Bayesian and Penalized Regression

Justin Silverman

Penn State University

02/02/2021

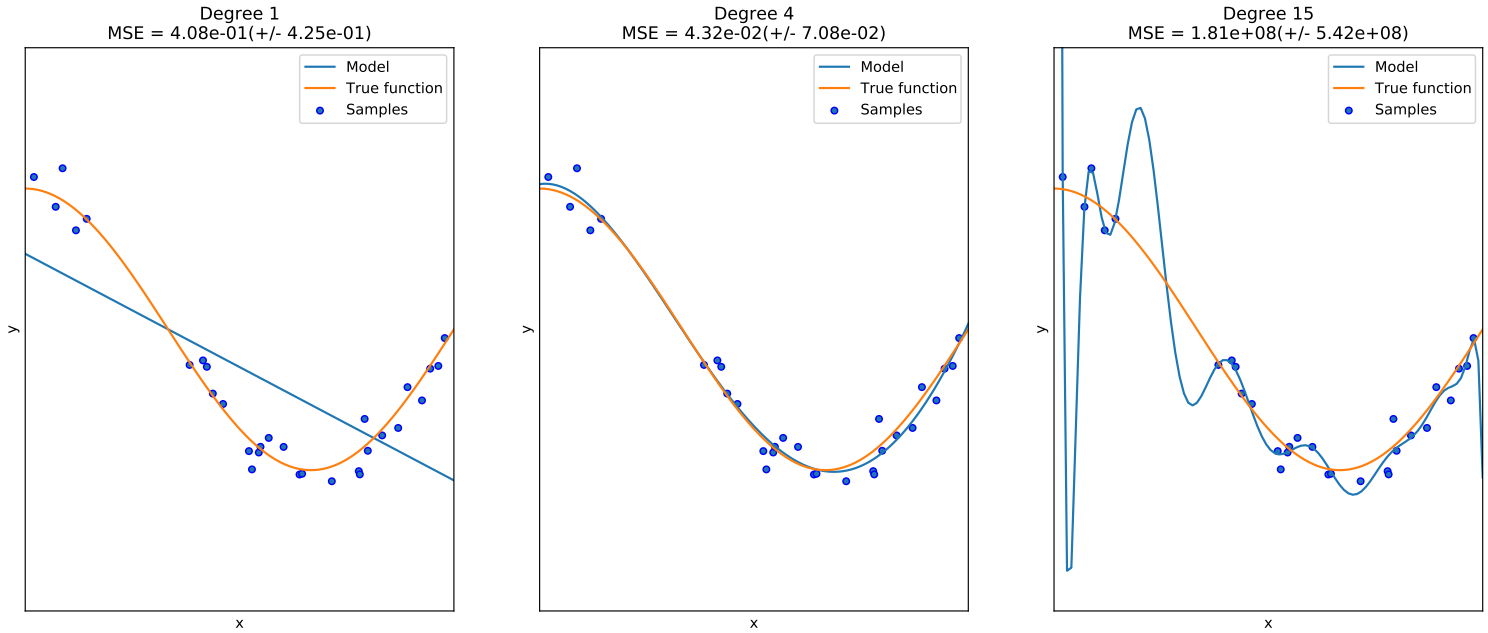
Table of Contents

Section 1

Motivation

Overfitting

So why don't we always just include a lot of higher order polynomial functions of x?



More likely to have overfitting if you have too many model parameters and not enough data $p >> n$.

$p >> n$

For a linear regression problem with p covariates and n samples we have a design matrix X that is a $n \times p$. **So what happens when we have more covariates than samples?** $p > n$**.**

Recall that the Maximum Likelihood estimator for linear regression (aka the OLS estimator) is given by:

$$\beta = (X^T X)^{-1} X^T y$$

$X^T X$ (a $p \times p$ covariance matrix) is rank deficient and there are multiple possible inverses $X^T X$, *i.e.*, $X^T X$ is a non-invertable matrix.

You could solve with Moore-Penrose pseudo-inverse but this just gives an arbitrary answer for β and not necessarily a good answer.[1]

[1] *i.e.*, inferential models may not make sense and predictive models may not generalize.

Sparsity and Variable Selection

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

Sparsity *e.g.*, I believe that most of the $\beta_i = 0$

Variable Selection *e.g.*, I want to know which of the x_i are useful.

Section 2

Penalized Regression (aka Shrinkage Models)

Penalized Regression (aka Shrinkage Models)

One solution is to assume that β is small. We can do this by adding a penalty (aka shrinkage) term to the loss-function.

Penalized Linear Regression

For $\lambda > 0$

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i + x_i^T \beta)^2 + g(\beta)$$

Vector p-norms

Think of a norm as the length of the vector. We have the standard Euclidean (aka ℓ_2 norm)

$$||x||_2 = \sqrt{\sum_i x_i^2}.$$

We also have the city-block norm (aka ℓ_1 norm)

$$||x||_1 = \sum_i |x_i|.$$

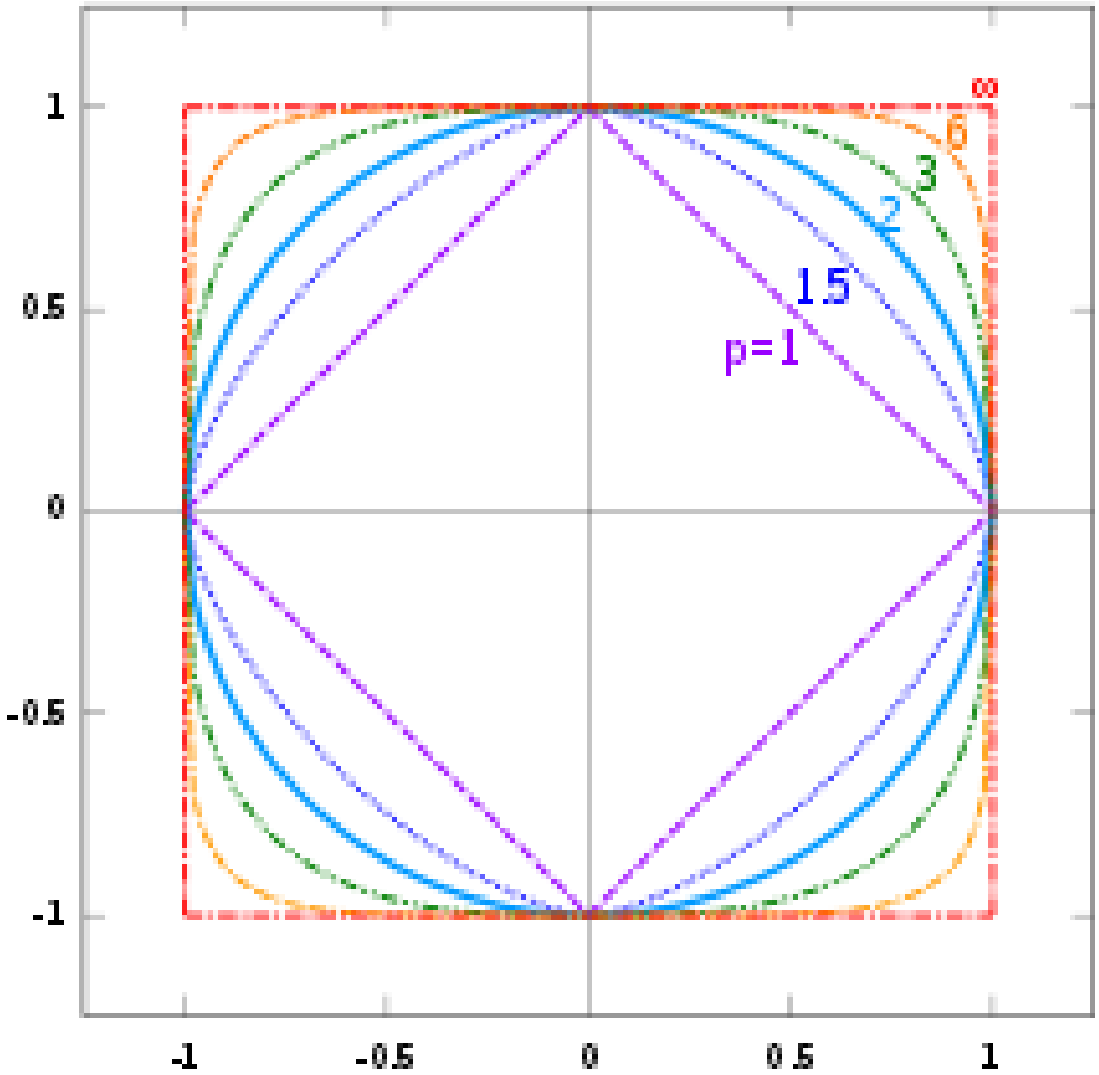
p-norms

For $p \geq 1$ the p-norm (also called the ℓ_p-norm) of a vector x is given by

$$||x||_p = \left(\sum_i |x_i|^p \right)^{1/p}$$

Visualizing p-norms

There are an infinite number of p-norms and we can visualize them by thinking about the areas of equal "length" in a 2D plane [2].



Intuition of p-norms and penalization

Back to penalized regression

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i + x_i^T \beta)^2 + g(\beta)$$

For $\lambda, \lambda_1, \lambda_2 > 0$

$g(\beta) = \lambda ||\beta||_2 = \sqrt{\sum_j \beta_j^2}$: **Ridge Regression** – Penalizes the length of β (aka we believe that overall the length of β is close to zero.)

$g(\beta) = \lambda ||\beta||_1 = \sum_j |\beta_j|$: **Lasso** – Penalizes non-zero elements in β (aka we believe that many elements β_i in β are zero)

$g(\beta) = \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2$: **Elastic Net** – A mixture of Ridge and Lasso

Solutions to penalized regression

Most penalty functions require numeric optimization algorithms to solve for $\hat{\beta}$. In the special case Ridge penalty for linear regression (aka Ridge regression) there is a closed form given by:

$$\hat{\beta} = (X^T X + \lambda n I)^{-1} X^T Y$$

This can be derived in a similar manner to how we derived the OLS/Maximum Likelihood estimator.

Solutions to penalized regression

Most penalty functions require numeric optimization algorithms to solve for $\hat{\beta}$. In the special case Ridge penalty for linear regression (aka Ridge regression) there is a closed form given by:

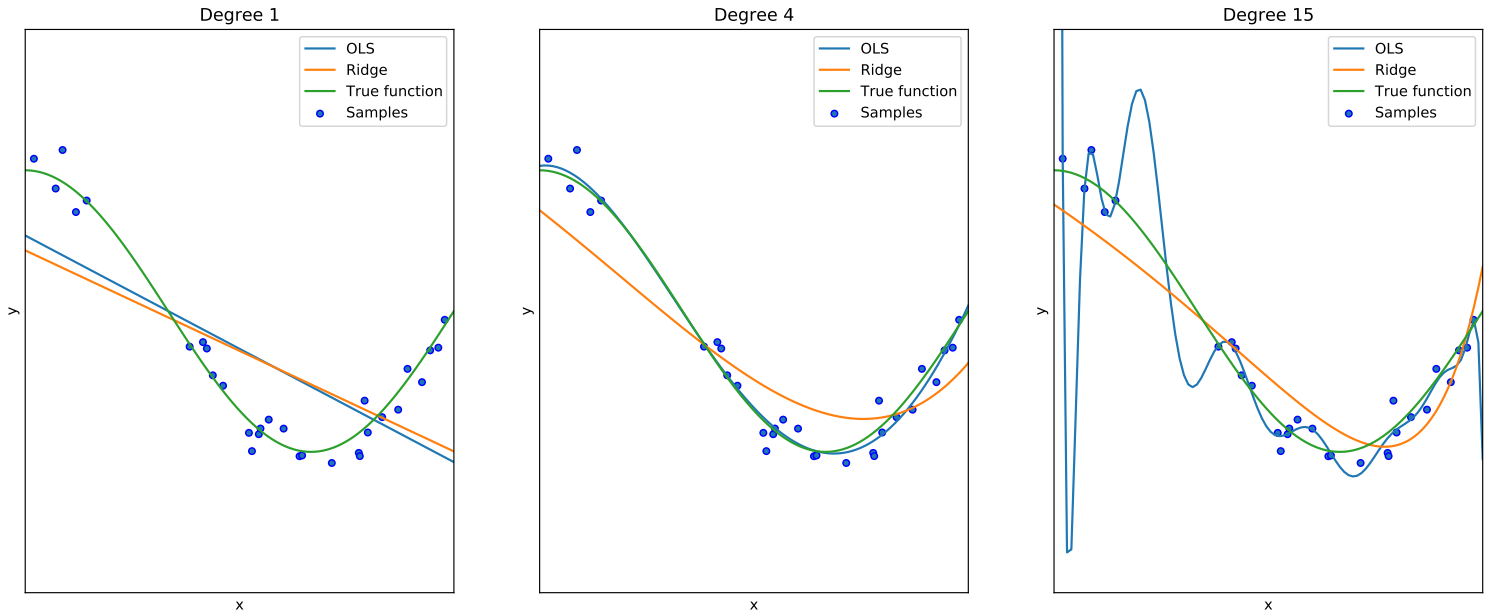
$$\hat{\beta} = (X^T X + \lambda n I)^{-1} X^T Y$$

This can be derived in a similar manner to how we derived the OLS/Maximum Likelihood estimator.

Notice how the additional term $\lambda n I$ now guarantees that the matrix $(X^T X + \lambda n I)^{-1}$ is invertable even in the $p >> n$ case.

As we will see, these penalized regression models can be seen as a special case of Bayesian regression. In particular, Ridge Regression is a special case of conjugate Bayesian linear regression.

Ridge Regression Example



Section 3

Introduction to Bayesian Statistics

Frequentist Statistics and Bayesian Statistics

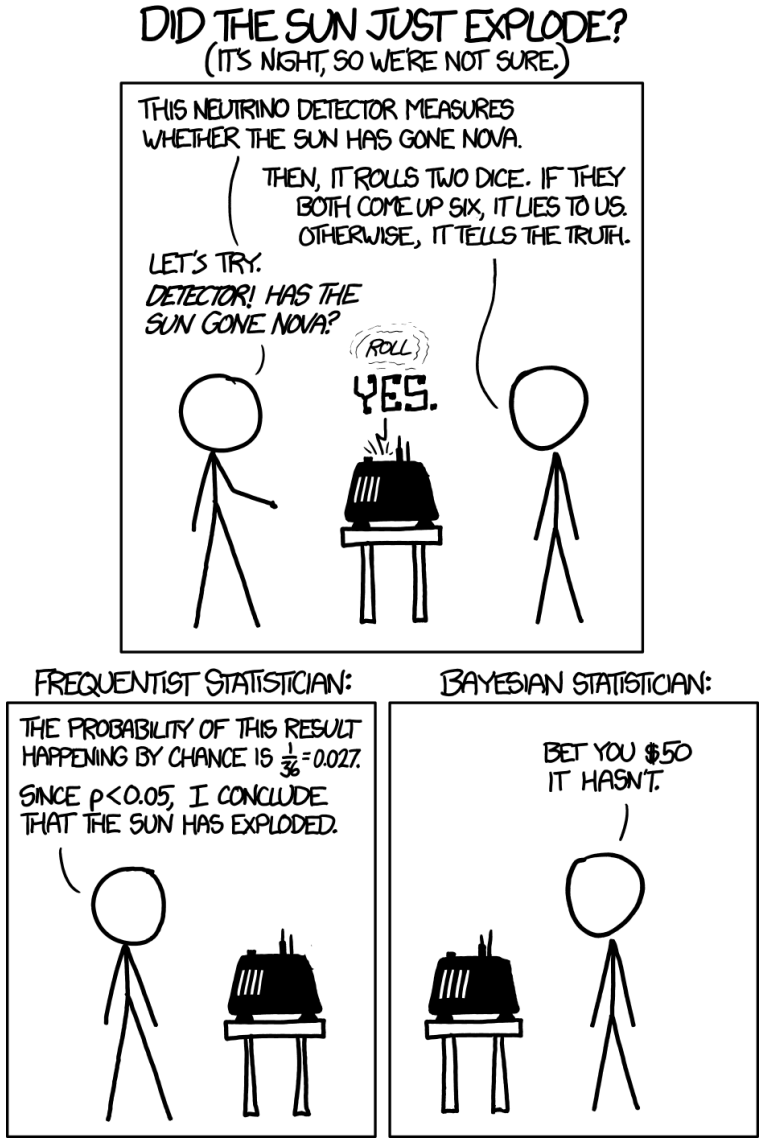
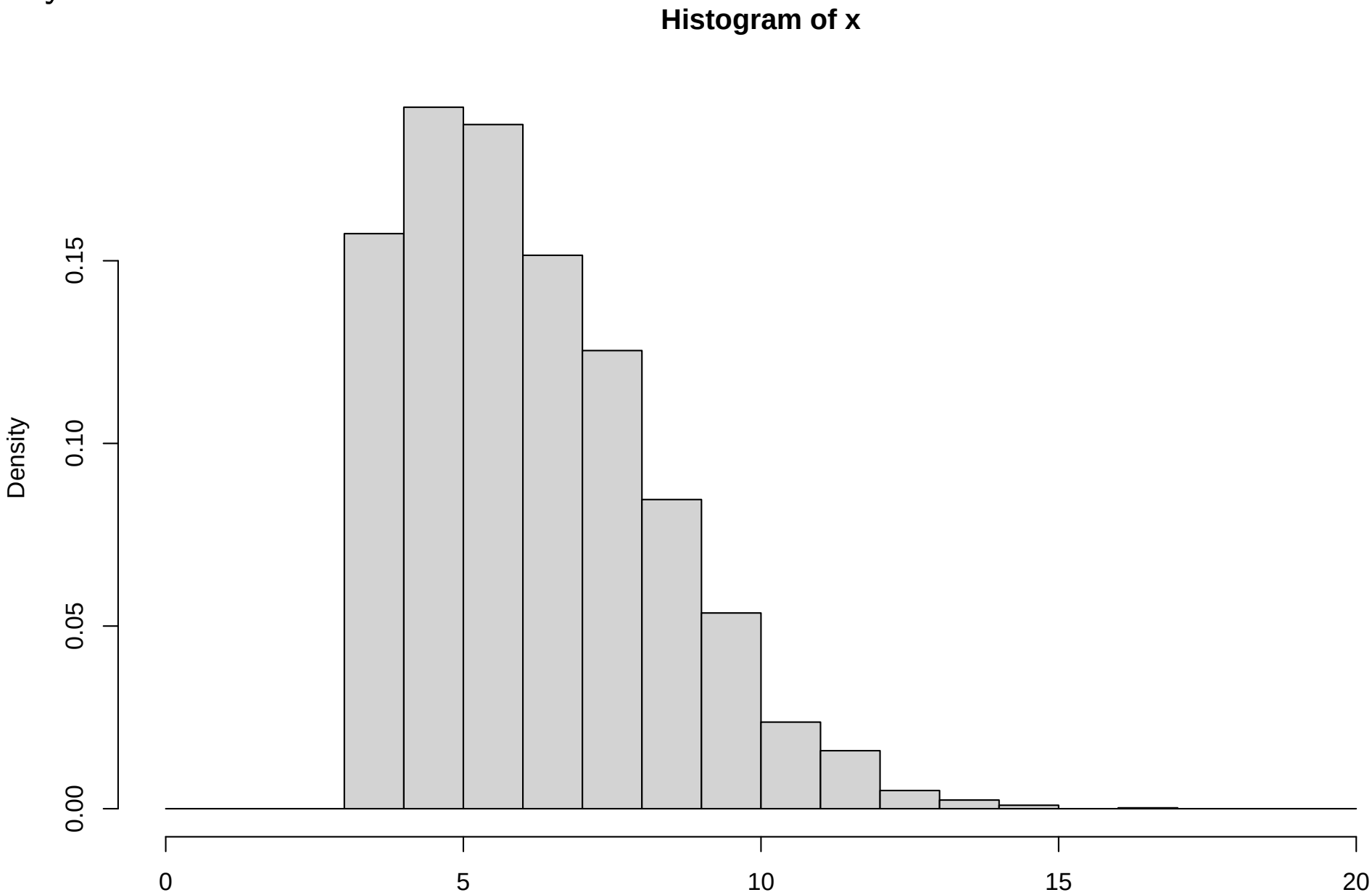


Figure 1: Image from xkcd

Using Probabilty to Express Beliefs

My belief in the maximum number of sodas I can drink in one hour:



Bayesian Updating

$$p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{\int p(D|\Theta)p(\Theta)d\Theta}$$

$p(\Theta)$ **The Prior** – Our belief regarding the "true value" of Θ **before** seeing the data

$p(D|\Theta)$ **The Likelihood** – How likely is our data given a particular value of Θ

$p(D) = \int p(D|\Theta)p(\Theta)d\Theta$ **The Marginal Likelihood** (aka model evidence) How likely is the data integrated over our prior beliefs

$p(\Theta|D)$ The Posterior – Our belief regarding "true value" of Θ **after** seeing the data

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Bayesian Updating

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

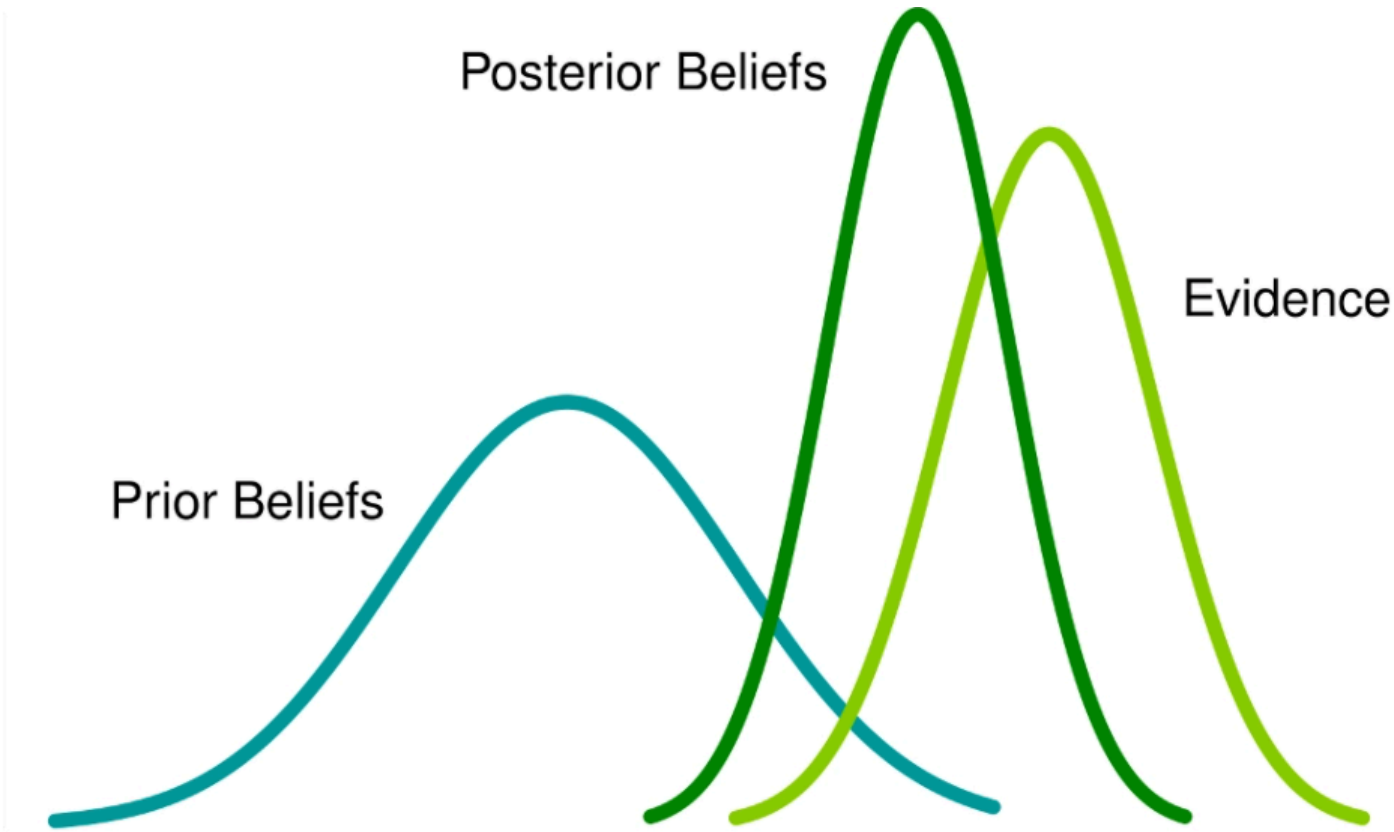


Figure 2: Image from NSS on analyticsvidhya.com

An Example

$$y \sim N(\beta x, \sigma^2)$$
$$\beta \sim N(\mu, \tau^2)$$

Crutial Points

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

1. Bayesian statistics can be computationally challenging because $\int p(D|\Theta)p(\Theta)d\Theta$ is difficult to calculate.
2. The two things you need to do to specify a Bayesian model is the prior and the likelihood.
3. Bayesian statistics allows you do perform analyses while incorporating prior beliefs
   - This is particularly powerful when you have small amounts of data (*e.g.*, when $p >> n$)
   - The more data you have, the larger / "stronger" the likelihood and the more the data can overcome potential bias introduced by the prior.

Conjugate Priors

There are certain pairings of likelihoods and priors that are special in that they induce a posterior distribution with the same functional form as the prior. For a given likelihood, we call these priors **Conjugate Priors**.

Conjugate Priors

There are certain pairings of likelihoods and priors that are special in that they induce a posterior distribution with the same functional form as the prior. For a given likelihood, we call these priors **Conjugate Priors**.

Normal-Normal Example

For a set of *iid* observations $(x_1, \ldots, x_n)$ with likelihood

$$x_i \sim N(\mu, \sigma^2)$$

And a prior

$$\mu \sim N(\alpha, \nu^2)$$

We have a posterior that is still a normal distribution and the following parameterization:

$$\mu | x_1, \ldots, x_n \sim N\left(\frac{1}{\frac{1}{\nu^2} + \frac{n}{\sigma^2}} \left(\frac{\alpha}{\nu^2} + \frac{\sum_i x_i}{\sigma^2} \right), \frac{1}{\frac{1}{\nu^2} + \frac{n}{\sigma^2}} \right)$$

Common Conjugate Likelihood-Prior Pairings

Likelihood	Prior
Discrete Likelihoods	—
Poisson	Gamma
Binomial	Beta
Multinomial	Dirichlet
Continuous Likelihoods	—
Normal (mean parameter)	Normal
Normal (variance parameter)	Inverse Gamma
Multivariate Normal (mean parameter)	Multivariate Normal
Multivariate Normal (variance parameter)	Inverse Wishart
Gamma	Gamma

Model Inference

Except when using conjugate priors, typically we cannot find a closed form solution for $p(\Theta|D)$ and so we instead focus on drawing samples from $p(\Theta|D)$ using various computational methods

Asymptotically Exact Methods

- Markov Chain Monte Carlo
  - e.g., Gibbs Sampling, Metropolis Hastings, Metropolis-within-Gibbs, Hamiltonian Monte Carlo
- Importance Sampling

Approximate Methods

- Laplace Approximation
- Variational Methods

MAP Estimation

MAP = Maximum a Posteriori

In some cases it becomes too difficult to obtain samples from $p(\Theta|D)$ then we sometimes just settle for taking a point estimate of Θ as

MAP Estimation

$$\begin{aligned}\hat{\Theta} &= \underset{\Theta}{\operatorname{argmax}} p(\Theta|D) \\ &= \underset{\Theta}{\operatorname{argmax}} p(D|\Theta)p(\Theta)\end{aligned}$$

This is easier because we don't need to know $p(D)$ to find the maximum.

Posterior Predictive Distribution

This is one method for making predictions using Bayesian models.

$$p(Y_*|Y) = \int p(Y_*|\Theta)p(\Theta|Y)d\Theta$$

$p(Y_*|\Theta)$ just refers to the likelihood, however, here we are now predicting new data Y_* rather than analyzing observed data Y.

Posterior Predictive Distribution

This is one method for making predictions using Bayesian models.

$$p(Y_*|Y) = \int p(Y_*|\Theta)p(\Theta|Y)d\Theta$$

$p(Y_*|\Theta)$ just refers to the likelihood, however, here we are now predicting new data Y_* rather than analyzing observed data Y.

This integral may seem daunting but its pretty easy to sample from $p(Y_*|Y)$ when you have samples $\{\Theta^{(1)}, \ldots \Theta^{(S)}\}$ from the posterior $p(\Theta|Y)$:

$$Y_*^{(s)} \sim p(Y_*|\Theta^{(s)}).$$

That is, you just simulate from the likelihood using posterior samples $\Theta^{(s)}$. We will see how to do this for linear regression later.

Section 4

Bayesian Linear Regression

Mean Unknown, Variance Known: Conjugate Priors

$$y_i \sim N(\beta^T x_i, \sigma^2)$$
$$\beta \sim N(0, \tau^2 I)$$

Mean Unknown, Variance Known: Conjugate Priors

$$y_i \sim N(\beta^T x_i, \sigma^2)$$
$$\beta \sim N(0, \tau^2 I)$$

After a bunch of algebra or using Multivariate Normal theory you can then obtain the posterior $\beta | Y, X, \sigma^2 \sim N(\mu, \Sigma)$ and with

$$\Sigma = (\tau^{-2} I + \sigma^{-2} X^T X)^{-1}$$
$$\mu = \sigma^{-2} \Sigma X^T Y$$

Mean Unknown, Variance Known: Conjugate Priors

$$y_i \sim N(\beta^T x_i, \sigma^2)$$
$$\beta \sim N(0, \tau^2 I)$$

After a bunch of algebra or using Multivariate Normal theory you can then obtain the posterior $\beta | Y, X, \sigma^2 \sim N(\mu, \Sigma)$ and with

$$\Sigma = (\tau^{-2} I + \sigma^{-2} X^T X)^{-1}$$
$$\mu = \sigma^{-2} \Sigma X^T Y$$

As the normal distribution is a symmetric function the MAP estimate is simply the mean (μ). Note the similarity of this to the Ridge regression estimator – they are the same (with $n\lambda = \sigma^2/\tau^2$).

That is, ridge regression is equivalent to the MAP estimate of Bayesian Conjugate Linear Regression.

Link between Bayesian Linear Regression and Ridge Regression

The MAP estimate for the following model

$$y_i \sim N(\beta^T x_i, \sigma^2)$$
$$\beta \sim N(0, \tau^2 I)$$

is identical to

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i + x_i^T \beta)^2 + \lambda ||\beta||_2$$

with $n\lambda = \sigma^2/\tau^2$.

Posterior Predictive Distribution

Given posterior samples for β (e.g., $\{\beta^{(i)}, \ldots, \beta^{(S)}\}$) we now want to predict y_* given covariates x_*.

We can obtain simulations of potential values of y_* given our posterior uncertainty in β using

$$y_*^{(s)} \sim N([\beta^{(s)}]^T x_*, \sigma^2)$$

Unlike Ridge Regression where we can only make point predictions (e.g., a single estimated y_*) here we can make probabilistic predictions in the range of possible values y_* could take given our prior believes and our observed data.

The value of probabilistic predictions

Mean Unknown, Variance Unknown: Conjugate Priors

$$y_i \sim N(\beta^T x_i, \sigma^2)$$
$$\beta \sim N(0, \sigma^2 I)$$
$$\sigma^2 \sim \text{Inverse-Gamma}(a, b)$$

This has a closed form and the solution is given on wikipedia: https://en.wikipedia.org/wiki/Bayesian_linear_regression

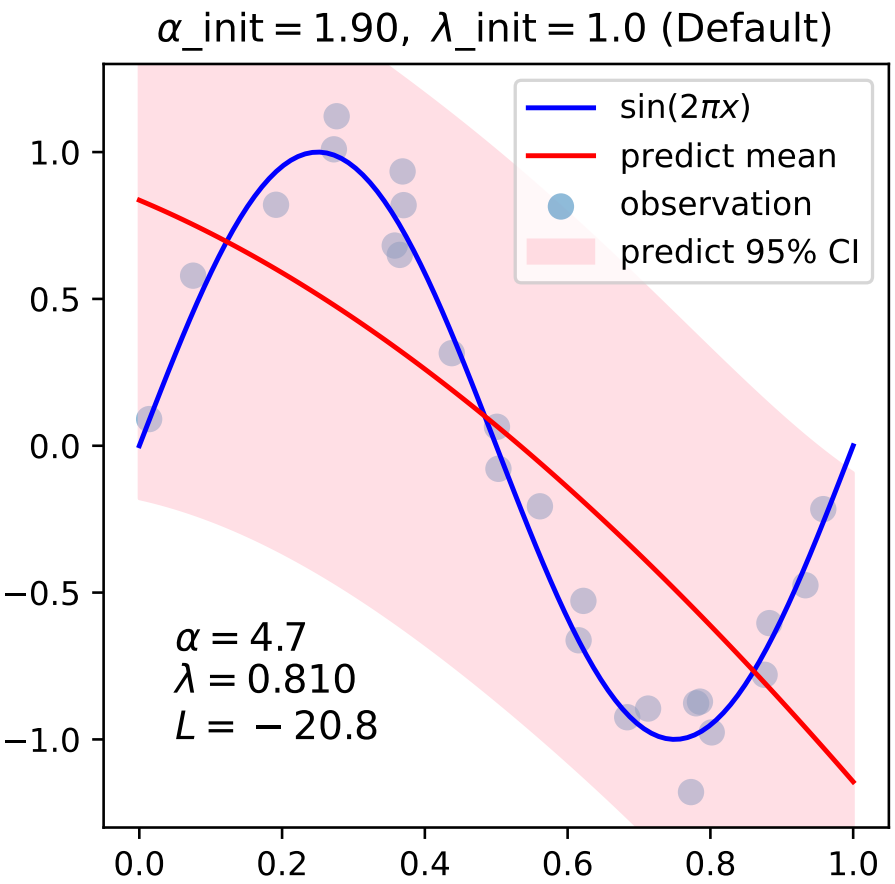
Note the requirement that y_i and β share the variance term σ^2

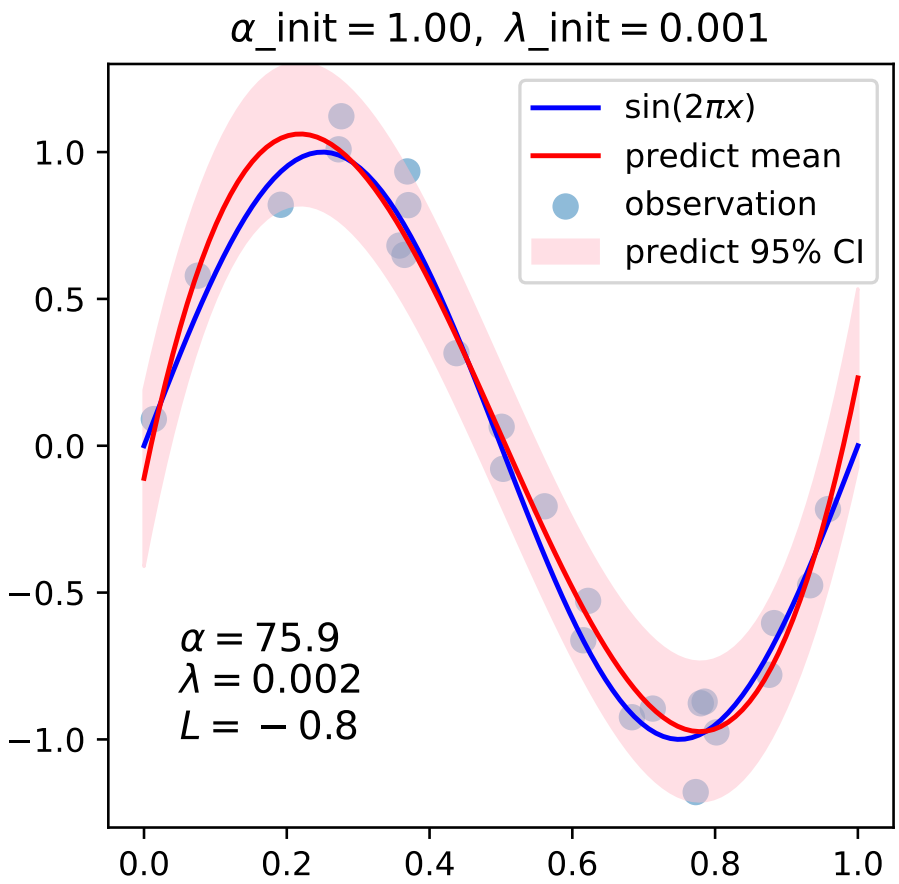
Mean Unknown, Variance Unknown: Semi-Conjugate Priors

Here we allow y_i and β to have different variances but this is no longer fully conjugate and requires MCMC methods to sample from the posterior.

$$
\begin{aligned}
y_i &\sim N(\beta^T x_i, \sigma^2) \\
\beta &\sim N(0, \Lambda) \\
\sigma^2 &\sim \text{Inverse-Gamma}(a, b)
\end{aligned}
$$

Bayesian Regression Example





Subsection 1

Bayesian Linear Regression Variants

Automatic Relevance Determination (ARD)

$$y_i \sim N(\beta^T x_i, \sigma^2)$$
$$\beta \sim N(0, \text{diag}(\lambda_1, \ldots, \lambda_p))\lambda_i \qquad \sim p(\phi)$$

Robust Regression

$$y_i \sim \mathrm{t}(\upsilon, \beta^T x_i, \sigma^2)$$
$$\beta \sim N(0, \tau^2 I)$$

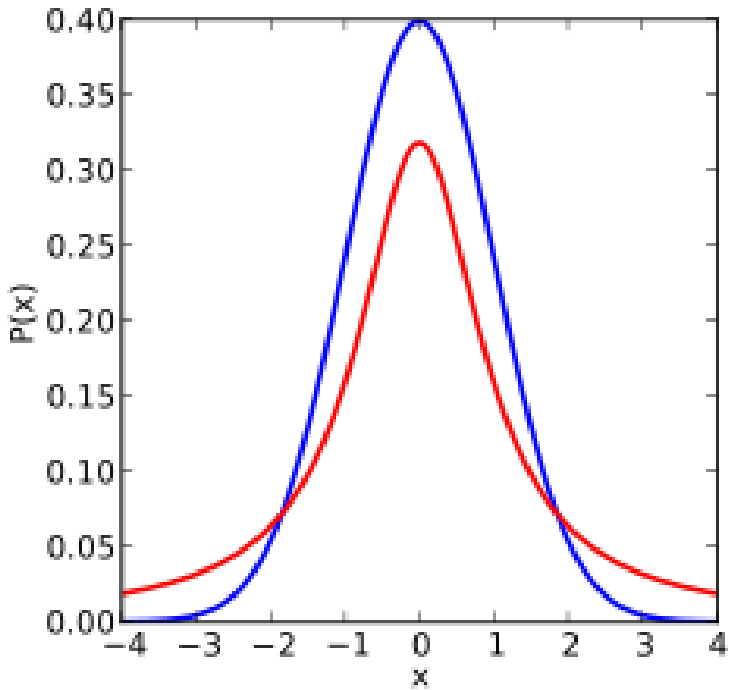


Figure 3: t distribution (red) versus the normal distribution(blue), Image from Wikipedia

Sparse Priors

For $\alpha \in [0, 1]$:

$$y_i \sim N(\beta^T x_i, \sigma^2)$$
$$\beta_j \sim \alpha N(0, \tau^2) + (1 - \alpha)\delta_0$$

where δ_0 refers to the delta function at $\beta_j = 0$.

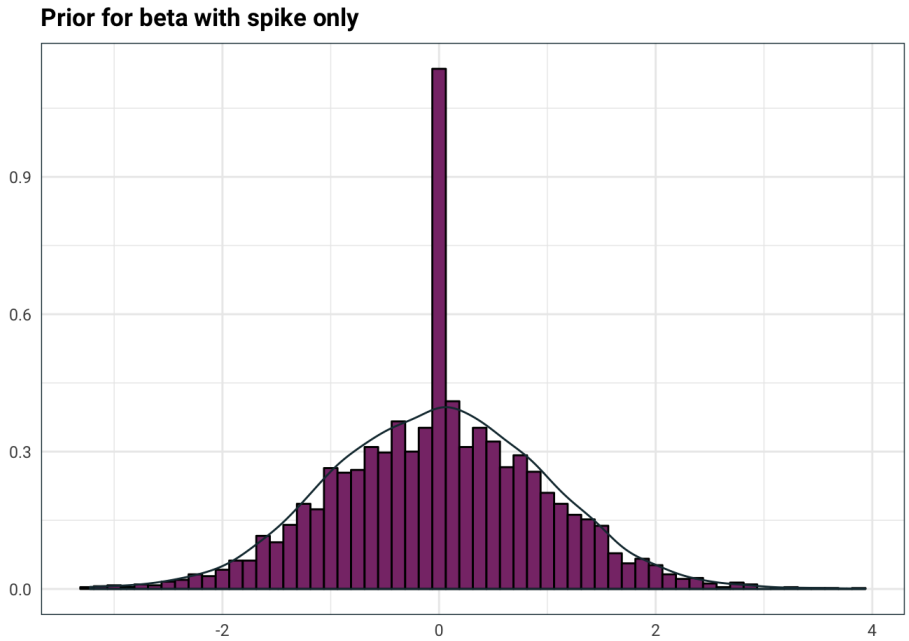


Figure 4: Spike and Slab Prior, Image from Bati Sengul

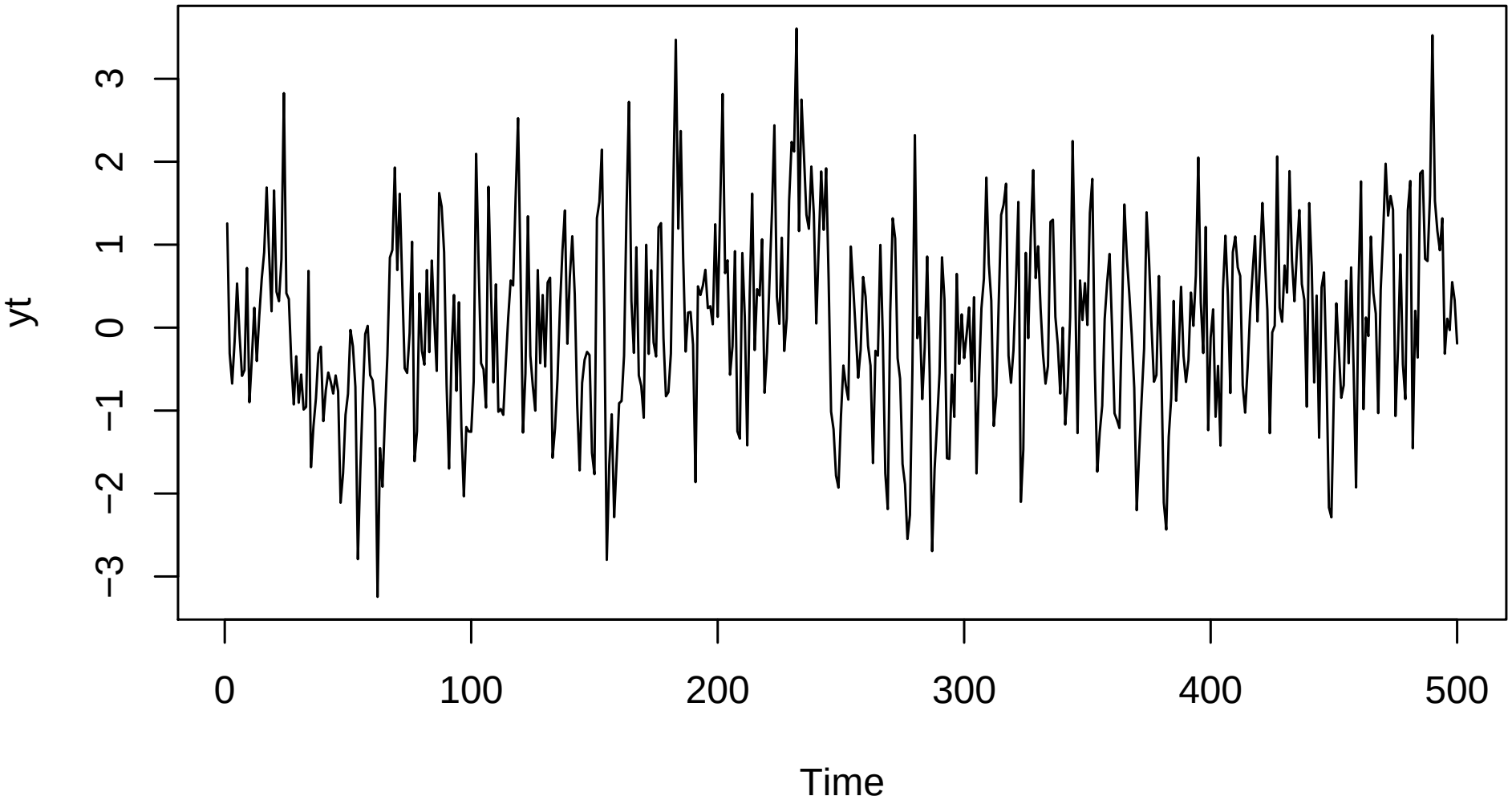
Heteroskedastic Models

$$y_i \sim N(\beta^T x_i, [\sigma z_i]^2)$$
$$\beta \sim N(0, \tau^2 I)$$
$$\sigma \sim N(0, \gamma^2 I)$$

Autoregressive Models

Let $y_1, \ldots, y_T$ be a time-series.

$$y_t \sim N(\beta y_{t-1}, \sigma^2)$$
$$\beta \sim N(0, \tau^2)$$



More complicated "Linear" Models

$$y_i \sim N(\beta_i^T x_i, \sigma^2)$$
$$\beta_i \sim N(\mu^T z_i, \tau^2 I)$$
$$\mu_i \sim N(0, \gamma^2 I)$$

Section 5

Practical Concerns

Practical Concerns

- Center and Scale Your Data

Practical Concerns

- Center and Scale Your Data
- Pick your priors wisely - if you put zero probability on something the likelihood can't get over that.

Practical Concerns

- Center and Scale Your Data
- Pick your priors wisely - if you put zero probability on something the likelihood can't get over that.
- These are not magic bullets - they are assumptions.