Bayesian modeling and inference

Sinead Williamson
TA: Evan Ott

github.com/sinead/DS32019
sinead.williamson@mccombs.utexas.edu

About me

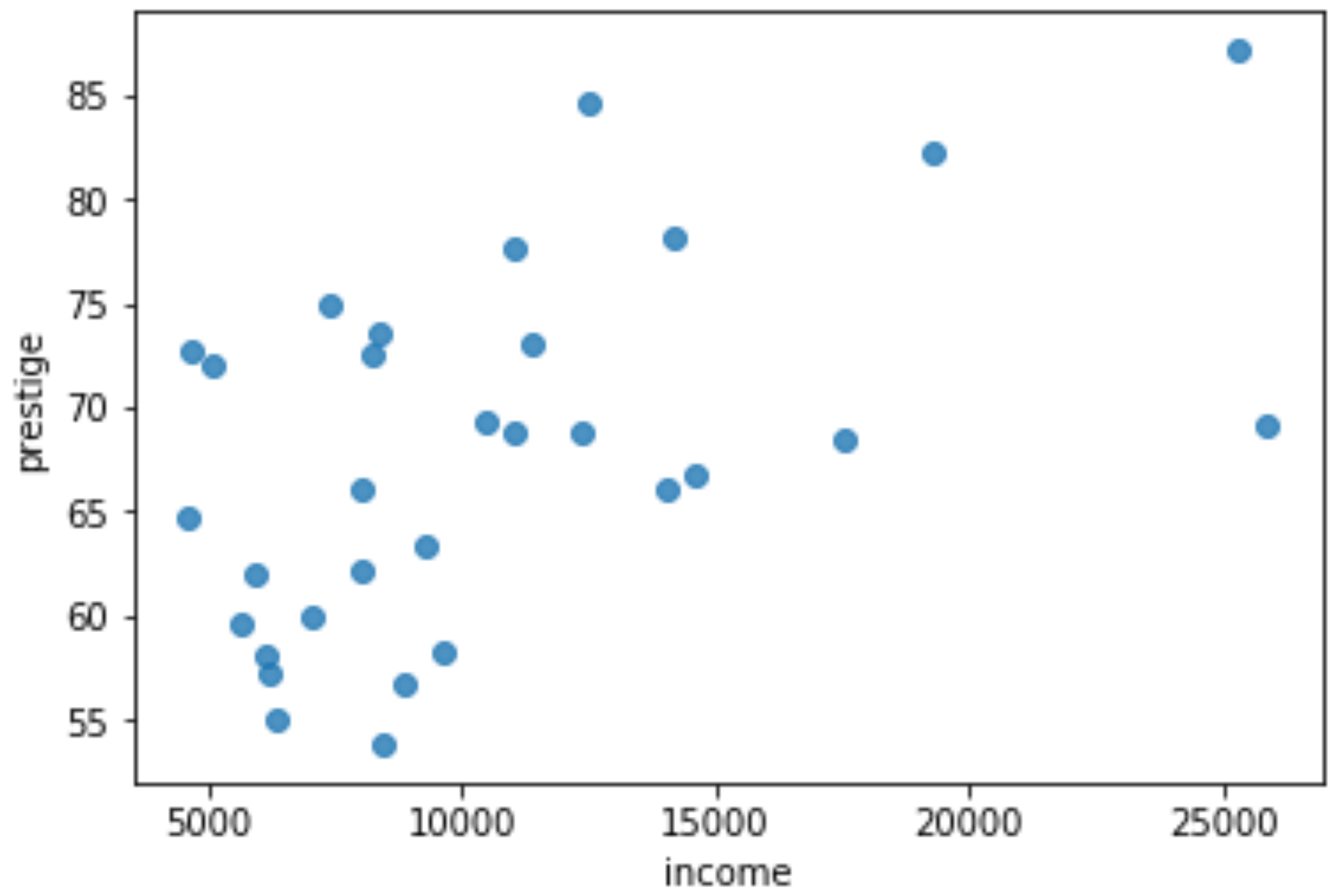
- PhD in Machine Learning, focusing on Bayesian methods (particularly Bayesian nonparametrics)
- Currently: Assistant professor of Statistics at UT Austin/Lead research scientist at CognitiveScale
- Research interests: Bayesian modeling, scalable Bayesian inference, random graphs, private ML, fair ML.
- Non-research interests: Bouldering, rollerskating, hanging out with my dog Fritz.

About Evan

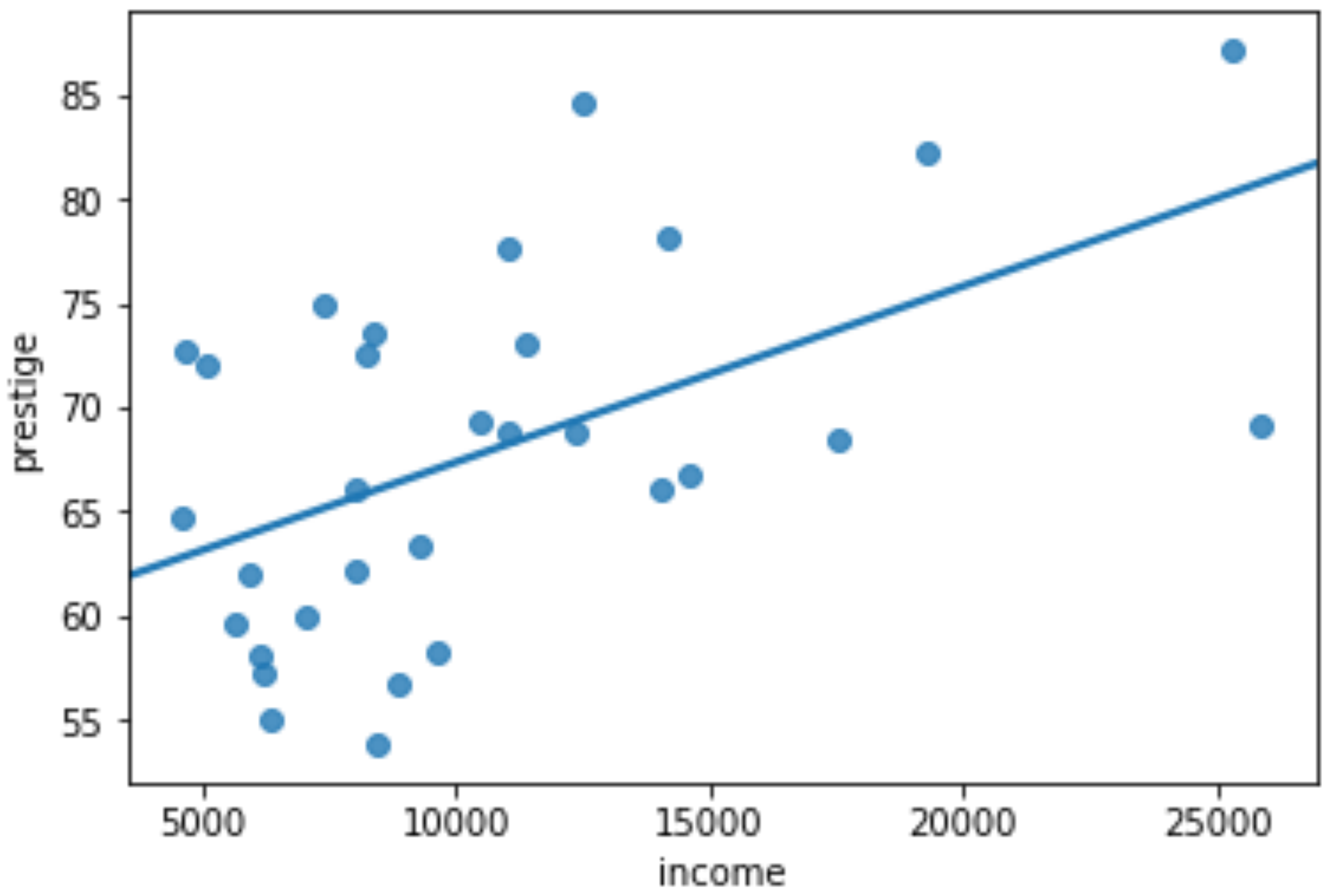
- Statistics PhD Student at UT Austin
- Research interests: Bayesian neural networks, approximating distributions
- Non-research interests: baking bread, volunteering at church, hanging out with my cat Caffrey

About you?

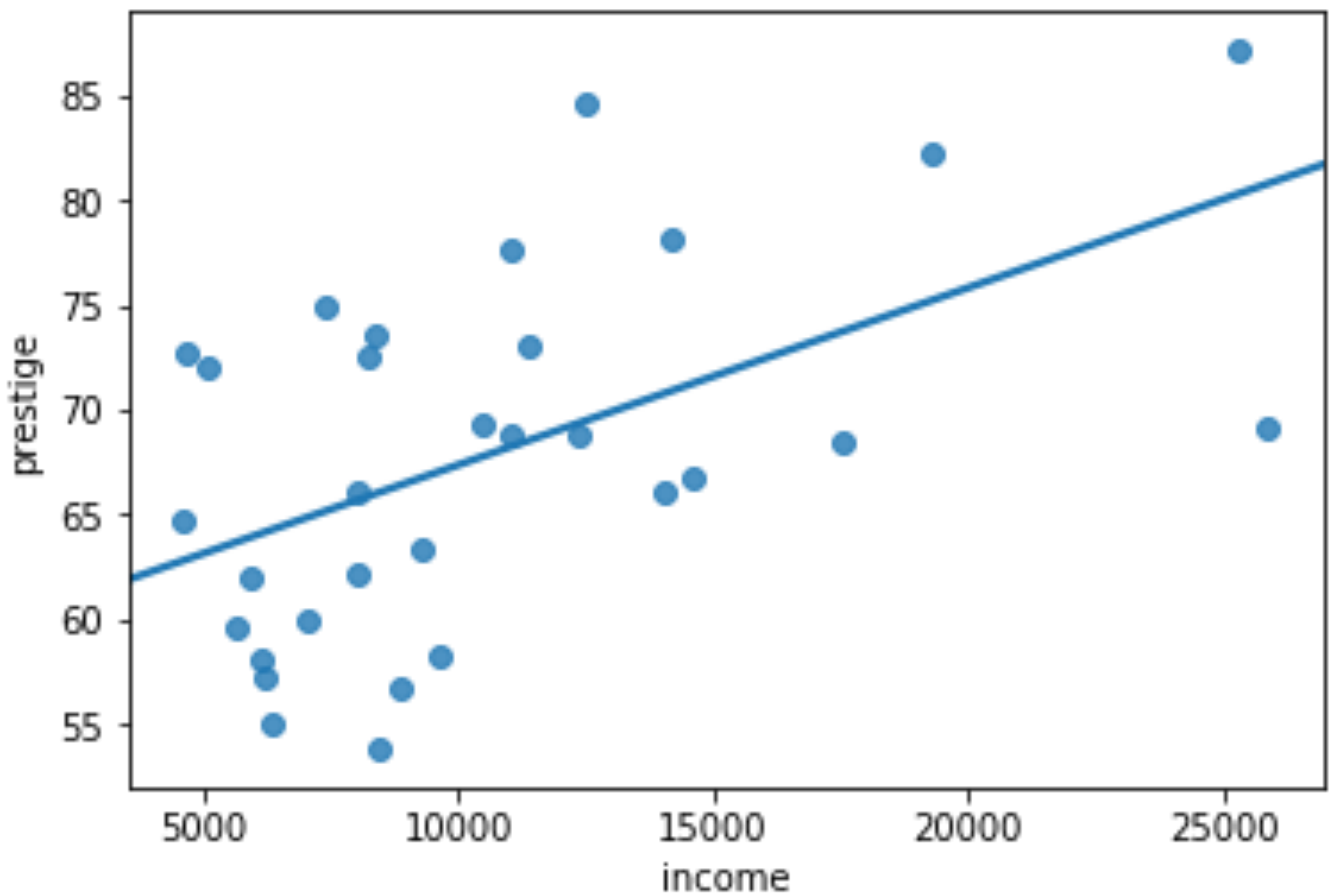
A familiar example



A familiar example



A familiar example



How did I calculate the line?

Linear regression as maximum likelihood

- Standard regression assumption: $y = X\beta + \epsilon$
- Additional assumption: $\epsilon \sim \text{Normal}(0, \sigma^2)$
- Likelihood:

$$p\left(y_i \,|\, x_i, \beta, \sigma\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}\left(y_i - x_i^T\beta\right) \right\}$$

$$p\left(y \,|\, X, \beta, \sigma\right) = (2\pi\sigma^2)^{-n/2} \, \exp\left\{ -\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) \right\}$$

Linear regression as maximum likelihood

$$\hat{\beta} = \arg\max_{\beta} p\left(y \,|\, X, \beta, \sigma\right)$$

Linear regression as maximum likelihood

$$\begin{aligned}\hat{\beta} &= \arg\max_{\beta} p\left(y \,|\, X, \beta, \sigma\right) \\ &= \arg\max_{\beta} \log p\left(y \,|\, X, \beta, \sigma\right)\end{aligned}$$

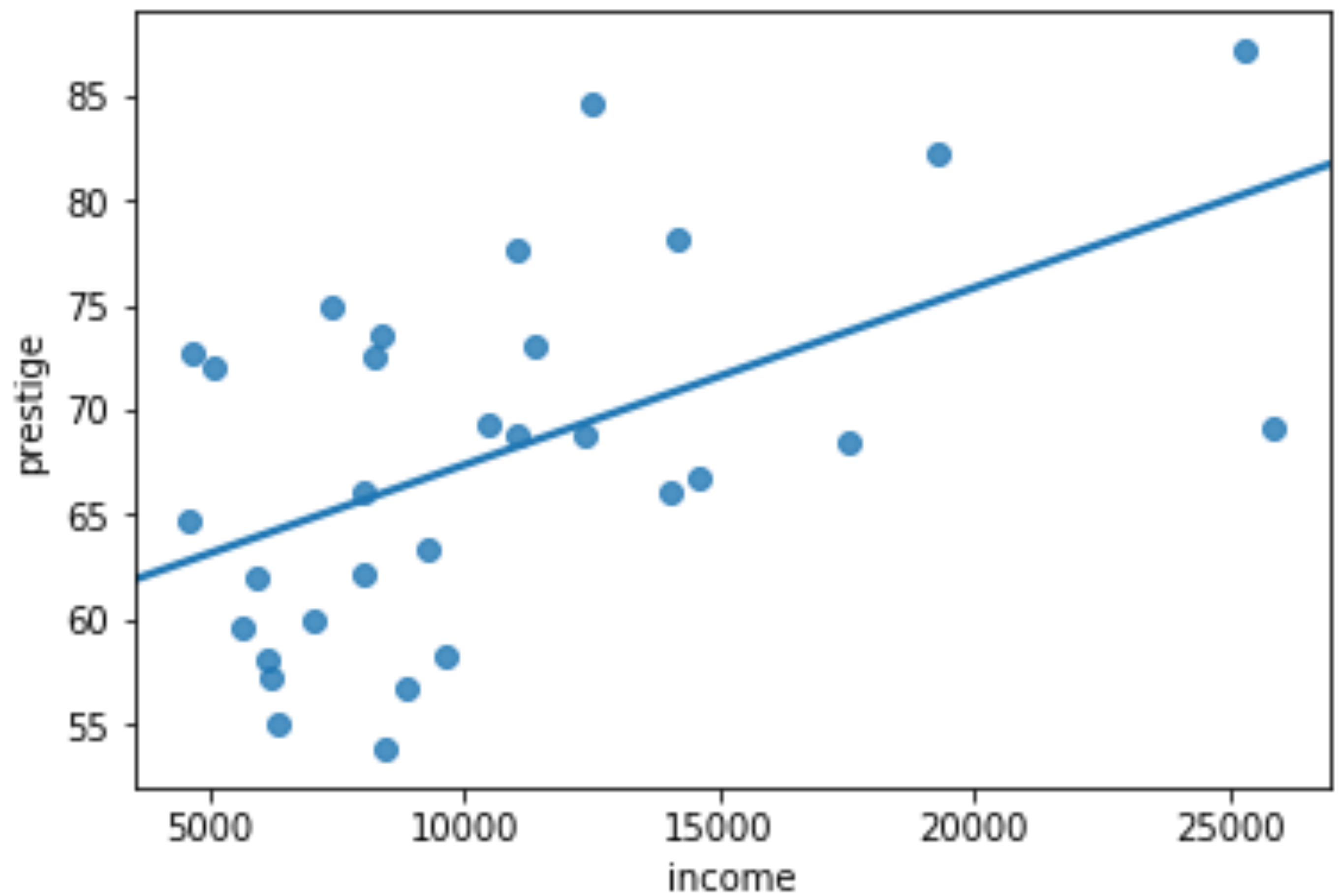
Linear regression as maximum likelihood

$$
\begin{aligned}
\hat{\beta} &= \arg\max_{\beta} p\left(y \,|\, X, \beta, \sigma\right) \\
&= \arg\max_{\beta} \log p\left(y \,|\, X, \beta, \sigma\right) \\
&= \arg\min_{\beta} \, (y - X\beta)^T (y - X\beta)
\end{aligned}
$$

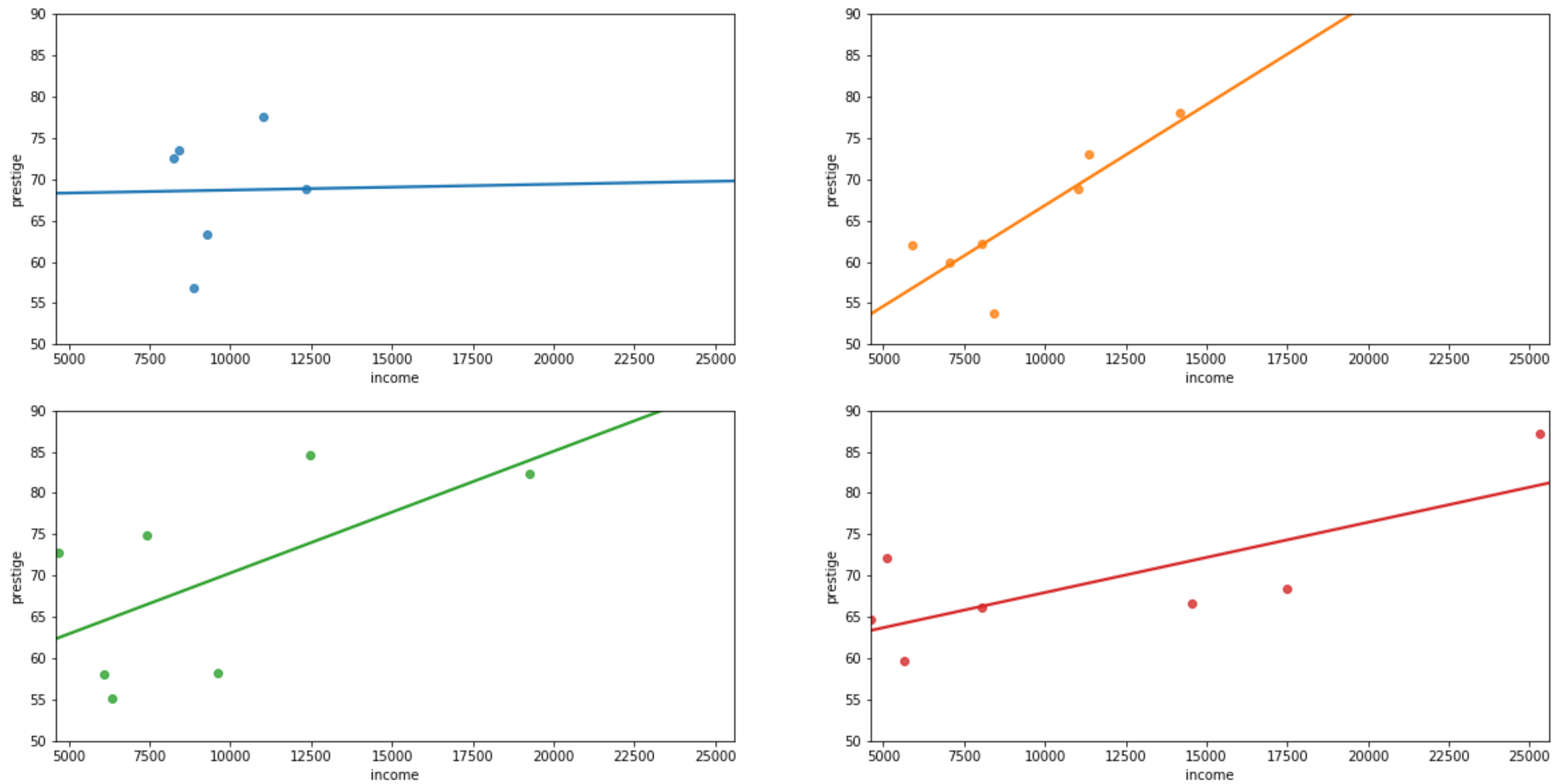
Linear regression as maximum likelihood

$$
\begin{aligned}
\hat{\beta} &= \arg\max_{\beta} p\left(y \,|\, X, \beta, \sigma\right) \\
&= \arg\max_{\beta} \log p\left(y \,|\, X, \beta, \sigma\right) \\
&= \arg\min_{\beta} \, (y - X\beta)^T (y - X\beta) \\
&= (X^T X)^{-1} X^T y
\end{aligned}
$$

Looks pretty good!

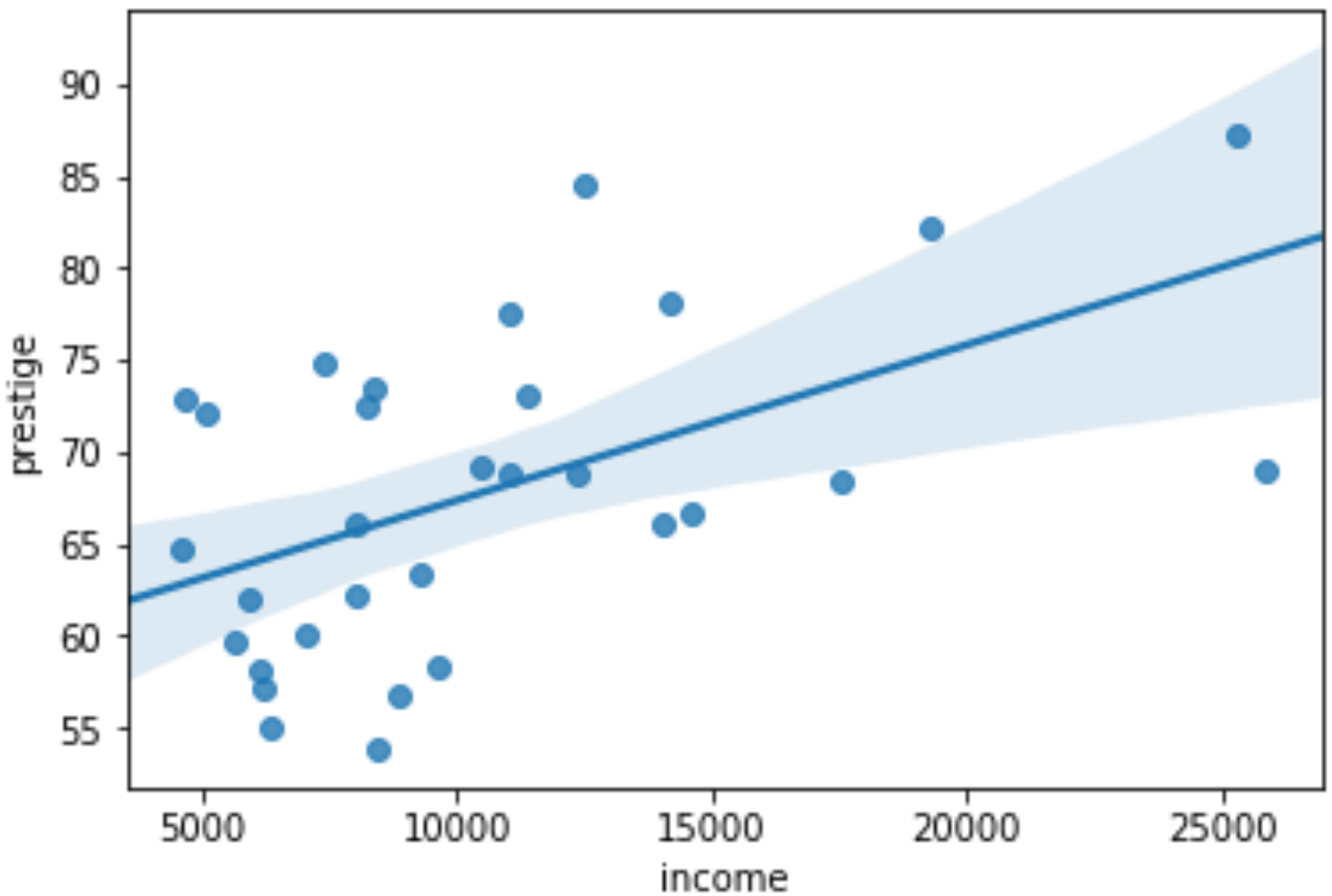


But what if we have fewer data?



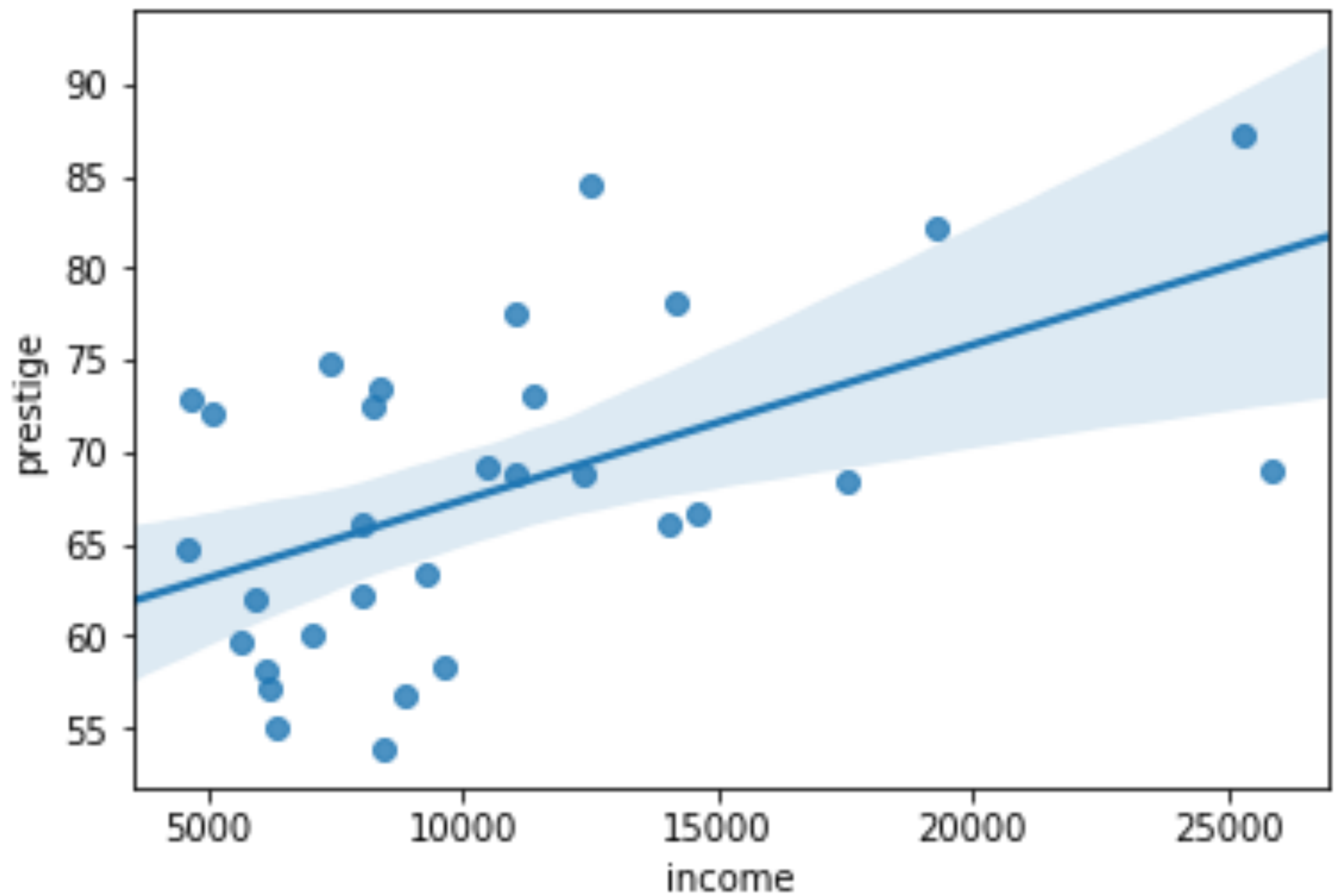
- When we have few datapoints relative to the number of parameters, we tend to overfit.
- What to do about this??

Confidence intervals: expressing uncertainty about β



If the true parameter is β, and the data is randomly generated, β is in the 95% c.i. 95% of the time.

Confidence intervals: expressing uncertainty about β



Regularization:
Pulling β towards something "reasonable"

$$\hat{\beta}_{LS} = \arg\min_{\beta} (y - X\beta)^T (y - X\beta) = (X^T X)^{-1} X^T y$$

$$\hat{\beta}_{ridge} = \arg\min_{\beta} (y - X\beta)^T (y - X\beta) \qquad \text{s.t.} \quad \beta^T \beta \leq t$$

Regularization: Pulling β towards something "reasonable"

$$\hat{\beta}_{LS} = \arg\min_{\beta} (y - X\beta)^T (y - X\beta) = (X^T X)^{-1} X^T y$$

$$\hat{\beta}_{ridge} = \arg\min_{\beta} (y - X\beta)^T (y - X\beta) \qquad \text{s.t.} \quad \beta^T \beta \leq t$$

Rewrite using Lagrangian!

$$= \arg\min_{\beta} \left((y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta \right)$$

Regularization:
Pulling β towards something "reasonable"

$$\hat{\beta}_{LS} = \arg\min_{\beta} (y - X\beta)^T (y - X\beta) = (X^T X)^{-1} X^T y$$

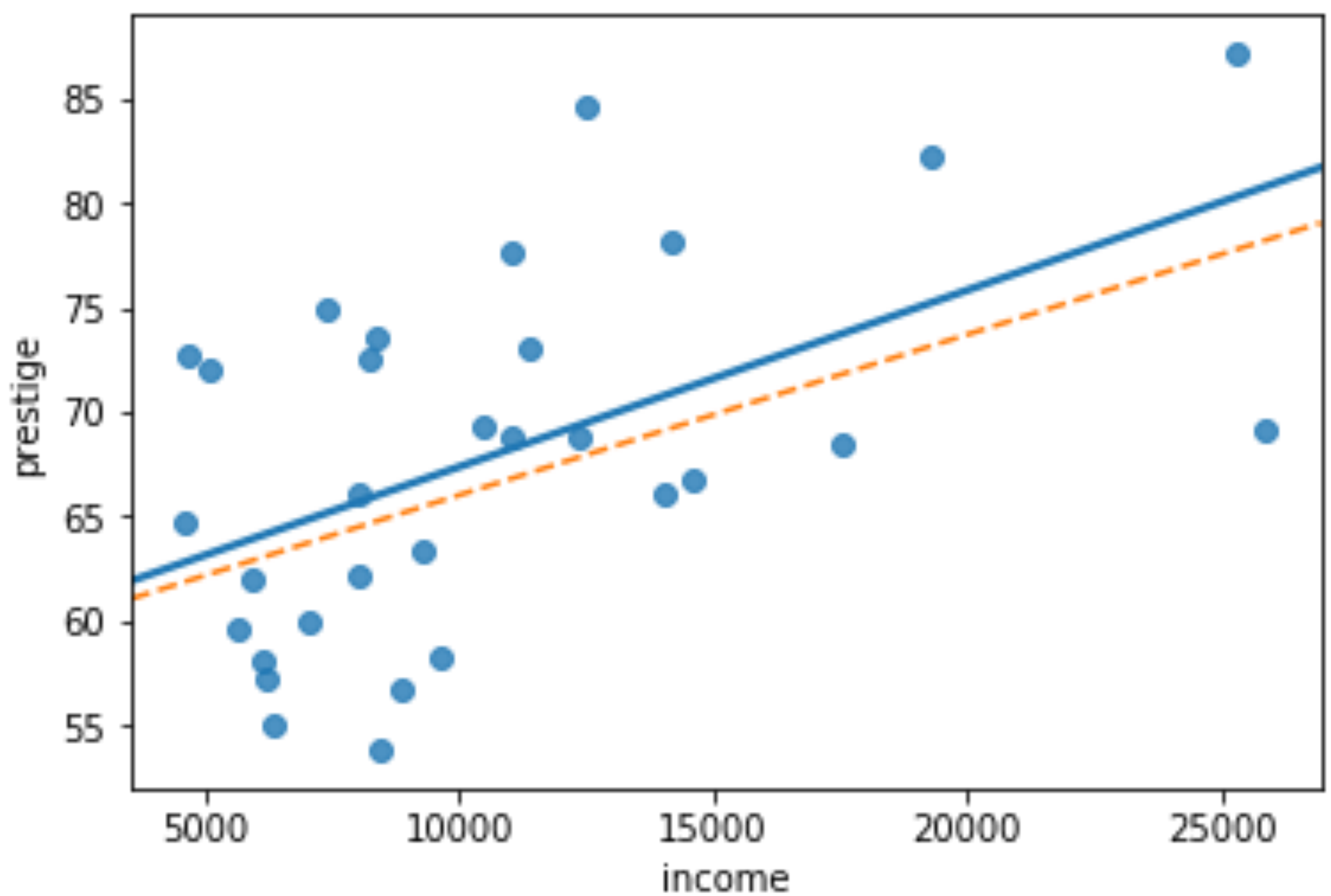
$$\hat{\beta}_{ridge} = \arg\min_{\beta} (y - X\beta)^T (y - X\beta) \qquad \text{s.t.} \quad \beta^T \beta \leq t$$

Rewrite using Lagrangian!

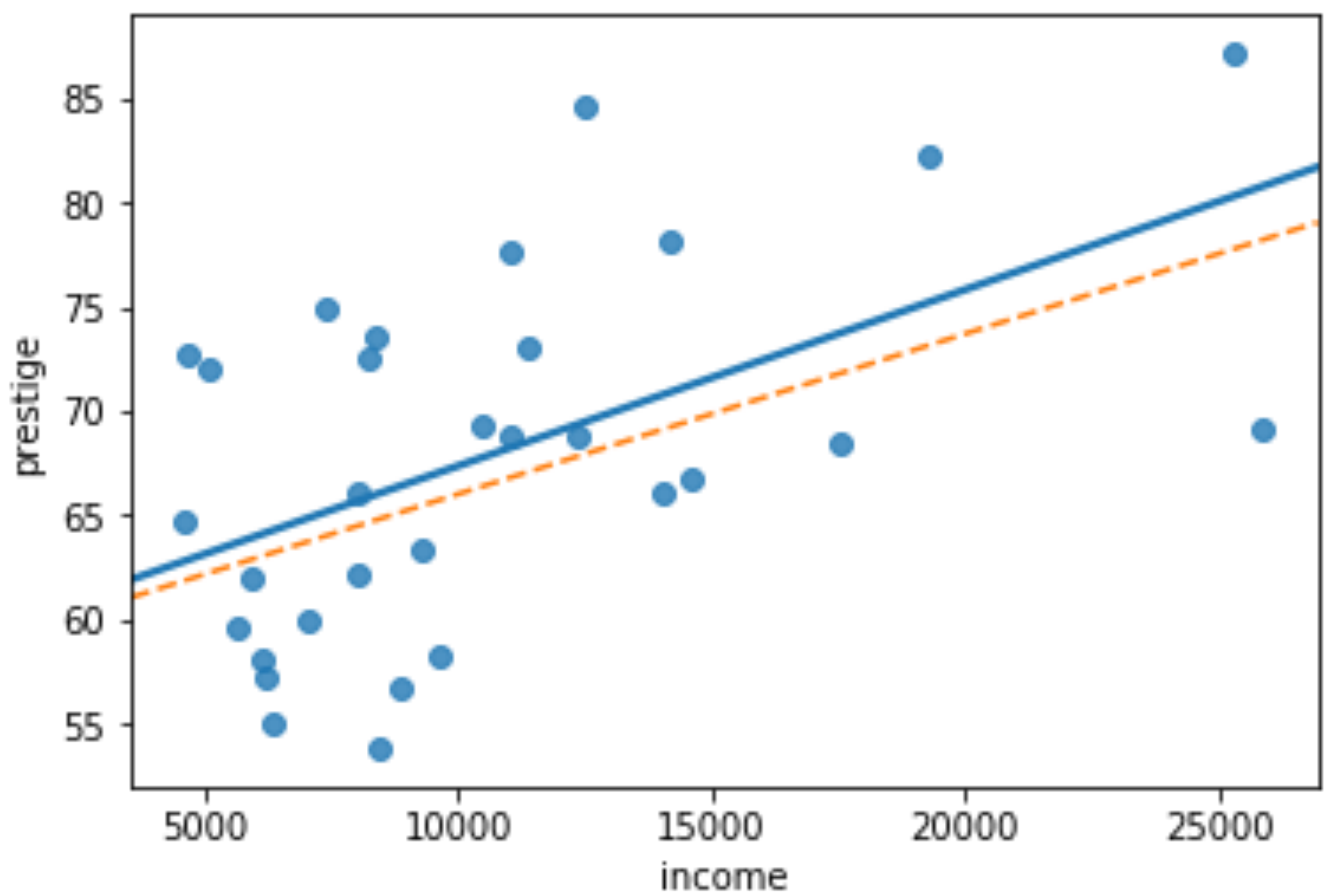
$$= \arg\min_{\beta} \left((y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta \right)$$

$$= (X^T X + \lambda \mathbf{I})^{-1} X^T y$$

Regularization:
Pulling β towards something "reasonable"



Regularization:
Pulling β towards something "reasonable"



β is pulled towards (0, 0)

Bayesian methods

- What do we mean by Bayesian?

Bayesian methods

- What do we mean by Bayesian?
- Back up: What do we mean by probability?

Bayesian methods

- What do we mean by Bayesian?
- Back up: What do we mean by probability?

“This coin returns heads with probability 0.4”

Bayesian methods

- What do we mean by Bayesian?
- Back up: What do we mean by probability?

"This coin returns heads with probability 0.4"

… if I repeatedly toss the coin, the proportion of heads will tend to 0.4

Bayesian methods

- What do we mean by Bayesian?
- Back up: What do we mean by probability?

"This coin returns heads with probability 0.4"

… if I repeatedly toss the coin, the proportion of heads will tend to 0.4

"The probability of alien life existing in our solar system is 0.4"

Bayesian methods

- What do we mean by Bayesian?
- Back up: What do we mean by probability?

"This coin returns heads with probability 0.4"

… if I repeatedly toss the coin, the proportion of heads will tend to 0.4

"The probability of alien life existing in our solar system is 0.4"

… this is a measure of **belief**, or **degree of certainty**

Bayesian methods

- When looking at confidence intervals, we thought of **data** as **random** and **parameters** as **fixed.**
  - This ties in with the frequentist view: randomness implies repeatability.
- The Bayesian viewpoint interprets probabilities as (un)certainties…
- In this framework, **data** are **fixed** (certain) and **parameters** are **random** (uncertain).

Bayes' Law

- We want to quantify $p(\beta | D)$ - our beliefs about β given our data D.
- Bayes' Law allows us to write

$$p(\beta | D) = \frac{p(D | \beta) p(\beta)}{p(D)}$$

Bayes' Law

- We want to quantify $p(\beta \mid D)$ - our beliefs about β given our data D.
- Bayes' Law allows us to write

$$\underset{\text{posterior}}{p(\beta \mid D)} = \frac{\overset{\text{likelihood}}{p(D \mid \beta)}\,\overset{\text{prior}}{p(\beta)}}{\underset{\text{evidence}}{p(D)}}$$

Bayesian linear regression

- Let's assume σ is known, and $\beta \sim \text{Normal}(\mu_0, \Sigma_0)$

$$p(\beta \,|\, X, y, \sigma) \propto \overbrace{(2\pi\sigma^2)^{-n/2} \ \exp\left\{ -\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) \right\}}^{p(D|\beta,\sigma)}$$

$$\times \underbrace{(2\pi)^{-k/2} \,|\, \Sigma_0 \,|^{-1/2} \ \exp\left\{ -\frac{1}{2}(\beta - \mu_0)^T \Sigma_0^{-1} (\beta - \mu_0) \right\}}_{p(\beta)}$$

Bayesian linear regression

- Let's assume σ is known, and $\beta \sim \text{Normal}(\mu_0, \Sigma_0)$

$$p(\beta \,|\, X, y, \sigma) \propto \exp\left\{ -\frac{1}{2}\left(\frac{1}{\sigma^2}(y - X\beta)^T(y - X\beta) + (\beta - \mu_0)^T\Sigma_0^{-1}(\beta - \mu_0)\right)\right\}$$

Bayesian linear regression

- Let's assume σ is known, and $\beta \sim \text{Normal}(\mu_0, \Sigma_0)$

$$p(\beta \,|\, X, y, \sigma) \propto \exp\left\{ -\frac{1}{2}\left(\frac{1}{\sigma^2}(y - X\beta)^T(y - X\beta) + (\beta - \mu_0)^T\Sigma_0^{-1}(\beta - \mu_0)\right)\right\}$$

$$\propto \exp\left\{ -\frac{1}{2}(\beta - \mu_n)^T\Sigma_n(\beta - \mu_n)\right\}$$

$$\mu_n = \left(\Sigma_0^{-1} + X^T X/\sigma^2\right)^{-1}\left(\Sigma_0^{-1}\mu_0 + X^T y/\sigma^2\right)$$

$$\Sigma_n = \left(\Sigma_0^{-1} + X^T X/\sigma^2\right)^{-1}$$

Bayesian linear regression

- Let's assume σ is known, and $\beta \sim \text{Normal}(\mu_0, \Sigma_0)$

$$p(\beta \,|\, X, y, \sigma) \propto \exp\left\{ -\frac{1}{2}\left(\frac{1}{\sigma^2}(y - X\beta)^T(y - X\beta) + (\beta - \mu_0)^T \Sigma_0^{-1}(\beta - \mu_0) \right) \right\}$$

$$\propto \exp\left\{ -\frac{1}{2}(\beta - \mu_n)^T \Sigma_n (\beta - \mu_n) \right\}$$

Normal(μ_n, Σ_n)

$$\mu_n = \left(\Sigma_0^{-1} + X^T X/\sigma^2\right)^{-1} \left(\Sigma_0^{-1}\mu_0 + X^T y/\sigma^2\right)$$

$$\Sigma_n = \left(\Sigma_0^{-1} + X^T X/\sigma^2\right)^{-1}$$

$\hat{\beta}_{ridge}$ when $\mu_0 = 0$ and $\Sigma_0 = \frac{1}{\lambda^2}\mathbf{I}$

Prior selection

- Sometimes we have reasonable intuition that we can use
  - Doubling salary probably doubles prestige? Reasonable prior mean for slope = 1?
  - Increasing salary unlikely to decrease prestige… perhaps prior standard deviation for slope = 0.5?
- If not, standard is to go for vague priors that capture all reasonable settings
  - In the lab next, we're going to standardize the data to be zero-mean, unit variance - and use zero-mean, unit-variance priors.
- We'll explore what effect the prior has in the lab.

Lab 1

- `github.com/sinead/DS32019`
- 3 partially complete notebooks (we'll do 1 now)

Lab 1 discussion

- What difference did the prior specification make?
- How does the amount of data change the posterior?
- In what ways was the model misspecified?

Changing our assumptions

The Bayesian linear model makes a number of assumptions:

- Mean variation is linear in the covariates; prior captures our beliefs.
- Observations are i.i.d. Gaussian, given that mean.

Changing our assumptions

The Bayesian linear model makes a number of assumptions:

- Mean variation is linear in the covariates; prior captures our beliefs.

Add transformed covariates - e.g. x^2, x^3

Combine multiple regressions in a neural net

Choose a different prior

Gaussian processes

- Observations are i.i.d. Gaussian, given that mean.

Choose a different likelihood

Pick a different model - HMM, AR(1) process…

Changing our assumptions

The Bayesian linear model makes a number of assumptions:

- Mean variation is linear in the covariates; prior captures our beliefs.

Add transformed covariates - e.g. x^2, x^3

Combine multiple regressions in a neural net

Gaussian processes

Choose a different prior

- Observations are i.i.d. Gaussian, given that mean.

Choose a different likelihood

Pick a different model - HMM, AR(1) process…

Conjugate priors

- Previously, we assumed

$$\beta \sim \text{Normal}(\mu_0, \Sigma_0) \qquad\qquad y_i \sim \text{Normal}(x_i^T \beta, \sigma^2)$$

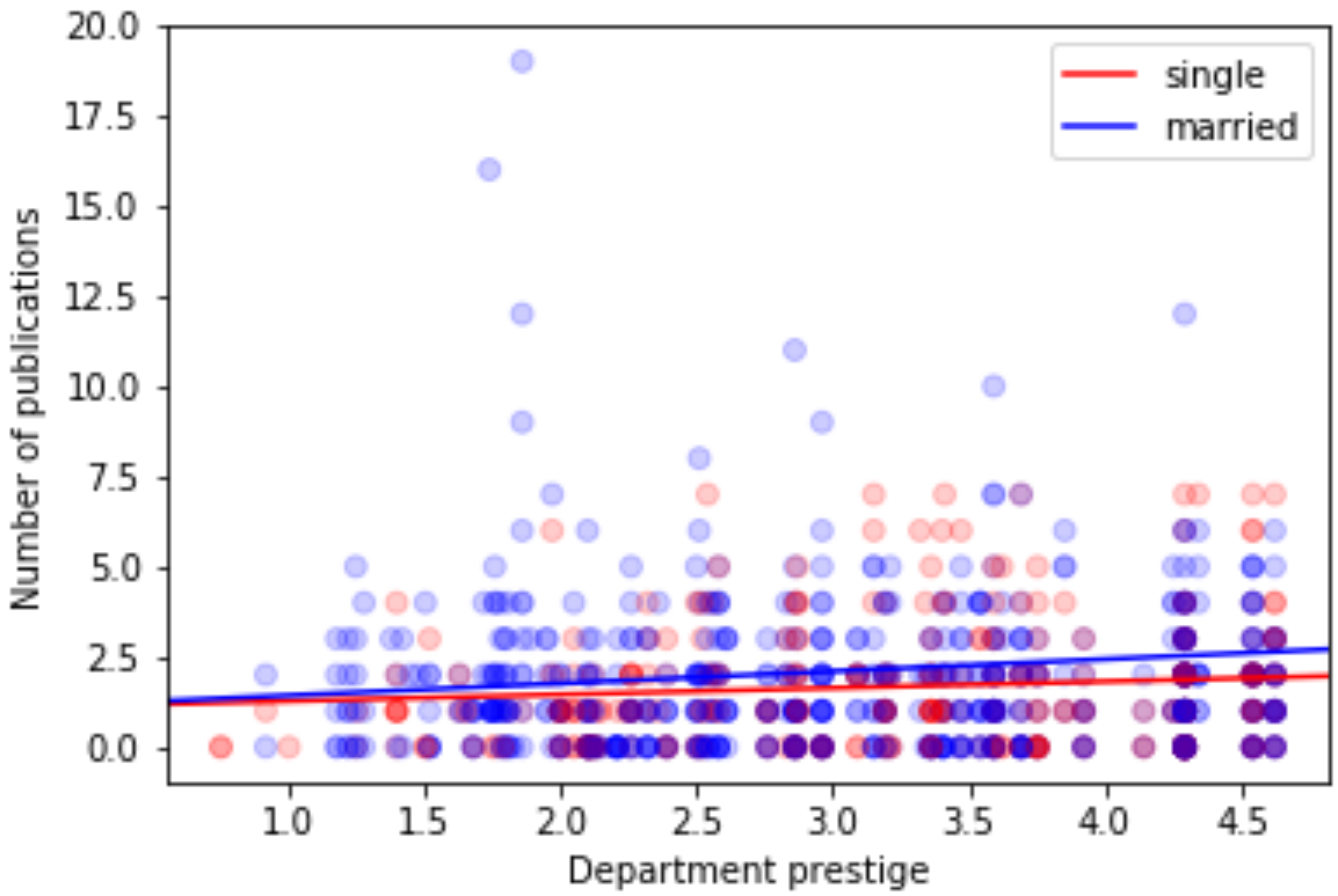
- Our posterior was $\beta \,|\, y_1, \ldots, y_n \sim \text{Normal}(\mu_n, \Sigma_n)$, where

$$\mu_n = \left(\Sigma_0^{-1} + X^T X/\sigma^2\right)^{-1} \left(\Sigma_0^{-1}\mu_0 + X^T y/\sigma^2\right)$$
$$\Sigma_n = \left(\Sigma_0^{-1} + X^T X/\sigma^2\right)^{-1}$$

- This is an example of *conjugacy* - the posterior has the same form as the prior.
  - This makes everything easy… but if we change the prior or the likelihood, the posterior might be intractable

From the lab… publication dataset

- Predict number of publications for biochem students

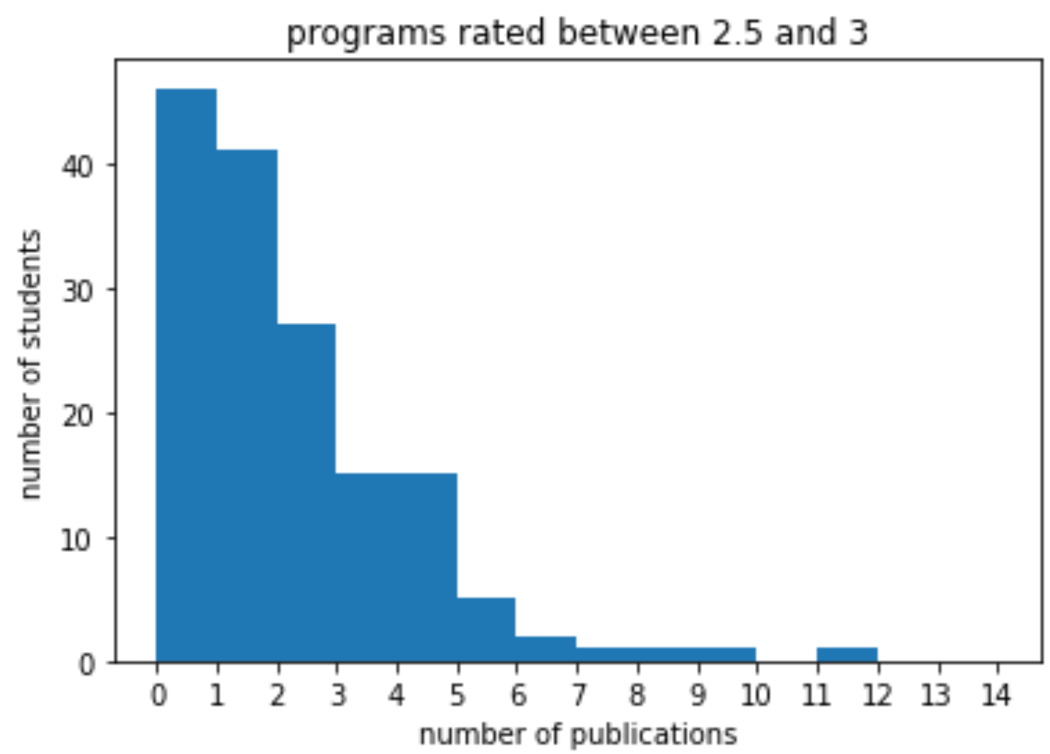


Regressions for count data

- Gaussian likelihood not really appropriate for count data
  - Poisson is a better choice

Regressions for count data

- Gaussian likelihood not really appropriate for count data
    - Poisson is a better choice



    - Maybe over dispersed - negative Binomial?

Bayesian Poisson regression

- Previously: $\beta \sim \text{Normal}(\mu_0, \Sigma_0)$ $\quad y_i \sim \text{Normal}(x_i^T \beta, \sigma^2)$
- Count data: $\beta \sim \text{Normal}(\mu_0, \Sigma_0)$ $\quad y_i \sim \text{Poisson}(?)$

Bayesian Poisson regression

- Previously: $\beta \sim \text{Normal}(\mu_0, \Sigma_0)$ $y_i \sim \text{Normal}(x_i^T \beta, \sigma^2)$
- Count data: $\beta \sim \text{Normal}(\mu_0, \Sigma_0)$ $y_i \sim \text{Poisson}(?)$

- Poisson is parametrized by a positive integer, but $x_i^T\beta$ can be negative…

Bayesian Poisson regression

- Previously: $\beta \sim \text{Normal}(\mu_0, \Sigma_0)$ $y_i \sim \text{Normal}(x_i^T \beta, \sigma^2)$
- Count data: $\beta \sim \text{Normal}(\mu_0, \Sigma_0)$ $y_i \sim \text{Poisson}(?)$

- Poisson is parametrized by a positive integer, but $x_i^T\beta$ can be negative…
- Transform it! $y_i \sim \text{Poisson}\left(\exp\{x_i^T \beta\}, \sigma^2\right)$

Bayesian inference methods for intractable posteriors

- Posterior:

$$p(\beta \,|\, D) \propto \exp\left\{ -\frac{1}{2}(\beta - \mu_0)^T \Sigma_0^{-1} (\beta - \mu_0) \right\} \prod_i \frac{\exp\left\{ y_i x_i^T \beta - \exp\{x_i^T \beta\} \right\}}{y_i\,!}$$

Bayesian inference methods for intractable posteriors

- Posterior:

$$p(\beta \,|\, D) \propto \exp\left\{-\frac{1}{2}(\beta-\mu_0)^T\Sigma_0^{-1}(\beta-\mu_0)\right\}\prod_i \frac{\exp\left\{y_i x_i^T\beta - \exp\{x_i^T\beta\}\right\}}{y_i\,!}$$

  - Doesn’t simplify to something nice!

Bayesian inference methods for intractable posteriors

- Posterior:

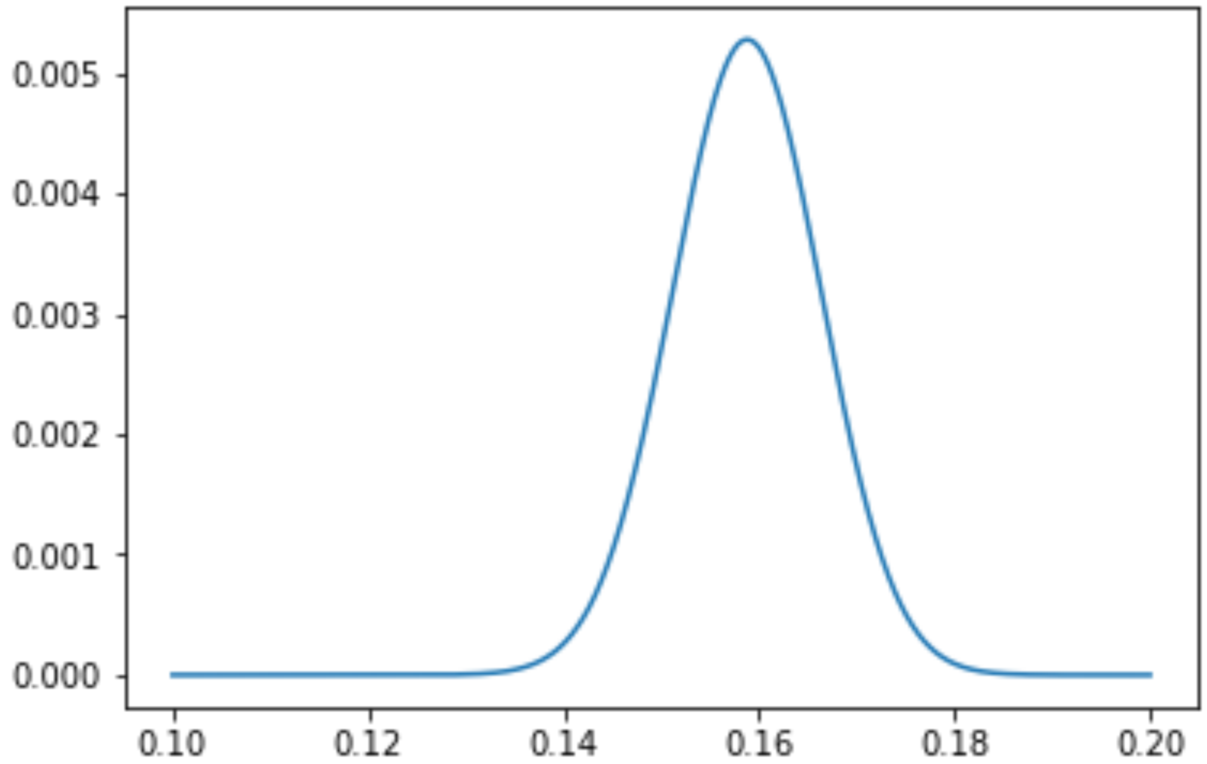
$$p(\beta \mid D) \propto \exp\left\{ -\frac{1}{2}(\beta - \mu_0)^T \Sigma_0^{-1} (\beta - \mu_0) \right\} \prod_i \frac{\exp\left\{ y_i x_i^T \beta - \exp\{x_i^T \beta\} \right\}}{y_i!}$$

  - Doesn't simplify to something nice!

- In general, we are going to have to use **approximate inference** when working with Bayesian methods.
  - Sampling methods (Monte Carlo methods such as MCMC, SMC)
  - Approximating using simpler distributions (Laplace, Variational Bayes, Belief Propagation, Method of Moments…)

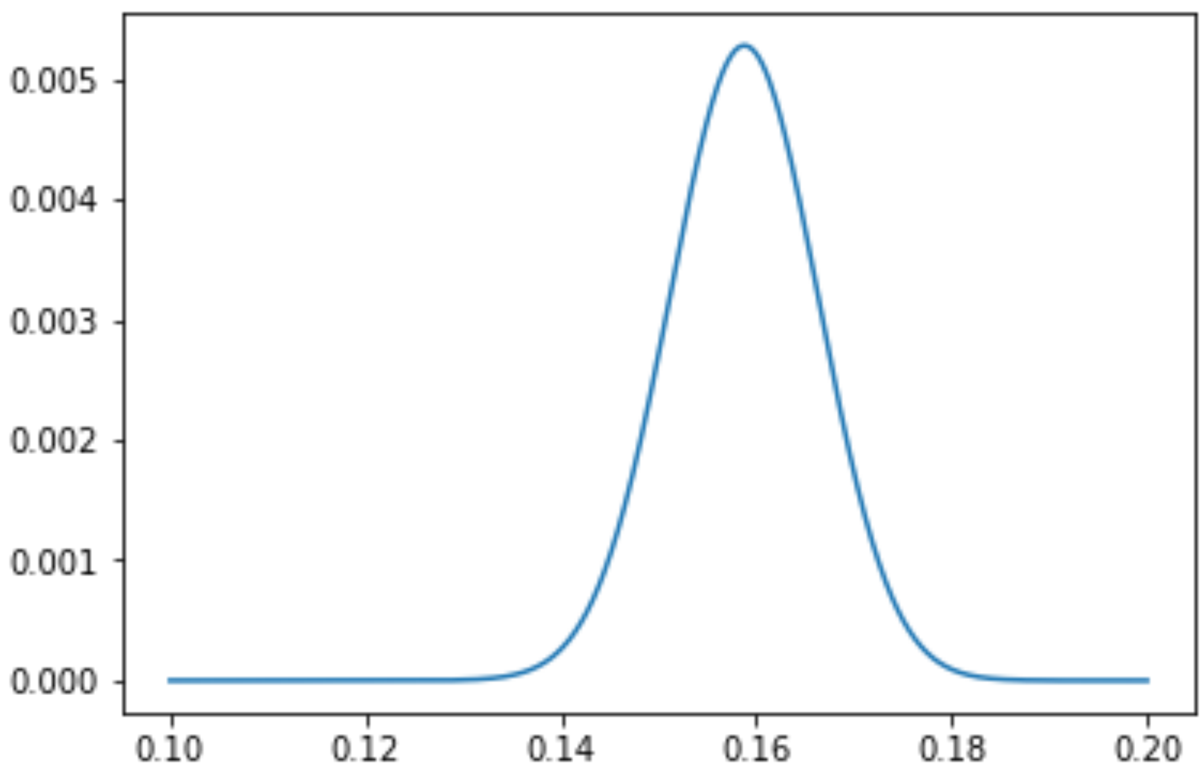
Laplace's method

- We know that $p(\beta | D) \propto \text{Normal}(\beta; \mu_0, \Sigma_0) \prod_i \text{Poisson}\left(y_i | \exp\{x_i^T \beta\}\right)$
  - We can't sample from this easily, but we can plot it (up to a constant).
  - Let's look at just number of publications vs prestige, ignoring intercept



Laplace's method

- We know that $p(\beta \mid D) \propto \text{Normal}(\beta; \mu_0, \Sigma_0) \prod_i \text{Poisson}\left(y_i \mid \exp\{x_i^T \beta\}\right)$
    - We can't sample from this easily, but we can plot it (up to a constant).
    - Let's look at just number of publications vs prestige, ignoring intercept



Looks kind of Gaussian!

Laplace’s method

- Let $P^*(\beta)$ be our (unnormalized) posterior
- Let $\hat{\beta}$ be the value that maximizes $P^*(\beta)$

Laplace's method

- Let $P^*(\beta)$ be our (unnormalized) posterior
- Let $\hat{\beta}$ be the value that maximizes $P^*(\beta)$
- Take a **Taylor expansion** of the log unnormalized posterior…

$$\log P^*(\beta) \approx \log P^*(\hat{\beta}) + (\beta - \hat{\beta}) \frac{d}{d\beta} \log P^*(\beta) \bigg|_{\beta=\hat{\beta}} + \frac{(\beta - \hat{\beta})^2}{2} \frac{d^2}{d\beta^2} \log P^*(\beta) \bigg|_{\beta=\hat{\beta}}$$

$$= \log P^*(\hat{\beta}) + \frac{(\beta - \hat{\beta})^2}{2} \frac{d^2}{d\beta^2} \log P^*(\beta) \bigg|_{\beta=\hat{\beta}}$$

Laplace's method

- Let $P^*(\beta)$ be our (unnormalized) posterior
- Let $\hat{\beta}$ be the value that maximizes $P^*(\beta)$
- Take a **Taylor expansion** of the log unnormalized posterior…

$$\log P^*(\beta) \approx \log P^*(\hat{\beta}) + (\beta - \hat{\beta}) \frac{d}{d\beta} \log P^*(\beta) \bigg|_{\beta=\hat{\beta}} + \frac{(\beta - \hat{\beta})^2}{2} \frac{d^2}{d\beta^2} \log P^*(\beta) \bigg|_{\beta=\hat{\beta}}$$

$$= \log P^*(\hat{\beta}) + \frac{(\beta - \hat{\beta})^2}{2} \frac{d^2}{d\beta^2} \log P^*(\beta) \bigg|_{\beta=\hat{\beta}}$$

- Looks kind of like a log normal!

$$\log \mathrm{Normal}(\beta \,|\, \mu, \sigma) = -\frac{(\beta - \mu)^2}{2} \frac{1}{\sigma^2} + C$$

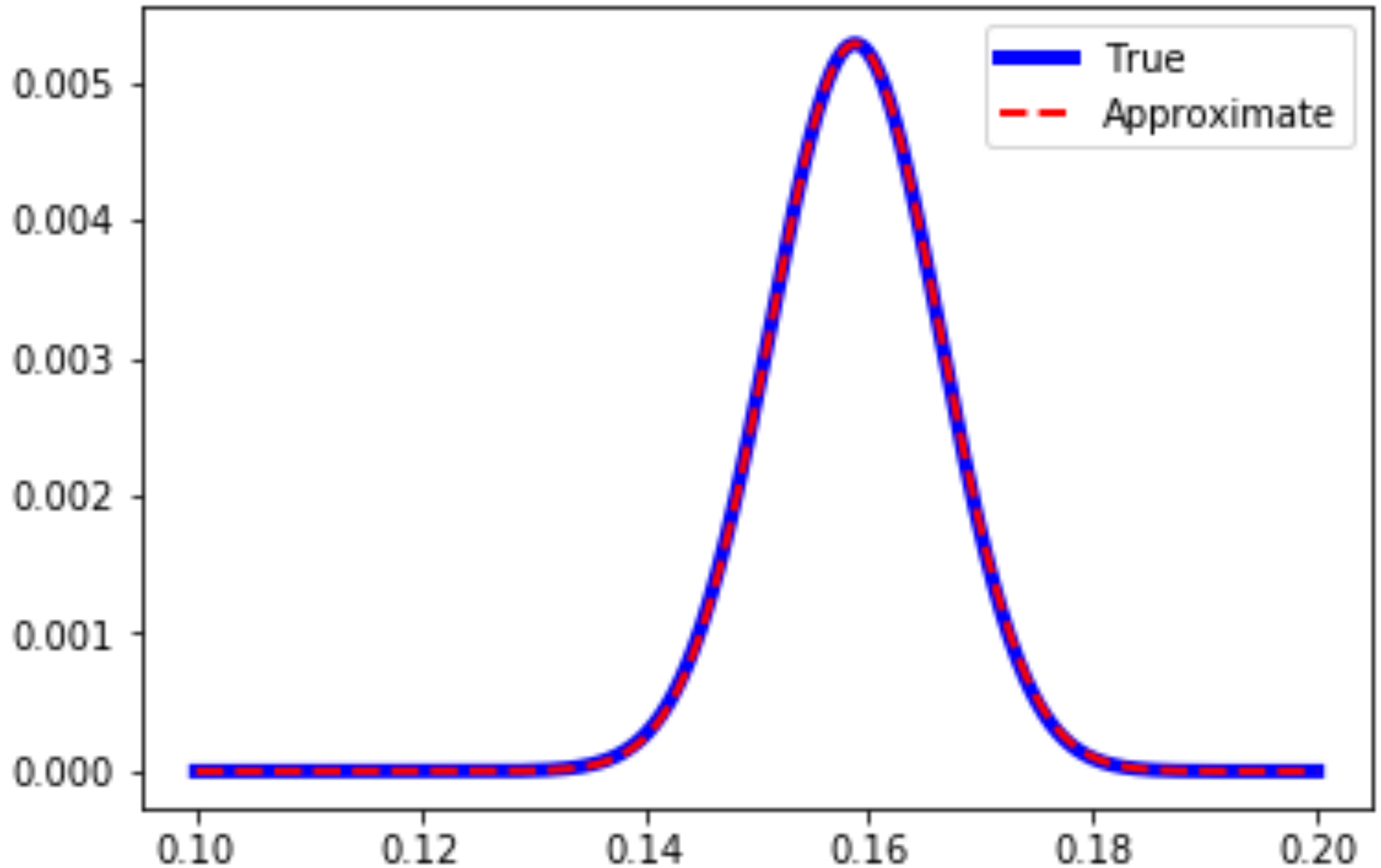
Laplace's method

- We can find $\hat{\beta}$ by optimization (in our case, 0.158)
- We have $\log P^*(\beta) = -\dfrac{(\beta - \mu_0)^2}{2\sigma_0^2} + \sum_i y_i x_i \beta - \exp\{x_i \beta\}$

- First derivative: $\dfrac{\mu_0 - \beta}{\sigma_0^2} + \sum_i y_i x_i - x_i \exp\{x_i \beta\}$
- Second derivative: $-\dfrac{1}{\sigma_0^2} - \sum_i x_i^2 \exp\{x_i \beta\}$ (in our case, -17478.45)

- So, approximate with Normal(0.158, 1/17478.45)

Laplace's method



Laplace’s method

- We can easily extend this to higher dimensions, by using the Hessian in place of the second derivative.
- Again, can find $\hat{\beta}$ by optimization
- Approximate covariance with inverse Hessian
    - Can calculate analytically, use autodiff, use numerical approx

ADD SLIDE HERE

Variational inference

- Laplace's approximation approximates distributions with a **Gaussian**, and matches the mode.

- We can choose other approximating distributions!

- **Variational methods** choose a family of approximate distributions, and find the **closest distribution** in that family.

  - Here, closest is defined in terms of KL divergence

$$\mathrm{KL}(q\,||\,p) = \mathbb{E}_q\left[\log \frac{q(\theta)}{p(\theta)}\right]$$

Variational inference

- Hidden variables θ, data x.
- Want: $p(\theta \mid x) = \dfrac{p(\theta, x)}{p(x)}$

evidence $p(x)$ is often intractable

Variational inference

- Hidden variables θ, data x.
- Want: $p(\theta | x) = \dfrac{p(\theta, x)}{p(x)}$

evidence $p(x)$ is often intractable

- Approximating distribution: $q(\theta)$

Variational inference

- Hidden variables θ, data x.
- Want: $p(\theta \mid x) = \dfrac{p(\theta, x)}{p(x)}$

evidence $p(x)$ is often intractable

- Approximating distribution: $q(\theta)$
- $\mathrm{KL}(q(\theta) \,||\, p(\theta \mid x) = \mathbb{E}_q\left[\log q(\theta)\right] - \mathbb{E}_q\left[\log p(\theta \mid x)\right]$

Variational inference

- Hidden variables θ, data x.
- Want: $p(\theta \mid x) = \dfrac{p(\theta, x)}{p(x)}$

evidence $p(x)$ is often intractable

- Approximating distribution: $q(\theta)$
- $$\begin{aligned} \mathrm{KL}(q(\theta) \,||\, p(\theta \mid x) &= \mathbb{E}_q \left[\log q(\theta)\right] - \mathbb{E}_q \left[\log p(\theta \mid x)\right] \\ &= \mathbb{E}_q \left[\log q(\theta)\right] - \mathbb{E}_q \left[\log p(\theta, x)\right] + \mathbb{E}_q \left[\log p(x)\right] \end{aligned}$$

Variational inference

- Hidden variables θ, data x.
- Want: $p(\theta | x) = \dfrac{p(\theta, x)}{p(x)}$

evidence $p(x)$ is often intractable

- Approximating distribution: $q(\theta)$
- $$\begin{aligned} \mathrm{KL}(q(\theta) \,||\, p(\theta | x)) &= \mathbb{E}_q \left[\log q(\theta)\right] - \mathbb{E}_q \left[\log p(\theta | x)\right] \\ &= \mathbb{E}_q \left[\log q(\theta)\right] - \mathbb{E}_q \left[\log p(\theta, x)\right] + \mathbb{E}_q \left[\log p(x)\right] \\ &\geq -\Big(\underbrace{\mathbb{E}_q \left[\log p(\theta, x)\right] - \mathbb{E}_q \left[\log q(\theta)\right]}_{\text{evidence lower bound (ELBO)}}\Big) \end{aligned}$$

Variational inference

- Hidden variables θ, data x.
- Want: $p(\theta | x) = \dfrac{p(\theta, x)}{p(x)}$

evidence $p(x)$ is often intractable

- Approximating distribution: $q(\theta)$
- $$\begin{aligned} \mathrm{KL}(q(\theta) || p(\theta | x)) &= \mathbb{E}_q \left[\log q(\theta)\right] - \mathbb{E}_q \left[\log p(\theta | x)\right] \\ &= \mathbb{E}_q \left[\log q(\theta)\right] - \mathbb{E}_q \left[\log p(\theta, x)\right] + \mathbb{E}_q \left[\log p(x)\right] \\ &\geq -\underbrace{\left(\mathbb{E}_q \left[\log p(\theta, x)\right] - \mathbb{E}_q \left[\log q(\theta)\right]\right)}_{\text{evidence lower bound (ELBO)}} \end{aligned}$$

- Minimizing KL equivalent to maximizing ELBO

Calculating the ELBO

The old way:

- Calculate the expectations analytically - time-consuming, complicated, not generalizable.

Calculating the ELBO

The old way:

- Calculate the expectations analytically - time-consuming, complicated, not generalizable.

The new way

- Estimate ELBO using samples from $q(\theta)$

$$\text{ELBO} \approx \frac{1}{S} \sum_s \left[\log q(\theta) - \log p(\theta) - \sum_{i=1}^{n} p(x_i \mid \theta) \right]$$

Calculating the ELBO

The old way:

- Calculate the expectations analytically - time-consuming, complicated, not generalizable.

The new way

- Estimate ELBO using samples from $q(\theta)$

$$\text{ELBO} \approx \frac{1}{S} \sum_s \left[\log q(\theta) - \log p(\theta) - \sum_{i=1}^{n} p(x_i \mid \theta) \right]$$

- Stochastic estimate using size-M minibatch of data:

$$\text{ELBO} \approx \frac{1}{S} \sum_s \left[\log q(\theta) - \log p(\theta) - \frac{N}{M} \sum_{x_i \in \mathscr{M}} p(x_i \mid \theta) \right]$$

Variational inference: Logistic regression

- For binary data, the obvious likelihood is a Bernoulli… so, we need to transform $x_i^T\beta$ to be between 0 and 1.
  - Logistic function is an obvious choice:

$$\beta \sim \text{Normal}(\mu_p, \Sigma_p) \qquad\qquad y_i \sim \text{Bernoulli}\underbrace{\left(\frac{1}{1+\exp - x_i^T\beta}\right)}_{\sigma(x_i^T\beta)}$$

  - Let's let $q(\theta) = \text{Normal}(\mu_q, \Sigma_q)$
- ELBO is then

$$\mathbb{E}_{\beta\sim\mathcal{N}(\mu_q,\Sigma_q)}\Big[\log\mathcal{N}(\beta;\mu_q,\Sigma_q) - \log\mathcal{N}(\beta;\mu_p,\Sigma_p)$$
$$- \sum_i \big(y_i \log\sigma(x_i^T\beta) + (1-y_i)\log(1-\sigma(x_i^T\beta)\big)\Big]$$

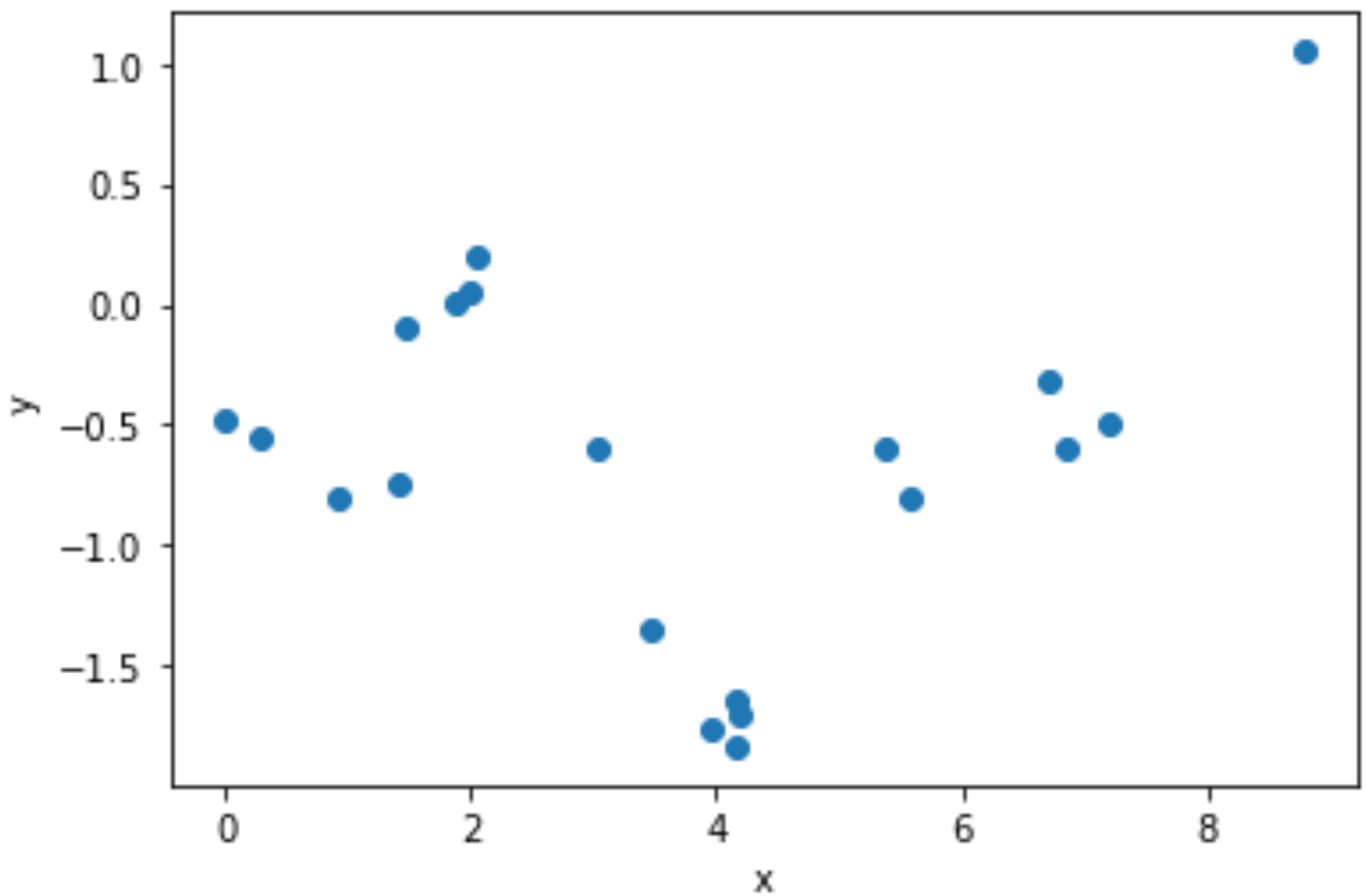
Lab 2

- `github.com/sinead/DS32019`
- We will implement both the Laplace approximation, and variational inference

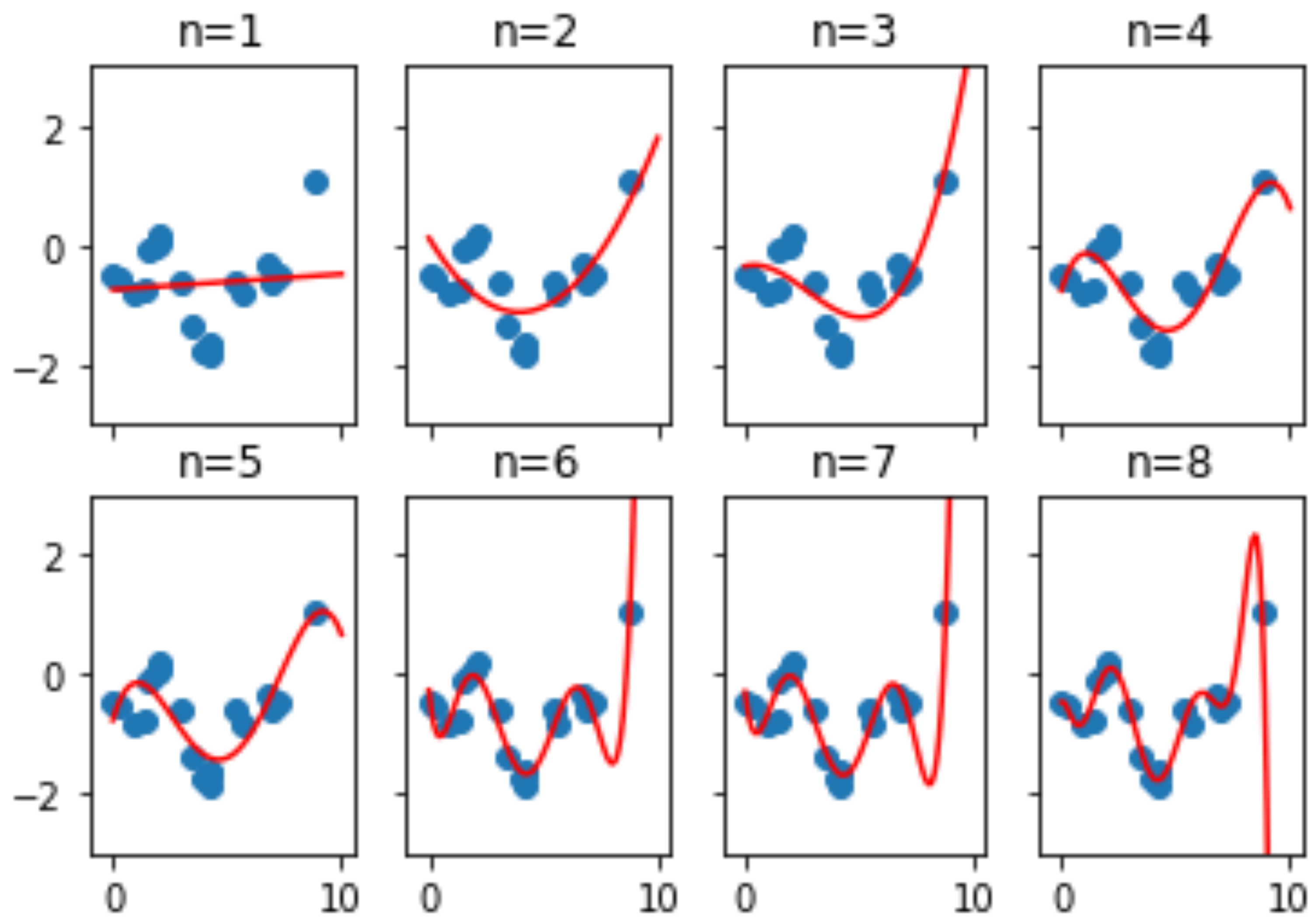
Lab 2 discussion

- How do the two approximations differ?
- What advantages might there be in using variational inference vs the Laplace approximation?
- How could we extend the variational methods we have looked at?

Non-linear regression



Polynomial regression?



Too few degrees of freedom → can't capture variation
Too many degrees of freedom → overfitting

Bayesian Neural network?

- In the lab, we implemented Variational inference for logistic regression.
- If we stack multiple logistic regressions, we can build a Bayesian neural network.
- We can use variational inference (or another approach) to infer the posterior distribution over weights.
- We're not going to explore this now… but check out the tutorials for TensorFlow Probability

Multivariate Gaussian distribution

- Covariance captures correlation between dimensions
  - Much like regression captures correlation between observations!

Multivariate Gaussian distribution

- Covariance captures correlation between dimensions
  - Much like regression captures correlation between observations!
- Marginal distributions are Gaussian:

$$\text{if } \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} k_{1,1} & k_{1,2} \\ k_{1,2}^T & k_{2,2} \end{pmatrix} \right) \quad \text{then } y_1 \sim \mathcal{N}(\mu_1, k_{1,1})$$

Multivariate Gaussian distribution

- Covariance captures correlation between dimensions
  - Much like regression captures correlation between observations!
- Marginal distributions are Gaussian:

$$\text{if } \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} k_{1,1} & k_{1,2} \\ k_{1,2}^T & k_{2,2} \end{pmatrix} \right) \quad \text{then } y_1 \sim \mathcal{N}(\mu_1, k_{1,1})$$

- Conditional distributions are Gaussian:

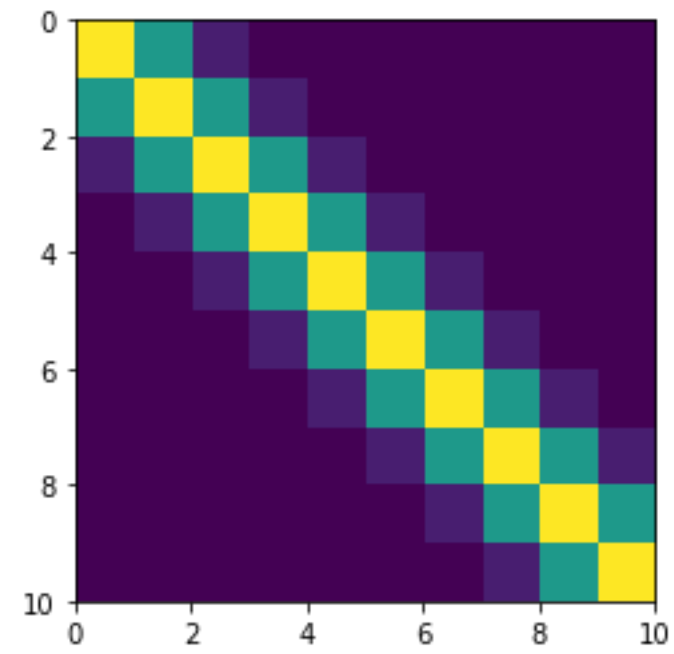
$$y_1 \,|\, y_2 \sim \mathcal{N}\left(\mu_1 + k_{1,2} k_{2,2}^{-1} \left(y_2 - \mu_2\right),\, k_{1,1} - k_{1,2} k_{2,2}^{-1} k_{1,2}^T\right)$$

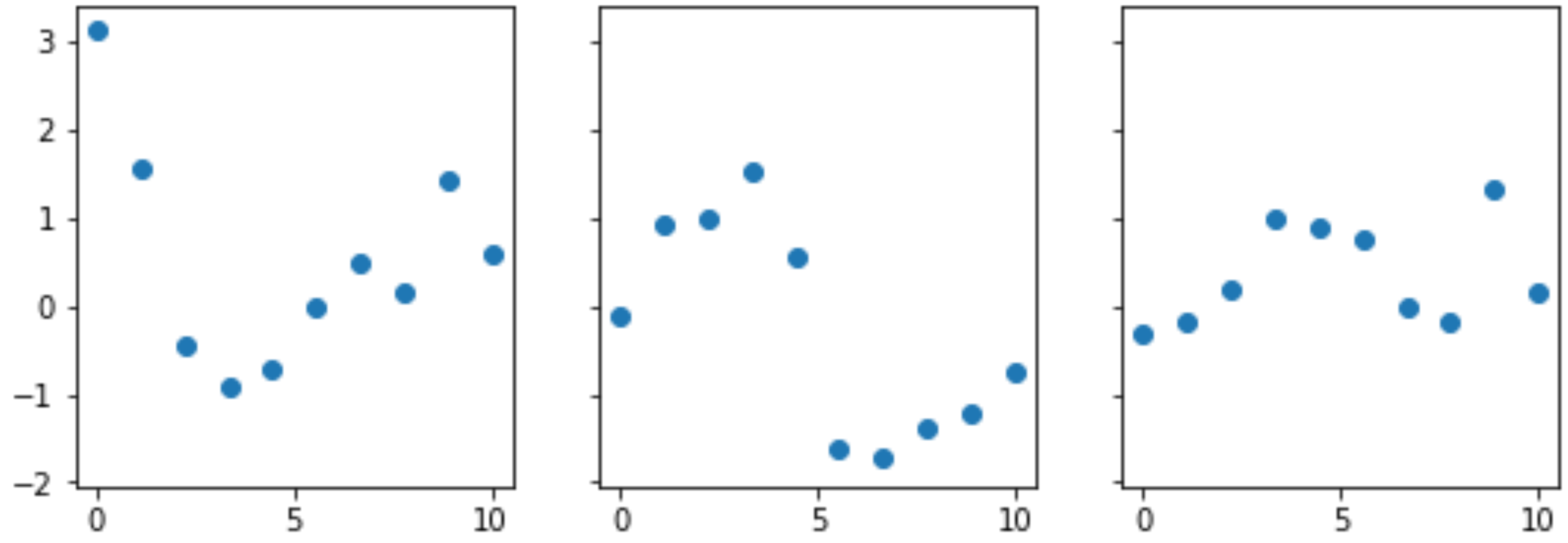
Multivariate Gaussian distribution

- Let each dimension of our Gaussian be a value of x

- Pick a covariance matrix K s.t. close values are highly correlated

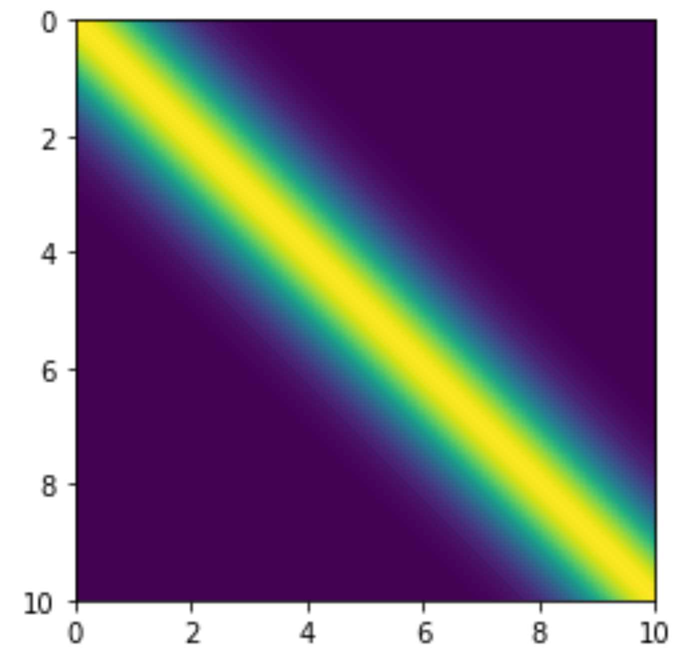
- Samples from Normal(0, K) look like functions:

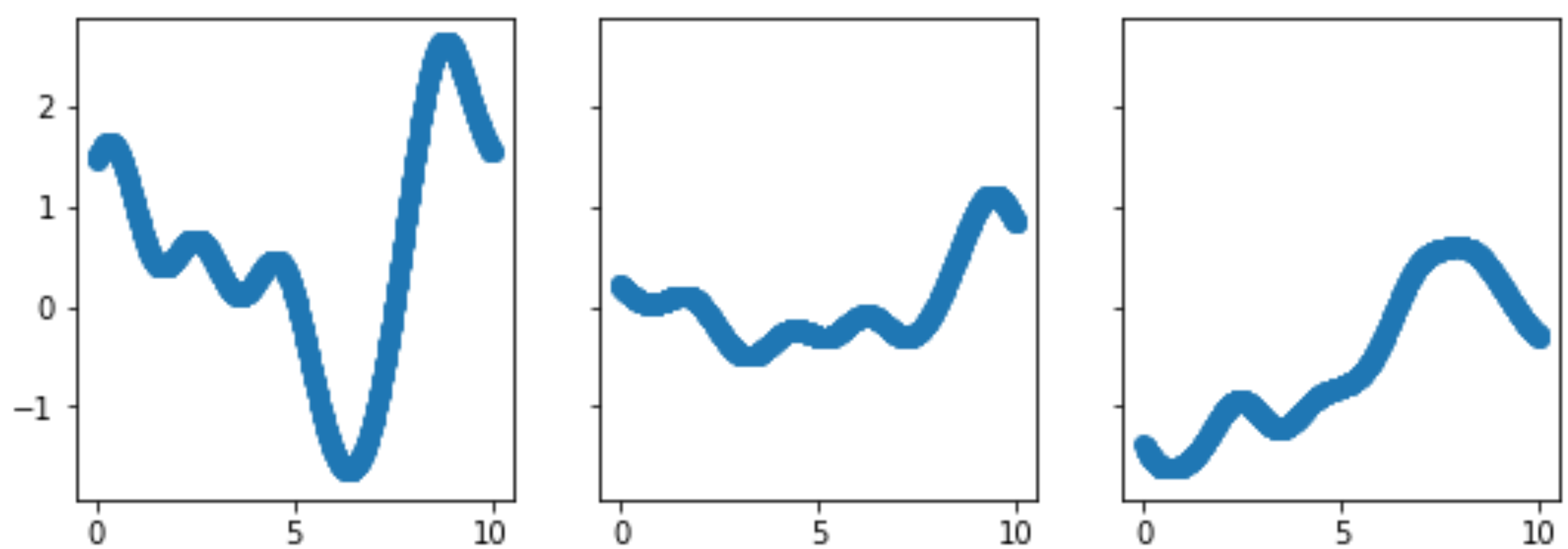




Multivariate Gaussian distribution

- Let each dimension of our Gaussian be a value of x
- Pick a covariance matrix K s.t. close values are highly correlated
- Samples from Normal(0, K) look like functions:



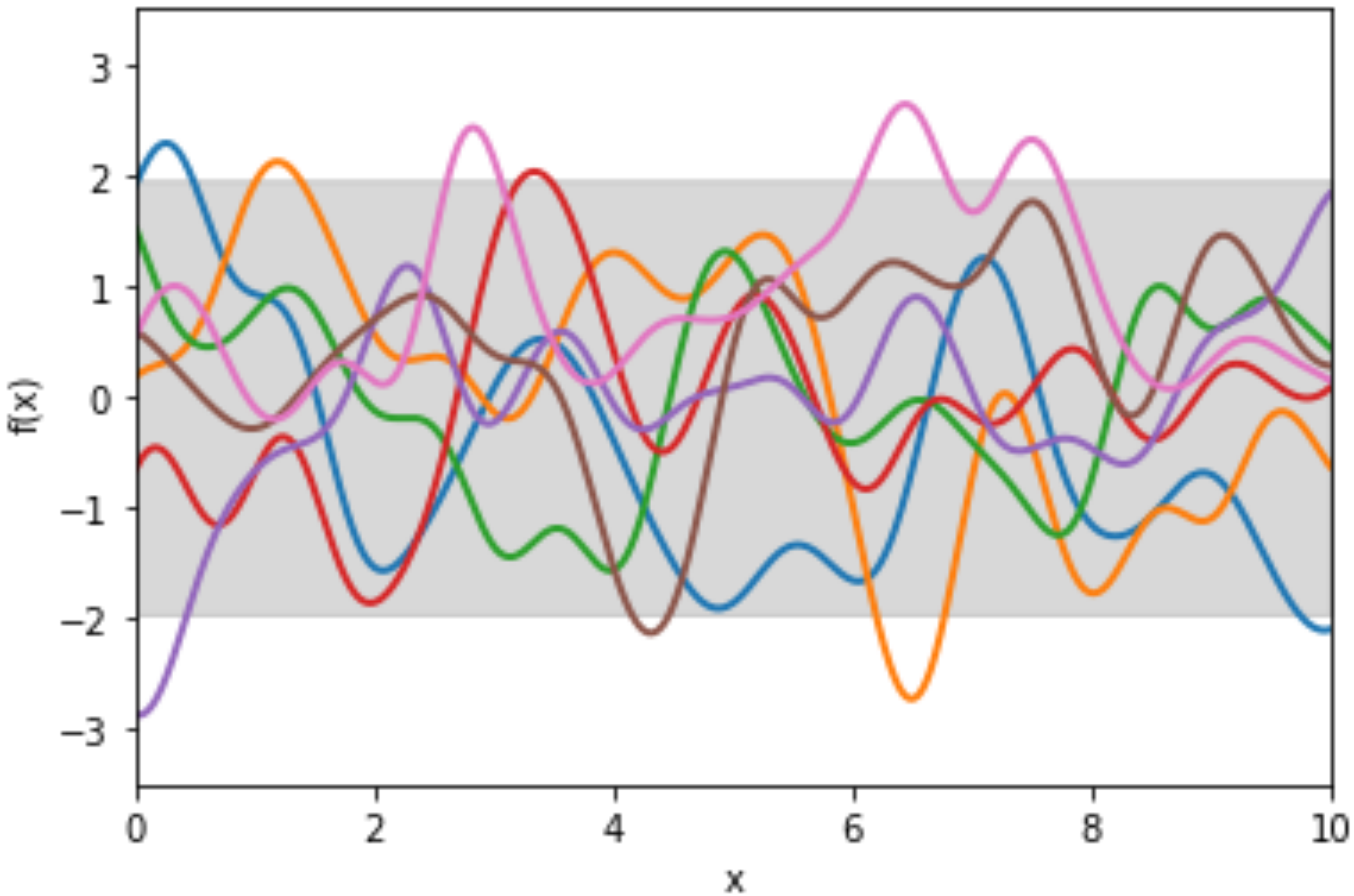


Gaussian process

- We can think of a Gaussian process (GP) as an "infinitely large" multivariate Gaussian.
- Mean and covariance replaced by functions $m(x)$, $k(x, x')$ that can be evaluated at any point.
- Marginal property means finite-dimensional instantiations are multivariate Gaussian:

$$\begin{bmatrix} y(x_1) \\ y(x_2) \\ y(x_3) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x_1) \\ m(x_2) \\ m(x_3) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & k(x_1, x_3) \\ k(x_2, x_1) & k(x_2, x_2) & k(x_2, x_3) \\ k(x_3, x_1) & k(x_3, x_2) & k(x_3, x_3) \end{bmatrix} \right)$$

Samples from a Gaussian process



Grey region = 95% credible interval

Posterior distribution

- Conditional property means that, conditioned on observed data (x, y), predictions at new locations x^* are Gaussian.

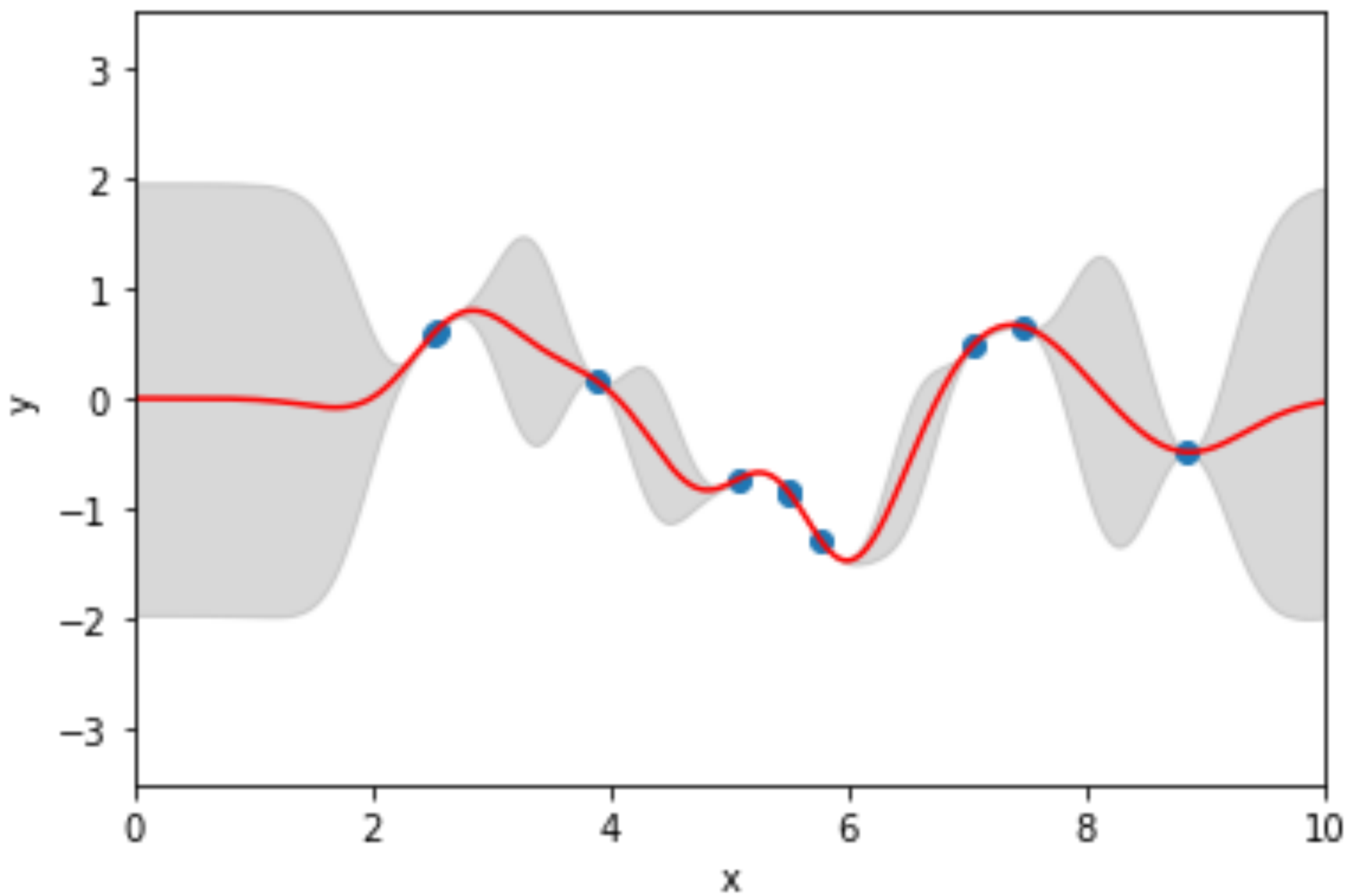
$$f(x^*) \,|\, x, f(x) \sim \mathcal{N}\left(\widetilde{m}(x^*), \widetilde{k}(x^*)\right)$$

$$\widetilde{m}(x^*) = k(x^*, x)k(x, x)^{-1}f$$

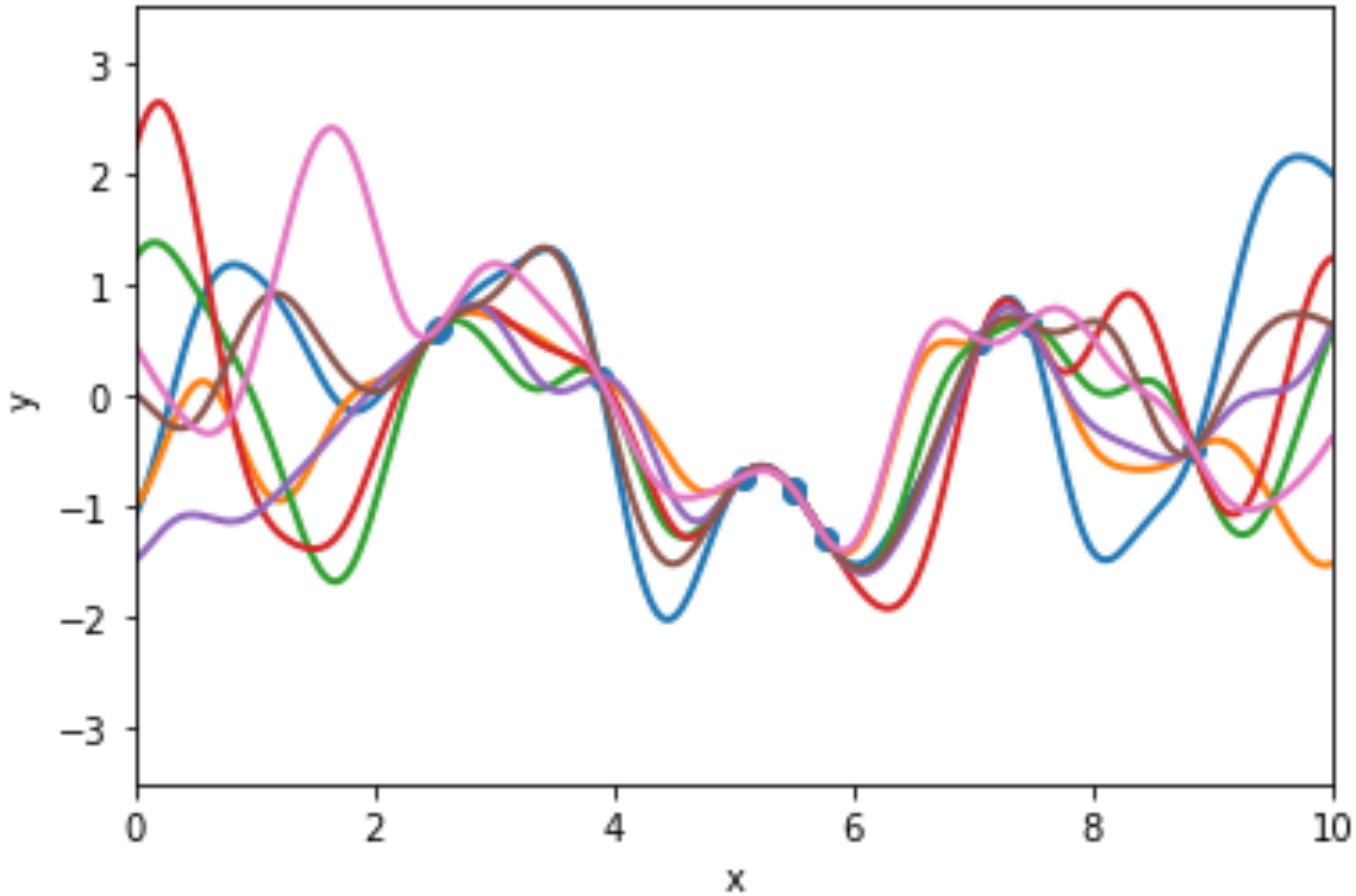
$$\widetilde{k}(x^*) = k(x^*, x^*) - k(x^*, x)k(x, x)^{-1}k(x, x^*)$$

- i.e., posterior is a Gaussian process

Posterior distribution



Posterior distribution



Lab 3

- `github.com/sinead/DS32019`
- For our final lab, we will implement Gaussian process regression

Closing remarks

- Bayesian methods are appropriate when…
  - We care about uncertainty
  - We have information we can incorporate into priors - either explicitly, or via sharing information between parts of the model.
- Inference is typically slower than optimization-based methods, but we have a lot of tools
  - We looked at Laplace approximations and Variational Bayes
  - Other tools exist such as MCMC
  - Software such as Tensorflow Probability and STAN allow us to automate inference