10-602 *Statistical Approaches to Learning and Discovery*
Spring 2002
Z. Ghahramani and T. Seidenfeld

# Homework #1

**Due**: Wednesday, Feb. 13 at 1:00

*Instructions*: Please submit written answers to each question. We prefer to receive these in printed (as opposed to handwritten) form. You are encouraged to study in groups; however, all of your written work that you submit individually, including homework, is taken by us as your individual effort. If, nonetheless, your written answers reflect collaborative effort, please indicate that to us in an initial footnote. In addition, you are encouraged to use outside references in preparing your answers; however, as with all of your work at the University, you should give citations for your sources.

## *Question 1* (on sufficiency and MLEs):

As a further qualification of sufficiency, consider the notion of a *minimal* sufficient statistic.
*Definition*: Let $\mathbf{t}(X) = Y$ be a sufficient statistic of the data $X$ for the parameter $\theta$. $Y$ is a minimial sufficient statistic if, for each sufficient statistics $\mathbf{s}(X) = Z$ for $\theta$, $Y = \mathbf{h}(Z)$ for some function $\mathbf{h}$.

The following theorem, due to Lehmann and Scheffe (1950), gives a helpful characterization of a minimal sufficient statistic:

Let $\mathbf{P}(X \mid \theta)$ be the probability distribution for a discrete data $X$, given the parameter $\theta$. (In the case of continuous data, let it be the probability density function given $\theta$.)
Then**,** $\mathbf{t}(X)$ is minimally sufficient for $\theta$ (with respect to $X$) if and only if it has the property that:
$\mathbf{t}(X = x') = \mathbf{t}(X = x)$ precisely when the likelihood function $\mathbf{P}(x' \mid \theta)/\mathbf{P}(x \mid \theta)$ is constant as a function of $\theta$.

Let $X = (X_1, \ldots, X_n)$ be a sample of $n$ observations, *iid* given $\theta$ according a statistical model.

Consider three (one-dimensional) statistics of the data $X$:
let $\mathbf{t}_{min}(X) = X_{min}$ be the **min**[X];,
let $\mathbf{t}_{max}(X) = X_{max}$ be the **max**[X];
and let $\mathbf{t}_{med}(X) = X_{med}$ be the **median**[X],

For each of the three statistical models, below,
(1.1) Identify the MLE(s) for the parameter(s).
(1.2) Establish which of the following three reductions of $X$ are sufficient for $\theta$.
(1.3) Show whether any of these three is minimally sufficient.

$Y = X_{max}$
$W = (X_{min}, X_{max})$
and $Z = (X_{min}, X_{med})$

Model$_1$: $X_i$ are *iid* **U**niform $(0, \theta]$, with $\theta > 0$.
Model$_2$: $X_i$ are *iid* **U**niform $[\theta_1, \theta_2]$, with $\theta_2 > \theta_1$ and $(\theta_1, \theta_2) \in \Re^2$.
Model$_3$: $X_i$ are *iid* **U**niform $[\theta, \theta+1]$, with $\theta > 0$.

## *Question 2* (on some conjugate and "improper" priors):

For each of the following five, one-parameter statistical models we offer the conjugate family of prior distributions for the parameter of that model. Assume that the data are a sample of $n$ iid observations from the model.
(2.1) Calculate the posterior (conditional probability) for the statistical parameter of the model and verify the conjugacy of the family.

(2.2) Explain what happens to the posterior distribution for the parameter as the conjugate prior approaches the boundary or boundaries of the family of prior distributions. Indicate whether or not the limit of the conjugate priors is a "proper" probability distribution, or whether it is "improper."

In each case below, you can find brief descriptions of the respective distributions either in the appendix to chapter 1 of Tanner's book, or (particularly for the Negative Binomial) in the text books that we placed on reserve in the E&S Library.

| | *Statistical Model* | *Conjugate Family of Prior distributions* | |
|---|---|---|---|
| 1. | Bernouilli (θ) $0 \leq \theta \leq 1$ | Beta (α,β) | $0 < \alpha, \beta$ |
| 2. | Negative Binomial (m, θ) with fixed m > 0, and $0 < \theta < 1$ | Beta (α,β) | $0 < \alpha, \beta$ |
| 3. | Poisson (λ) $0 < \lambda$ | Gamma (α,β) | $0 < \alpha, \beta$ |
| 4. | Normal (μ, $k^2$) with fixed k > 0, and $\mu \in \Re$ | Normal (θ, $\tau^2$). | $\mu \in \Re, \tau > 0.$ |
| 5. | Normal (m, $\sigma^2$) with fixed $m \in \Re$ and $\sigma > 0$ | Inverse Gamma (α,β) | $0 < \alpha, \beta$ |

*Question 3* (Delta method & Confidence Intervals for a ratio of means; see Tanner's Ch.2, problem 11)
*Let* $X = (X_1, \ldots, X_n)$ be an *iid* sample from $\boldsymbol{N}(\theta,1)$ and let $Y = (Y_1, \ldots, Y_n)$ be an independent *iid* sample from $\boldsymbol{N}(\phi,1)$. Denote by $\overline{X}$ and $\overline{Y}$, respectively, the two sample averages.

(3.1) Use the multivariate delta method to derive the asymptotic distribution of $(\theta/\phi - \overline{X}/\overline{Y})$ for $\phi, \overline{Y} \neq 0$.
Hint: See the discussion of the multivariate delta method in Casella and Berger's book, *Statistical Inference*, 2nd ed., p. 244-245, which is on the Reserve Shelf in the E&S Library.

(3.2) ***OPTIONAL***. In a contrast with this *asymptotic* solution about $(\theta/\phi - \overline{X}/\overline{Y})$, consider the following *exact* solution for inference about θ/φ using $\overline{X}$ and $\overline{Y}$. In an application of Tanner's equation (9.3) from the Appendix to chapter 1, p. 10, use the linear transformations $a(\theta - \overline{X}) + b(\phi - \overline{Y})$ with $a = -1$ and $b = \theta/\phi$. Use the resulting distribution to give a 95% Confidence Interval for θ/φ. Compute this 95% "interval" for data $\overline{X} = \overline{Y} = 1.0$ and n = 1. Are you surprised in any way? How does this solution compare to what you get using the asymptotic delta method?

*Question 4* (on the Neyman-Scott puzzle):
Recall the problem discussed in class on Jan. 30, where the pair $(X_i, Y_i)$ is conditionally *iid* $\mathbf{N}(\mu_i,\sigma^2)$, pairs are independent but not identically distributed given $(\sigma^2, \mu_1, \mu_{2, \ldots})$, where $\sigma^2$ is the parameter of interest – the common variance, and the $\mu_i$ are nuisances -- the unknown means. Recall, that the MLE for the parameter of interest converges to $\sigma^2/2$, almost surely.

For a quasi-Bayesian approach to the problem, known as an "Empirical" Bayes approach, consider the following statistical model for the nuisance parameters.
Let the $\mu_i$ be conditionally *iid* $\boldsymbol{N}(\theta, \sigma^2/\lambda)$. Now,

(4.1) Calculate the likelihood function $P(\langle X,Y\rangle \mid \sigma^2, \theta, \lambda)$. Hint: use the equivalent pair $(Z, W)$, where $Z_i = X_i - Y_i$ and $W_i = X_i + Y_i$.

(4.2) Calculate the MLE for θ and the MLE for $\sigma^2$ (as a function of λ).

(4.3) What is the Fisher Information in these data <X,Y> about the parameter of interest, $\sigma^2$ (as a function of λ)? And how does this compare with the Fisher Information using only the reduced data *Z*, where the $Z_i$ conditionally *iid* $\boldsymbol{N}(0, 2\sigma^2)$, though Z insufficient for $\sigma^2$ with respect to the full data $(X,Y)$.

*Question 5.* Beta-Binomial Models for Word Distributions (with thanks to John Lafferty) [1]

Binomial and multinomial distributions are often used in information retrieval and text processing applications to model word distributions. In this problem you will explore the use of binomial and beta-binomial models for text.

A simple model to predict the number of times a specific word w appears in a document is the binomial, $p_{bin}(k \mid n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$ for observing k occurrences of w in a document of size n. However the parameter $\theta$ may vary across documents, even those that are on the same topic. In this case we can use a mixture of binomials, using a beta prior. Given a collection of documents $d_i$ with lengths $n_i$ and observed counts $k_i$ for word w, the evidence for this data is given by

$$p_{bin-beta}\left(\mathbf{k} \mid \mathbf{n}, \alpha_w, \beta_w\right) = \prod_i \int_0^1 p_{bin}(k_i \mid n_i, \theta)\, p(\theta \mid \alpha_w, \beta_w) d\theta$$

We've made available a corpus of news documents on the course web site and in the file /afs/cs/academic/class/10602-s00/data/a094.tkntext.gz. This is a collection of data from the Topic Detection and Tracking (TDT) project, containing news stories from a variety of media. The collection is broken up into ``document sets," each containing stories from a particular source. Many of the stories are annotated with topic labels, which you will use for this problem. A more detailed description of the format of the file is given in a README in the same directory.

Your task is to estimate and compare binomial and beta-binomial models for specific words. Topic numbers 1, 2, 13 and 15 contain the most documents. Choose three words that are representative of each of these topics. Then, estimate the hyperparameters $\left(\alpha_w, \beta_w\right)$ for each of the words you chose for a topic by maximizing the evidence $p_{bin-beta}\left(\mathbf{k} \mid \mathbf{n}, \alpha_w, \beta_w\right)$ using only documents on that topic. You can use any appropriate numerical procedure, such as Newton's method or conjugate gradient, to estimate the parameters. If you use Matlab, you may find the functions `minimize` and `checkgrad` useful; these can be found at: `http://www.gatsby.ucl.ac.uk/~edward/code`

Generate a plot that compares the fit of the beta-binomial model to the maximum likelihood binomial model for each word you estimate parameters for. What do you conclude about the binomial versus the beta-binomial models? What models might give a better fit? Note: To carry out the estimation, you may find it useful to use simple properties of the digamma function $\Psi(x) = \frac{\partial}{\partial x} \log \Gamma(x)$. For example, $\Psi(x+1) = \Psi(x) + \frac{1}{x}$ and thus

$$\Psi(k+\alpha) - \Psi(\alpha) = \frac{1}{k-1+\alpha} + \frac{1}{k-2+\alpha} + \cdots + \frac{1}{\alpha}$$

Matlab code for evaluating the `digamma` function is available on the course web site.

[1] Reference: S. Lowe, ``The beta-binomial mixture model for word frequencies in documents with applications to information retrieval," Proceedings of EuroSpeech, 1999.