CALD 10–602

Statistical Approaches to Learning and Discovery

Assignment 4

March 29, 2000

Due 12:00 (noon), Thursday, April 13.

This assignment includes two problems involving the Markov chain Monte Carlo method. You will need to design simple MCMC algorithms and implement them for some small data sets. Write up your solutions and also pass in source code (C, C++, or S) into the directory `/afs/andrew/course/10/602/studentir/<id>/ass1`, where `<id>` is your andrew or cs userid.

1. *Posterior Distributions of Log-Odds Ratios.* In this problem you will investigate sampling methods for estimating the posterior distribution of a log-odds ratio. Consider a 2×2 table for two populations:

	Population 1	Population 2
Success	x	y
Failure	$m - x$	$n - y$
	m	n

Assume that x is binomial with parameters (m, π) and that y is an independent binomial with parameters (n, ρ). The *log-odds ratio*

$$\alpha = \log\left(\frac{\pi\,(1-\rho)}{\rho\,(1-\pi)}\right)$$

is a natural statistic for measuring the difference between the two populations (partly because the normal approximations tend to be more accurate than for other statistics such as $\pi - \rho$, which can be highly skewed).

(a) Assume *a priori* that $\log\left(\frac{\pi}{1-\pi}\right)$ and $\log\left(\frac{\rho}{1-\rho}\right)$ are independent and uniform. What prior $p(\pi, \rho)$ does this correspond to? For this prior, write down the joint posterior distribution $p(\pi, \rho \,|\, \text{data})$ for the 2×2 table.

Unfortunately, the posterior distribution of the log-odds ratio for the uniform prior considered above does not have a nice form, and to work with it computationally it is best to use sampling or approximation methods. One approach is to first broaden the problem to consider the joint distribution of *two* statistics rather than a single statistic.

(b) The *log-odds product* is defined as

$$\eta = \log\left(\frac{\pi}{1-\pi}\frac{\rho}{1-\rho}\right).$$

Assuming the prior $p(\pi, \rho)$ above, derive an expression (up to a proportionality constant) for the *joint* posterior distribution $p(\alpha, \eta \,|\, \text{data})$. Since, as you will see, this distribution has a nice form, it can be used to compute the marginal distribution of the log-odds ratio.

As a concrete example to compute with, consider a study of the association between smoking and lung cancer, for which the following data was collected:

	Cancer	Control
Smokers	83	72
Non-smokers	3	14
	86	86

(c) Generate a 2-dimensional contour plot of the joint posterior distribution $p(\alpha, \eta \,|\, \text{data})$ for the cancer data above, together with the normal approximation.

(d) Using the covariance matrix Σ of the normal approximation, a Metropolis algorithm can be defined using candidate distribution $q(\theta, \theta') = \theta + \gamma N_\Sigma(\theta - \theta')$, where N_Σ is a normal distribution with mean zero and covariance matrix Σ, and γ is a control parameter. Use this algorithm to generate samples from the posterior $p(\alpha, \eta \,|\, data)$. Make a scatter plot of the resulting points and compare them to a contour plot of the true distribution. Generate a histogram plot of the marginal distribution on the log-odds ratio α, and compare this to the normal approximation.

(e) Repeat the above using the method of composition.