Bayesian Regression

MLAI: Week 5

Neil D. Lawrence

Department of Computer Science
Sheffield University

29th October 2013

Outline

Prior Distribution

- Bayesian inference requires a prior on the parameters.
- The prior represents your belief *before* you see the data of the likely value of the parameters.
- For linear regression, consider a Gaussian prior on the intercept:

$$c \sim \mathcal{N}(0, \alpha_1)$$

Posterior Distribution

- Posterior distribution is found by combining the prior with the likelihood.
- Posterior distribution is your belief *after* you see the data of the likely value of the parameters.
- The posterior is found through **Bayes' Rule**

$$p(c|y) = \frac{p(y|c)p(c)}{p(y)}$$

Bayes Update

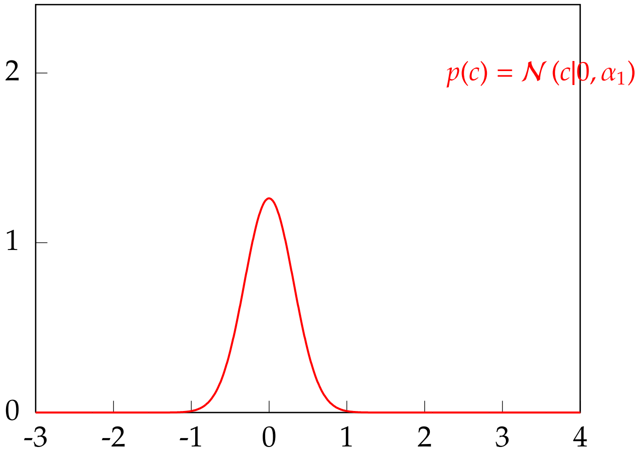


Figure : A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Bayes Update

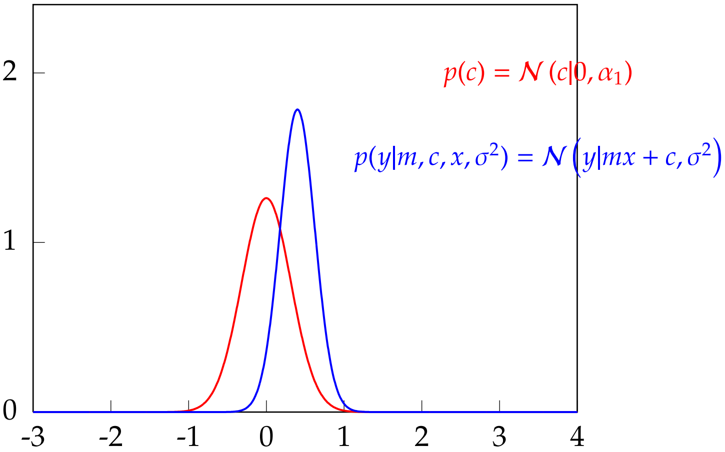


Figure : A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Bayes Update

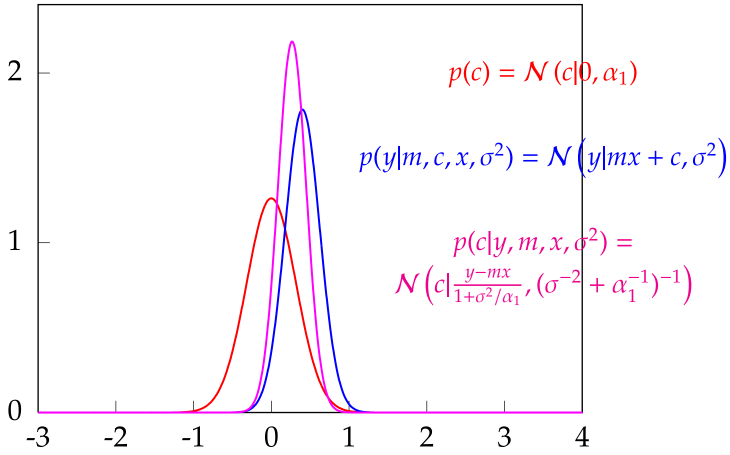


Figure : A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Stages to Derivation of the Posterior

- Multiply likelihood by prior
  - they are "exponentiated quadratics", the answer is always also an exponentiated quadratic because $\exp(a^2)\exp(b^2) = \exp(a^2 + b^2)$.
- Complete the square to get the resulting density in the form of a Gaussian.
- Recognise the mean and (co)variance of the Gaussian. This is the estimate of the posterior.

Main Trick

$$p(c) = \frac{1}{\sqrt{2\pi\alpha_1}} \exp\left(-\frac{1}{2\alpha_1}c^2\right)$$

$$p(\mathbf{y}|\mathbf{x}, c, m, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - mx_i - c)^2\right)$$

Main Trick

$$p(c) = \frac{1}{\sqrt{2\pi\alpha_1}} \exp\left(-\frac{1}{2\alpha_1}c^2\right)$$

$$p(\mathbf{y}|\mathbf{x}, c, m, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - mx_i - c)^2\right)$$

$$p(c|\mathbf{y}, \mathbf{x}, m, \sigma^2) = \frac{p(\mathbf{y}|\mathbf{x}, c, m, \sigma^2)p(c)}{p(\mathbf{y}|\mathbf{x}, m, \sigma^2)}$$

Main Trick

$$p(c) = \frac{1}{\sqrt{2\pi\alpha_1}} \exp\left(-\frac{1}{2\alpha_1}c^2\right)$$

$$p(\mathbf{y}|\mathbf{x}, c, m, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - mx_i - c)^2\right)$$

$$p(c|\mathbf{y}, \mathbf{x}, m, \sigma^2) = \frac{p(\mathbf{y}|\mathbf{x}, c, m, \sigma^2)p(c)}{\int p(\mathbf{y}|\mathbf{x}, c, m, \sigma^2)p(c)\mathrm{d}c}$$

Main Trick

$$p(c) = \frac{1}{\sqrt{2\pi\alpha_1}} \exp\left(-\frac{1}{2\alpha_1}c^2\right)$$

$$p(\mathbf{y}|\mathbf{x}, c, m, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - mx_i - c)^2\right)$$

$$p(c|\mathbf{y}, \mathbf{x}, m, \sigma^2) \propto p(\mathbf{y}|\mathbf{x}, c, m, \sigma^2)p(c)$$

$$\begin{aligned}\log p(c|\mathbf{y},\mathbf{x},m,\sigma^2) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - c - mx_i)^2 - \frac{1}{2\alpha_1}c^2 + \text{const}\\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - mx_i)^2 - \left(\frac{n}{2\sigma^2} + \frac{1}{2\alpha_1}\right)c^2\\ &\quad + c\frac{\sum_{i=1}^{n}(y_i - mx_i)}{\sigma^2},\end{aligned}$$

complete the square of the quadratic form to obtain

$$\log p(c|\mathbf{y},\mathbf{x},m,\sigma^2) = -\frac{1}{2\tau^2}(c-\mu)^2 + \text{const},$$

where $\tau^2 = \left(n\sigma^{-2} + \alpha_1^{-1}\right)^{-1}$ and $\mu = \frac{\tau^2}{\sigma^2}\sum_{n=1}^{N}(y_i - mx_i)$.

The Joint Density

- Really want to know the *joint* posterior density over the parameters c *and* m.
- Could now integrate out over m, but it's easier to consider the multivariate case.

Two Dimensional Gaussian

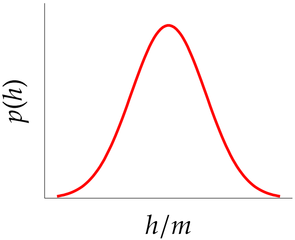
- Consider height, h/m and weight, w/kg.
- Could sample height from a distribution:

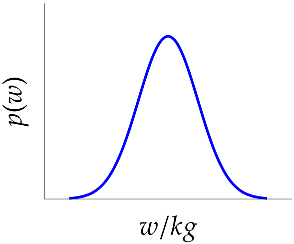
$$p(h) \sim \mathcal{N}(1.7, 0.0225)$$

- And similarly weight:

$$p(w) \sim \mathcal{N}(75, 36)$$

Height and Weight Models

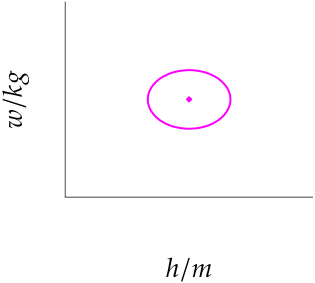




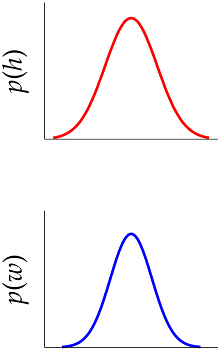
Gaussian distributions for height and weight.

Sampling Two Dimensional Variables

Joint Distribution



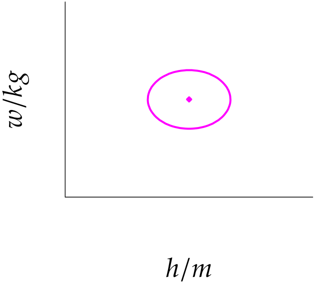
Marginal Distributions



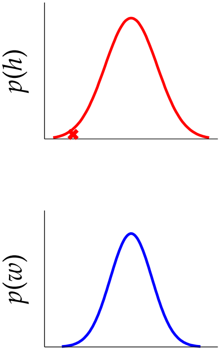
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



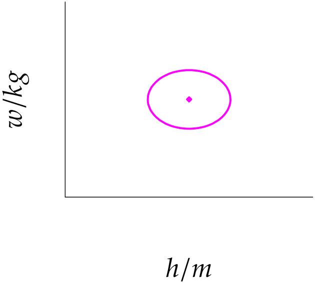
Marginal Distributions



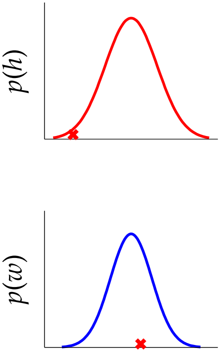
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



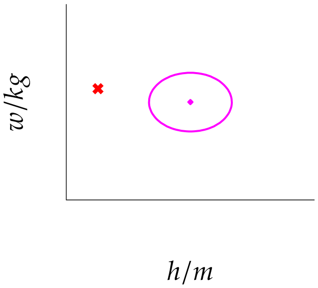
Marginal Distributions



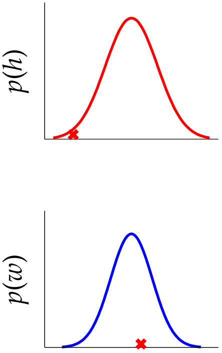
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



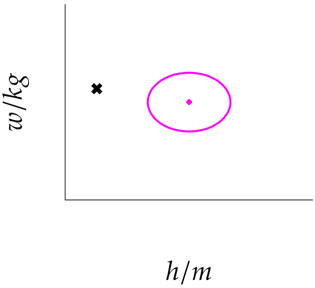
Marginal Distributions



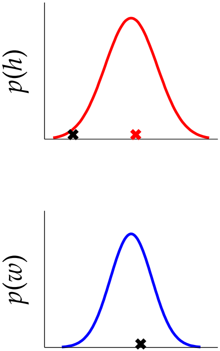
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



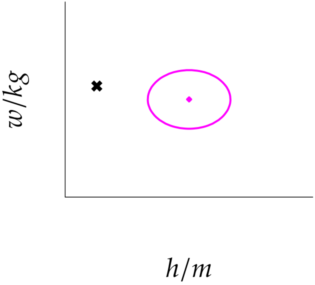
Marginal Distributions



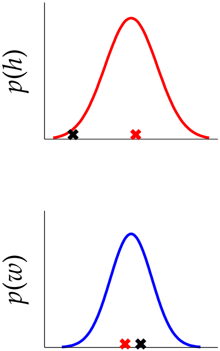
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



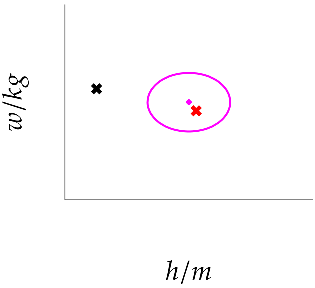
Marginal Distributions



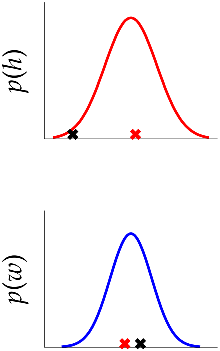
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



Marginal Distributions

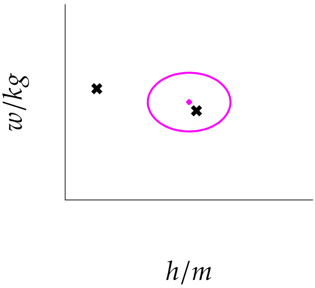


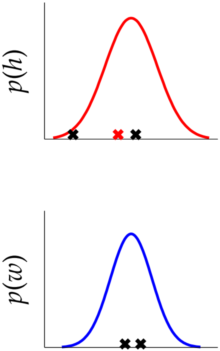
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



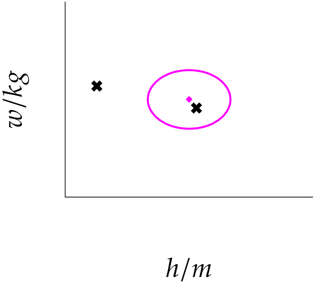


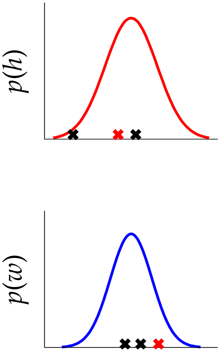
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution



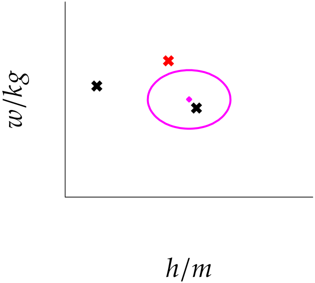


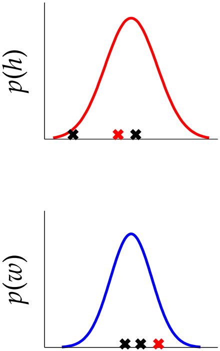
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution

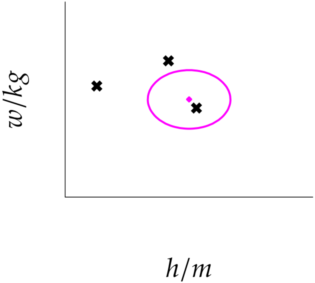




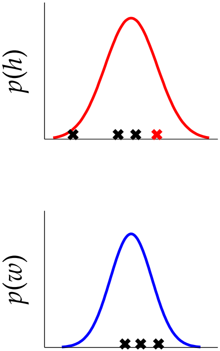
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



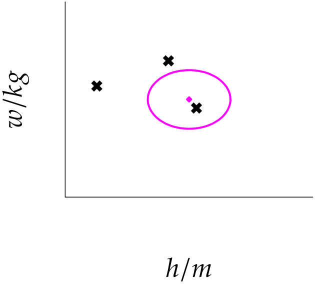
Marginal Distributions



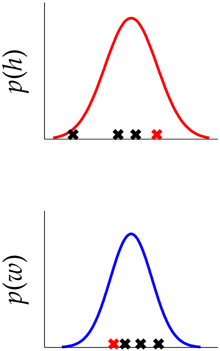
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



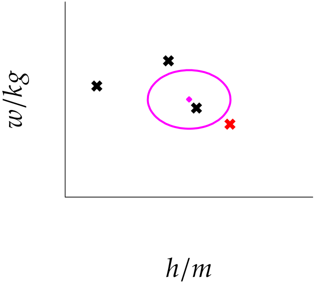
Marginal Distributions



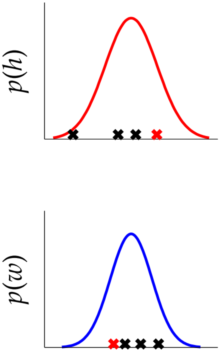
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



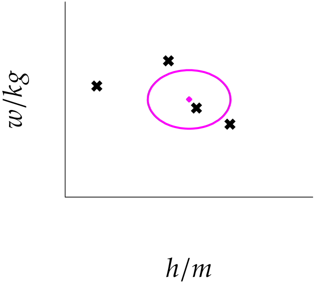
Marginal Distributions



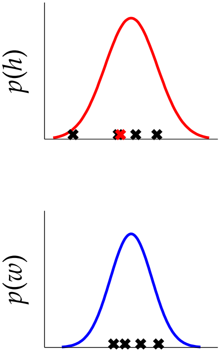
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution

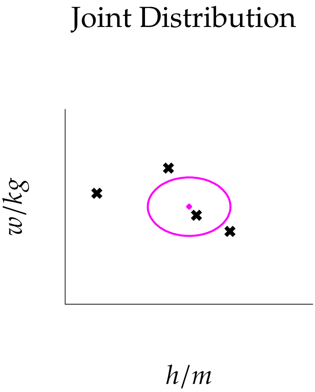


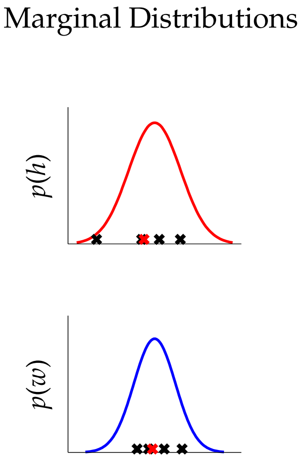
Marginal Distributions



Samples of height and weight

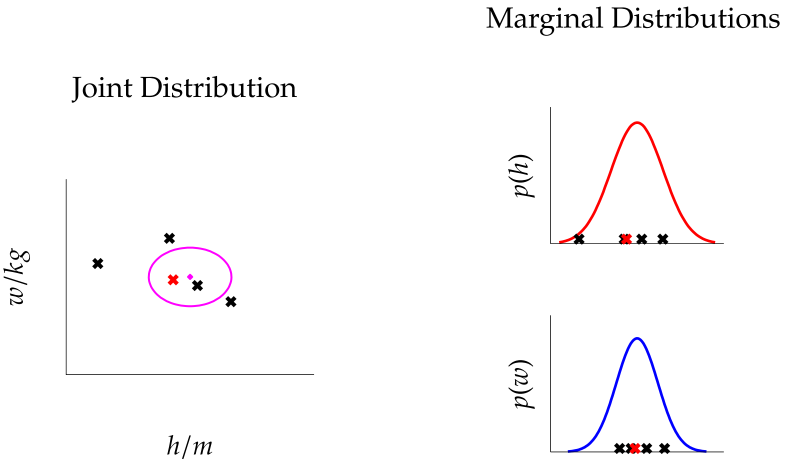
Sampling Two Dimensional Variables





Samples of height and weight

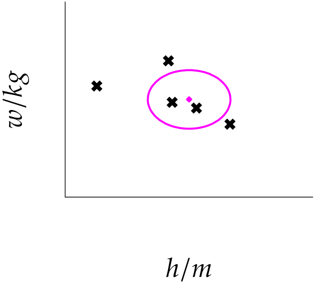
Sampling Two Dimensional Variables



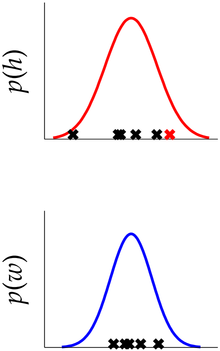
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



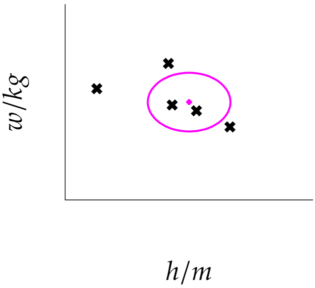
Marginal Distributions



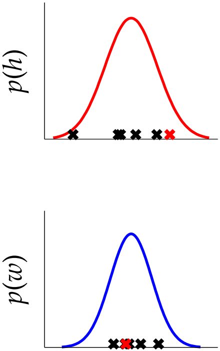
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



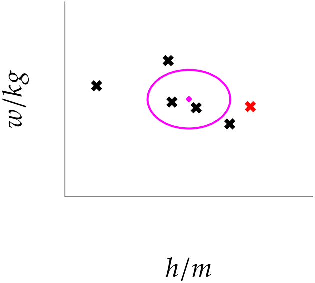
Marginal Distributions



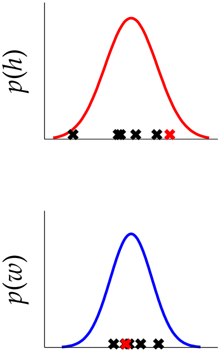
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



Marginal Distributions

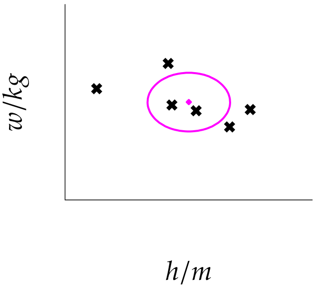


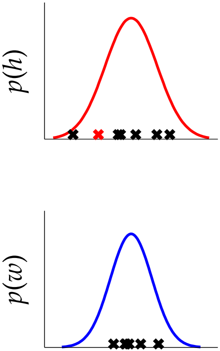
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution

Marginal Distributions



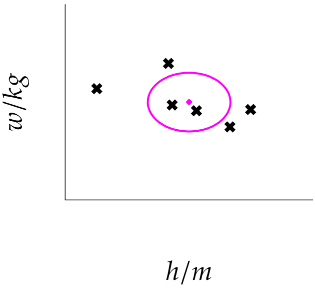


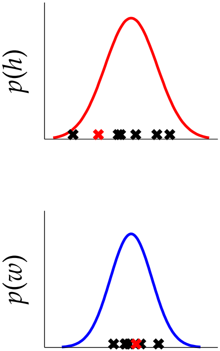
Samples of height and weight

Sampling Two Dimensional Variables

Marginal Distributions

Joint Distribution

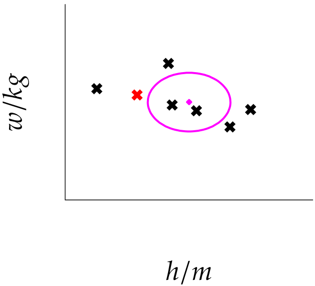




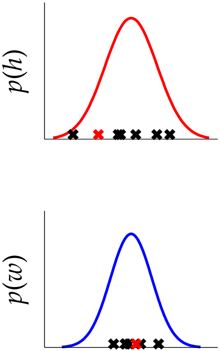
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



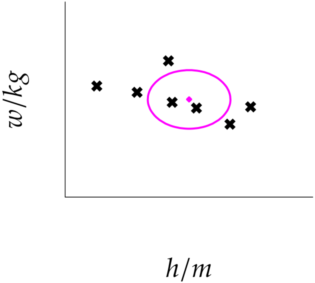
Marginal Distributions



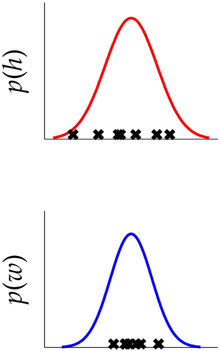
Samples of height and weight

Sampling Two Dimensional Variables

Joint Distribution



Marginal Distributions



Samples of height and weight

Independence Assumption

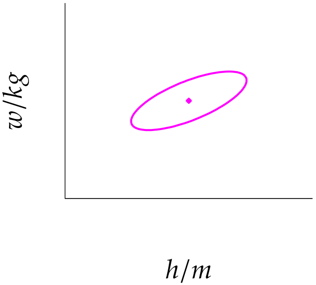
- This assumes height and weight are independent.

$$p(h, w) = p(h)p(w)$$

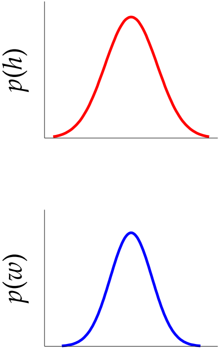
- In reality they are dependent (body mass index) $= \frac{w}{h^2}$.

Sampling Two Dimensional Variables

Joint Distribution



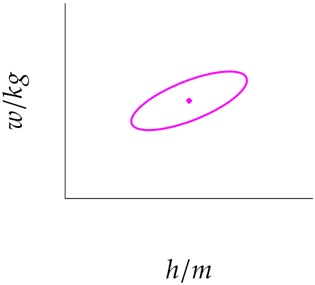
Marginal Distributions

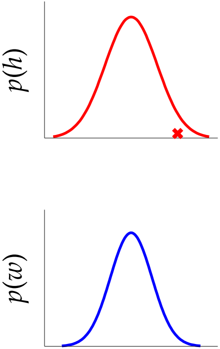


Sampling Two Dimensional Variables

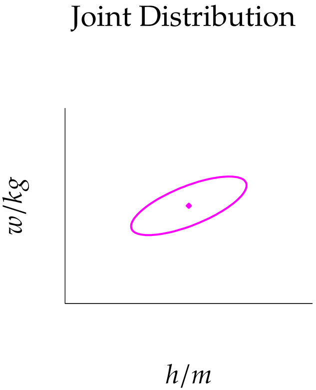
Marginal Distributions

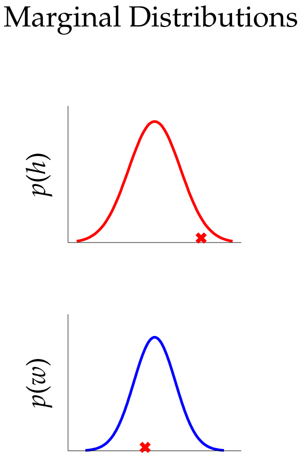
Joint Distribution



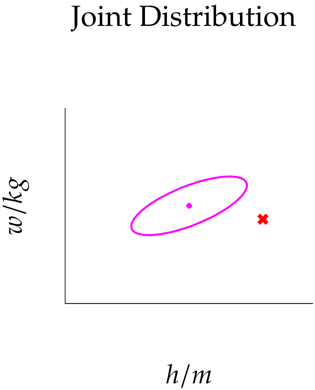


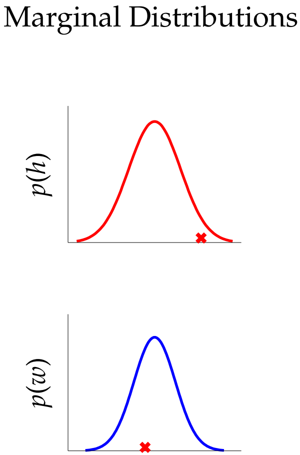
Sampling Two Dimensional Variables





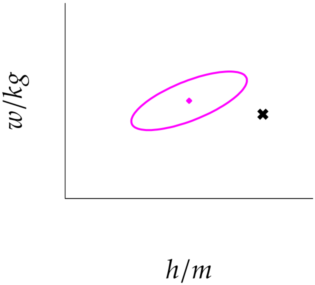
Sampling Two Dimensional Variables



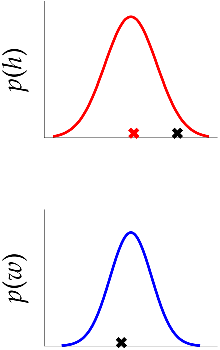


Sampling Two Dimensional Variables

Joint Distribution

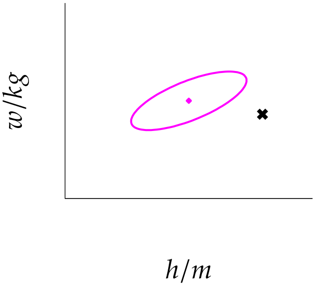


Marginal Distributions

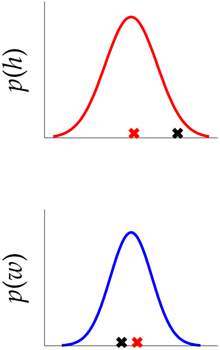


Sampling Two Dimensional Variables

Joint Distribution

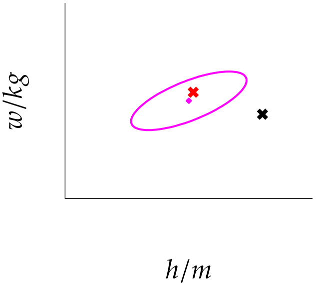


Marginal Distributions

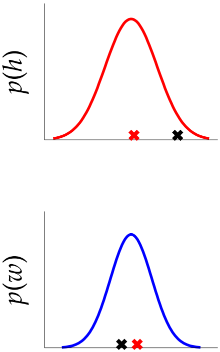


Sampling Two Dimensional Variables

Joint Distribution

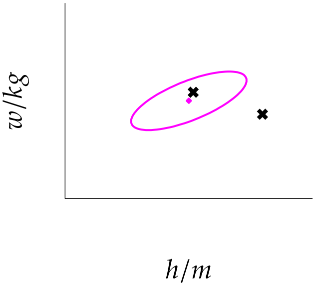


Marginal Distributions

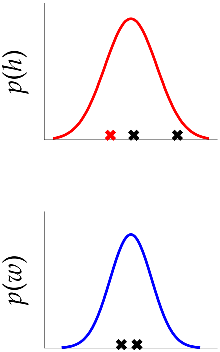


Sampling Two Dimensional Variables

Joint Distribution



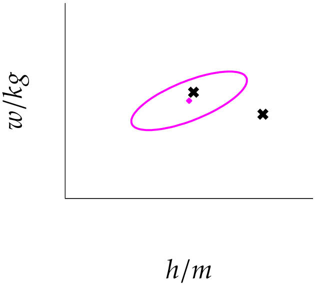
Marginal Distributions

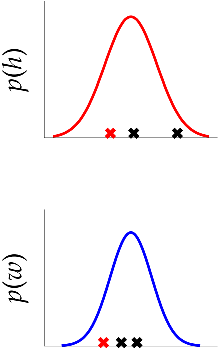


Sampling Two Dimensional Variables

Marginal Distributions

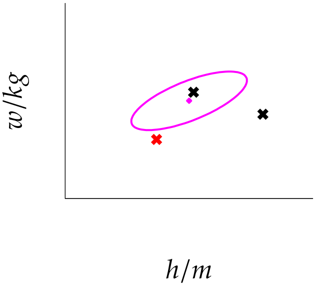
Joint Distribution



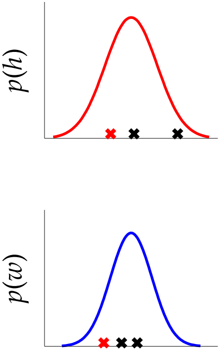


Sampling Two Dimensional Variables

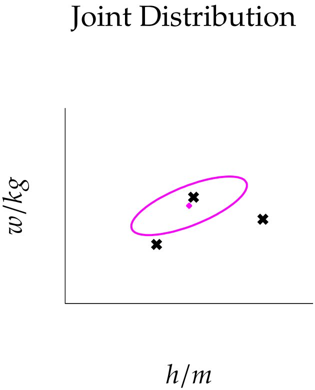
Joint Distribution

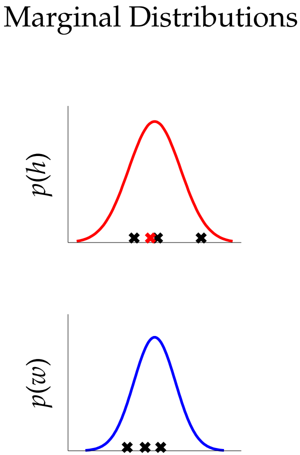


Marginal Distributions



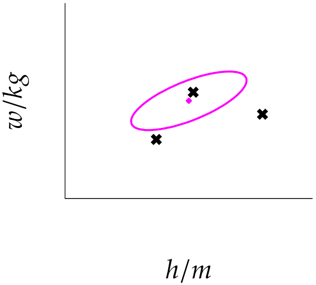
Sampling Two Dimensional Variables



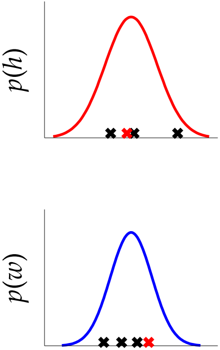


Sampling Two Dimensional Variables

Joint Distribution

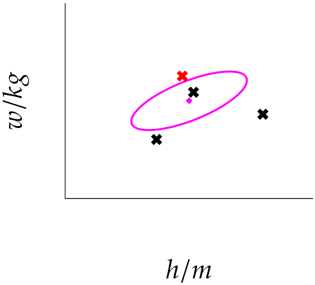


Marginal Distributions

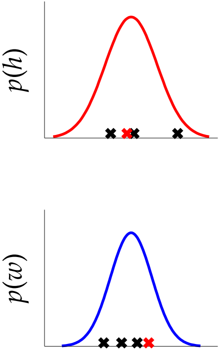


Sampling Two Dimensional Variables

Joint Distribution

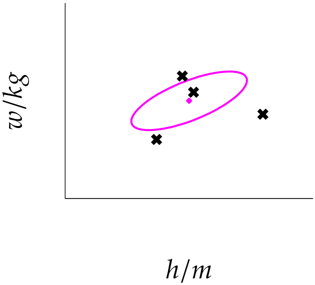


Marginal Distributions

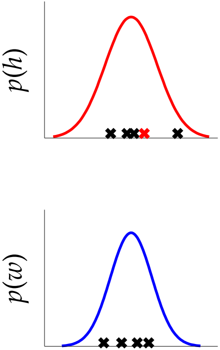


Sampling Two Dimensional Variables

Joint Distribution

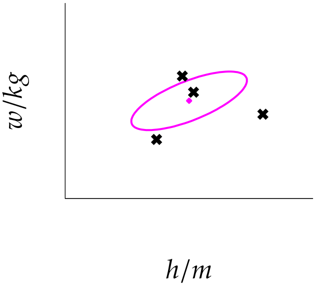


Marginal Distributions

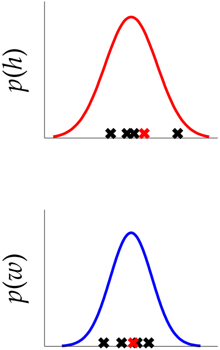


Sampling Two Dimensional Variables

Joint Distribution

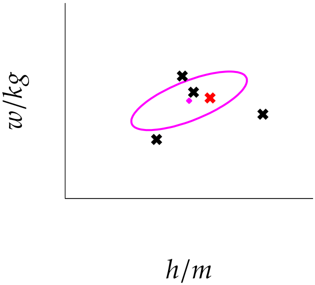


Marginal Distributions

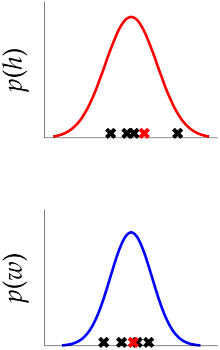


Sampling Two Dimensional Variables

Joint Distribution

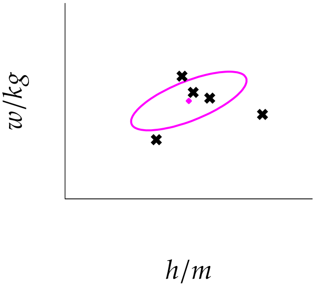


Marginal Distributions

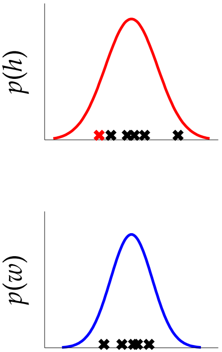


Sampling Two Dimensional Variables

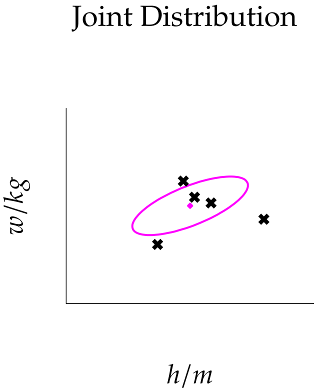
Joint Distribution

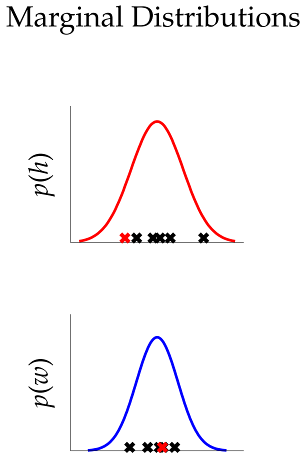


Marginal Distributions

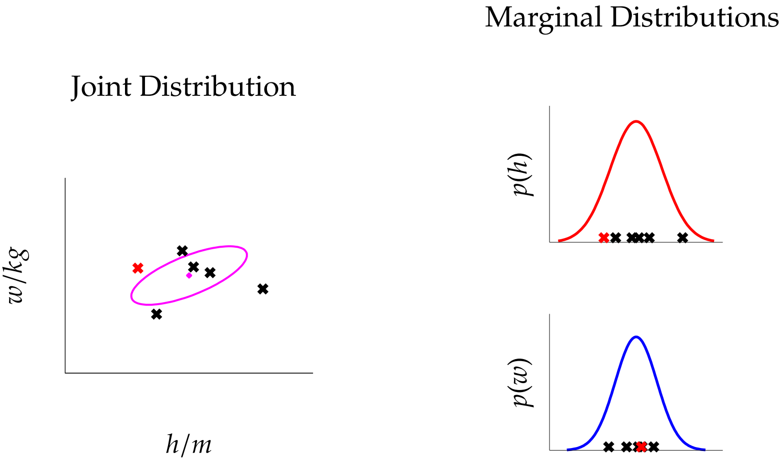


Sampling Two Dimensional Variables



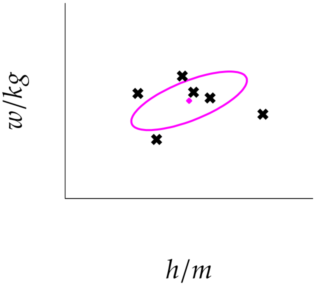


Sampling Two Dimensional Variables

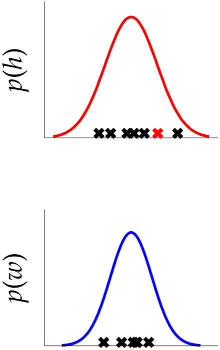


Sampling Two Dimensional Variables

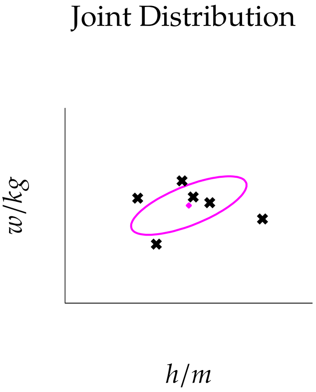
Joint Distribution

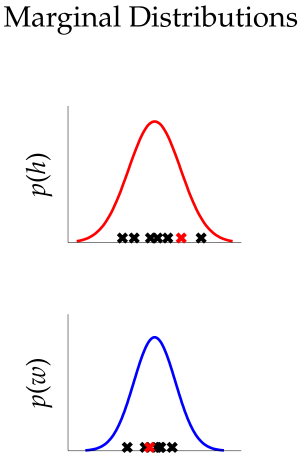


Marginal Distributions



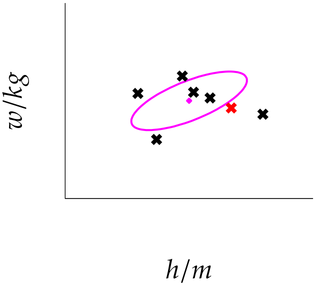
Sampling Two Dimensional Variables



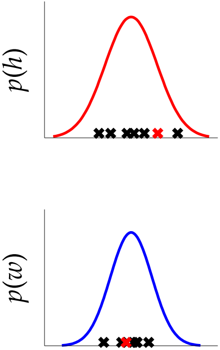


Sampling Two Dimensional Variables

Joint Distribution

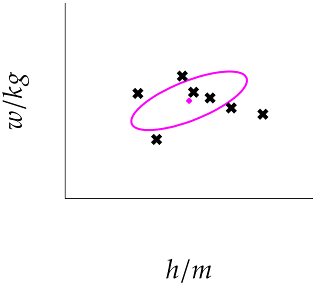


Marginal Distributions

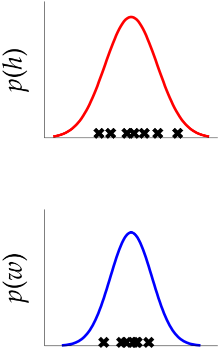


Sampling Two Dimensional Variables

Joint Distribution



Marginal Distributions



Independent Gaussians

$$p(w, h) = p(w)p(h)$$

Independent Gaussians

$$p(w,h) = \frac{1}{\sqrt{2\pi\sigma_1^2}\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{1}{2}\left(\frac{(w-\mu_1)^2}{\sigma_1^2} + \frac{(h-\mu_2)^2}{\sigma_2^2}\right)\right)$$

Independent Gaussians

$$p(w,h) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2}}\exp\left(-\frac{1}{2}\left(\begin{bmatrix}w\\h\end{bmatrix}-\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}\right)^\top\begin{bmatrix}\sigma_1^2 & 0\\0 & \sigma_2^2\end{bmatrix}^{-1}\left(\begin{bmatrix}w\\h\end{bmatrix}-\begin{bmatrix}\mu_1\\\mu_2\end{bmatrix}\right)\right)$$

Independent Gaussians

$$p(\mathbf{y}) = \frac{1}{2\pi\left|\mathbf{D}\right|}\exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\top}\mathbf{D}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix $\mathbf{R}$.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{D}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{D}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right)$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix **R**.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{D}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{R}^\top\mathbf{y} - \mathbf{R}^\top\boldsymbol{\mu})^\top \mathbf{D}^{-1}(\mathbf{R}^\top\mathbf{y} - \mathbf{R}^\top\boldsymbol{\mu})\right)$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix **R**.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{D}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\top}\mathbf{R}\mathbf{D}^{-1}\mathbf{R}^{\top}(\mathbf{y}-\boldsymbol{\mu})\right)$$

this gives a covariance matrix:

$$\mathbf{C}^{-1} = \mathbf{R}\mathbf{D}^{-1}\mathbf{R}^{\top}$$

Correlated Gaussian

Form correlated from original by rotating the data space using matrix **R**.

$$p(\mathbf{y}) = \frac{1}{2\pi \left|\mathbf{C}\right|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\top}\mathbf{C}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$$

this gives a covariance matrix:

$$\mathbf{C} = \mathbf{R}\mathbf{D}\mathbf{R}^{\top}$$

Outline

Multivariate Regression Likelihood

- Noise corrupted data point

$$y_i = \mathbf{w}^\top \mathbf{x}_{i,:} + \epsilon_i$$

Multivariate Regression Likelihood

- Noise corrupted data point

$$y_i = \mathbf{w}^\top \mathbf{x}_{i,:} + \epsilon_i$$

- Multivariate regression likelihood:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mathbf{w}^\top \mathbf{x}_{i,:}\right)^2\right)$$

Multivariate Regression Likelihood

- Noise corrupted data point

$$y_i = \mathbf{w}^\top \mathbf{x}_{i,:} + \epsilon_i$$

- Multivariate regression likelihood:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{w}^\top \mathbf{x}_{i,:}\right)^2\right)$$

- Now use a multivariate Gaussian prior:

$$p(\mathbf{w}) = \frac{1}{(2\pi\alpha)^{\frac{p}{2}}} \exp\left(-\frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w}\right)$$

Posterior Density

- Once again we want to know the posterior:

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})$$

- And we can compute by completing the square.

Posterior Density

- Once again we want to know the posterior:

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})$$

- And we can compute by completing the square.

$$\log p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n} y_i \mathbf{x}_{i,:}^\top \mathbf{w} - \frac{1}{2\sigma^2}\sum_{i=1}^{n} \mathbf{w}^\top \mathbf{x}_{i,:}\mathbf{x}_{i,:}^\top \mathbf{w} - \frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w} + \text{const.}$$

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}\left(\mathbf{w}|\boldsymbol{\mu}_w, \mathbf{C}_w\right)$$

$$\mathbf{C}_w = (\sigma^{-2}\mathbf{X}^\top\mathbf{X} + \alpha^{-1})^{-1} \text{ and } \boldsymbol{\mu}_w = \mathbf{C}_w \sigma^{-2}\mathbf{X}^\top \mathbf{y}$$

Bayesian vs Maximum Likelihood

- Note the similarity between posterior mean

$$\boldsymbol{\mu}_w = (\sigma^{-2}\mathbf{X}^\top\mathbf{X} + \alpha^{-1})^{-1}\sigma^{-2}\mathbf{X}^\top\mathbf{y}$$

- and Maximum likelihood solution

$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Marginal Likelihood is Computed as Normalizer

$$p(\mathbf{w}|\mathbf{y}, \mathbf{X})p(\mathbf{y}|\mathbf{X}) = p(\mathbf{y}|\mathbf{w}, \mathbf{X})p(\mathbf{w})$$

Marginal Likelihood

- Can compute the marginal likelihood as:

$$p(\mathbf{y}|\mathbf{X}, \alpha, \sigma) = \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \alpha\mathbf{X}\mathbf{X}^\top + \sigma^2\mathbf{I}\right)$$

Reading

- Section 2.3 of Bishop up to top of pg 85 (multivariate Gaussians).
- Section 3.3 of Bishop up to 159 (pg 152–159).

Outline

Univariate Bayesian Linear Regression

Multivariate Bayesian Linear Regression

Bayesian Polynomials

Revisit Olympics Data

- Use Bayesian approach on olympics data with polynomials.
- Choose a prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha\mathbf{I})$ with $\alpha = 1$.
- Choose noise variance $\sigma^2 = 0.01$

Sampling the Prior

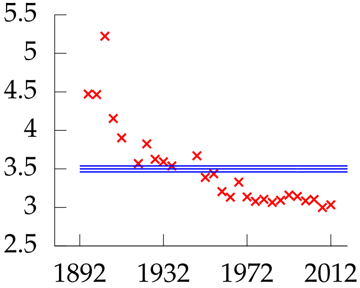
- Always useful to perform a 'sanity check' and sample from the prior before observing the data.
- Since $\mathbf{y} = \mathbf{\Phi w} + \boldsymbol{\epsilon}$ just need to sample

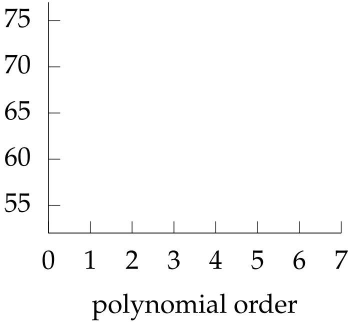
$$w \sim \mathcal{N}\left(0, \alpha\right)$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\right)$$

with $\alpha = 1$ and $\boldsymbol{\epsilon} = 0.01$.

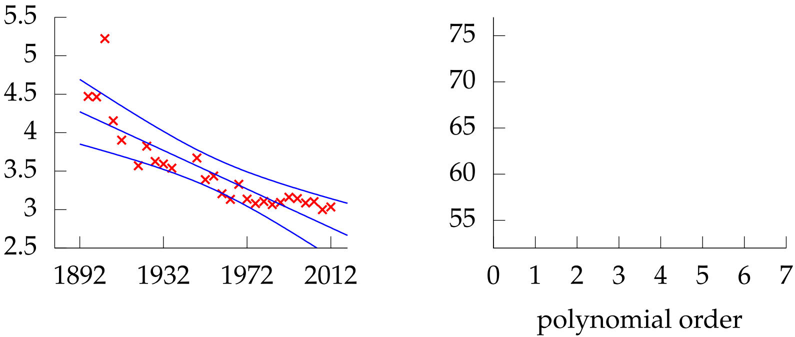
Polynomial Fits to Olympics Data





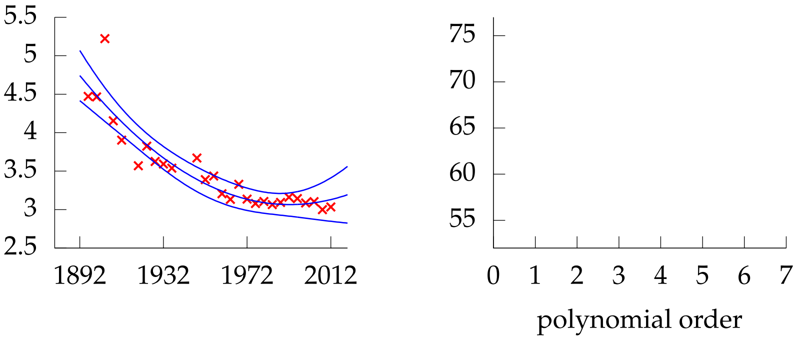
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 0, model error 29.757, $\sigma^2 = 0.286$, $\sigma = 0.535$.

Polynomial Fits to Olympics Data



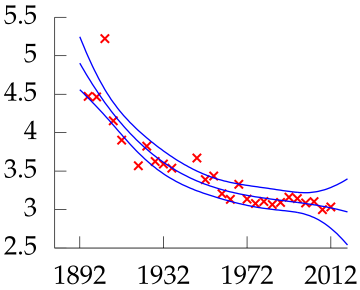
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 1, model error 14.942, $\sigma^2 = 0.0749$, $\sigma = 0.274$.

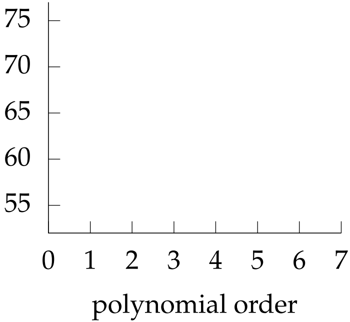
Polynomial Fits to Olympics Data



Left: fit to data, *Right*: marginal log likelihood. Polynomial order 2, model error 9.7206, $\sigma^2 = 0.0427$, $\sigma = 0.207$.

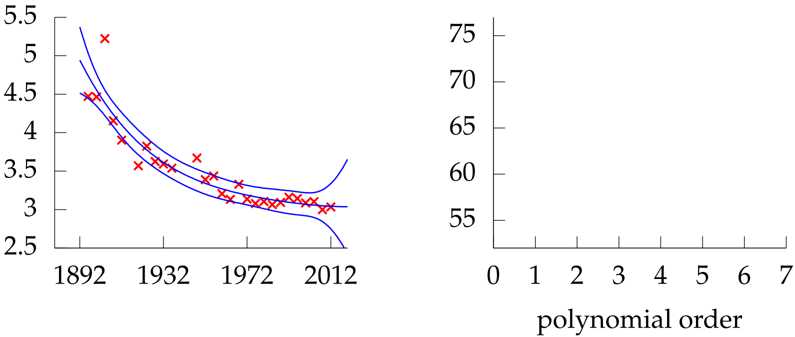
Polynomial Fits to Olympics Data





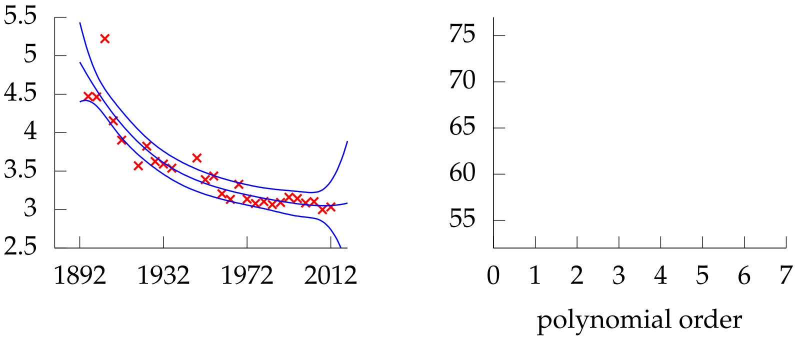
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 3, model error 10.416, $\sigma^2 = 0.0402$, $\sigma = 0.200$.

Polynomial Fits to Olympics Data



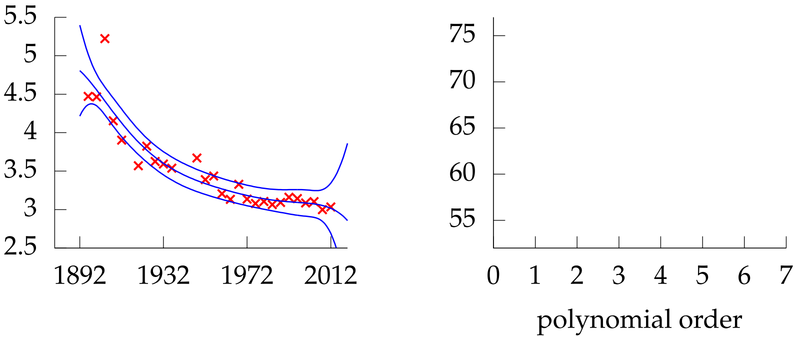
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 4, model error 11.34, $\sigma^2 = 0.0401$, $\sigma = 0.200$.

Polynomial Fits to Olympics Data



Left: fit to data, *Right*: marginal log likelihood. Polynomial order 5, model error 11.986, $\sigma^2 = 0.0399$, $\sigma = 0.200$.

Polynomial Fits to Olympics Data

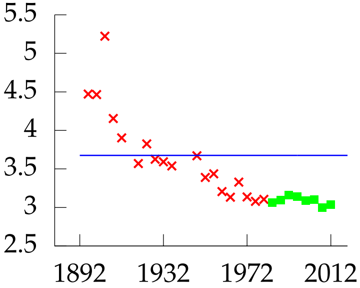


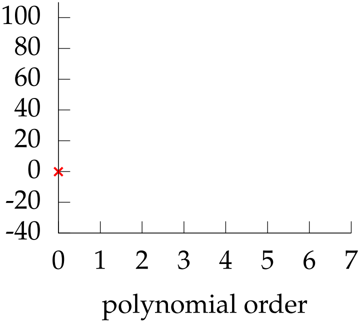
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 6, model error 12.369, $\sigma^2 = 0.0384$, $\sigma = 0.196$.

Model Fit

- Marginal likelihood doesn't always increase as model order increases.
- Bayesian model always has 2 parameters, regardless of how many basis functions (and here we didn't even fit them).
- Maximum likelihood model over fits through increasing number of parameters.
- Revisit maximum likelihood solution with validation set.

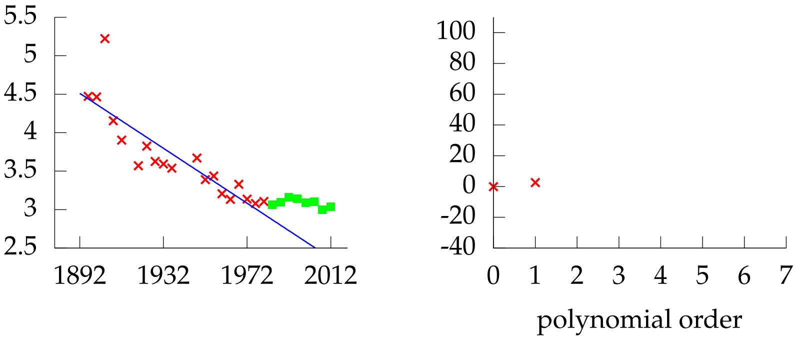
Recall: Validation Set for Maximum Likelihood





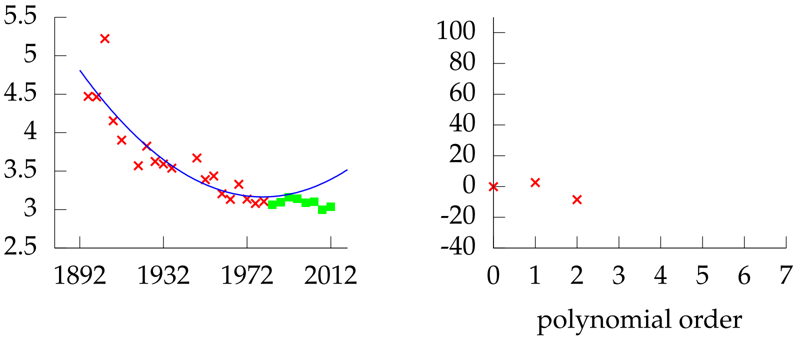
Left: fit to data, *Right*: model error. Polynomial order 0, training error -1.8774, validation error -0.13132, $\sigma^2 = 0.302$, $\sigma = 0.549$.

Recall: Validation Set for Maximum Likelihood



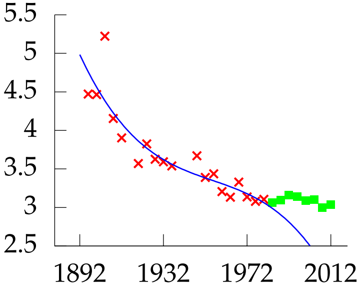
Left: fit to data, *Right*: model error. Polynomial order 1, training error -15.325, validation error 2.5863, $\sigma^2 = 0.0733$, $\sigma = 0.271$.

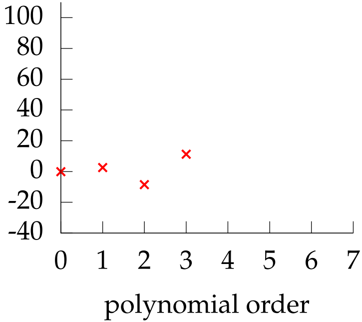
Recall: Validation Set for Maximum Likelihood



Left: fit to data, *Right*: model error. Polynomial order 2, training error -17.579, validation error -8.4831, $\sigma^2 = 0.0578$, $\sigma = 0.240$.

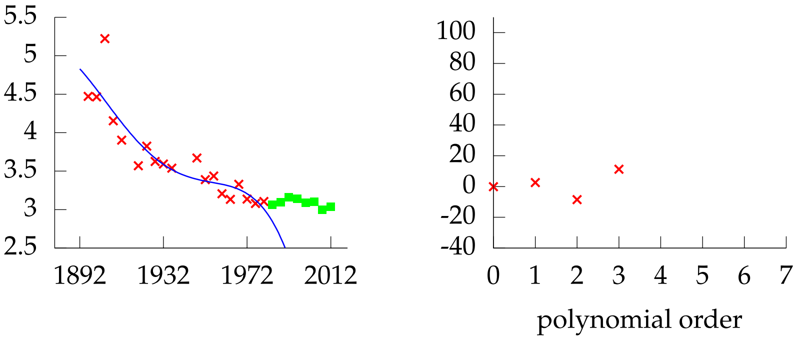
Recall: Validation Set for Maximum Likelihood





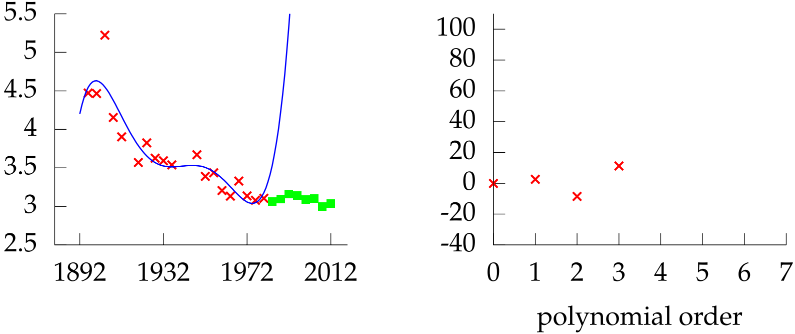
Left: fit to data, *Right*: model error. Polynomial order 3, training error -18.064, validation error 11.27, $\sigma^2 = 0.0549$, $\sigma = 0.234$.

Recall: Validation Set for Maximum Likelihood



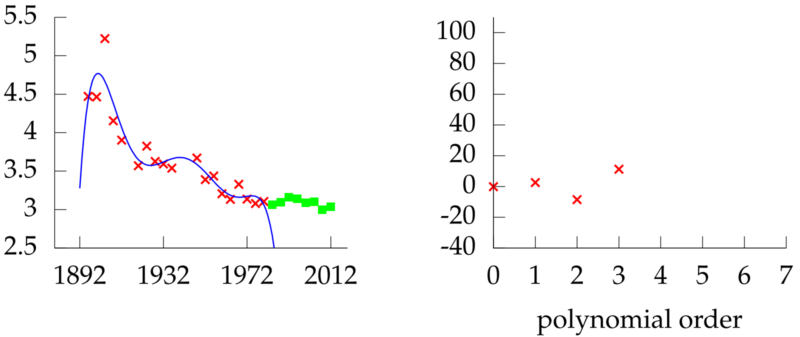
Left: fit to data, *Right*: model error. Polynomial order 4, training error -18.245, validation error 232.92, $\sigma^2 = 0.0539$, $\sigma = 0.232$.

Recall: Validation Set for Maximum Likelihood



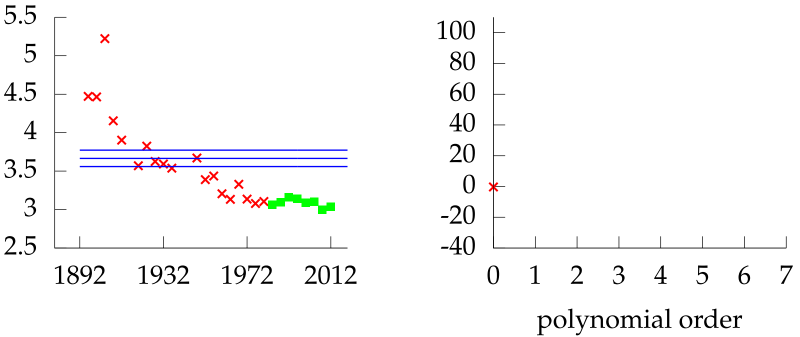
Left: fit to data, *Right*: model error. Polynomial order 5, training error -20.471, validation error 9898.1, $\sigma^2 = 0.0426$, $\sigma = 0.207$.

Recall: Validation Set for Maximum Likelihood



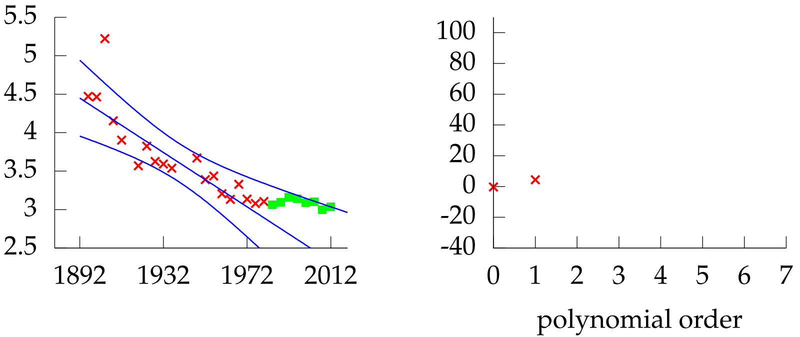
Left: fit to data, *Right*: model error. Polynomial order 6, training error -22.881, validation error 67775, $\sigma^2 = 0.0331$, $\sigma = 0.182$.

Validation Set



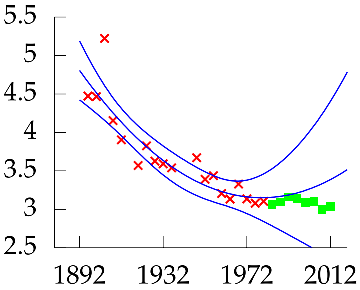
Left: fit to data, *Right*: model error. Polynomial order 0, training error 29.757, validation error -0.29243, $\sigma^2 = 0.302$, $\sigma = 0.550$.

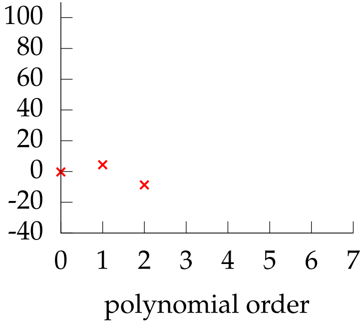
Validation Set



Left: fit to data, *Right*: model error. Polynomial order 1, training error 14.942, validation error 4.4027, $\sigma^2 = 0.0762$, $\sigma = 0.276$.

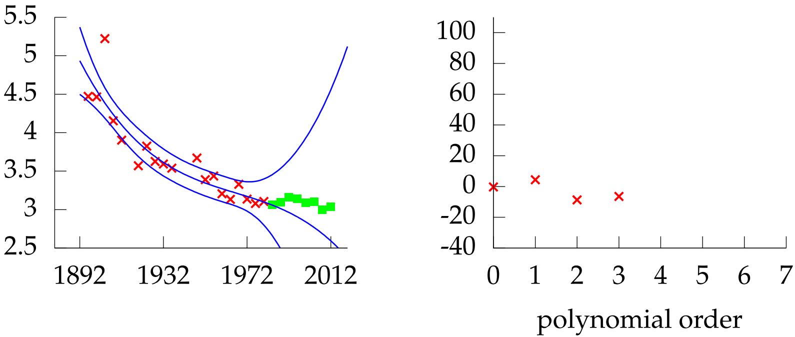
Validation Set





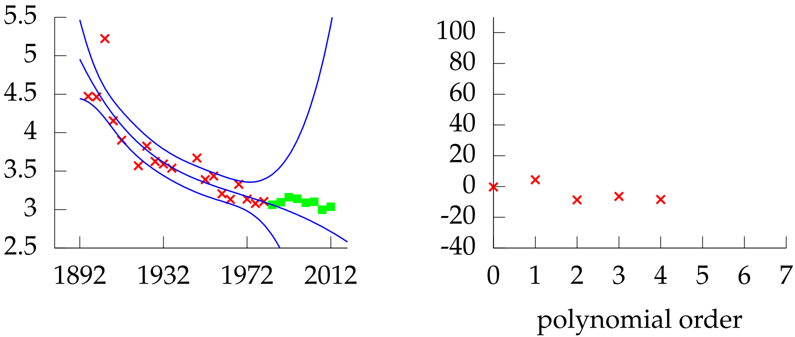
Left: fit to data, *Right*: model error. Polynomial order 2, training error 9.7206, validation error -8.6623, $\sigma^2 = 0.0580$, $\sigma = 0.241$.

Validation Set



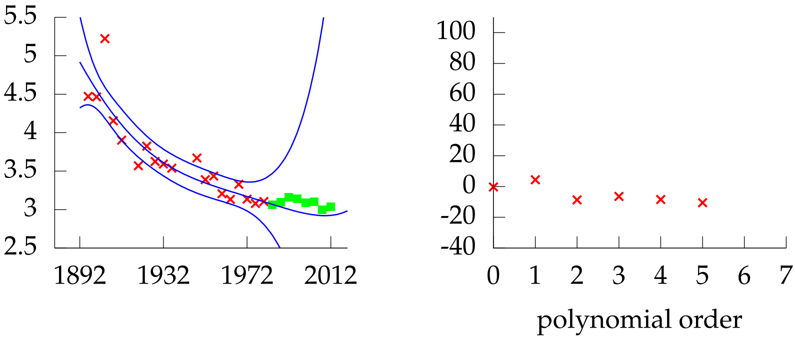
Left: fit to data, *Right*: model error. Polynomial order 3, training error 10.416, validation error -6.4726, $\sigma^2 = 0.0555$, $\sigma = 0.236$.

Validation Set



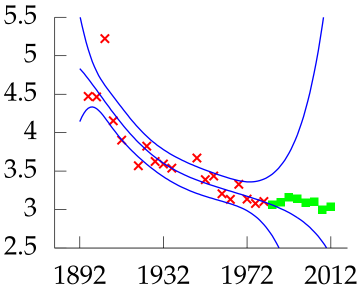
Left: fit to data, *Right*: model error. Polynomial order 4, training error 11.34, validation error -8.431, $\sigma^2 = 0.0555$, $\sigma = 0.236$.

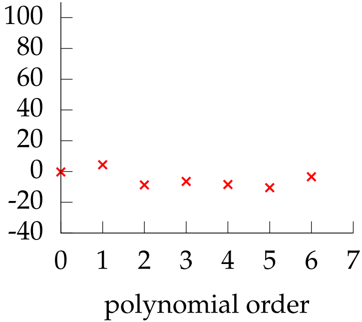
Validation Set



Left: fit to data, *Right*: model error. Polynomial order 5, training error 11.986, validation error -10.483, $\sigma^2 = 0.0551$, $\sigma = 0.235$.

Validation Set





Left: fit to data, *Right*: model error. Polynomial order 6, training error 12.369, validation error -3.3823, $\sigma^2 = 0.0537$, $\sigma = 0.232$.

Regularized Mean

- Validation fit here based on mean solution for $\mathbf{w}$ only.
- For Bayesian solution

$$\boldsymbol{\mu}_w = \left[\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \alpha^{-1}\mathbf{I}\right]^{-1}\sigma^{-2}\boldsymbol{\Phi}^\top\mathbf{y}$$

  instead of

$$\mathbf{w}^* = \left[\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right]^{-1}\boldsymbol{\Phi}^\top\mathbf{y}$$

- Two are equivalent when $\alpha \to \infty$.
- Equivalent to a prior for $\mathbf{w}$ with infinite variance.
- In other cases $\alpha\mathbf{I}$ *regularizes* the system (keeps parameters smaller).

Sampling the Posterior

- Now check samples by extracting $\mathbf{w}$ from the *posterior*.
- Now for $\mathbf{y} = \mathbf{\Phi}\mathbf{w} + \boldsymbol{\epsilon}$ need

$$w \sim \mathcal{N}\left(\boldsymbol{\mu}_w, \mathbf{C}_w\right)$$

with $\mathbf{C}_w = \left[\sigma^{-2}\mathbf{\Phi}^\top\mathbf{\Phi} + \alpha^{-1}\mathbf{I}\right]^{-1}$ and $\boldsymbol{\mu}_w = \mathbf{C}_w\sigma^{-2}\mathbf{\Phi}^\top\mathbf{y}$

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\right)$$

with $\alpha = 1$ and $\epsilon = 0.01$.

Marginal Likelihood

- The marginal likelihood can also be computed, it has the form:

$$p(\mathbf{y}|\mathbf{X}, \sigma^2, \alpha) = \frac{1}{(2\pi)^{\frac{n}{2}} |\mathbf{K}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}\right)$$

  where $\mathbf{K} = \alpha \mathbf{\Phi}\mathbf{\Phi}^\top + \sigma^2 \mathbf{I}$.
- So it is a zero mean n-dimensional Gaussian with covariance matrix $\mathbf{K}$.

Computing the Expected Output

- Given the posterior for the parameters, how can we compute the expected output at a given location?
- Output of model at location $\mathbf{x}_i$ is given by

$$f(\mathbf{x}_i; \mathbf{w}) = \boldsymbol{\phi}_i^\top \mathbf{w}$$

- We want the expected output under the posterior density, $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \sigma^2, \alpha)$.
- Mean of mapping function will be given by

$$\begin{aligned}\left\langle f(\mathbf{x}_i; \mathbf{w})\right\rangle_{p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \sigma^2, \alpha)} &= \boldsymbol{\phi}_i^\top \left\langle \mathbf{w}\right\rangle_{p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \sigma^2, \alpha)} \\ &= \boldsymbol{\phi}_i^\top \boldsymbol{\mu}_w\end{aligned}$$

Variance of Expected Output

- Variance of model at location $\mathbf{x}_i$ is given by

$$\begin{aligned}\mathrm{var}(f(\mathbf{x}_i;\mathbf{w})) &= \left\langle (f(\mathbf{x}_i;\mathbf{w}))^2 \right\rangle - \left\langle f(\mathbf{x}_i;\mathbf{w}) \right\rangle^2 \\ &= \boldsymbol{\phi}_i^\top \left\langle \mathbf{w}\mathbf{w}^\top \right\rangle \boldsymbol{\phi}_i - \boldsymbol{\phi}_i^\top \left\langle \mathbf{w} \right\rangle \left\langle \mathbf{w} \right\rangle^\top \boldsymbol{\phi}_i \\ &= \boldsymbol{\phi}_i^\top \mathbf{C}_i \boldsymbol{\phi}_i\end{aligned}$$

where all these expectations are taken under the posterior density, $p(\mathbf{w}|\mathbf{y}, \mathbf{X}, \sigma^2, \alpha)$.

Reading

- Section 3.7–3.8 of Rogers and Girolami (pg 122–133).
- Section 3.4 of Bishop (pg 161–165).

References I

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006. [Google Books] .

S. Rogers and M. Girolami. *A First Course in Machine Learning*. CRC Press, 2011. [Google Books] .