CS281 Section #2 Notes, v 1.0

Due: never

1 Beta-Binomial Distribution

1.1 Review:

$X \sim Bin(n, \theta)$ and $\theta \sim Beta(\alpha, \beta)$. Given $X = k$, the posterior of θ is $Beta(\alpha + k, \beta + n - k)$. The marginal distribution of X is $\frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \frac{\Gamma(k+\alpha)\Gamma(n-k+\beta)}{\Gamma(n+\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$.

1.2 Setting the beta hyper-parameters

([Murphy] Extended Exercise 3.15)
Often, the direct estimation of means and variances is easier than estimation of distribution parameters (although, these can be the same, like for Gaussians). Suppose $\theta \sim Beta(\alpha, \beta)$ and we believe that $E[\theta] = m$ and $var[\theta] = v$. Using Equation 2.62, solve for α and β in terms of m and v. What values do you get if $m = 0.7$ and $v = 0.2^2$?

Equation 2.62:

$$m = \frac{\alpha}{\alpha + \beta}$$
$$v = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

2 Gaussian Distribution

2.1 Review

Let X be a D-dimensional MVN random vector with mean μ and covariance matrix $\mathbf{\Sigma}$, denoted $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$. Then the pdf of X is

$$p(x) = (2\pi)^{-D/2} |\mathbf{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(x - \mu)^T \mathbf{\Sigma}^{-1}(x - \mu)\right]$$

MLE of MVN

$$\hat{\mu} = \bar{\mathbf{x}}(\text{sample mean})$$
$$\hat{\mathbf{\Sigma}} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{x})^T$$

If $\mathbf{\Sigma}$ is known, suppose prior $p(\mu) = N(\mu|\mathbf{m}_0, \mathbf{V}_0)$, the posterior distribution of μ:

$$p(\mu|\mathbf{x}, \mathbf{\Sigma}) = N(\mu|\mathbf{m}_N, \mathbf{V}_N)$$

where $\mathbf{m}_N = \mathbf{V}_N(\mathbf{\Sigma}^{-1}(N\bar{\mathbf{x}}) + \mathbf{V}_0^{-1}\mathbf{m}_0)$ and $\mathbf{V}_N^{-1} = \mathbf{V}_0^{-1} + N\mathbf{\Sigma}^{-1}$.

School	Estimated treatment effect, y_j	Standard error of effect estimate, σ_j
A	28	15
B	8	10
C	−3	16
D	7	11
E	−1	9
F	1	11
G	18	10
H	12	18

Figure 1: Data for 8 schools.

If $\mathbf{\Sigma}$ unknown, suppose prior

$$\begin{aligned} p(\mu, \mathbf{\Sigma}) &= p(\mu|\mathbf{\Sigma})p(\mathbf{\Sigma}) = N(\mu|\mathbf{m}_0, \frac{1}{\kappa_0}\mathbf{\Sigma})IW(\mathbf{\Sigma}|\mathbf{S}_0, \nu_0) \\ &= \frac{|\mathbf{S}_0|^{\nu_0/2}\kappa_0^{D/2}}{(2\pi)^{D/2}2^{\nu_0 D/2}\Gamma_D(\nu_0/2)}|\mathbf{\Sigma}|^{-\frac{v_0+D+2}{2}}e^{-\frac{\kappa_0}{2}(\mu-\mathbf{m}_0)^T\mathbf{\Sigma}^{-1}(\mu-\mathbf{m}_0)-\frac{1}{2}tr(\mathbf{\Sigma}^{-1}\mathbf{S}_0)} \end{aligned}$$

follows Normal-Inverse-wishart distribution. As $\mathbf{m}_0 = 0, \mathbf{S}_0 = 0, \mathbf{v}_0 = \kappa_0 = 0$, $p(\mu, \mathbf{\Sigma}) \propto |\mathbf{\Sigma}|^{-(D/2+1)}$. The posterior distribution of $(\mu, \mathbf{\Sigma})$ given x will be

$$NIW(\frac{\kappa_0}{\kappa_0+N}\mathbf{m}_0 + \frac{N}{\kappa_0+N}\bar{x}, \kappa_0+N, \nu_0+N, S_0+\mathbf{S}_{\bar{x}} + \frac{\kappa_0 N}{\kappa_0+N}(\bar{x}-m_0)(\bar{x}-\mathbf{m}_0)^T)$$

2.2 Hierarchial Gaussian Model

([Bayesian Data Analysis] Chapter 5.) Consider J experiments, each with n_j datapoints generated from normal with mean parameter θ_j and known variance σ^2:

$$y_{ij}|\theta_j \sim N(\theta_j, \sigma^2)$$

Sufficient statistic for θ_j is $\bar{y}_{.j}$:

$$\bar{y}_{.j}|\theta_j \sim N(\theta_j, \sigma_j^2 = \sigma^2/n_j)$$

1. Consider each experiment indepently. With noninformative prior, $p(\theta_j) \propto 1$, what is the posterior of θ_j?

2. Consider J experiment jointly, suppose $\theta_j|\mu, \tau^2 \sim N(\mu, \tau^2)$, $j \in \{1, 2, .., J\}$, what is the posterior of θ_j? What if $\tau^2 = 0$ and $\tau^2 = \infty$?

3. How should we set μ and τ^2? A full Bayesian approach treats μ and τ^2 as unknown and having their own prior. This is called hierarchical model and (μ, τ^2) are hyperparameters whose prior is called hyperprior. Usually we don't have much information about the hyperparameters and assign a diffuse prior distribution, i.e. $p(\mu, \tau) \propto 1$. What is the posterior distribution $p(\mu, \tau^2|y)$?

4. In order to get posterior samples of (μ, τ), we could rewrite the posterior as $p(\mu, \tau^2|y) = p(\mu|\tau^2, y)p(\tau^2|y)$, thus first sample τ^2 from $p(\tau^2|y)$ and then sample μ from $p(\mu|\tau^2, y)$. What are $p(\tau^2|y)$ and $p(\mu|\tau^2, y)$?

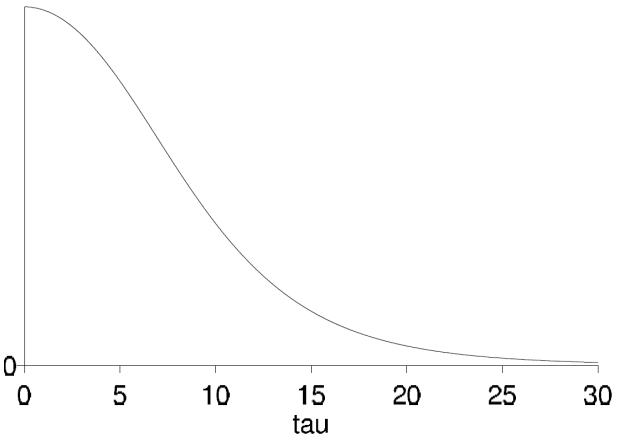

Figure 2: Posterior distribution $p(\tau^2|y)$.

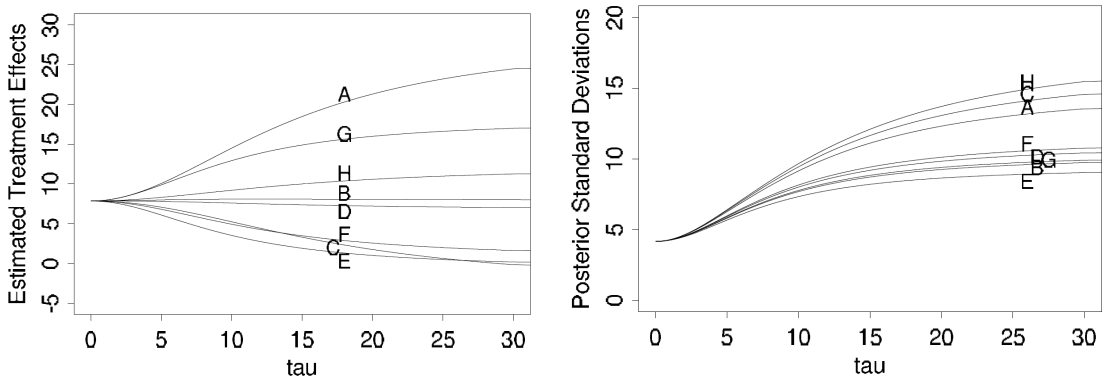

Figure 3: Posterior mean and standard deviation of θ_j at different values of τ^2.

School	Posterior quantiles				
	2.5%	25%	median	75%	97.5%
A	−2	7	10	16	31
B	−5	3	8	12	23
C	−11	2	7	11	19
D	−7	4	8	11	21
E	−9	1	5	10	18
F	−7	2	6	10	28
G	−1	7	10	15	26
H	−6	3	8	13	33

Figure 4: Posterior distribution $p(\tau^2|y)$.

2.3 Solution

1. The posterior of θ_j is $\theta_j|y_{ij} \sim N(\bar{y}_{.j}, \sigma_j^2)$.

2. The posterior of θ_j is $\theta_j|y_{ij},\mu,\tau \sim N(\frac{\frac{n_j\bar{y}_{.j}}{\sigma^2}+\frac{\mu}{\tau^2}}{\frac{1}{\tau}^2+\frac{n_j}{\sigma^2}}, \frac{1}{\frac{1}{\tau^2}+\frac{n_j}{\sigma^2}})$. When $\tau^2 = 0$, θ_js are equal, that is, every school has the same mean μ; when $\tau^2 = \infty$, θ_js are independent and posterior of θ_j only depends on $\bar{y}_{.j}$.

3. Integrating out θ_j, the marginal distribution of $\bar{y}_{.j}$ is

$$\bar{y}_{.j}|\mu,\tau^2 \sim N(\mu, \sigma_j^2+\tau^2)$$

Thus the marginal posterior density is

$$p(\mu,\tau|y) \propto p(\mu,\tau)\prod_{j=1}^{J} N(\bar{y}_{.j}|\mu,\sigma_j^2+\tau^2) \propto \prod_{j=1}^{J} N(\bar{y}_{.j}|\mu,\sigma_j^2+\tau^2)$$

4. First

$$\mu|\tau^2,y \sim N(\hat{\mu} = \frac{\sum_{j=1}^{J}\frac{\bar{y}_{.j}}{\sigma_j^2+\tau^2}}{\sum_{j=1}^{J}\frac{1}{\sigma_j^2+\tau^2}}, \hat{V}_\mu = \frac{1}{\sum_{j=1}^{J}\frac{1}{\sigma_j^2+\tau^2}})$$

Since $p(\tau|y) = \frac{p(\mu,\tau|y)}{p(\mu|\tau,y)}$ holds for any μ, then

$$p(\tau|y) \propto \frac{p(\tau)\prod_{j=1}^{J} N(\bar{y}_{.j}|\mu,\sigma_j^2+\tau^2)}{N(\mu|\hat{\mu},\hat{V}_\mu)} = \frac{p(\tau)\prod_{j=1}^{J} N(\bar{y}_{.j}|\hat{\mu},\sigma_j^2+\tau^2)}{N(\hat{\mu}|\hat{\mu},\hat{V}_\mu)} \propto \hat{V}_\mu^{1/2}\prod_{j=1}^{J}(\sigma_j^2+\tau^2)^{-1/2}e^{-\frac{(\bar{y}_{.j}-\hat{\mu})^2}{2(\sigma_j^2+\tau^2)}}$$

3 Linear Regression

3.1 Review

Recall our regression model: we are given "fixed" inputs $\mathbf{x}$, these are used to "predict" outputs y, the prediction is linear. Recall that solving for weights in linear regression corresponds to an orthogonal projection of the known outputs y_i onto the column space of $\mathbf{x}$, that is, the linear combination of $\mathbf{x}$'s features that reduce the distance to the true y. For our set of parameters θ:

$$p(\mathbf{y}|\mathbf{X},\theta) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w},\sigma^2\mathbf{I})$$

where $\mathbf{w}^\top\mathbf{x}$ is the linear term and σ^2 observation noise that is considered to be known (for now). Make sure you are comfortable switching between single data point and whole dataset notation. Note that Murphy uses $\mathcal{D}$ (for data) and $(\mathbf{X},\mathbf{y})$ interchangeably. The log-likelihood for $\{(\mathbf{x}_i,y_i)\}_{i=1}^N$:

$$\log p(\mathcal{D}|\theta) = \sum_{i=1}^{N}\log p(y_i|\mathbf{x}_i,\theta) = -\big[\sum_{i=1}^{N}(y_i-\mathbf{w}^\top\mathbf{x}_i)^2 + \log(const)\big]$$

Maximum Likelihood Estimate for the weights ($\mathbf{w}_{MLE}$):

$$\arg\max_w -\big[\sum_{i=1}^{N}(y_i-\mathbf{w}^\top\mathbf{x}_i)^2\big] = \arg\max_w -\big[(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w})\big] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Being Bayesian, we now put a (normal) prior over the weights $\mathbf{w}$:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0,\mathbf{S}_0)$$

Our likelihood is:

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \sigma^2\mathbf{I}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$$

The posterior over the weights (which is a distribution, not a single point-estimate) is proportional to the product of the prior and the likelihood. Writing out this product and completing the square is hard, but remember that we have our closed form solutions for the parameters of the posterior of a linear-gaussian:

$$\begin{aligned}
\mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \mathbf{A}^\top \boldsymbol{\Sigma}_y^{-1} \mathbf{A} \\
\mathbf{m}_N &= \mathbf{S}_N \left[\mathbf{S}_0^{-1} \mathbf{m}_0 + \mathbf{A}^\top \boldsymbol{\Sigma}_y^{-1} (\mathbf{y} - \mathbf{b}) \right]
\end{aligned}$$

Looks familiar. Recall the general case:

$$\begin{aligned}
\mathbf{x} &\sim \mathcal{N}(\mathbf{m}_0, \boldsymbol{\Sigma}_0) \\
\mathbf{y}|\mathbf{x} &\sim \mathcal{N}(\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_y) \\
p(\mathbf{x}|\mathbf{y}) &\propto p(\mathbf{x}) p(\mathbf{y}|\mathbf{x}) \\
\mathbf{S}_N^{-1} &= \boldsymbol{\Sigma}_0^{-1} + \mathbf{A}^\top \boldsymbol{\Sigma}_y^{-1} \mathbf{A} \\
\mathbf{m}_N &= \mathbf{S}_N \left[\boldsymbol{\Sigma}_0^{-1} \mathbf{m}_0 + \mathbf{A}^\top \boldsymbol{\Sigma}_y^{-1} (\mathbf{y} - \mathbf{b}) \right]
\end{aligned}$$

Our case is the same with $\mathbf{b} = \mathbf{0}, \mathbf{A} = \mathbf{X}, \mathbf{y} = \mathbf{y}, \boldsymbol{\Sigma}_y = \sigma^2\mathbf{I}$. To connect our notation with Murphy, $\mathbf{m}_0 = \mathbf{m}_x, \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_x, \mathbf{S}_N = \boldsymbol{\Sigma}_{x|y}, \mathbf{m}_N = \mathbf{m}_{x|y}$.

Returning to regression, we can now state the posterior distribution:

$$p(\mathbf{w}|\mathbf{x},\mathbf{y},...) = \mathcal{N}(\mathbf{w}|\mathbf{S}_N\big[\mathbf{S}_0^{-1}\mathbf{m}_0 + \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{y}\big], \big[\mathbf{S}_0^{-1} + \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}\big]^{-1})$$

If you multiply the $\mathbf{S}_N$ in the mean out, and replace $\mathbf{S}_N$ for its definition in terms of $\mathbf{X}$, the MLE for $\mathbf{w}$ comes out as one of the terms. Finally, the posterior predictive:

$$p(\tilde{y}|\tilde{\mathbf{x}},\mathbf{X},\mathbf{y}) = \int \mathcal{N}(\tilde{y}|\tilde{\mathbf{x}}^\top\mathbf{w},\sigma^2)\mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)d\mathbf{w} = \mathcal{N}(\tilde{y}|\tilde{\mathbf{x}}^\top\mathbf{m}_N,\sigma^2 + \tilde{\mathbf{x}}^\top\mathbf{S}_N\tilde{\mathbf{x}})$$

where the variance now depends on the data point to predict on $\mathbf{x}_*$. This means that we can keep track of how certain (or uncertain) we are on new predictions based on what we saw (or didn't see) in our data.

3.2 Unknown Observation Noise σ^2

Following Murphy p.234, we now consider the case where the observation noise σ^2 is unknown:

$$p(\mathbf{w},\sigma^2|\mathcal{D})$$

Our likelihood term looks the same:

$$p(\mathbf{y}|\mathbf{X},\mathbf{w},\sigma^2) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w},\sigma^2\mathbf{I})$$

One choice of a conjugate prior is the Normal-inverse Gaussian:

$$p(\mathbf{w},\sigma^2) = NIG(\mathbf{w},\sigma^2|\mathbf{w}_0,\mathbf{S}_0,a_0,b_0) = \mathcal{N}(\mathbf{w}|\mathbf{w}_0,\sigma^2\mathbf{S}_0)IG(\sigma^2|a_0,b_0)$$

where a_0 is a tail heaviness parameter and b_0 is an asymmetry parameter. The posterior has the form:

$$\begin{aligned}
p(\mathbf{w},\sigma^2|\mathcal{D}) &= NIG(\mathbf{w},\sigma^2|\mathbf{w}_N,\mathbf{S}_N,a_N,b_N)\\
\mathbf{w}_N &= \mathbf{S}_N\big[\mathbf{S}_0^{-1}\mathbf{w}_0 + \mathbf{X}^\top\mathbf{y}\big]\\
\mathbf{S}_N &= \big[\mathbf{S}_0^{-1} + \mathbf{X}^\top\mathbf{X}\big]^{-1}\\
a_N &= a_0 + n/2\\
b_N &= b_0 + \frac{1}{2}\big[\mathbf{w}_0^\top\mathbf{S}_0^{-1}\mathbf{w}_0 + \mathbf{y}^\top\mathbf{y} - \mathbf{w}_N^\top\mathbf{S}_N^{-1}\mathbf{w}_N\big]
\end{aligned}$$

The expressions for $\mathbf{w}_N$ and $\mathbf{S}_N$ are similar to the case where σ^2 is known. a_N updates counts and b_N is made up of b_0 (prior sum of squares), $\mathbf{y}^\top\mathbf{y}$ (empirical sum of squares), and error from the prior on $\mathbf{w}$.
The marginals are:

$$\begin{aligned}
p(\sigma^2|\mathcal{D}) &= IG(a_N,b_N)\\
p(\mathbf{w}|\mathcal{D}) &= \mathcal{T}(\mathbf{w}_N,\frac{b_N}{a_N}\mathbf{S}_N,2a_N)
\end{aligned}$$

Finally, when prediciting on m new data points $\tilde{\mathbf{X}}$, the posterior predictive takes the form:

$$p(\tilde{\mathbf{y}}|\tilde{\mathbf{X}},\mathcal{D}) = \mathcal{T}(\tilde{\mathbf{y}}|\tilde{\mathbf{X}}\mathbf{w}_N,\frac{b_N}{a_N}(\mathbf{I}_m + \tilde{\mathbf{X}}\mathbf{S}_N\tilde{\mathbf{X}}^\top),2a_N)$$

The variance of the distribution (middle of three parameters) is made up of two components. The left is due to measurement noise and the right is due to uncertainty in $\mathbf{w}$. The latter term varies depending on how similar $\tilde{\mathbf{X}}$ are to $\mathbf{X}$ (captured by $\mathbf{S}_N$).