Fundamental Theory of Statistical Inference

G. Alastair Young

Department of Mathematics
Imperial College London

LTCC, 2017

Fundamental ideas

In classical statistics Y is random, with a density given by $f(y;\theta)$, but θ is a fixed unknown parameter.

In Bayesian statistics Y and θ are both regarded as random variables, with joint density given by $\pi(\theta)f(y;\theta)$ where $\pi(\theta)$ represent the prior density of θ and $f(y;\theta)$ is the conditional density of Y given θ.

Posterior density

The posterior density of θ, conditional on the observed value $Y = y$, is derived by applying Bayes' rule:

$$\pi(\theta|y) = \frac{\pi(\theta)f(y;\theta)}{\int_{\Omega_\theta} \pi(\theta')f(y;\theta')d\theta'}.$$

We write

$$\pi(\theta|y) \propto \pi(\theta)f(y;\theta)$$

where the constant of proportionality does not depend on θ.

So:

$$\text{posterior} \;\propto\; \text{prior} \;\times\; \text{likelihood}$$

since $f(y;\theta)$, treated as a function of θ for fixed y, is called the likelihood function.

An example

Suppose $Y \sim \mathrm{Bin}(n, \theta)$ for known n and unknown θ. Suppose the prior density is a Beta density on (0,1),

$$\pi(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}, \quad 0 < \theta < 1,$$

where $a > 0, b > 0$. For the density of Y, we have

$$f(y;\theta) = \binom{n}{y}\theta^{y}(1-\theta)^{n-y}.$$

Ignoring all components of π and f which do not depend on θ, we have

$$\pi(\theta|y) \propto \theta^{a+y-1}(1-\theta)^{n-y+b-1}.$$

This is also of Beta form, with the parameters a and b replaced by $a+y$ and $b+n-y$, so the full posterior density is

$$\pi(\theta|y) = \frac{\theta^{a+y-1}(1-\theta)^{n-y+b-1}}{B(a+y, b+n-y)}.$$

A general property

A very general property of Bayesian statistical procedures is that in large samples they give answers which are very similar to the answers provided by classical statistics.

Nevertheless, in small samples the procedures do lead to different answers, and the Bayesian solution does depend on the prior adopted.

Conjugate Priors

In the example, by adopting a prior density of Beta form, we obtained a posterior density which was also a member of the Beta family, but with different parameters.

When this happens, the common parametric form of the prior and posterior are called a conjugate prior family for the problem.

Often very convenient, because it avoids having to integrate to find the normalising constant in the posterior density.

Another example

Suppose $Y_1, ..., Y_n$ are IID $N(\theta, \sigma^2)$ where the mean θ is unknown and the variance σ^2 is known. Assume that the prior density for θ is $N(\mu_0, \sigma_0^2)$ with μ_0, σ_0^2 known.

Denote by Y the vector $(Y_1, ..., Y_n)$ and let its observed value be $y = (y_1, ..., y_n)$. Ignoring all quantities that do not depend on θ, the prior × likelihood can be written in the form

$$\pi(\theta) f(y; \theta) \propto \exp\left\{ -\frac{(\theta - \mu_0)^2}{2\sigma_0^2} - \sum_{i=1}^{n} \frac{(y_i - \theta)^2}{2\sigma^2} \right\}.$$

Completing the square shows that

$$\begin{aligned} &\tfrac{(\theta-\mu_0)^2}{\sigma_0^2} + \textstyle\sum_{i=1}^{n} \tfrac{(y_i-\theta)^2}{\sigma^2} \\ &= \theta^2 \left(\tfrac{1}{\sigma_0^2} + \tfrac{n}{\sigma^2}\right) - 2\theta \left(\tfrac{\mu_0}{\sigma_0^2} + \tfrac{n\bar{y}}{\sigma^2}\right) + \text{ Const} \\ &= \tfrac{1}{\sigma_1^2}(\theta - \mu_1)^2 + \text{ Const}, \end{aligned}$$

where $\bar{y} = \sum y_i/n$, "Const" denotes a quantity which does not depend on θ.

Here μ_1 and σ_1^2 are defined by

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}, \quad \mu_1 = \sigma_1^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{y}}{\sigma^2} \right).$$

Therefore the posterior density is the normal density with mean μ_1 and variance σ_1^2. Another example of a conjugate prior family.

The general form of Bayes rule

How to solve Bayesian decision problems.

Writing the joint density $f(y;\theta)\pi(\theta)$ in a different way as $f(y)\pi(\theta|y)$, where $f(y)=\int f(y;\theta)\pi(\theta)d\theta$ is the marginal density of Y, we have

$$\begin{aligned} r(\pi,d) &= \int_{\Omega_\theta} R(\theta,d)\pi(\theta)d\theta \\ &= \int_{\Omega_\theta}\int_{\mathcal{Y}} L(\theta,d(y))f(y;\theta)\pi(\theta)dyd\theta \\ &= \int_{\Omega_\theta}\int_{\mathcal{Y}} L(\theta,d(y))f(y)\pi(\theta|y)dyd\theta \\ &= \int_{\mathcal{Y}} f(y)\left\{\int_{\Omega_\theta} L(\theta,d(y))\pi(\theta|y)d\theta\right\}dy. \end{aligned}$$

Because $f(y) \geq 0$ for all y, we see that to find the Bayes rule $d(y)$ for any y, it suffices to minimise the expression inside the brackets: for each y we choose $d(y)$ to minimise

$$\int_{\Omega_\theta} L(\theta, d(y))\pi(\theta|y)d\theta,$$

the expected posterior loss associated with the observed y.

Result highlights natural property of Bayesian procedures: in order to decide what to do based on a particular observed y, only have to think about the losses that follow from one value $d(y)$. Don't have to worry (as with, say, minimax procedure) about all the other values of y that might have occurred, but did not.

Case 1: Hypothesis testing

Testing the hypothesis $H_0 : \theta \in \Theta_0$ against the hypothesis $H_1 : \theta \in \Theta_1 \equiv \Omega_\theta \setminus \Theta_0$, the complement of Θ_0. Now the action space $\mathcal{A} = \{a_0, a_1\}$, where a_0 denotes 'accept H_0' and a_1 denotes 'accept H_1'.

Assume the following form of loss function:

$$L(\theta, a_0) = \begin{cases} 0 & \text{if } \theta \in \Theta_0, \\ 1 & \text{if } \theta \in \Theta_1, \end{cases}$$

and

$$L(\theta, a_1) = \begin{cases} 1 & \text{if } \theta \in \Theta_0, \\ 0 & \text{if } \theta \in \Theta_1. \end{cases}$$

The Bayes decision rule is: accept H_0 if

$$\Pr(\theta \in \Theta_0 | y) > \Pr(\theta \in \Theta_1 | y).$$

Since $\Pr(\theta \in \Theta_1 | y) = 1 - \Pr(\theta \in \Theta_0 | y)$, this is equivalent to accepting H_0 if $\Pr(\theta \in \Theta_0 | y) > 1/2$.

Case 2: Point estimation

Squared error: $L(\theta, d) = (\theta - d)^2$. For observed $Y = y$, the Bayes estimator chooses $d = d(y)$ to minimise

$$\int_{\Omega_\theta} (\theta - d)^2 \pi(\theta|y) d\theta.$$

Differentiating with respect to d, we find

$$\int_{\Omega_\theta} (\theta - d)\pi(\theta|y)d\theta = 0.$$

The posterior density integrates to 1, so this becomes

$$d = \int_{\Omega_\theta} \theta\pi(\theta|y)d\theta,$$

the posterior mean of θ. Bayes estimator is the mean of the posterior distribution.

Case 3: Point estimation

Suppose $L(\theta, d) = |\theta - d|$. The Bayes rule will minimise

$$\int_{-\infty}^{d} (d - \theta)\pi(\theta|y)d\theta + \int_{d}^{\infty} (\theta - d)\pi(\theta|y)d\theta.$$

Differentiating with respect to d, we must have

$$\int_{-\infty}^{d} \pi(\theta|y)d\theta - \int_{d}^{\infty} \pi(\theta|y)d\theta = 0$$

or

$$\int_{-\infty}^{d} \pi(\theta|y)d\theta = \int_{d}^{\infty} \pi(\theta|y)d\theta = \frac{1}{2}.$$

The Bayes rule is the posterior median of θ.

Case 4: Interval estimation

Suppose

$$L(\theta, d) = \begin{cases} 0 & \text{if } |\theta - d| \le \delta, \\ 1 & \text{if } |\theta - d| > \delta, \end{cases}$$

for prescribed $\delta > 0$.

The expected posterior loss is the posterior probability that $|\theta - d| > \delta$.

Motivated as a Bayesian version of interval estimation: we want to find the best interval of the form $(d - \delta, d + \delta)$ (of predetermined length 2δ). Best here means the interval that maximises the posterior probability that θ is within the interval specified.

The resulting interval is often called the HPD (for highest posterior density) interval.

Shrinkage and Stein's Paradox

Let Y have a p-dimensional ($p \geq 3$) normal distribution with mean vector μ and known covariance matrix equal to the identity I, so that $Y_i \sim N(\mu_i, 1)$, independently, $i = 1, \ldots, p$.

Consider estimation of μ, with loss function $L(\mu, d) = \|\mu - d\|^2 = \sum_{i=1}^{p}(\mu_i - d_i)^2$ equal to the sum of squared errors.

James-Stein estimator

Consider the class of estimators of the form

$$d^a(Y) = \left(1 - \frac{a}{Y^T Y}\right) Y,$$

indexed by $a \geq 0$, which shrink Y towards 0

Risks

$Y \equiv d^0(Y)$ has risk

$$\begin{aligned} R(\mu, d^0(Y)) &= E\|\mu - Y\|^2 \\ &= \sum_{i=1}^{p} E(\mu_i - Y_i)^2 = \sum_{i=1}^{p} \mathrm{var}\, Y_i \\ &= p, \end{aligned}$$

for all μ.

Stein's lemma

Integration by parts shows that, for each i, for suitably behaved real-valued functions h,

$$E\{(Y_i - \mu_i)h(Y)\} = E\left\{\frac{\partial h(Y)}{\partial Y_i}\right\}.$$

Then

$$\begin{aligned} R\big(\mu, d^a(Y)\big) &= E\|\mu - d^a(Y)\|^2 \\ &= E\|\mu - Y\|^2 - 2aE\left[\frac{Y^T(Y-\mu)}{Y^TY}\right] \\ &\quad +a^2E\left[\frac{1}{Y^TY}\right]. \end{aligned}$$

We have

$$\begin{aligned}
E\left[\frac{Y^T(Y-\mu)}{Y^TY}\right] &= E\left[\sum_{i=1}^{p}\frac{Y_i(Y_i-\mu_i)}{\Sigma_j Y_j^2}\right] \\
&= \sum_{i=1}^{p} E\left[\frac{\partial}{\partial Y_i}\left\{\frac{Y_i}{\Sigma_j Y_j^2}\right\}\right] \\
&= \sum_{i=1}^{p} E\left[\frac{\Sigma_j Y_j^2 - 2Y_i^2}{(\Sigma_j Y_j^2)^2}\right] \\
&= E\left[\frac{p-2}{Y^TY}\right],
\end{aligned}$$

so

$$R\big(\mu, d^a(Y)\big) = p - \left[2a(p-2) - a^2\right] E\left(\frac{1}{Y^TY}\right).$$

Discussion

The obvious estimator of μ is Y. But we then note that $R(\mu, d^a(Y)) < p = R(\mu, d^0(Y))$ provided $2a(p-2) - a^2 > 0$. For such a, $d^a(Y)$ strictly dominates $d^0(Y)$, so that the obvious estimator Y is inadmissible!

Note also that the risk of $d^a(Y)$ is minimised for $a = p - 2$. When $\mu = 0$, $Y^T Y \sim \chi^2_p$, so that $E[1/(Y^T Y)] = 1/(p-2)$, by direct calculation. Hence $d^{p-2}(Y)$ has risk $p - [(p-2)^2/(p-2)] = 2$ when $\mu = 0$.

The result seems incredible. The components of the vector Y are independent, and the loss function is ordinary squared error loss: there is no apparent tying together of the losses in different components yet the obvious estimator is not admissible.

It is now known that this is a very general phenomenon when comparing three or more populations.

Further discussion

- $d(Y) = Y$ is admissible if $p = 1$ or 2.
- Estimator $d^{p-2}(Y)$ is actually inadmissible!

Empirical Bayes

In a standard Bayesian analysis, there will usually be parameters in the prior distribution that have to be specified.

For example, consider the model in which $Y \mid \theta \sim N(\theta, 1)$ and θ has the prior distribution $\theta \mid \tau^2 \sim N(0, \tau^2)$. If a value is specified for the parameter τ^2 of the prior, a standard Bayesian analysis can be carried out.

What if τ^2 is not specified?

It is readily shown that the marginal distribution of Y is $N(0, \tau^2 + 1)$, and can therefore provide information on τ^2.

Empirical Bayes analysis is characterised by the estimation of prior parameter values from marginal distributions of data. Having estimated prior parameter values, proceed as if these values are fixed.

Stein's paradox revisited

The estimator $d^{p-2}(Y)$ may be viewed as an empirical Bayes estimator of μ: the Bayes rule with prior parameter values replaced by estimates constructed from the marginal distribution of the Y_i.

Let $Y_i \mid \mu_i$ be distributed as $N(\mu_i, 1)$, independently $i = 1, \ldots, p$, and suppose $\mu_1, \ldots, \mu_p$ are IID $N(0, \tau^2)$. If τ^2 is known, the Bayes estimator, for the given squared errors loss, of $\mu = (\mu_1, \ldots, \mu_p)^T$ is the posterior mean $\frac{\tau^2}{\tau^2+1} Y$.

Empirical Bayes interpretation

Marginally the Y_i are IID $N(0, \tau^2 + 1)$, so that

$$E\left[1 - \frac{(p-2)}{Y^T Y}\right] = \frac{\tau^2}{\tau^2 + 1},$$

if $p \geq 3$. Estimating $\tau^2/(\tau^2+1)$ by $1 - (p-2)/(Y^T Y)$ yields the Stein estimator $d^{p-2}(Y)$.

Hierarchical Modelling

An alternative way of dealing with the specification of prior parameter values is with a hierarchical specification. The prior parameter values are themselves given a (second-stage) prior.

For example, in the normal model we might specify $Y \mid \theta \sim N(\theta, 1)$, $\theta \mid \tau^2 \sim N(0, \tau^2)$ and $\tau^2 \sim \mathrm{uniform}(0, \infty)$, an 'improper' prior.

Inference on θ is based on the marginal posterior of θ, after integrating out τ^2 from the joint posterior of θ and τ^2:

$$\pi(\theta \mid y) = \int \pi(\theta, \tau^2 \mid y) d\tau^2,$$

where the joint posterior $\pi(\theta, \tau^2 \mid y) \propto f(y; \theta)\pi(\theta \mid \tau^2)\pi(\tau^2)$.

Very effective practical tool and usually yields answers that are reasonably robust to misspecification of the model. Often, answers from a hierarchical analysis are quite similar to those obtained from an empirical Bayes analysis.

Predictive distributions

We may not be interested directly in parameter θ, but in some independent future observation depending on θ.

Possible to obtain the conditional distribution of the value of a future observation $Y^\dagger$, given the data y, from the posterior $\pi(\theta \mid y)$.

Details

Suppose that $y = (y_1, \ldots, y_n)$, with the y_i independent from $f(y; \theta)$. Since, given θ, $Y^\dagger$ and y are independent and $Y^\dagger$ has density $f(y^\dagger; \theta)$, the posterior joint distribution of $Y^\dagger$ and θ is $f(y^\dagger; \theta)\pi(\theta \mid y)$. Integrating out θ gives the posterior predictive distribution as

$$g(Y^\dagger \mid y) = \int f(Y^\dagger; \theta)\pi(\theta \mid y) d\theta.$$

If a point prediction of $Y^\dagger$ is required, we might use the mean, median or other function of this distribution, depending on our loss function.

Discussion

In Bayesian inference, predictive inference is (in principle) straightforward: future observation $Y^\dagger$ and parameter θ have same logical status, as random variables.

Choice of prior distributions

Main approaches to the selection of prior distributions:

(a) physical reasoning (Bayes) – too restrictive for most practical purposes;

(b) flat or uniform priors, including improper priors (Laplace, Jeffreys) – the most widely used method in practice, but theoretical justification is source of argument;

(c) subjective priors (de Finetti, Savage) – used in certain specific situations such as weather forecasting and for certain kinds of business application where it is worthwhile to go to the trouble of trying to elicit the client's true subjective opinions, but hardly used at all for routine statistical analysis;

(d) prior distributions for convenience, e.g. conjugate priors – in practice these are very often used just to simplify the calculations.

Computational techniques

Bayesian methods often applied in very complicated situations where both Y and θ are very high-dimensional. Main computational problem is to compute numerically the normalising constant that is required to make the posterior density a proper density function.

Monte Carlo algorithms

Simulate samples from posterior distribution.

Main algorithms:

- Gibbs sampler;
- Hastings-Metropolis.