STAT 339
Bayesian Inference III

6 March 2017

Colin Reimer Dawson

Reflections

What is the lateness policy on HW?

[In the context of the linear model] I understand that ε_n should be distributed around $\hat{t}_n$, but not how.

How does the normalization of $\mathbf{X}$ [in the derivation of bias and variance in the linear model] work on a practical level?

Reflections

Could one say that the variance will increase as the number of features increases because the probability space grows as the number of features goes up?

If the bias is created by the data then wouldn't the bias scale with the size of the data[set]?

Summary: Inferring a Normal Mean

So, for an i.i.d. Normal sample

$$Y_1, \ldots, Y_N \overset{i.i.d.}{\sim} \mathcal{N}(\mu, h^{-1})$$

with a Normal prior on μ

$$\mu \sim \mathcal{N}(\mu_0, h_0^{-1})$$

the posterior is

$$\mu_0 \mid Y_1, \ldots, Y_N \sim \mathcal{N}(\alpha\mu_0 + (1-\alpha)\bar{y}, h_0 + Nh)$$

where $\alpha = \frac{h_0}{h_0+Nh} = \frac{h_0/h}{h_0/h+N}$. Then, h_0/h is a measure of the effective number of data points that the prior substitutes for.

Conjugate Priors and Exponential Families

Most common likelihoods, specifically, those whose PMF/PDF can be factored as

$$p(y \mid \theta) = h(y) \exp\{-A(\theta)\} \prod_{d=1}^{D} \exp\{T_d(y)\eta_d(\theta)\}$$

for some functions h, A, T_d, and η_d (where D is the dimension of θ) have priors that are conjugate for i.i.d. data. The joint likelihood for $\mathbf{y} = (y_1, \ldots, y_N)$ is

$$p(\mathbf{y} \mid \theta) \propto \exp\{-N A(\theta)\} \prod_{d=1}^{D} \exp\{\eta_d(\theta) \sum_{n=1}^{N} T_d(y_n)\}$$

Conjugate Priors and Exponential Families

The joint likelihood for $\mathbf{y} = (y_1, \ldots, y_N)$ is

$$p(\mathbf{y} \mid \theta) \propto \exp\{-NA(\theta)\} \prod_{d=1}^{D} \exp\{\eta_d(\theta) \sum_{n=1}^{N} T_d(y_n)\}$$

so if we choose as the prior a distribution of the form

$$p(\theta \mid \tau, n_0) \propto \exp\{-n_0 A(\theta)\} \prod_{d=1}^{D} \exp\{\tau_d \eta_d(\theta)\}$$

then the posterior will be

$$p(\theta \mid \mathbf{y}, \tau, n_0) \propto \exp\{-(n_0+N)A(\theta)\} \prod_{d=1}^{D} \exp\{(\tau_d + \sum_{n=1}^{N} T_d(y_n))\eta_d(\theta)\}$$

which corresponds to the parameter update

$$\tau_d \to \tau_d + \sum_{n=1}^{N} T_d(y_n)$$

$$n_0 \to n_0 + N$$

Example: Gamma-Poisson

Suppose $Y \mid \lambda \sim \text{Poisson}(\lambda)$. Then

$$p(y \mid \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$$

Is this of the form:

$$p(y \mid \lambda) = h(y)\exp\{-A(\lambda)\}\prod_{d=1}^{D}\exp\{T_d(y)\eta_d(\lambda)\}?$$

We can write

$$p(y \mid \lambda) = (y!)^{-1}e^{-\lambda}e^{y\log(\lambda)},$$

that is, choose

$$\begin{aligned} h(y) &= (y!)^{-1} & A(\lambda) &= \lambda \\ T_1(y) &= y & \eta_1(\lambda) &= \log(\lambda) \end{aligned}$$

Finding the Conjugate Prior

For N i.i.d. observations, the likelihood is

$$p(\mathbf{y} \mid \lambda) = \left(\prod_{n=1}^{N} (y!)^{-1} \right) e^{-N\lambda} e^{(\sum_{n=1}^{N} y) \log(\lambda)},$$

The conjugate prior should be of the form

$$p(\lambda \mid \tau, n_0) = c(\tau, n_0) e^{-n_0 \lambda} e^{\tau \log(\lambda)}$$

The posterior is then

$$\begin{aligned} p(\lambda \mid \mathbf{y}, \tau, n_0) &= c(\tau, n_0, \mathbf{y}) e^{-(n_0+N)\lambda} e^{(\tau + \sum_{n=1}^{N} y_n) \log(\lambda)} \\ &= c(\tau, n_0, \mathbf{y}) \lambda^{\tau + \sum n=1^{N} y_n} e^{-(n_0+N)\lambda} \end{aligned}$$

which is a Gamma$(\tau + 1, n_0)$ distribution.

Example: Normal

The Normal density (for $h = (\sigma^2)^{-1}$ fixed) is

$$p(y \mid \mu)(\frac{h}{2\pi})^{1/2} \exp -\frac{h}{2}(y-\mu)^2$$

Is this of the form:

$$p(y \mid \mu) = h(y) \exp\{-A(\mu)\} \prod_{d=1}^{D} \exp\{T_d(y)\eta_d(\mu)\}?$$

Yes. Rewrite as

$$(\frac{h}{2\pi})^{1/2} \exp\{-\frac{y^2 h}{2}\} \exp\{-\frac{1}{2}\mu^2 h\} \exp\{h\mu^2 y\}$$

choosing

$$h(y) = (\frac{h}{2\pi})^{1/2} \exp\{-\frac{h}{2}\} \qquad A(\mu) = \frac{\mu^2 h}{2}$$

$$T_1(y) = y \qquad \eta_1(\mu) = h\mu$$

Example: Normal

The likelihood for an i.i.d. sample is

$$(\frac{h}{2\pi})^{N/2} \exp\{-h \sum_{n=1}^{N} y^2\} \exp\{-N\frac{h\mu^2}{2}\} \exp\{h\mu \sum_{n=1}^{N} y\}$$

so the conjugate prior should be

$$\begin{aligned} p(\mu) &= c(\tau, n_0) \exp\{-n_0\frac{h\mu^2}{2}\} \exp\{\tau h\mu\} \\ &= c(\tau, n_0) \exp\{-\frac{hn_0}{2}(\mu^2 - 2\frac{\tau}{n_0}\mu)\} \end{aligned}$$

which is a $\mathcal{N}(\tau n_0^{-1}, (hn_0)^{-1})$ density.

Some Important (1-D) Conjugate Pairs

Likelihood	Prior	n_0	τ	$T(y)$
Bernoulli$(y \mid \theta)$	Beta$(\theta \mid a, b)$	$a+b$	a	y
Poisson$(y \mid \theta)$	Gamma$(\theta \mid a, b)$	b	$a+1$	y
Exponential$(y \mid \theta)$	Gamma$(\theta \mid a, b)$	a	b	y
$\mathcal{N}(y \mid \mu, h^{-1})$	$\mathcal{N}(\mu \mid \mu_0, h_0^{-1})$	h_0/h	$\frac{h_0}{h}\mu_0$	y
	Gamma$(h \mid \mu_0, h_0^{-1})$	$2a$	$2b$	$(y-\mu)^2$

where the Gamma and Exponential densities are parameterized as

$$\text{Gamma}(\theta \mid a, b) \propto \theta^{a-1} \exp\{-b\theta\}$$

$$\text{Exponential}(y \mid \theta) \propto \theta \exp\{-\theta y\}$$

In all cases, the posterior mean is equal to $\alpha\mu_0 + (1-\alpha)\hat{\mu}$, where μ_0 is the prior mean, $\hat{\mu}$ is the MLE of the mean, and $\alpha = n_0/(n_0+N)$

Predicting a New Value

- We often care about the parameters of models mainly to to make predictions. How can we go about this?
- Option 1: Take a point estimate of θ from the posterior (e.g., the mode) and plug that in to produce a predictive model.
- However, this discards our uncertainty, which was the main point of being Bayesian.
- Option 2: Compute the posterior predictive distribution:

$$p(y_{new} \mid \mathbf{y}) = \int p(y_{new} \mid \theta) p(\theta \mid \mathbf{y}) \, d\theta$$

in other words, take a weighted average over possible worlds of the predictive probability/density.

Example: Beta-Bernoulli Model

Suppose the y_n are i.i.d. Bernoulli(μ), where the true $\mu = 0.25$, and we observe $12/40$ heads. Suppose a Beta$(1, 1)$ prior: then $\mu \mid \mathbf{y} \sim$ Beta$(12 + 1, 28 + 1)$

- MLE: $\hat{\mu} = 0.3 \to \hat{P}(Y_{new} = 1) = 0.3$.
- Posterior mode: $\frac{12+1-1}{40+2-2} = 0.3$.
- Predictive probability:

$$
\begin{aligned}
p(y_{new} = 1 \mid \mathbf{y}) &= \int_0^1 p(y_{new} = 1 \mid \mu) p(\mu \mid \mathbf{y}) \, d\mu \\
&= \int_0^1 \mu \, p(\mu \mid \mathbf{y}) \, d\mu \\
&= \mathbb{E}\{\mu \mid \mathbf{y}\} \\
&= \frac{12 + 1}{40 + 2} = 0.31
\end{aligned}
$$

Marginal Likelihood

Another situation where we often want to average over possible worlds is in computing the **marginal likelihood**:

Marginal Likelihood

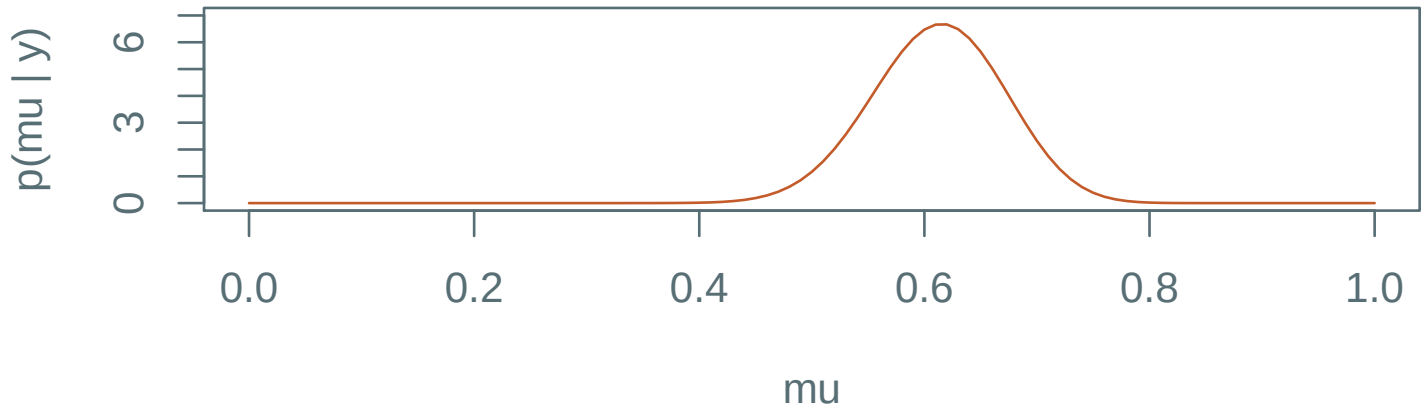
The **marginal likelihood** for a dataset $\mathbf{y}$ and a model (includes likelihood and prior on parameters, θ) is

$$p(\mathbf{y}) = \int p(y \mid \theta) p(\theta) \, d\theta$$

and can be used (as an alternative to, e.g., cross-validation) as a metric to select among competing model classes / competing priors.

Example: Fair or Biased Coin?

Suppose we don't know whether a coin is fair or not. After 40 flips, we see 25 heads. With a uniform prior on μ, the posterior is Beta$(25+1, 15+1)$



Example: Fair or Biased Coin?

The marginal likelihood of 25 heads under the uniform prior is

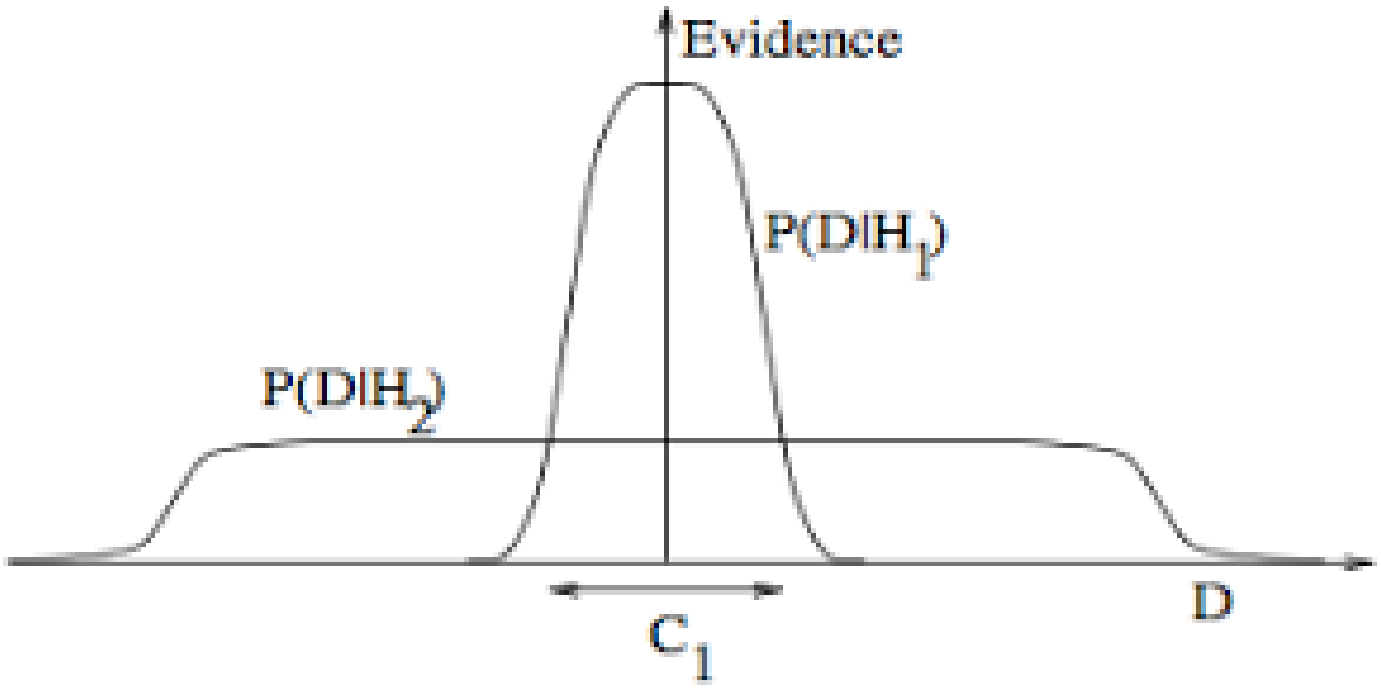
$$
\begin{aligned}
p(y) &= \int_0^1 p(y \mid \mu) p(\mu)\, d\mu \\
&= \binom{40}{25} \frac{\Gamma(1+1)}{\Gamma(1)\Gamma(1)} \int_0^1 \mu^{26-1}(1-\mu)^{16-1} \\
&= \binom{40}{25} \frac{\Gamma(1+1)}{\Gamma(1)\Gamma(1)} \frac{\Gamma(26)\Gamma(16)}{\Gamma(42)} \\
&= \frac{40!}{25!15!} \frac{1!}{0!0!} \frac{25!15!}{41!} = 1/41 = 0.0243
\end{aligned}
$$

If the coin is fair (i.e., $\mu = 0.5$ with probability 1), then the marginal likelihood is just

$$
p(y) = \binom{40}{25} (1/2)^2 5 (1/2)^{15} = 0.0366
$$

and so the "fair coin hypothesis" yields a higher marginal likelihood than the "Bayesian alternative" with a uniform prior.

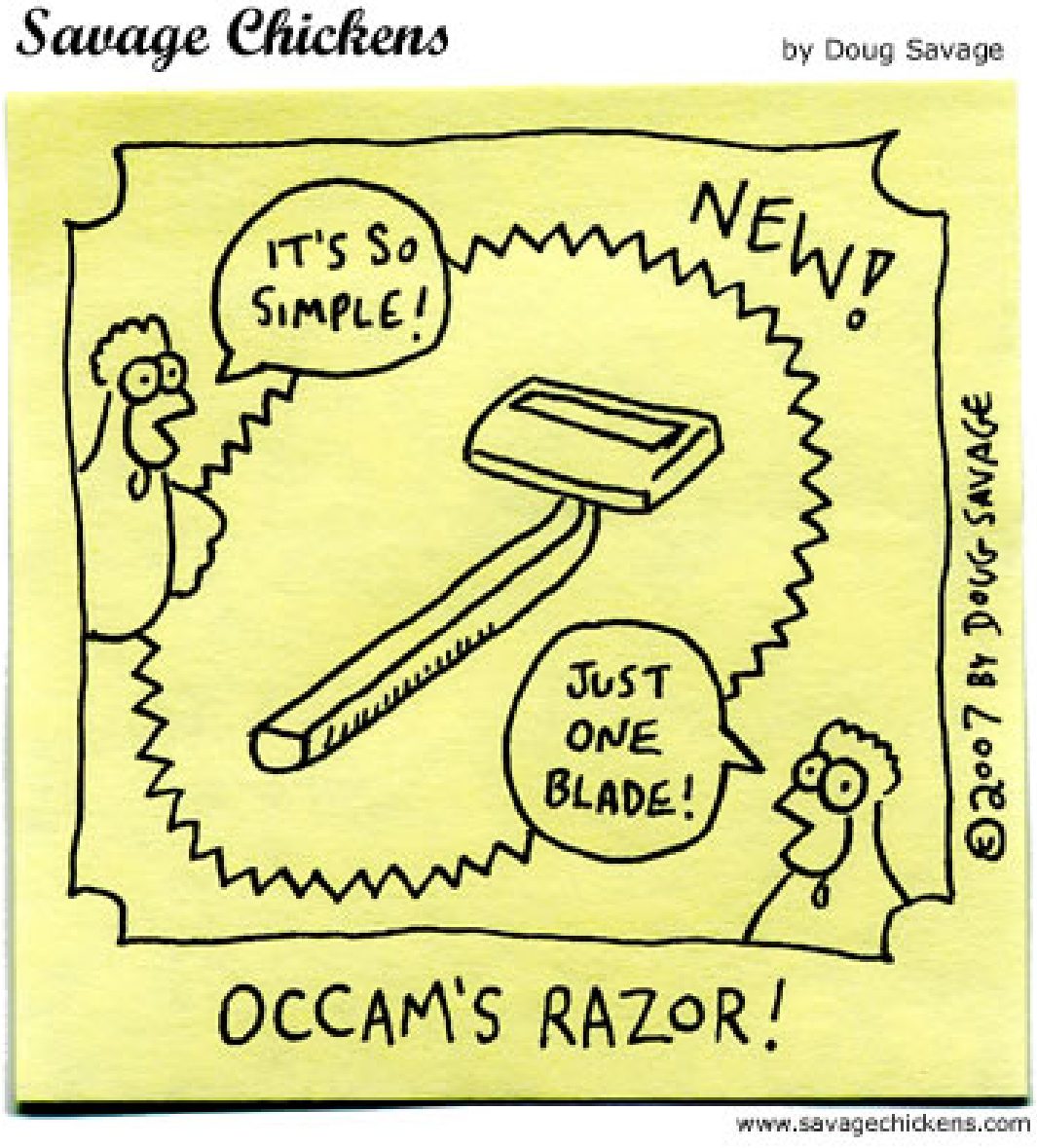
Conservation of Explanatory Power



Marginal likelihood "rewards" specific predictions

Conservation of Explanatory Power

Probabilistic Occam's Razor



Bayesian Occam's Razor

A "possible world" consists of a model $\mathcal{M}$, along with a (possibly trivial) parameter-setting, θ

$$p(\mathcal{M}|\mathbf{y}) = \int \frac{p(\mathcal{M},\theta)p(\mathbf{y}|\mathcal{M},\theta)}{p(\mathbf{y})} d\theta$$
$$= \int \frac{p(\mathcal{M})p(\theta|\mathcal{M})p(\mathbf{y}|\mathcal{M},\theta)}{p(\mathbf{y})} d\theta$$

$p(\mathbf{y}\|\mathcal{M},\theta)$	Rewards specific predictions by $(\mathcal{M},\theta)$
$p(\theta\|\mathcal{M})$	Penalizes flexibility of the model class