SOME FURTHER NOTES ON BAYESIAN STATISTICS

1. The $U[0,1]$ prior for a binomial probability p

This prior dates back to Laplace (1774) or earlier to Bayes (1764), which I have not seen. Suppose we assume this prior and in n independent trials with probability p of success on each, we observe X successes. Then the likelihood function is $\binom{n}{X}p^X(1-p)^{n-X}$. In forming the posterior distribution, the factor $\binom{n}{X}$ not depending on p will cancel. It's clear that as a function of p, the posterior is proportional to a $\mathrm{Beta}(X+1, n-X+1)$ density and so must equal such a density, by unique normalization of probability densities.

The Bayes estimate of p for squared-error loss is the integral of p times the posterior density. As the expectation for a $\mathrm{Beta}(a,b)$ density is $a/(a+b)$, in this case we get the Bayes estimate

$$\hat{p}_B = T(X) = \int_0^1 p\pi_X(p)dp = \frac{X+1}{n+2}. \tag{1}$$

If for the true $p = p_0$, $0 < p_0 \leq 1$, then as $n \to \infty$, since $X/n \to p_0$, also $X \to +\infty$, and $\hat{p}_B$ in (1) will be asymptotic to $\hat{p} = X/n$, the usual maximum likelihood estimate of p. If the true $p_0 = 0$, then $X \equiv 0$, $X/n \equiv 0$, and $\hat{p}_B = 1/(n+2) \to 0$. On the other hand for $n = 0$, $X = 0$ and $\hat{p}_B = 1/2$, the expectation for the prior $U[0,1]$ distribution.

2. Comparisons with unbiased estimation

We've already seen some examples earlier in the course showing that unbiased estimation leads to some undesirable or non-optimal estimates. One was Yatracos's proof that the usual sample variance s_X^2, an unbiased estimator of the true variance σ^2 when it is finite, is strongly inadmissible for mean-square error, for i.i.d. variables X_j having a finite fourth moment $E(X_j^2)$, being not as good as an estimator in which the factor $1/(n-1)$ is replaced by a slightly smaller factor. Another case was estimating the function $g(p) = p^2$ of a binomial parameter p for $n = 2$, which has a unique unbiased estimator T with $T(1) = 0$; that was paradoxical since if $p^2 = 0$ then $p = 0$ and the probability of 1 success would be 0.

Date: 18.650, Dec. 4, 2015.

Recently, in Theorem 4 of our handout "Some topics in Bayesian statistics," we saw that an estimator $T(X)$ of a function $g(\theta)$ which is Bayes for mean-squared error can be unbiased only if $T(X) = g(\theta)$ with probability 1, which is hardly ever possible. So, if being Bayes and being unbiased are virtually incompatible, which property, if either, is to be preferred? This gives statisticians preferring a Bayesian viewpoint a reason to look for examples where unbiased estimation works badly.

The textbook by DeGroot and Schervish (2002,2012) takes predominantly a Bayesian viewpoint. (It has been adopted here at MIT for 18.443, predecessor of 18.650, in the fall many times in the past.) The book gives an interesting example (which I saw in the 2002 edition) of unbiased estimation of the parameter p in a subgeometric distribution, which works equally well for a geometric distribution. Recall that for $0 < p \le 1$ and $q = 1 - p$, in a sequence of independent trials with probability p of success on each, the number X of trials needed to get the first success has $\Pr(X = k) = q^{k-1}p$ for $k = 1, 2, ...$, a geometric distribution, and $Y \equiv X - 1$, the number of failures before the first success, has the distribution $\Pr(Y = j) = q^j p$ for $j = 0, 1, ...$, called in this course a subgeometric distribution. Let $T(X)$ be an unbiased estimator of p. Then $\sum_{k=1}^{\infty} T(k)q^{k-1}p = p$ for $0 < p \le 1$ which implies $\sum_{k=1}^{\infty} T(k)q^{k-1} = 1$. Since a power series in the variable q has uniquely determined coefficients, we must have $T(1) = 1$ and $T(k) = 0$ for $k \ge 2$. Likewise, an unbiased estimator $V(Y)$ of p must have $V(0) = 1$ and $V(j) = 0$ for $j \ge 1$. These estimates are unreasonable. A reasonable estimator of p is $1/X = 1/(Y + 1)$, which in a problem we found to be the MLE from two viewpoints (geometric and binomial) and the method-of-moments estimator.

Another example DeGroot and Schervish give is, for the Poisson parameter λ, estimating $g(\lambda) = e^{-2\lambda}$. Let T be an unbiased estimator of g. Then

$$\sum_{k=0}^{\infty} T(k)e^{-\lambda}\frac{\lambda^k}{k!} = e^{-2\lambda}$$

for $0 < \lambda < \infty$, which implies by Taylor series

$$\sum_{k=0}^{\infty} T(k)\frac{\lambda^k}{k!} = e^{-\lambda} = \sum_{k=0}^{\infty} \frac{(-\lambda)^k}{k!}$$

which gives $T(k) = (-1)^k$ for all $k = 0, 1, ...$ by unique coefficients of power series, an absurd estimator. A similar case was the problem of unbiased estimation of $e^{-\lambda}$ given an observation X of a Poisson(λ) variable conditional on $X \ge 1$, which by a slightly longer calculation gives $T(k) = (-1)^{k-1}$ for $k \ge 1$. This case is perhaps more motivated,

in that if one is not seeing certain observations (with value 0) one could be interested in the probability of not seeing one.

If the function being estimated is positive, as $e^{-2\lambda}$ and $e^{-\lambda}$ are, then for any prior, a Bayes estimator T of it could not take negative values such as -1, because replacing T by $\max(T, 0)$ would reduce the mean-square error.

Overall then, it seems that unbiased estimation is looking not very good.

3. Maximum likelihood estimation

We've found maximum likelihood estimates (MLEs) to be useful in several situations. For example, in χ^2 tests of composite hypotheses, MLEs based on ungrouped data are often easy to compute. Sometimes, as with contingency tables, MLEs based on grouped data are also easy to compute. Whenever they can be computed, even with difficulty, they give a well-defined number of degrees of freedom for χ^2.

When comparing models of different dimensions, one cannot simply maximize the likelihood. For example, suppose we have two models $H_0 \subset H_1$ of respective dimensions $d_0 < d_1$. The likelihood maximized over H_1 is nearly always going to be larger than the maximum over H_0. But in the Wilks test, H_0 is not rejected if twice the difference in maximum log likelihoods, $W = 2(MLL_1 - MLL_0)$, is not too large, where under H_0 for n large enough it has approximately a χ^2 distribution with $d_1 - d_0$ degrees of freedom. Thus, we use maximum likelihood within each model, but use something more when comparing models.

When we have more than two models, of different dimensions, and want to choose a best model, then the Bayes Information Criterion (BIC) gives a method, described in the corresponding handout. The leading terms in the logarithms of the posterior probabilities of the models give $BIC_i = MLL_i - \frac{1}{2} d_i \log n$ (equation (1) of the BIC handout). Here again we maximize the likelihood within each model, but in this case we subtract a penalty for dimension d_i, a larger penalty the higher the dimension is. The leading terms don't depend on the specifics of the prior distributions as long as they satisfy some mild conditions such as strictly positive prior densities and prior probabilities for each model.

For comparing models two at a time, with one included in the other, it can be interesting to see if the BIC preference between the models agrees with the result of the Wilks test, as will be done in a problem in PS10.

REFERENCES

Bayes, T. (1764), An essay toward solving a problem in the doctrine of chances, *Philos. Trans. Roy. Soc. London* **53**, 370-418; repr. in *Biometrika* **45** (1958) 293-315.

DeGroot, M. H., and Schervish, M. J., *Probability and Statistics*, 3d. ed. 2002, 4th ed. 2012, Addison-Wesley.

Dudley, R. M. (2003). *Mathematical Statistics*, 18.466 lecture notes, Spring 2003. On MIT OCW (OpenCourseWare) website, 2004.

Laplace, P. S. (1774), "Memoir on the probability of the causes of events," *Mémoires de Math. et de Physique, Présentés à l'Académie Royale des Sciences, par divers Savans & lûs dans ses Assemblées*, **6**, 621-646; transl. by S. M. Stigler in Stigler (1986), pp. 364–378.

Stigler, S. M. (1986), Laplace's 1774 memoir on inverse probability, *Statistical Science* **1**, 359-378.