Characteristic functions, inequalities, models, likelihood, SNLM

October 1, 2013

Characteristic functions of distributions

The characteristic function of a random variable X is a function $f : \mathbb{R} \to \mathbb{C}$ defined by

$$f(\omega) = E[e^{-i\omega X}] \, .$$

Properties of characteristic functions

- Every distribution on the real line has a characteristic function, which is always bounded in absolute value by 1.
- A characteristic function f satisfies $f(\omega) = \overline{f(-\omega)}$, where the bar over the right-hand side represents complex conjugation.
- The sum of two independent random variables with characteristic functions f and g has characteristic function $f \cdot g$.
- A distribution that is symmetric about zero has a characteristic function whose values are all real — no imaginary part.

- A characteristic function f always has $f(0) = 1$.

- Lévy's continuity theorem: If the sequence of random variables $\{X_n\}$ have characteristic functions $\{f_n\}$, and if $f_n(\omega) \to f(\omega)$ for every ω, then f is a characteristic function for some random variable X, and $X_n \xrightarrow{D} X$.

Markov and Chebyshev inequalities

Markov $P[X \geq 0] = 1$ and $E[X] = M \Rightarrow P[X > c] \leq \frac{M}{c}$.

Chebyshev $E[X^2] = \sigma^2 \Rightarrow P[|X| > c] \leq \frac{\sigma^2}{c^2}$.

Markov and Chebyshev inequalities

Markov $P[X \geq 0] = 1$ and $E[X] = M \Rightarrow P[X > c] \leq \frac{M}{c}$.

Chebyshev $E[X^2] = \sigma^2 \Rightarrow P[|X| > c] \leq \frac{\sigma^2}{c^2}$.

These inequalities give very weak bounds in many cases, but they can't be improved without further restrictions on the distributions, because there are distributions for which the inequalities are equalities.

A catalog of useful distributions

$$\begin{aligned}
\text{Gamma}(n,\alpha): \quad & p(x) \propto \alpha^n x^{n-1} e^{-\alpha x} \text{ on } (0,\infty) && EX = n/\alpha, \text{Var}(X) = n/\alpha^2 \\
\text{Beta}(n,m): \quad & p(x) \propto x^{n-1}(1-x)^{m-1} \text{ on } (0,1) && EX = n/(n+m) \\
\text{Poisson}(\lambda): \quad & p(x) = e^{-\lambda}\frac{\lambda^x}{x!} \text{ on } \mathbb{Z}^+ && EX = \lambda, \text{Var}(x) = \lambda
\end{aligned}$$

$$\text{Wishart}(\nu, S): p(x) \propto |S|^{-\nu/2} |x|^{(\nu-k-1)/2} \exp(-\tfrac{1}{2}\operatorname{tr}(Sx^{-1}))$$

on $k \times k$ positive definite matrices x, $\quad EX = \nu S$

See, e.g. Appendix A of Gelman, Carlin, Stern and Rubin for more distributions, more detail, including normalizing constants.

The multivariate normal

$N(\mu, \Sigma)$ $\qquad$ $\phi(x \mid \mu, \Sigma) = (2\pi)^{-k/2} \, |\Sigma|^{-k/2} \exp\big((x-\mu)'\Sigma^{-1}(x-\mu)\big)$ on $\mathbb{R}^k$

A convention: Φ for normal cdf, ϕ for normal pdf.

Take any $n \times k$ real matrix A. If $X \sim N(\mu, \Sigma)$, $AX \sim N(A\mu, A\Sigma A')$. The transformations of mean and variance matrix holds for any random vector. What's special about the normal is that the linear transformations always leave the distribution normal.

Statistics

- Statistics, or inference, is learning from data.
- Probability-based statistics is learning from data when we treat the data as drawn from some probability distribution.
- Bayesian statistics is learning from data when we treat all unknown quantities as drawn from some probability distribution, whether we are going to see them ("data") or not ("parameters", "latent variables").

Models?

- Some radical Bayesians grumble about the very notion of a model. In this purist view, there are only decision problems.

Models?

- Some radical Bayesians grumble about the very notion of a model. In this purist view, there are only decision problems.

- The problem begins with some joint distribution over X, Y, we then discover the realized values of X, use these observations to generate the $Y \mid X$ conditional distribution, and solve our decision problem.

Models

- We postulate that we know a conditional distribution for Y with pdf $p(Y \mid \beta)$. Usually we assume we are going to see Y and that we don't know, and will not see, β. Y: "data"; β: "parameter".

- This model is taken to be "objective". For frequentists: grounded in some accepted theory.

- For Bayesians an "objective" probability distribution is simply one that is widely accepted as correct, or at least interesting to explore, in some community that is sharing research results or working on a decision problem.

Likelihood principle

For Bayesian inference, the likelihood principle is obvious. Given a model $p(Y \mid \beta)$ and a prior $\pi(\beta)$, we form the posterior by Bayes' rule as

$$\frac{p(Y \mid \beta)\pi(\beta)}{\int p(Y \mid \beta)\pi(\beta)\, d\beta},$$

And this object, is everything we learn from the data. It depends on the data only through $p(Y \mid \beta)$, the likelihood function.

Implications of the likelihood principle: sufficient statistics

If $\{x_1, \ldots, x_n\}$ is an i.i.d. sample from the exponential distribution ($\mathrm{Gamma}(1, \alpha)$), the likelihood of the sample is

$$\alpha^n e^{\alpha \sum_j x_j} .$$

The likelihood thus depends on the data only via its sum, $\sum_j x_j$. When this happens — the likelihood depends on the data only through a low-dimensional set of functions of the data — the functions of the data that determine the likelihood are **sufficient statistics**. When there is a set of sufficient statistics, Bayesian inference depends only on them.

A frequentist result about sufficient statistics

Theorem 1. [Rao-Blackwell] *If $\hat{\beta}$ is an unbiased estimate of β based on a sample $\vec{x}$ whose likelihood has a sufficient statistic $S(\vec{x})$, then if $\hat{\beta}$ is not a function of $S(\vec{x})$ alone, $\hat{\hat{\beta}} = E[\hat{\beta} \mid S(\vec{x})]$ is an unbiased estimate of β with lower variance than $\hat{\beta}$.*

Samples we didn't see

- Suppose we have a sample vector $\vec{x}$ consisting of n i.i.d. draws from a truncated exponential distribution: For each observation j, x_j^* is drawn from the exponential distribution with pdf $\alpha e^{-\alpha x_j^*}$, and $x_j = \min\left\{100, x_j^*\right\}$.

Samples we didn't see

- Suppose we have a sample vector $\vec{x}$ consisting of n i.i.d. draws from a truncated exponential distribution: For each observation j, x_j^* is drawn from the exponential distribution with pdf $\alpha e^{-\alpha x_j^*}$, and $x_j = \min\left\{100, x_j^*\right\}$.

- Suppose further that the particular sample at hand consists of 10 x_j values, all less than 100. This kind of thing can occur in economic data when some items are "top-coded" — income, e.g., might be recorded only as "over $200,000" if it is very high.

Bayesian interpretation of this sample

The likelihood function for this sample, because no truncation has actually occurred in the sample, is just

$$\alpha^{10} e^{-\alpha \sum_j x_j} ,$$

just as if there were no truncation possible. A Bayesian, or anyone else believing in the likelihood principle, should then treat this sample as having exactly the same implications for the unknown parameter α as any other sample with the same $\sum_j x_j$ drawn from a setting where there is no truncation. If the problem is to estimate α, the posterior expectation of α under a flat prior is, since the likelihood has the form of a $\mathrm{Gamma}(11, \sum_j x_j)$ pdf, $\hat{\alpha} = 11/\sum x_j$.

Frequentist interpretation

For a frequentist, the problem is harder. The Bayesian does not have to specify what his or her decision or estimator would be in other samples, but the frequentist must. If the problem is to estimate α, an unbiased estimator without the truncation would be

$$\hat{\hat{\alpha}} = \frac{9}{\sum x_j} .$$

However, because of the truncation, this estimator is obviously biased upwards in a frequentist sense. I believe there is no way to correct it to make it unbiased, though I could be wrong.

The stopping rule principle

Suppose we are told that in a sample of size 1253 classrooms, SAT scores in class rooms where the teacher had a graduate degree were higher on average by 8 points, and that the standard error of this estimate was 4, so the difference was "statistically significant" by the usual measures. The conventional conclusion would be that this means there is an effect of the graduate degree training that is statistically meaningful.

But now we discover that actually the study director herself has a graduate degree and planned to continue adding classrooms to her sample until she found a t statistic (ratio of the estimated score difference to its standard error) of 2, at which point she stopped her data collection. That the result would be "statistically significant" was therefore determined from the start. A frequentist would therefore say that the result is not at all statistically significant if this is the way the study was conducted.

Implications of the likelihood principle for the stopping rule example

Suppose that the pdf of the data for a given classrom is $p(S-\mu-g\theta)q(g)$, where $g = 1$ if the teacher has a graduate degree and is 0 otherwise, θ is the effect of the teacher's graduate degree on S, and S is the class average test score. And suppose the sample size is determined by the "t statistic of 2 on θ" rule described above.

Stopping rule example, continued

The joint pdf of the sample of n S_j, g_j values is

$$\prod_{1}^{n-1} p(S_j - \mu - \theta g_j)(1 - q_j(S_1, \ldots, S_j, g_1, \ldots, g_j))$$

$$\cdot\, p(S_n - \mu - \theta g_n) q_n(S_1, \ldots, S_n, g_1, \ldots, g_n)\,,$$

where $q_j(S_1, \ldots, S_j, g_1, \ldots, g_j)$ is the probability of the sampling being terminated after the j'th draw. In the case we are discussing, the q_j's are all zero or (at $j = n$) one, so they drop out of the likelihood. Since in any case they do not depend on the unknown parameters μ and θ, they affect the likelihood only as a scale factor and drop out when it, or the posterior pdf, is normalized.

Bayesian inference with the "t statistic of 2" stopping rule

We will come back to a formal analysis of this case later. For now, we just note that Bayesian inference in a situation like this, where there is non-zero prior probability on a single point, but also probability on a continuous interval of points, treats t statistics differently according to whether they come from a large or small sample size. It makes decisions a function of the likelihood function, but not of the t statistic.

The stopping rule principle, which says that inference is the same whether the sample size was fixed in advanced or determined by the "sample until the t-statistic is 2" rule, is unreasonable if one insists on making decisions based on the size of the t statistic alone. But Bayesian inference does not do that.

Priors: Ignorance

- One way to proceed with Bayesian inference is simply to normalize the likelihood to integrate to one and treat that as a posterior pdf.

Priors: Ignorance

- One way to proceed with Bayesian inference is simply to normalize the likelihood to integrate to one and treat that as a posterior pdf.

- If the parameter space is bounded, this is equivalent to using a prior with a constant pdf — a "flat prior".

Priors: Ignorance

- One way to proceed with Bayesian inference is simply to normalize the likelihood to integrate to one and treat that as a posterior pdf.
- If the parameter space is bounded, this is equivalent to using a prior with a constant pdf — a "flat prior".
- This looks appealing as reflecting "ignorance": all parameter values are treated as equally likely.
- But note that it doesn't work if the likelihood is not integrable.

Ignorance?

- A flat prior on θ is not flat on $f(\theta)$, where f is some monotone function, unless f is linear.

- For example, what is an "ignorance" prior on a variance parameter?

Likelihood dominance, reporting to diverse audience

Jeffreys priors

Conjugate priors

$$p(y_1, \ldots, y_n \mid \theta)\pi(\theta) \propto p(y_1, \ldots, y_n, y^*_{n+1}, y^*_{n+2}, \ldots, y^*_{n+m} \mid \theta)$$

Examples:

- Exponential, Gamma
- Normal with known σ^2.
- with unknown σ^2?

- Beta
- Double exponential

SNLM