STA 4503, Spring 2013 — Theory Assignment #4

Due April 10 by 5pm (slip it under my door). Please hand it in on 8 1/2 by 11 inch paper, stapled in the upper left, with no other packaging.

This assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own. In particular, you should not leave any discussion with someone else with any written notes (either on paper or in electronic form).

In this assignment, you will evaluate the variance of two estimators of marginal likelihood or its inverse. For the first estimator — what we will here call the "mean likelihood estimator" — the marginal likelihood is estimated by the average likelihood for parameter values drawn from the prior. For the second estimator — the infamous "'harmonic mean estimator" — the inverse (reciprocal) of the marginal likelihood is estimated by the average of the inverse likelihood for parameter values drawn from the posterior.

The mean likelihood estimator is simply the sample version of the definition of the marginal likelihood, which is

$$Z \;=\; P(x) \;=\; \int P(x|\theta)P(\theta)d\theta$$

Given $\theta_1, \ldots, \theta_N$ sampled independently from the prior, $P(\theta)$, the mean likelihood estimate is

$$\widehat{Z}_m \;=\; \frac{1}{N}\sum_{i=1}^{N} P(x|\theta_i)$$

This is an unbiased estimate of Z, and from the Law of Large Numbers, it is consistent for any proper Bayesian model for which Z is well-defined.

The harmonic mean estimator is simply the sample version of the following formula for the inverse of the marginal likelihood:

$$Z^{-1} \;=\; \int \frac{1}{P(x|\theta)} P(\theta|x)d\theta$$

This formula is valid provided that $P(x|\theta)$ is non-zero for all θ for which $P(\theta)$ is non-zero. Given $\theta_1, \ldots, \theta_N$ sampled independently from the posterior, $P(\theta|x)$, the harmonic mean estimate for Z^{-1} is

$$\widehat{Z}_h^{-1} \;=\; \frac{1}{N}\sum_{i=1}^{N} \frac{1}{P(x|\theta_i)}$$

Given the assumption above, this is an unbiased estimate of Z^{-1}, and from the Law of Large Numbers, it is consistent. In practice, it may be difficult to sample independently from the posterior, but these properties will often still hold for parameter values sampled from the posterior using MCMC.

Consistency of an estimator is a weak property, however. For practical purposes, it is highly desirable for the expressions within the summations above to have finite variances, so that the Central Limit Theorem will apply (though further conditions are needed when the points are not independent, but come from an MCMC run). The accuracy of the estimator will then improve at the usual $1/\sqrt{N}$ rate, and one can then find a confidence interval based on the distribution of the estimator being (asymptotically) Gaussian. Of course, from a practical standpoint, a very large variance is almost as bad as an infinite variance.

In this assignment, you will evaluate the variance of these two estimators for a simple model in which binary observations, $x_1, \ldots, x_n$, are assumed, conditional on a parameter θ in $(0,1)$, to be independent, with all x_i having the Bernoulli(θ) distribution. The prior distribution for θ is Beta(a,b), where a and b are known positive constants.

You should look at the situation when the n observations consist of m zeros and $2m$ ones (so $n = 3m$). You should consider what happens with data sets of this sort for all positive integers m.

Specifically, for all m, a, and b, you should determine whether the variance of $\widehat{Z}_m$ is finite, and whether the variance of $\widehat{Z}_h^{-1}$ is finite. When the variance of one of these estimators is finite, you should find an explicit, and not terribly complex, formula for it, as a function of m, a, b, and N.

From these formulas, you should find formulas for N times the variance of $\log \widehat{Z}_m$ and N times the variance of $-\log \widehat{Z}_h^{-1}$ that are valid in the limit as N goes to infinity.

Finally, you should use theses formulas to investigate, either theoretically or empirically, which of the two methods is better (for data with m zeros and $2m$ ones). Which is better might of course depend on the values of a, b, and m.