STAT 339
Bayesian Regression

13 March 2017

Colin Reimer Dawson

Outline

Generative Model

Recall the standard generative linear regression model

$$\mathbf{t} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

We can do Bayesian inference by putting a prior on $\mathbf{w}$ (we'll treat σ^2 as fixed for now; we'll relax this soon!). E.g., conjugate

$$\mathbf{w} \mid \boldsymbol{\mu}_0, \sigma_0^2 \sim \mathcal{N}(\boldsymbol{\mu}_0, \sigma_0^2\mathbf{I})$$

(Note that this is equivalent to saying that each $w_d \sim \mathcal{N}(\mu_{0d}, \sigma_0^2)$) independent.)

Then, letting $\Sigma = \sigma^2\mathbf{I}$ and $\Sigma_0 = \sigma_0^2\mathbf{I}$, the posterior is proportional to

$$\exp\{-\frac{1}{2}(\mathbf{t} - \mathbf{X}\mathbf{w})^\mathsf{T}\Sigma^{-1}(\mathbf{t} - \mathbf{X}\mathbf{w}) - \frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_0)^\mathsf{T}\Sigma_0^{-1}(\mathbf{w} - \boldsymbol{\mu}_0)\}$$

Finding the posterior parameters

The posterior for $\mathbf{w}$ is proportional to

$$\exp\{-\frac{1}{2}(\mathbf{t}-\mathbf{X}\mathbf{w})^{\mathsf{T}}\Sigma^{-1}(\mathbf{t}-\mathbf{X}\mathbf{w})-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_0)^{\mathsf{T}}\Sigma_0^{-1}(\mathbf{w}-\boldsymbol{\mu}_0)\}$$

We can see this is going to be quadratic in $\mathbf{w}$; just need to find the mean and covariance matrix. Expand and group terms like in $\mathbf{w}$; the exponent becomes

$$\begin{aligned}
&-\frac{1}{2}(\mathbf{t}^{\mathsf{T}}\Sigma^{-1}\mathbf{t}-2\mathbf{t}^{\mathsf{T}}\Sigma^{-1}\mathbf{X}\mathbf{w}+\mathbf{w}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\Sigma^{-1}\mathbf{X}\mathbf{w} \\
&\qquad +\mathbf{w}^{\mathsf{T}}\Sigma_0^{-1}\mathbf{w}-2\boldsymbol{\mu}_0^{\mathsf{T}}\Sigma_0^{-1}\mathbf{w}+\boldsymbol{\mu}_0^{\mathsf{T}}\Sigma_0^{-1}\boldsymbol{\mu}_0) \\
&=-\frac{1}{2}\left(\mathbf{w}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\Sigma^{-1}\mathbf{X}+\Sigma_0^{-1})\mathbf{w}-2(\mathbf{X}^{\mathsf{T}}\Sigma^{-1}\mathbf{t}+\Sigma_0^{-1}\boldsymbol{\mu}_0)^{\mathsf{T}}\mathbf{w}+c\right)
\end{aligned}$$

Finding the posterior parameters

The exponent is

$$-\frac{1}{2}\left(\mathbf{w}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\Sigma^{-1}\mathbf{X}+\Sigma_0^{-1})\mathbf{w}-2(\mathbf{X}^{\mathsf{T}}\Sigma^{-1}\mathbf{t}+\Sigma_0^{-1}\boldsymbol{\mu}_0)^{\mathsf{T}}\mathbf{w}+c\right)$$

Let $\Sigma_{post}=(\mathbf{X}^{\mathsf{T}}\Sigma^{-1}\mathbf{X}+\Sigma_0^{-1})^{-1}$ and let $\mathbf{a}=\mathbf{X}^{\mathsf{T}}\Sigma^{-1}\mathbf{t}+\Sigma_0^{-1}\boldsymbol{\mu}_0$.
Then the above is

$$\begin{aligned}&-\frac{1}{2}\left(\mathbf{w}^{\mathsf{T}}\Sigma_{post}^{-1}\mathbf{w}-2(\Sigma_{post}^{-1}\Sigma_{post}\mathbf{a})^{\mathsf{T}}\mathbf{w}+c\right)\\&=-\frac{1}{2}(\mathbf{w}-\Sigma_{post}\mathbf{a})^{\mathsf{T}}\Sigma_{post}^{-1}(\mathbf{w}-\Sigma_{post}\mathbf{a})+c\end{aligned}$$

So

$$\boldsymbol{\mu}_{post}=\Sigma_{post}\mathbf{a}=\Sigma_{post}(\mathbf{X}^{\mathsf{T}}\Sigma^{-1}\mathbf{t}+\Sigma_0^{-1}\boldsymbol{\mu}_0)$$

Summary: Bayesian Inference for Weights

Beginning with the standard generative model

$$\mathbf{t} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

(where σ^2 is fixed and known) and putting independent $\mathcal{N}(\mu_{0d}, \sigma_0^2)$ priors on each w_d, the posterior is

$$\mathbf{w} \mid \mathbf{t}, \mathbf{X}, \sigma^2, \sigma_0^2 \sim \mathcal{N}(\boldsymbol{\mu}_{post}, \Sigma_{post})$$

where

$$\boldsymbol{\mu}_{post} = \left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X} + \frac{1}{\sigma_0^2}\mathbf{I}\right)^{-1}\left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{t} + \frac{1}{\sigma_0^2}\mathbf{I}\boldsymbol{\mu}_0\right)$$

$$\Sigma_{post} = \left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X} + \frac{1}{\sigma^2}\mathbf{I}\right)^{-1}$$

Illustration of Prior and Posterior

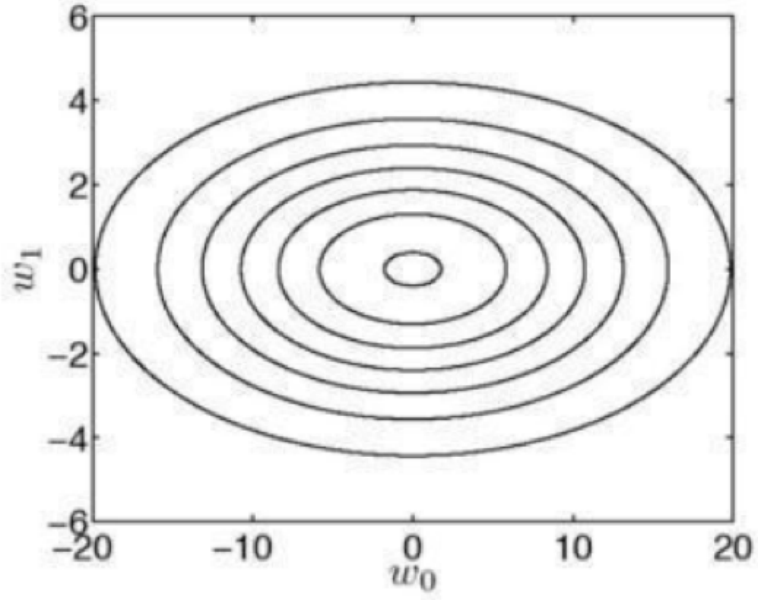


Figure: Example prior density for $\mathbf{w}$

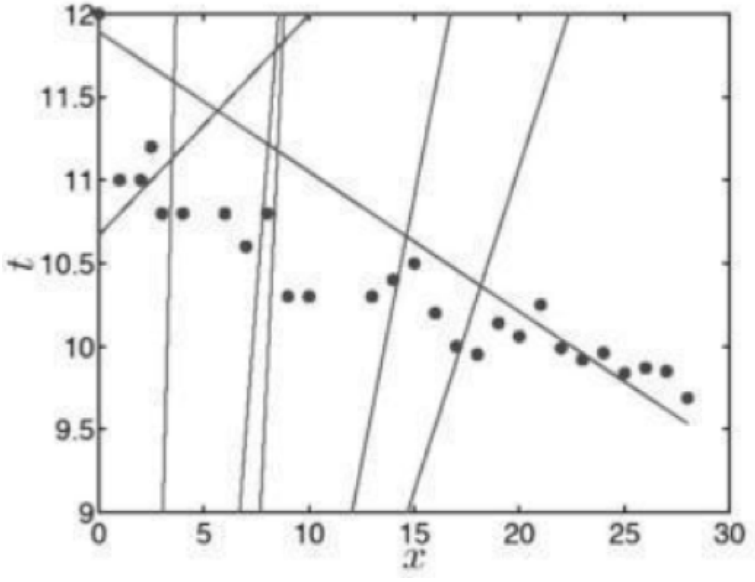
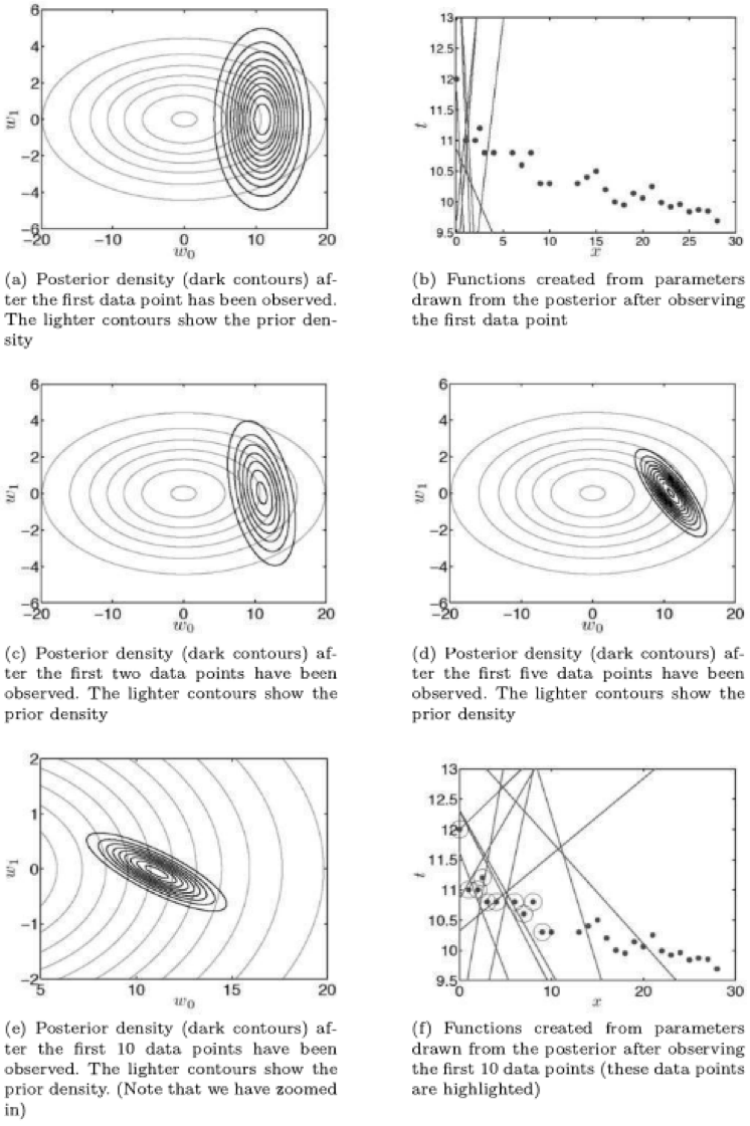


Figure: Samples of regression lines drawn from the prior

Iterative Updates to Posterior



(a) Posterior density (dark contours) after the first data point has been observed. The lighter contours show the prior density

(b) Functions created from parameters drawn from the posterior after observing the first data point

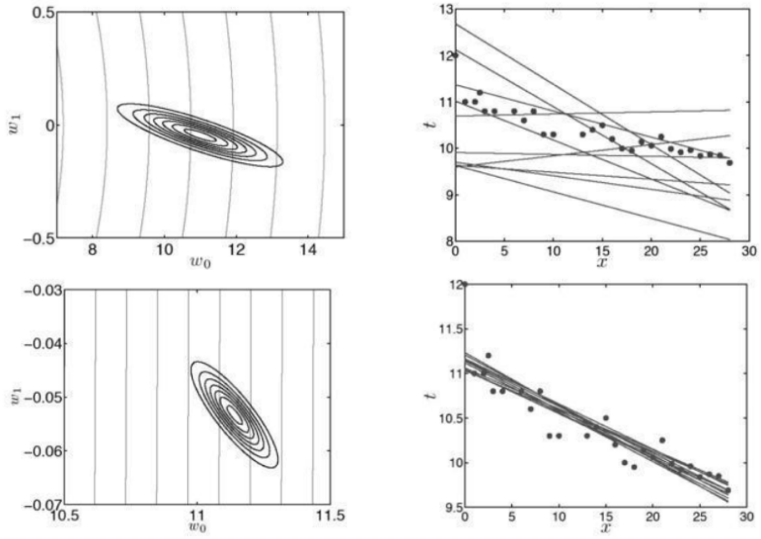
(c) Posterior density (dark contours) after the first two data points have been observed. The lighter contours show the prior density

(d) Posterior density (dark contours) after the first five data points have been observed. The lighter contours show the prior density

(e) Posterior density (dark contours) after the first 10 data points have been observed. The lighter contours show the prior density. (Note that we have zoomed in)

(f) Functions created from parameters drawn from the posterior after observing the first 10 data points (these data points are highlighted)

Figure: Illustration of posterior evolution with $\sigma^2 = 10$.



1. Top Left: Posterior and Samples after $N = 1$
2. Left middle: Posterior after $N = 2$ and $N = 5$
3. Bottom left: Posterior and samples after $N = 10$
4. Above Top: Posterior and samples after $N = 27$.
5. Above Bottom: Posterior and samples with $\sigma^2 = 0.05$.

Bayesian vs. Ridge Regression

The posterior mean for $\mathbf{w}$ is

$$\begin{aligned}\boldsymbol{\mu}_{post} &= \left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X} + \frac{1}{\sigma_0^2}\mathbf{I}\right)^{-1}\left(\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{t} + \frac{1}{\sigma_0^2}\mathbf{I}\boldsymbol{\mu}_0\right) \\ &= \left(\frac{\sigma^2}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X} + \frac{\sigma^2}{\sigma_0^2}\mathbf{I}\right)^{-1}\left(\frac{\sigma^2}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{t} + \frac{\sigma^2}{\sigma_0^2}\mathbf{I}\boldsymbol{\mu}_0\right) \\ &= \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \frac{\sigma^2}{\sigma_0^2}\mathbf{I}\right)^{-1}\left(\mathbf{X}^\mathsf{T}\mathbf{t} + \frac{\sigma^2}{\sigma_0^2}\mathbf{I}\boldsymbol{\mu}_0\right)\end{aligned}$$

Recall that the $\mathbf{w}$ that minimizes the ridge loss $L(\mathbf{w}) = (\mathbf{t} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{t} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w}^\mathsf{T}\mathbf{w}$ is

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}\mathbf{t}$$

Under what conditions are these equivalent?

When $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\lambda = \frac{\sigma^2}{\sigma_0^2}$

Posterior Predictive Distribution

We can integrate out $\mathbf{w}$ and make predictions based on the (posterior) predictive distribution

$$p(t_{new} \mid \mathbf{x}_{new}, \mathbf{X}, \mathbf{t}, \sigma^2, \sigma_0^2) = \int p(t_{new} \mid \mathbf{x}_{new}, \mathbf{w}, \sigma^2) p(\mathbf{w} \mid \mathbf{X}, \mathbf{t}, \sigma^2, \sigma_0^2) \; d\mathbf{w}$$

Since both the densities in the integrand are Normal, the result will be a Normal.

$$t_{new} \mid \mathbf{x}_{new}, \mathbf{X}, \mathbf{t}, \sigma^2, \sigma_0^2 \sim \mathcal{N}(\mathbf{x}_{new}^{\mathsf{T}} \boldsymbol{\mu}_{post}, \sigma^2 + \mathbf{x}_{new}^{\mathsf{T}} \Sigma_{post} \mathbf{x}_{new})$$

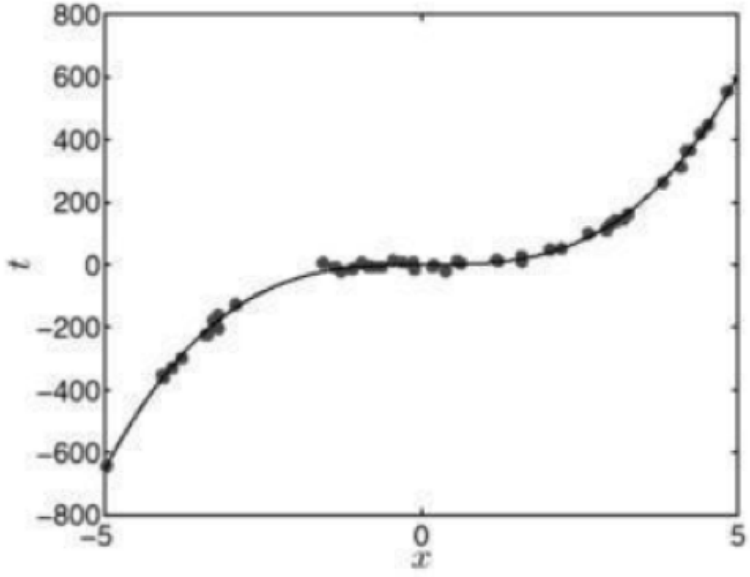
Model Selection via Marginal Likelihood

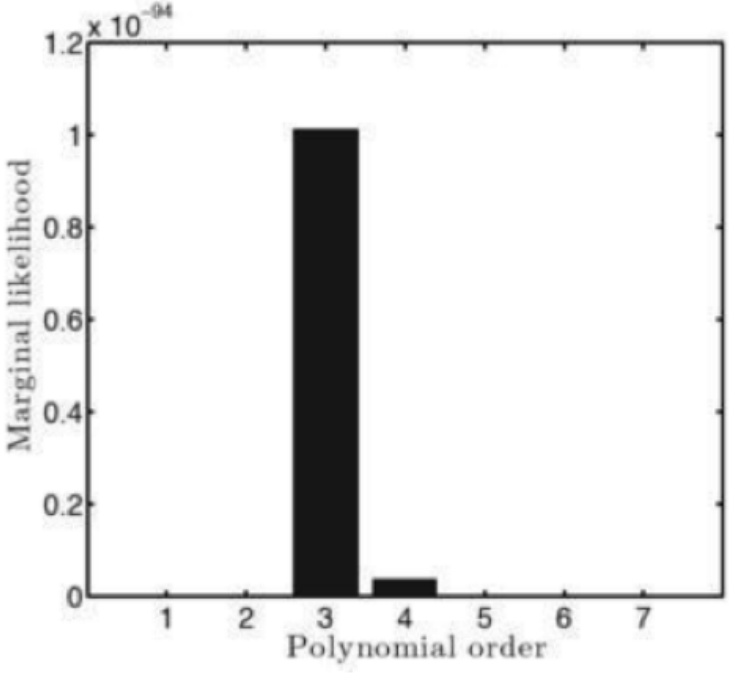
We can similarly find the *prior* predictive distribution, to evaluate the marginal likelihood for the whole model.

$$p(\mathbf{t} \mid \mathbf{X}, \sigma^2, \sigma_0^2) = \int p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \sigma^2) p(\mathbf{w} \mid \boldsymbol{\mu}_0, \sigma_0^2) \; d\mathbf{w}$$

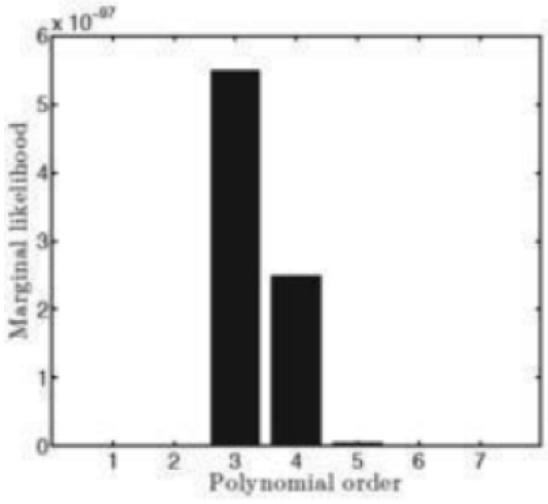
$$\mathbf{t} \mid \mathbf{X}, \sigma^2, \sigma_0^2 \sim \mathcal{N}(\mathbf{x}_{new}^\mathsf{T} \boldsymbol{\mu}_0, \sigma^2 \mathbf{I} + \mathbf{X}^\mathsf{T} \Sigma_0 \mathbf{X})$$

Comparing Polynomial Orders

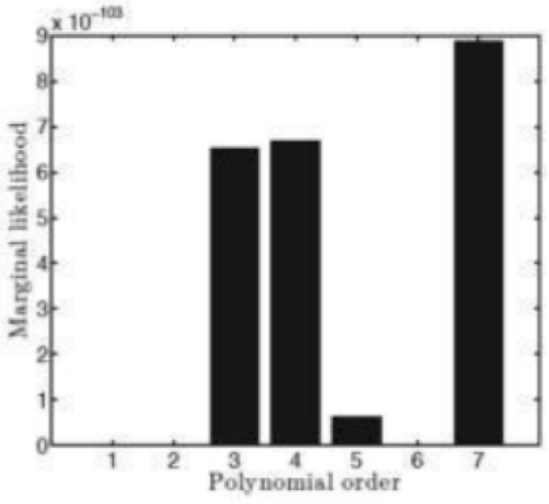




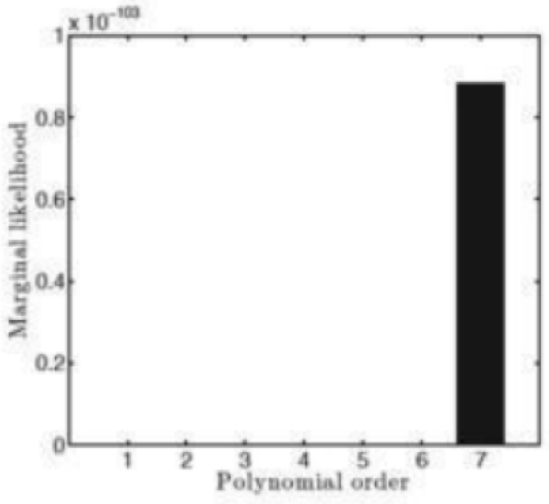
Results Depend on Priors



(a) $\sigma_0^2 = 0.7$



(b) $\sigma_0^2 = 0.4$



(c) $\sigma_0^2 = 0.3$