CS 340 Fall 2007: Homework 3

1 Marginal likelihood for the Beta-Bernoulli model

We showed that the marginal likelihood is the ratio of the normalizing constants:

$$p(D) = \frac{B(\alpha_1 + N_1, \alpha_0 + N_0)}{B(\alpha_1, \alpha_0)} \tag{1}$$

$$= \frac{\Gamma(\alpha_1 + N_1)\Gamma(\alpha_0 + N_0)}{\Gamma(\alpha_1 + \alpha_0 + N)} \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \tag{2}$$

We will now derive an alternative derivation of this fact. By the chain rule of probability,

$$p(x_{1:N}) = p(x_1)p(x_2|x_1)p(x_3|x_{1:2}) \ldots \tag{3}$$

We also showed that the posterior predictive distribution is

$$p(X = 1|D_{1:N}) = \frac{N_1 + \alpha_1}{N_1 + \alpha_1 + N_0 + \alpha_0} \stackrel{\text{def}}{=} \frac{N_1 + \alpha_1}{N + \alpha} \tag{4}$$

where $D_{1:N}$ is the data seen so far. Now suppose $D = H, T, T, H, H, H$ or $D = 1, 0, 0, 1, 1, 1$. Then

$$p(D) = \frac{\alpha_1}{\alpha} \cdot \frac{\alpha_0}{\alpha + 1} \cdot \frac{\alpha_0 + 1}{\alpha + 2} \cdot \frac{\alpha_1 + 1}{\alpha + 3} \cdot \frac{\alpha_1 + 2}{\alpha + 4} \tag{5}$$

$$= \frac{[\alpha_1(\alpha_1 + 1)(\alpha_1 + 2)]\,[\alpha_0(\alpha_0 + 1)]}{\alpha(\alpha + 1) \cdots (\alpha + 4)} \tag{6}$$

$$= \frac{[(\alpha_1) \cdots (\alpha_1 + N_1 - 1)]\,[(\alpha_0) \cdots (\alpha_0 + N_0 - 1)]}{(\alpha) \cdots (\alpha + N)} \tag{7}$$

Show how this reduces to Equation 2 by using the fact that, for integers, $(\alpha - 1)! = \Gamma(\alpha)$.

2 Beta updating from censored likelihood

Suppose we toss a coin $n = 5$ times. Let X be the number of heads. We observe that there are fewer than 3 heads. Let the prior probability of heads be $p(\theta) = Beta(\theta|1, 1)$. Compute the posterior $p(\theta|X < 3)$ up to normalization constants, i.e., derive an expression proportional to $p(\theta, X < 3)$. Plot the (unnormalized) posterior.

3 Fun with the Beta-binomial model

1. Let $\theta \sim Be(a, b)$. (In class, we called $a = \alpha_1$ and $b = \alpha_0$.) Sometimes our prior knowledge is not in the form of pseudo counts, so it is not immediately clear how to set a and b. For example, suppose you believe that $E\theta = m$ and Var $\theta = v$. Use the following properties of the Beta distribution to solve for a and b in terms of m and v.

$$E\theta = m = \frac{a}{a + b} \tag{8}$$

$$\text{Var } \theta = v = \frac{m(1 - m)}{a + b + 1} = \frac{ab}{(a + b)^2(a + b + 1)} \tag{9}$$

2. Let θ represent the proportion of adults in New York who support the death penalty. Suppose θ is beta with mean 0.7 and standard deviation 0.2. What are the values of the hyper-parameters a and b that correspond to this?

3. A random sample of 1000 adults in New York is taken, and 62% support the death penalty. Plot the prior $p(\theta)$ and posterior $p(\theta|D)$. What is the posterior mean and variance? What is the 95% posterior credible interval? i.e., find values $\theta_{2.5}$ and $\theta_{97.5}$ such that

$$p(\theta_{2.5} < \theta < \theta_{97.5}|D) = 0.95 \tag{10}$$

   Hint: use the function `betainv` in the statistics toolbox.

4 Gaussian posterior credible interval

Let $X \sim \mathcal{N}(\mu, \sigma^2 = 4)$ where μ is unknown but has prior $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2 = 4)$. The posterior after seeing n samples is $\mu \sim \mathcal{N}(\mu_n, \sigma_n^2)$. How big does n have to be to ensure

$$p(\mu \in I(\mu_n)|D) \geq 0.95 \tag{11}$$

where $I(\mu_n)$ is an interval (centered on μ_n) of width 1, and D is the data. Hint: recall that 95% of the probability mass of a Gaussian is within $\pm 1.96\sigma$ of the mean.

5 Gaussian sensor fusion

Suppose we have two sensors with known (and different) variances v_x and v_y, but unknown (and the same) mean μ. Suppose we observe n_x observations from the first sensor and n_y observations from the second sensor. Call these $\mathcal{D}_x$ and $\mathcal{D}_y$. Assume all distributions are Gaussian.

1. What is the posterior $p(\mu|\mathcal{D}_x, \mathcal{D}_y)$, assuming a non-informative prior for μ? Give an explicit expression for the posterior mean and variance. Hint: use Bayesian updating twice, once to get from $p(\mu) \rightarrow p(\mu|\mathcal{D}_x)$ (starting from a non-informative prior, which we can simulate using a precision of 0), and then again to get from $p(\mu|\mathcal{D}_x) \rightarrow p(\mu|\mathcal{D}_x, \mathcal{D}_y)$.

2. Suppose the y sensor is very unreliable. What will happen to the posterior mean estimate? Give a simplified approximate expression.

6 Estimation of σ^2 when μ is known

Suppose we sample $x_1, \ldots, x_N \sim \mathcal{N}(\mu, \sigma^2)$ where μ is a *known* constant. Derive an expression for the MLE for σ^2 in this case. Is it unbiased?

7 Detecting differentially expressed genes

Consider the problem of detecting which genes change when a certain treatment is applied. Load the (synthetic) data stored in `bayesFactorGeneData.mat`. You should see the following

```
Name           Size             Bytes  Class     Attributes
Xcontrol     100x2               1600  double
Xtreat       100x2               1600  double
truth          1x100              800  double
```

If you plot the data, you should see something like the top two panels of Figure 1. The goal is to figure out which of the 100 genes are different between treatment and control. We will apply a Bayesian and a frequentist approach to this.

Let $x_{g,r}$ be the r'th replicate (sample) of gene g under the control condition, and $y_{g,r}$ be the r'th replicate (sample) of gene g under the treatment condition. For each gene, we want to choose between model H_0, which says x_g and y_g were generated from a Gaussian with the same mean, and H_1, which says they were generated from two Gaussians with different means. We will treat each gene independently (more sophisticated models can capture the interdependence between genes; see e.g., [FD06].) The Bayes factor in favor of H_0 for gene g is

$$BF_g = \frac{p(D_g|H_0)}{p(D_g|H_1)} \tag{12}$$

where $D_g = (x_{g,1:n_x}, y_{g,1:n_y})$ is all the data for gene g. The provided function `bayesianTtest(x,y)` computes this. To apply this to the gene data, just type

```
bf(g) = bayesianTtest(Xtreat(g,:), Xcontrol(g,:));
```

The classical alternative to a Bayes factor is to compute the p-value, defined as

$$pval_g = p(t(D'_g) \geq t(D_g)|D'_g \sim M_g = 0) \tag{13}$$

where $t(D_g)$ is the value of the t-statistic on the observed data D_g and $t(D'_g)$ is the t-statistic for a fictitious dataset D'_g drawn from the null distribution. P-values are often described in words as "the probability, under the null hypothesis, of getting a test statistic as large or larger than the observed one".[1] The smaller the p-value, the less likely is H_0. In Matlab, to compute a p-value for a two-sample t-test using the statistics toolbox, just type

```
[hyptest(g),pval(g)] = ttest(Xtreat(g,:), Xcontrol(g,:));
```

(Here $hyptest(g) = 0$ if the null hypothesis cannot be rejected at the 5% significance level; this is just a thresholded version of a p-value, and can be ignored.)

Finally, we get to the tasks you have to perform.

1. Load the file `bayesFactorGeneData.mat`. For each gene, compute the Bayes factor in favor of model 0 and the p-value. Plot the data and $\log(1/BF_g)$ and $1/pval(g)$. You should get something like Figure 1. Turn in your plots and code.

2. Using the provided function `ROCcurve`, compute a ROC curve using $\log(1/BF_g)$ as the score and the binary vector `truth` as the true label. (Here `truth(g)=1` means that gene g is differentially expressed under the treatment.) Repeat this using $1/pval(g)$ for the score. You should get something like Figure 2. Turn in your plots and code. Notice that the area under the curve (AUC) is higher for the Bayes factor than for the p-value.

3. Use the `truth` vector to sort the data, so that the unaffected genes come first, and then the affected (differentially expressed) ones come last. Redo you plot from part 1, using the permuted indices. You should get something like Figure 3. Turn in your plots and code. Now it is clear that, on average, the Bayes factor scoring function more clearly separates the two groups than the pvalue. The benefits of the Bayesian approach are even greater when one models dependence between the genes (see e.g., [FD06]), which is hard to do with t-tests.

4. **Optional, for extra credit**. The Bayesian t-test requires a prior (see Section 8 for details). Try changing the prior (the third and fourth arguments to `bayesianTtest`) and see what effect it has. Summarize your conclusions.

[1] It should be obvious that p-values are very unintuitive quantities. For example, why should we care about the probability of getting a statistic larger than what we observed? See `http://www.stat.duke.edu/~berger/p-values.html` and [Goo99a, Goo99b] for more details on p-values, and why you should avoid using them.

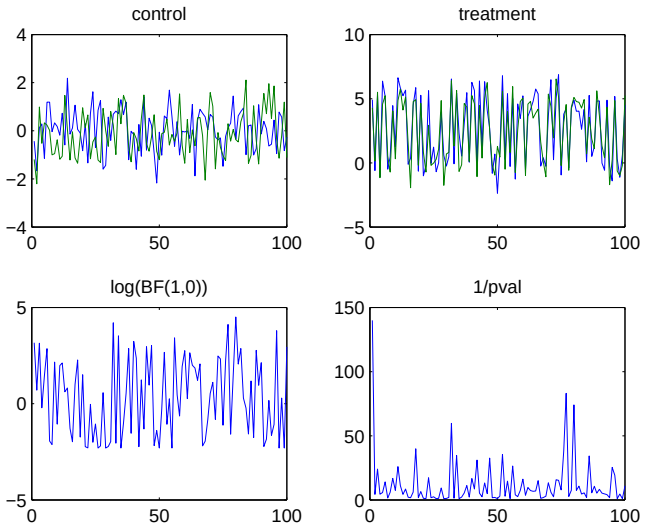

Figure 1: Unsorted data. We have 2 replicates for each of the 100 genes.

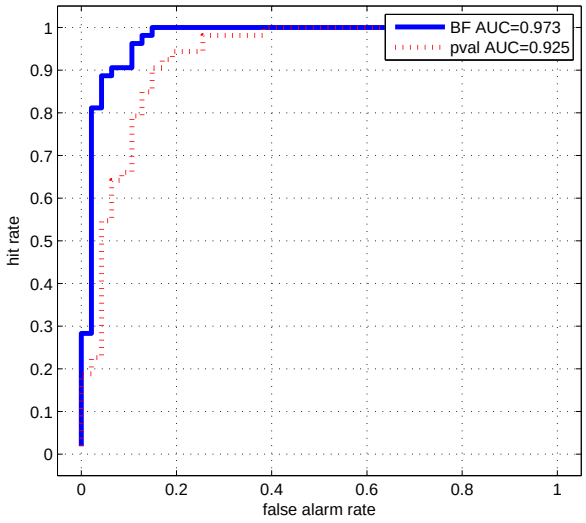

Figure 2: ROC curves

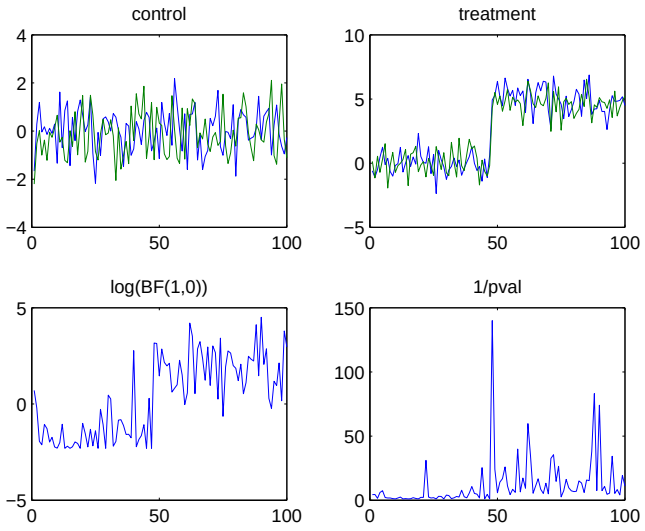

Figure 3: Sorted data

8 Appendix: details on the Bayesian T test

(This section will not be on the exam.) Since we model each gene separately, we will drop the g subscript. The data is $D = (x_{1:n_x}, y_{1:n_y})$ from two groups. The probability model is

$$p(D|M=0) = \int\int \left[\prod_{r=1}^{n_x} \mathcal{N}(x_r|\mu,\sigma^2)\right] \left[\prod_{r=1}^{n_y} \mathcal{N}(y_r|\mu,\sigma^2)\right] p(\mu,\sigma^2) d\mu d\sigma^2 \quad (14)$$

$$p(D|M=1) = \int \left[\int \prod_{r=1}^{n_x} \mathcal{N}(x_r|\mu_x,\sigma^2) p(\mu_x|\sigma^2) d\mu_x\right] \left[\int \prod_{r=1}^{n_y} \mathcal{N}(y_r|\mu,\sigma^2) p(\mu_y|\sigma^2) d\mu_y\right] p(\sigma^2) d\sigma^2 \quad (15)$$

Note that model 0 requires the prior $p(\mu,\sigma^2)$ and model 1 requires the prior $p(\mu_1,\mu_2,\sigma^2)$. Following [GJLW05], we will reparameterize the latter in terms of $\mu = \mu_1$ and $\delta = \mu_2 - \mu_1$, since it is δ that we care about. Let us write

$$p(\mu,\delta,\sigma^2) = p(\mu,\sigma^2)p(\delta|\mu,\sigma^2) \quad (16)$$

Then the marginal likelihood function becomes

$$p(D|M=1) = \int\int\int p(\mu,\sigma^2)p(\delta|\mu,\sigma^2) \left[\prod_{r=1}^{n_x} \mathcal{N}(x_r|\mu,\sigma^2)\right] \left[\prod_{r=1}^{n_y} \mathcal{N}(y_r|\mu+\delta,\sigma^2) p(\mu_y|\sigma^2)\right] d\mu d\sigma^2 d\delta \quad (17)$$

Since μ and σ^2 are common to both models, we can safely give them an uninformative prior

$$p(\mu,\sigma^2) \propto \sigma^{-2} \quad (18)$$

Rather than putting a prior on δ, we will put a prior on δ/σ, the standardized effect size, since this is a more interpretable quantity (dimensionless). Specifically, we will assume

$$p(\delta/\sigma|\sigma,\mu,M=1) \sim \mathcal{N}(\mu_\delta,\sigma_\delta^2) \quad (19)$$

If we do not know whether the difference will be positive or negative, we can set $\mu_\delta = 0$, and make the variance suitably large (say $\sigma_\delta^2 = 100$) to reflect our relative ignorance.

[GWJL05] proves that, under this model, the Bayes factor in favor of the null hypothesis is given by

$$BF(0,1) = \frac{p(D|M=0)}{p(D|M=1)} = \frac{T_\nu(t|0,1)}{T_\nu(t|n_\delta \mu_\delta, 1+n_\delta \sigma_\delta^2)} \tag{20}$$

$$\nu = n_x + n_y - 2 \tag{21}$$

$$n_\delta = \frac{1}{1/n_x + 1/n_y} \tag{22}$$

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{\sum_r (x_r - \overline{x})^2 + \sum_r (y_r - \overline{y})^2}{n_\delta (n_x + n_y - 2)}}} \tag{23}$$

Here t is the standard two-sample t-statistic and $T_\nu(x|a,b)$ is a noncentral t distribution with location a, scale b and dof ν. The provided function `bayesianTtest(x,y)` computes this expression, assuming $\mu_\delta = 0$ and $\sigma_\delta^2 = 100$.

References

[FD06] R. Fox and M. Dimmic. A two-sample Bayesian t-test for microarray data. *BMC Bioinformatics*, 7(126), 2006.

[GJLW05] M. Gonen, W. Johnson, Y. Lu, and P. Westfall. The Bayesian Two-Sample t Test. *The American Statistician*, 59(3):252–257, August 2005.

[Goo99a] S. Goodman. Toward evidence-based medical statistics. 1: The p-value fallacy. *Annals Internal Medicine*, 130(12):995–1004, 1999.

[Goo99b] S. Goodman. Toward evidence-based medical statistics. 2: The bayes factor. *Annals Internal Medicine*, 130(12):1005–1113, 1999.

[GWJL05] M. Gonen, P. Westfall, W. Johnson, and Y. Lu. The Bayesian two-sample t-Test. Technical report, Texas Tech University, 2005.