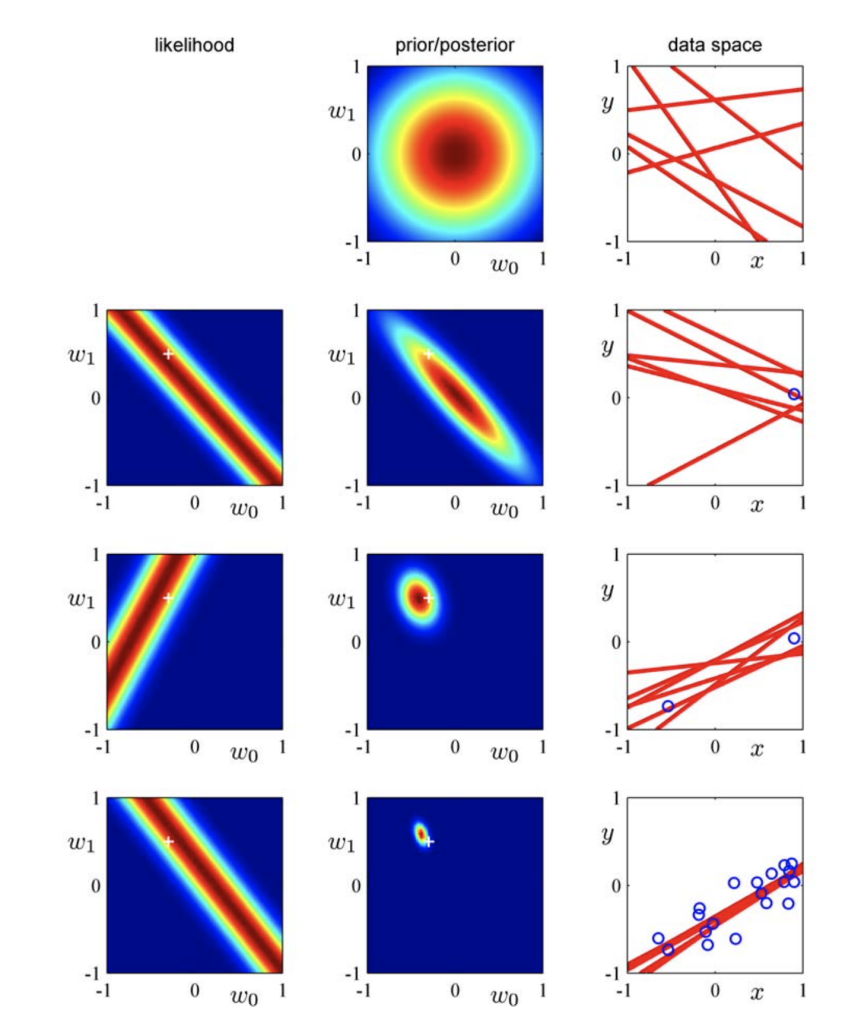
- Provides a *distribution* over regression lines
- Automatically supports model selection / complexity control.
- Easy access to nuanced inferential quantities.

Why Bayesian Linear Regression?

Distribution over regression lines

Bayesian Linear Regression

We learn a *distribution* over regression lines.



Sequential Bayesian learning for a simple linear model.

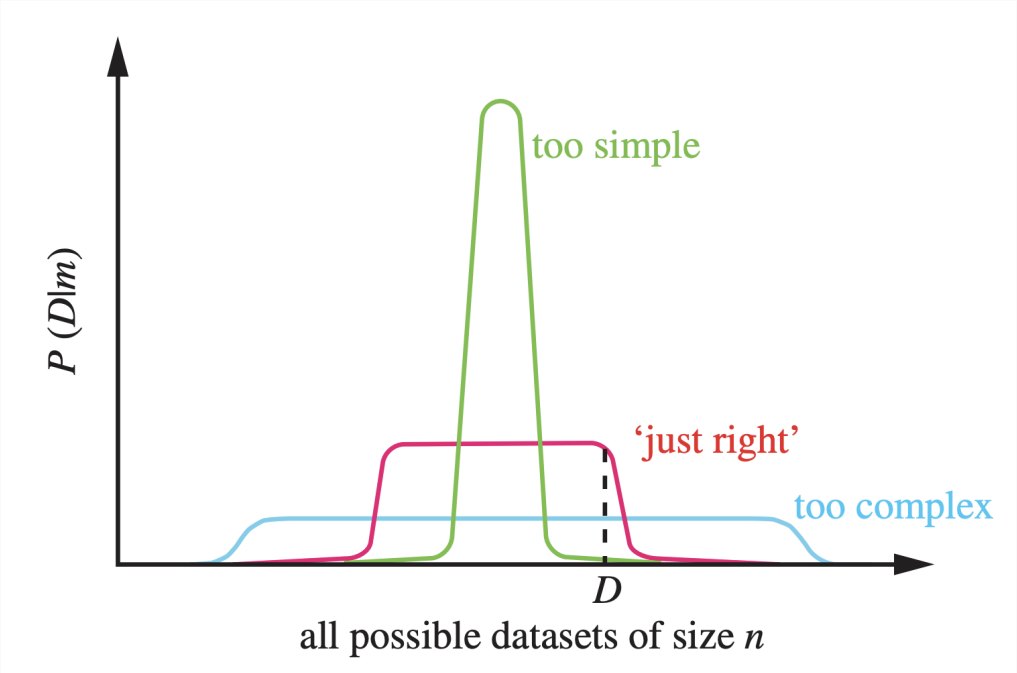
Image Credit: Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Why Bayesian Linear Regression?

Automatic model selection

Bayesian Occam's Razor

Remember this slide?

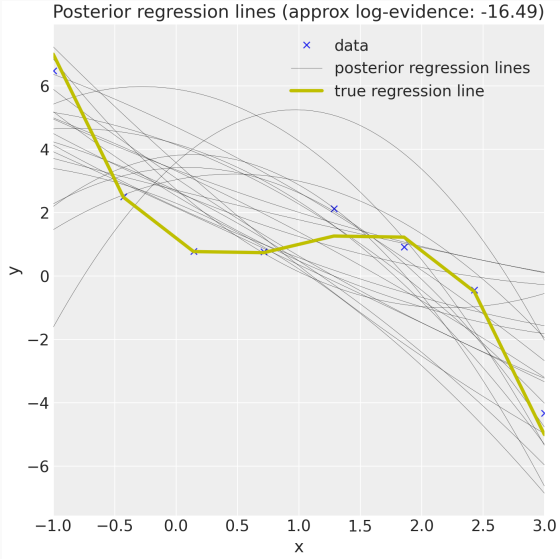


A *complex* model (shown in blue) spreads its mass over many more possible datasets, whereas a *simple* model (shown in green) concentrates its mass on a smaller fraction of possible data. Because probabilities have to sum to one, the complex model spreads its mass at the cost of not being able to model simple datasets as well as a simple model—this normalization is what results in an automatic Occam razor. Given any particular dataset, here indicated by the dotted line, we can use the marginal likelihood to reject both overly simple models, and overly complex models.

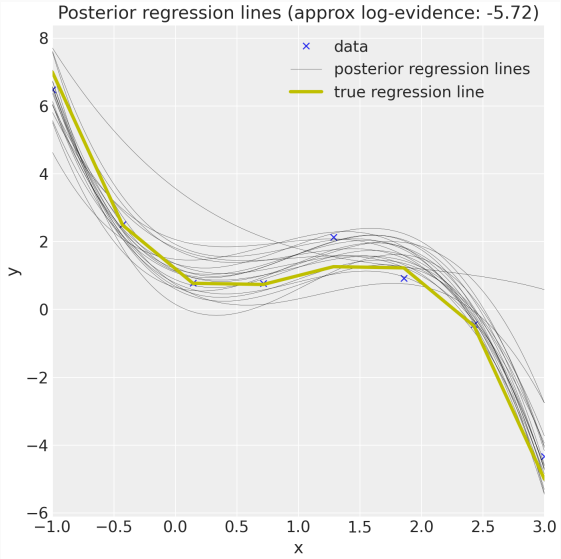
Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371(1984), 20110553.

Bayesian Occam's Razor

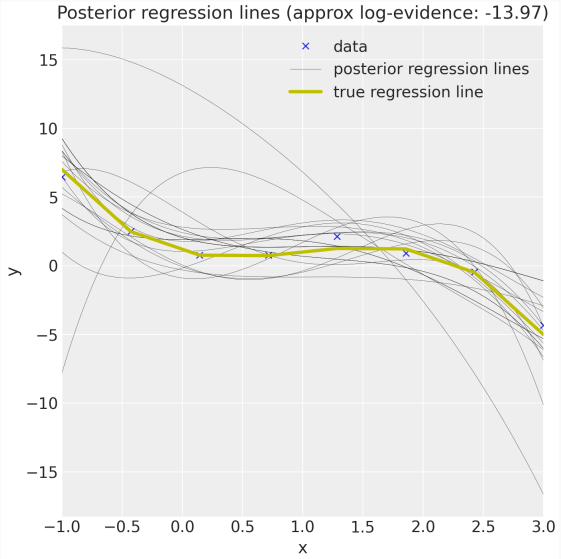
I generated $n = 8$ data points from a cubic distribution and used `NUTS` to fit Bayesian polynomials of various orders p.



Quadratic model ($p = 2$)



Cubic model ($p = 3$)

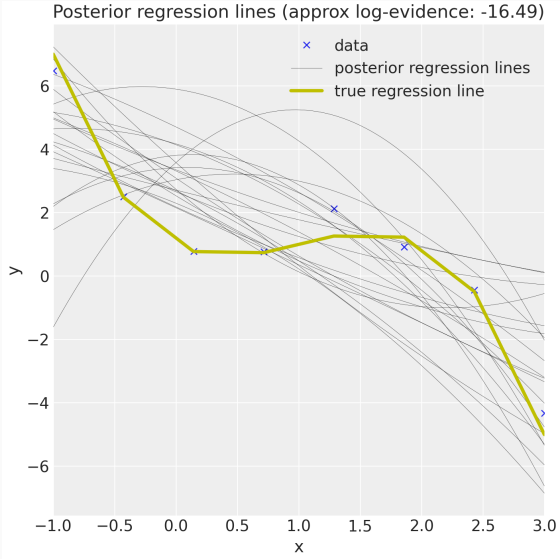


Quartic model ($p = 4$)

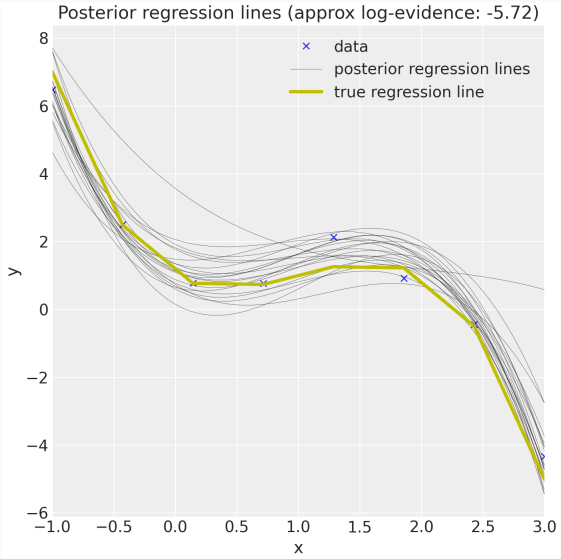
How does this align with the previous slide?

Bayesian Occam's Razor

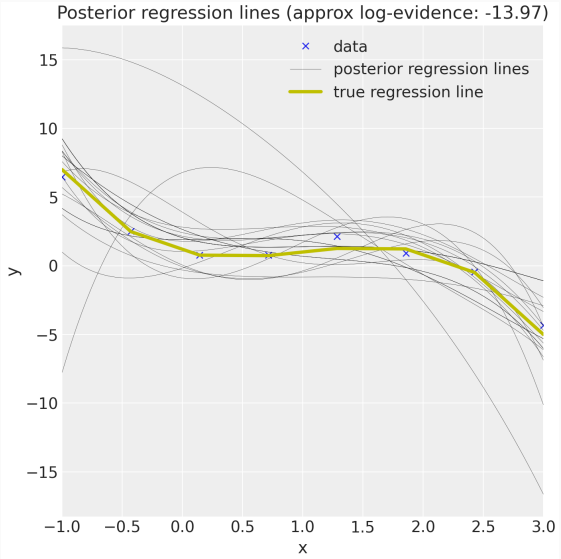
I generated $n = 8$ data points from a cubic distribution and used `NUTS` to fit Bayesian polynomials of various orders p.



Quadratic model ($p = 2$)



Cubic model ($p = 3$)



Quartic model ($p = 4$)

How does this align with the previous slide?

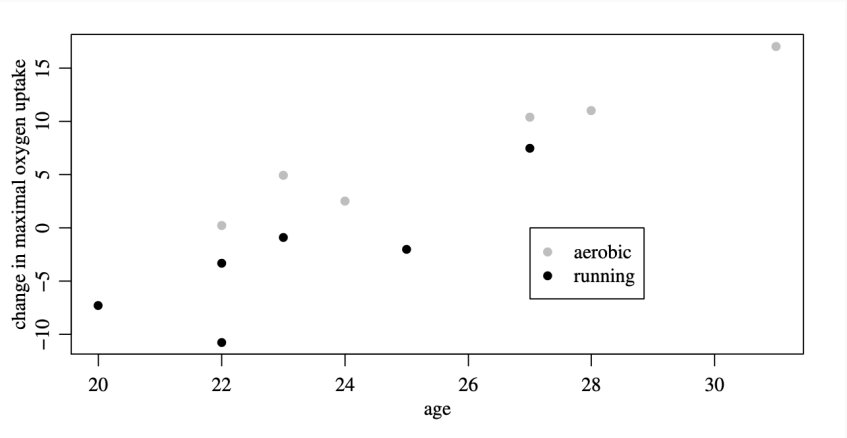
- Bayesian model selection works well here! The true (cubic) model has the highest evidence. The evidence is lower for models that are underfit (quadratic) or overfit (quartic).
- Posterior draws from the cubic model best match the true data generating process.
- Maximum likelihood doesn't do this. ML says: the higher the order, the *better* the fit.
- The ranking of models by evidence matches what would be expected from the previous slide.

Why Bayesian Linear Regression?

Easy access to nuanced inferential quantities

Exercise Program Dataset

(Hoff, A first course in Bayesian Statistical Methods)



Change in maximal O_2 uptake as a function of age and exercise program

Model

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i, \text{ where}$$

$x_{i,1} = 1$ for each subject i

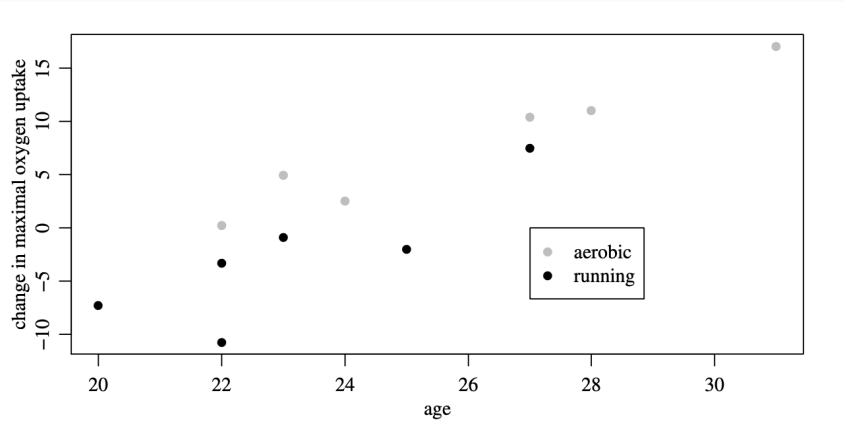
$x_{i,2} = 0$ if subject i is on the running program, 1 if on aerobic

$x_{i,3} =$ age of subject i

$x_{i,4} = x_{i,2} \times x_{i,3}$

Exercise Program Dataset

(Hoff, A first course in Bayesian Statistical Methods)



Change in maximal O_2 uptake as a function of age and exercise program

Model

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i, \text{ where}$$

$x_{i,1} = 1$ for each subject i

$x_{i,2} = 0$ if subject i is on the running program, 1 if on aerobic

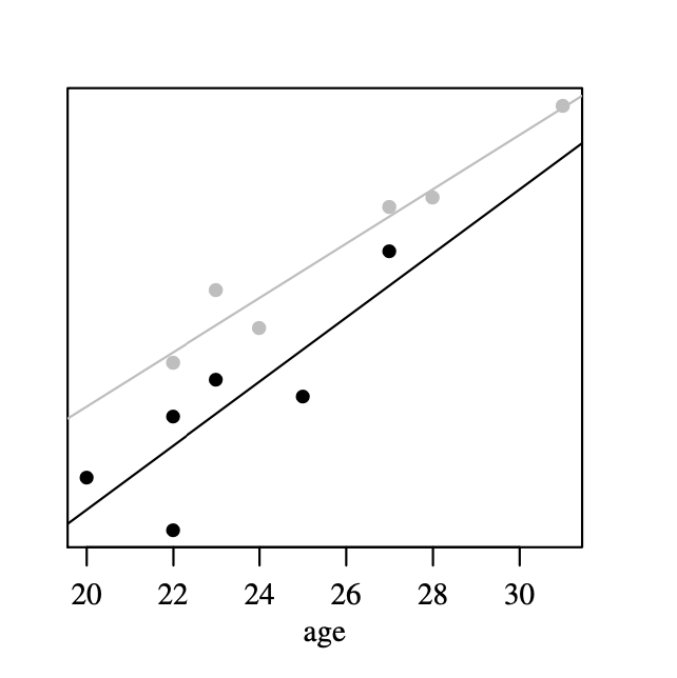
$x_{i,3} =$ age of subject i

$x_{i,4} = x_{i,2} \times x_{i,3}$

Under this model, the conditional expectations for Y are:

$$\mathbb{E}[Y|\boldsymbol{x}] = \beta_1 + \beta_3 \times \text{age if } x_1 = 0, \text{ and}$$
$$\mathbb{E}[Y|\boldsymbol{x}] = (\beta_1 + \beta_2) + (\beta_3 + \beta_4) \times \text{age if } x_1 = 1$$

Frequentist Inference



Maximum likelihood regression lines for the O_2 uptake data.

$$\widehat{\boldsymbol{\beta}}_{\mathsf{ML}} = (-51.29, 13.11, 2.09, -.32)^T$$
$$\mathsf{SE}(\widehat{\boldsymbol{\beta}}_{\mathsf{ML}}) = (12.25, 15.76, 0.53, 0.65)^T$$

Comparing the values of $\widehat{\boldsymbol{\beta}}_{\mathsf{ML}}$ to their standard errors suggests the evidence for differences between exercise programs is not very strong.

Bayesian Inference

Bayesian inference agrees with the ML estimate, showing only weak evidence of a difference between exercise programs.

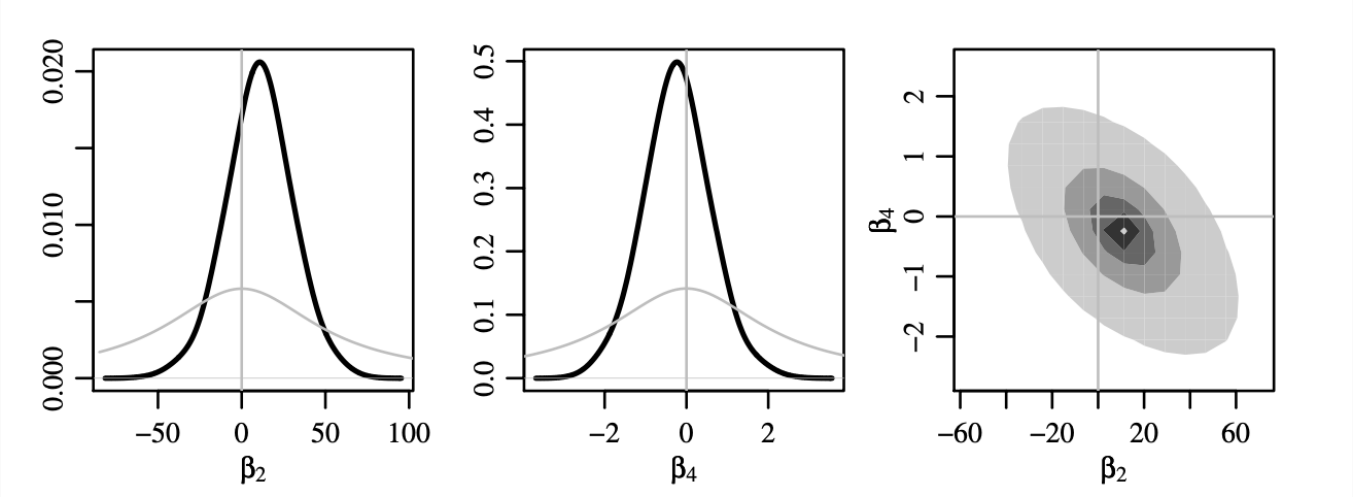


Figure 1: Posterior distributions for β_2 and β_4. The first two plots show the marginal prior distributions (grey) for comparison. The 95% posterior intervals for β_2 and β_4 both contain 0.

Bayesian Inference

But the parameters by themselves don't tell the whole story.

We can also look at the posterior distributions of $\beta_2 + \beta_4 x$ for each age x.

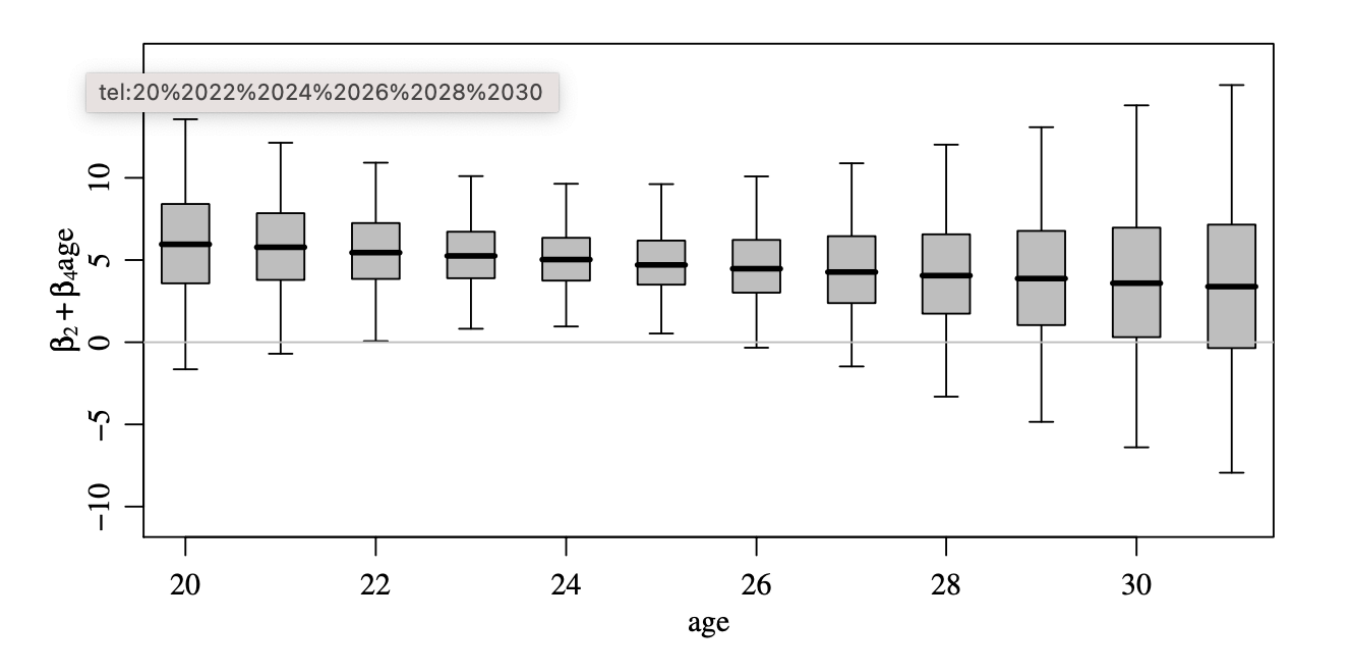


Figure 2: 95% confidence intervals for the difference in expected change scores between aerobics subjects and running subjects

This suggests reasonably strong evidence of a difference at young ages, and less evidence at older ones.

The model

The model

A basic Bayesian multiple linear regression model is given by

$$
\begin{aligned}
\boldsymbol{\beta} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0) \\
\sigma^2 &\sim \mathcal{IG}(\frac{\nu_0}{2}, \frac{\nu_0}{2}\sigma_0^2) \\
y_i \mid \boldsymbol{\beta}, \sigma^2 &\overset{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2)
\end{aligned}
\tag{1}
$$

Inference

Gibbs sampler

The model is conditionally conjugate. The complete conditionals are

$$\boldsymbol{\beta} \mid \boldsymbol{y}, \sigma^2 \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \tag{2}$$

where

$$\Sigma = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{X} \right)^{-1}$$

$$\boldsymbol{\mu} = \Sigma \left(\Sigma_0^{-1} \boldsymbol{\mu}_0 + \frac{1}{\sigma^2} \boldsymbol{X}^T \boldsymbol{y} \right)$$

and

$$\sigma^2 \mid \boldsymbol{y}, \boldsymbol{\beta} \sim \mathcal{IG} \left(\frac{\nu_0 + n}{2}, \frac{\nu_0 \sigma_0^2 + \text{SSR}(\boldsymbol{\beta})}{2} \right) \tag{3}$$

where

$$\text{SSR}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

is the sum of squared residuals.

Intepretation

Suppose the observation noise σ^2 is known. Then (2) is our posterior on $\boldsymbol{\beta}$. How can we interpret it?

Intepretation

Suppose the observation noise σ^2 is known. Then (2) is our posterior on $\boldsymbol{\beta}$. How can we interpret it?

- When the prior on the regression coefficients $\boldsymbol{\beta}$ is diffuse, the elements of the prior precision matrix Σ_0^{-1} will be small, and so the posterior mean satisfies

$$\boldsymbol{\mu} \approx (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} = \widehat{\boldsymbol{\beta}}_{\text{ML}}$$

  i.e. it approximately equals the standard least squares estimate.

- On the other hand, when the observation variance σ^2 is large, then the measurement precision is small, and the posterior mean satisfies

$$\boldsymbol{\mu} \approx \boldsymbol{\mu}_0$$

  i.e. it approximately equals the prior mean.

So the posterior mean is a tradeoff between the maximum likelihood estimate and the prior, with weights governed by prior variance vs. data variance

Proof (complete conditional for β)

First, we consider the likelihood $L(\boldsymbol{\beta}) := p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2)$, dropping terms proportional to $\boldsymbol{\beta}$.

$$
\begin{aligned}
p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) &\propto \exp\left\{ -\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right\} \\
&= \exp\left\{ -\frac{1}{2\sigma^2}(-2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}) \right\}
\end{aligned}
$$

Proof (complete conditional for β)

First, we consider the likelihood $L(\boldsymbol{\beta}) := p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2)$, dropping terms proportional to $\boldsymbol{\beta}$.

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) &\propto \exp\left\{ -\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right\} \\ &= \exp\left\{ -\frac{1}{2\sigma^2}(-2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}) \right\} \end{aligned}$$

Doing the same for the prior $p(\boldsymbol{\beta})$, we have

$$p(\boldsymbol{\beta}) \propto \exp\left\{ -\frac{1}{2}(-2\boldsymbol{\beta}^T\Sigma_0\boldsymbol{\mu}_0 + \boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}) \right\}$$

Thus, by Bayes rule

$$\begin{aligned} p(\boldsymbol{\beta} \mid \boldsymbol{y}, \sigma^2) &\propto p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) \times p(\boldsymbol{\beta}) \\ &\propto \exp\left\{ \boldsymbol{\beta}^T \underbrace{\left(\Sigma_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{y}\right)}_{:=\boldsymbol{b}} - \frac{1}{2}\boldsymbol{\beta}^T \underbrace{\left(\Sigma_0^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}\right)}_{:=\boldsymbol{A}} \boldsymbol{\beta} \right\} \end{aligned}$$

which reveals that the posterior is normal, along with the particular form of its parameter (covariance $\boldsymbol{A}^{-1}$ and mean $\boldsymbol{A}^{-1}\boldsymbol{b}$).

Application Notes

Basis functions

A linear regression model must be linear in its *regression weights*, $\boldsymbol{\beta}$, but need not be linear in its *covariates* $\boldsymbol{x}$.

More flexible models can be constructed by considering linear combinations of nonlinear functions of the covariates.

Examples

Basis functions

A linear regression model must be linear in its *regression weights*, $\boldsymbol{\beta}$, but need not be linear in its *covariates* $\boldsymbol{x}$.

More flexible models can be constructed by considering linear combinations of nonlinear functions of the covariates.

Examples

For example, for a single covariate, the linear predictor $\eta_i := \mathbb{E}[y_i \mid \boldsymbol{\beta}]$ could be given by

$$\eta_i = \beta_0 + \sum_{m=1}^{M-1} \beta_m \phi_m(x_i)$$

where

- polynomial basis functions take
$$\phi_m(x_i) = x_i^m$$
(and such a model is known as *polynomial regression*)
- gaussian basis functions take
$$\phi_m(x_i) = \exp\left\{ -\frac{1}{2} \frac{(x - \mu_m)^2}{s^2} \right\}$$

Radial basis function features

Example with radial basis function (RBF) features

$$\phi_m(x_i) = \exp\left\{ -\frac{1}{2}\frac{(x-\mu_m)^2}{s^2} \right\}$$

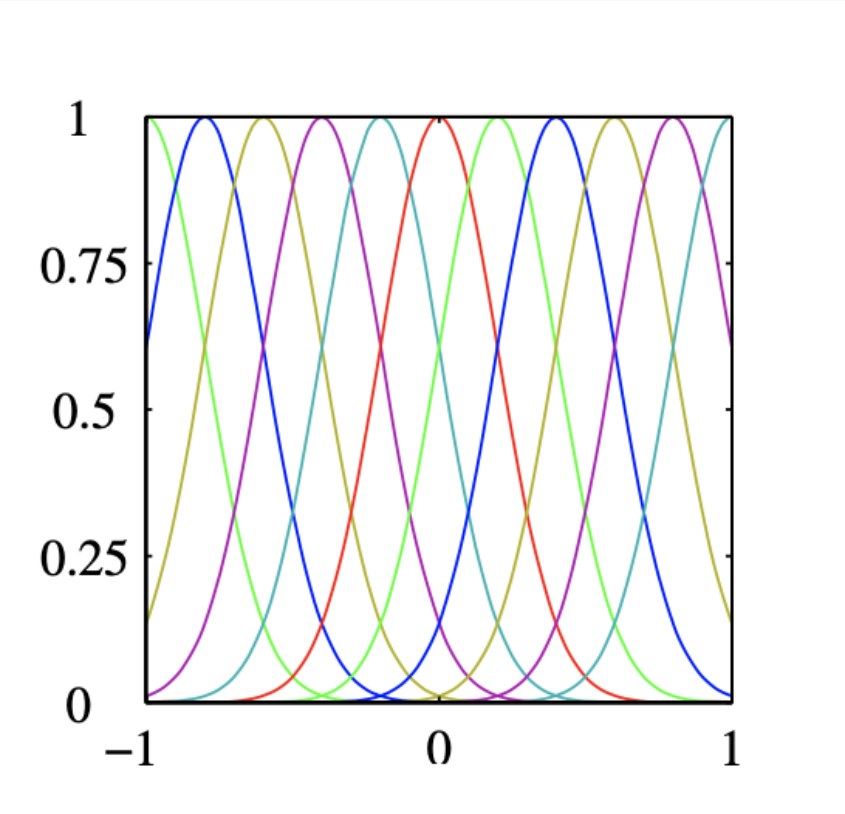


Image credit: Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Radial basis function features

Functions sampled from posterior

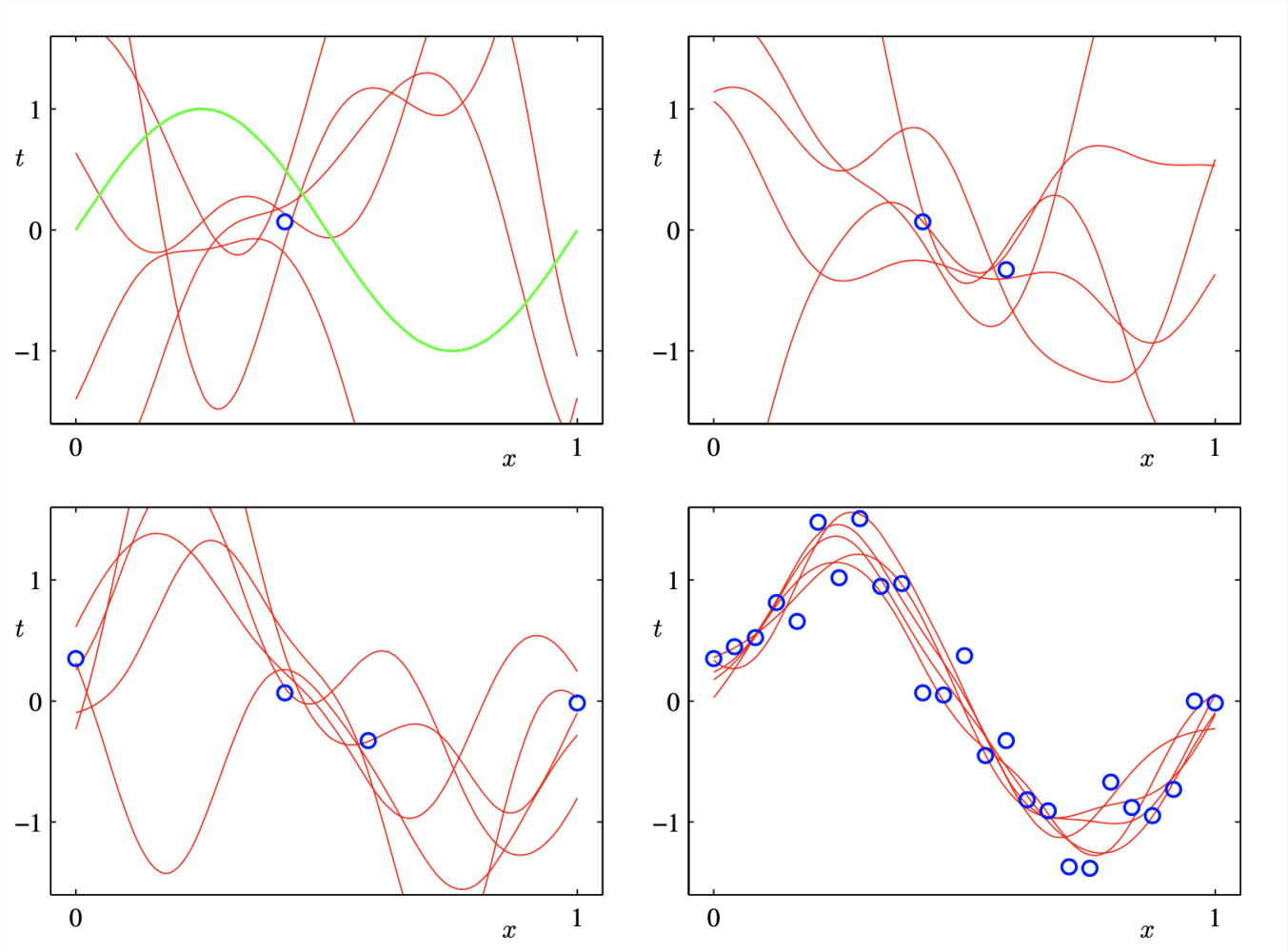


Image credit: Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Radial basis function features

Warning : The model becomes very confident outside of the basis function centers (clear from (2)).

This behavior is generally undesirable.

For flexible nonlinear regressions, it is more common to use *Gaussian process* models.

Prior specification

A common choice of prior is obtained by setting the hyperparameters as follows:

Prior on regression weights β

- Standardize the covariates $\boldsymbol{x}$.
- Set $\boldsymbol{\mu}_0 = 0, \Sigma_0 = \boldsymbol{I}$.

Prior on observation noise σ^2

The hyperparameters (σ_0^2, ν_0) can be interpreted as the sample variance and sample size of prior observations.

- Set $\nu_0 = 1$.
- Set σ_0^2 based on prior expectations.

Weakly informative priors

Idea: if prior is not going to represent real prior information about the parameters, make it as minimally informative as possible. Here are a couple of strategies:

Weakly informative priors

Idea: if prior is not going to represent real prior information about the parameters, make it as minimally informative as possible. Here are a couple of strategies:

The unit information prior

(Kass and Wasserman, 1995)

The precision of $\widehat{\boldsymbol{\beta}}_{\text{ML}}$ is $\text{Var}^{-1}[\widehat{\boldsymbol{\beta}}_{\text{ML}}] = (\boldsymbol{X}^T\boldsymbol{X})/\sigma^2$, and gives the amount of information in n observations. The unit information prior provides "one nth" as much information

- Set $\boldsymbol{\beta}_0 = \widehat{\boldsymbol{\beta}}_{\text{ML}}$.
- Set $\Sigma_0^{-1} = (\boldsymbol{X}^T\boldsymbol{X})/(n\sigma^2)$

Note: Strictly speaking, this is not a real *prior* distribution.

Weakly informative priors

Idea: if prior is not going to represent real prior information about the parameters, make it as minimally informative as possible. Here are a couple of strategies:

The unit information prior

(Kass and Wasserman, 1995)

The precision of $\widehat{\boldsymbol{\beta}}_{\text{ML}}$ is $\text{Var}^{-1}[\widehat{\boldsymbol{\beta}}_{\text{ML}}] = (\boldsymbol{X}^T\boldsymbol{X})/\sigma^2$, and gives the amount of information in n observations. The unit information prior provides "one nth" as much information

- Set $\boldsymbol{\beta}_0 = \widehat{\boldsymbol{\beta}}_{\text{ML}}$.
- Set $\Sigma_0^{-1} = (\boldsymbol{X}^T\boldsymbol{X})/(n\sigma^2)$

Note: Strictly speaking, this is not a real *prior* distribution.

The g-prior

(Zellner, 1986)

Parameter estimation should be invariant to changes in scale of the regressors. If $x_3 =$ age in years and $\widetilde{x}_3 =$ age in months, then the posterior for $12 \times \widetilde{\beta}_3$ should be the same as the posterior for β_3.

This condition is met if

- Set $\boldsymbol{\beta}_0 = 0$.
- Set $\Sigma_0^{-1} = k(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ for some k.

The g-prior specifies k in terms of the error variance, $k = g\sigma^2$ for some g. Michael Jordan recommends setting $p(\sigma^2) \propto \frac{1}{\sigma^2}$ and $g \sim \mathcal{IG}(\frac{1}{2}, \frac{n}{2})$ to avoid statistical paradoxes.