smfd10(13).tex
Day 10. 7.6.2013

3. Non-parametric likelihood

At first glance, 'non-parametric likelihood' seems a contradiction in terms (an oxymoron – 'square circle', etc.) But it turns out that maximum-likelihood estimation (MLE) can indeed be usefully combined with non-parametrics. First, we interpret the empirical F_n as a non-parametric MLE (NPMLE) for the unknown true distribution F. For, if the data is $\{x_1, \ldots, x_n\}$, the *likelihood* of F is $L(F) := \prod_1^n \Delta F(x_i)$ (where $\Delta F(x) := F(x) - F(x-)$ is the probability mass on x), $F(\{x\})$). It makes sense to restrict attention to distributions F with support in $\{x_1, \ldots, x_n\}$, that is, absolutely continuous wrt the empirical F_n: $F << F_n$, and F_n does indeed maximise the likelihood over these F (Kiefer & Wolfowitz, 1956). Then it makes sense to call $T(F_n)$ a NPMLE for $T(F)$, where T is some functional – the mean, for example.

Let $X, X_1, \ldots, X_n \ldots$ be iid random p-vectors, with mean $EX = \mu$ and covariance matrix Σ of rank q. In higher dimensions, the distribution function, $P(. \leq .)$, which leads to *confidence intervals*, is replaced by $P(. \in .)$, which leads to *confidence regions* (which covers the unknown parameter with some probability); convexity is a desirable property of such confidence regions. For $r \in (0, 1)$, let

$$C_{r,n} := \{\int X dF : F << F_n, L(F)/L(F_n) \geq r\}.$$

Then $C_{r,n}$ is a convex set, and

$$P(\mu \in C_{r,n}) \to P(\chi^2(q) \leq -2 \log r) \qquad (n \to \infty)$$

(the rate is $O(1/\sqrt{n})$ if $E[\|X\|^4] < \infty$). This is a non-parametric analogue of Wilks' Theorem (II.3 above) (A. Owen 1990; P. Hall 1990): "$-2 \log LR \sim \chi^2(q)$". For a monograph account, see Owen [O].

In view of results of this type, it is common practice, when we want the distribution of $T(F)$ when F is unknown, to use $T(F_n)$ as an approximation for it. This is commonly known as a *plug-in estimator* (just plug it in as an approximation when we need the exact answer but do not know it); 'empirical estimator', or 'NPMLE', would also be reasonable names.

Suppose we want to estimate an unknown density f, which is known to be *decreasing* on $[0, \infty)$(example: the exponential). A density is the derivative of a distribution; a concave function has a decreasing derivative (when

differentiable). The NPMLE f_n of such a density is the (left-hand) derivative of the *least concave majorant* of F_n (Grenander, 1956). This example is interesting in that a CLT is known for it, but with an unusual rate of convergence – *cube-root asymptotics*:

$$n^{1/3}(f_n(t) - f(t)) \to |4f'(t)f(t)|^{1/3}\text{argmax}_h(B(h) - h^2),$$

with B BM and argmax the argument (= point) at which the maximum is attained (Kim and Pollard 1990).

Semi-parametrics.

Consider the *elliptical model*, with multidimensional density

$$f(\mathbf{x}) = const.g(Q(\mathbf{x})), \qquad Q(\mathbf{x}) = (\mathbf{x} - \mu)^T\Sigma^{-1}(\mathbf{x} - \mu).$$

Here $g : \mathbf{R}_+ \to \mathbf{R}_+$ is a function, the *density generator*, to be estimated. This is the *non-parametric* part of the model; (μ, Σ) is as above, the *parametric* part of the model. The model as a whole is then called *semi-parametric*.

Such models are very suited to financial applications. Notice how they generalise the multivariate normal or Gaussian (recall Edgeworth's theorem of IV.3). The parametric part (μ, Σ) is clearly needed in financial modelling, because of Markowitz's work on risk (Σ) and return (μ), and diversification (Σ again) (I.5, Day 2). The non-parametric part g allows us to choose a g that reflects the tail-behaviour observed in the data. For instance, for financial return data, it turns out that the *return interval*, Δ is crucial. For Δ long (monthly returns, say – though the rule of thumb is that 16 trading days suffice), the Gaussian ($g(x) = e^{-\frac{1}{2}x}$) suffices. This is an instance of *aggregational Gaussianity* – in other words, the Central Limit Theorem (CLT – see e.g. SP). For intermediate Δ – daily returns, say – the *generalised hyperbolic (GH)* distributions have been found to fit well. For short Δ – high-frequency data (tick data), g decreasing like a power (*Pareto tails*, or *heavy tails* – e.g. Student t) is both observed and predicted theoretically (the renormalisation group in Physics). These models have been extensively studied; see e.g. [BKRW], and [BFK] for some applications. In some cases, ignorance of one part of the model imposes no loss of efficiency when estimating the other part. This is the case for the elliptic model above, essentially for reasons to do with invariance under the action of the affine group. See [BKRW], 4.2.3, 6.3.9, 7.2.4, 7.8.3 for the theory, [BFK] for some applications.

4. Limit theorems; Markov chains; MCMC

We quote (see e.g. SP, PfS):

1. Strong Law of Large Numbers (SLLN): if $X_1, X_2, \ldots$ are independent and identically distributed (iid), with each $X_n, X \sim F$, then

$$\frac{1}{n}\sum_1^n X_i \to E[X] = \mu := \int x dF(x) \qquad (n \to \infty) \qquad a.s.$$

This includes as a special case the Weak Law of Large Numbers (WLLN), with convergence in probability in place of convergence a.s.

2. Central Limit Theorem (CLT). If also the X_n have variance $\sigma^2 < \infty$, then

$$\frac{1}{\sigma\sqrt{n}}\sum_1^n (X_i - \mu) \to N(0,1) \qquad (n \to \infty) \qquad \text{in distribution.}$$

So if f is such that $f(X_n)$ also has (finite) mean and variance, then

$$\frac{1}{n}\sum_1^n f(X_i) \to E[f(X)] \quad a.s.; \qquad \frac{1}{\sqrt{n\ var\ X}}\sum_1^n (f(X_i) - E[f(X)]) \to N(0,1).$$

The mode of convergence here is convergence in distribution, also known as weak convergence. This is weaker than convergence in probability, but when the limit is a constant (as in WLLN), the two are equivalent.

The convergence in the Glivenko-Cantelli theorem is uniform a.s., which is very strong. Similarly for weak convergence: for bounded continuous f,

$$\int f dF_n \to \int f dF : \qquad \frac{1}{n}\sum_1^n f(X_i) \to E[f(X)] \quad a.s.,$$

as above. The CLT above follows similarly from Donsker's theorem.

All this can be generalised far beyond the setting above of the iid case. We can work with *Markov chains* (see e.g. PfS VII) (discrete time will suffice for us, but the theory can be developed in continuous time). In PfS VII Markov chains are developed for discrete state spaces (finite or countably infinite, so the states can be counted as $x_1, \ldots, x_n, \ldots$). The definition of the Markov property is that, for predicting the future, knowing where one is at the present is all that matters – if we know where we are, how we got there is irrelevant. This irrelevance of the past suggests that as time passes the past 'becomes forgotten', and the chain settles down to some sort of steady state or equilibrium distribution, π – even to a limit distribution π in favourable

cases. Some Markov chains have no limit distribution (e.g., the trivial chain on the integers, which just moves 1 to the right at each step). But many Markov chains do have an equilibrium distribution, and even (if periodicity complications are absent) a limit distribution. See e.g. PfS VII for details. In particular, we need the idea of *detailed balance* (DB). A Markov chain with transition probability matrix $P = (p_{ij})$ and limiting distribution $\pi = \pi_i$ satisfies the *detailed balance* condition if

$$\pi_i p_{ij} = \pi_j p_{ji} \qquad \forall\ i, j. \tag{DB}$$

We quote (Kolmogorov's theorem) that this is the same as *time-reversibility* – the chain being the same if run backwards in time.

When the Markov chain has suitably good properties (which ensure a limit distribution) – typically, appropriate *recurrence* properties, of returning repeatedly to its starting point – then the Markov chain satisfies a SLLN and a CLT as above. We shall not give details (see e.g. [MeyT] Ch. 17).

It turns out that all this carries over to continuous-state Markov chains (the case relevant to Statistics), subject to suitable restrictions on the chain, of which *Harris recurrence* is the best known.

Markov Chain Monte Carlo (MCMC); Hastings-Metropolis algorithm (HM)

We briefly sketch this; see VII.6 below for statistical applications.

The aim here is to sample from a distribution π. This may be straightforward (see IS); if not, we may proceed as follows. We construct a Markov chain $X = (X_n)$ for which π is the limit distribution (we assume this has a density, also written π). HM selects a transition density $q(x, .)$ (see below for choice of q), and then at each step, conditional on $X_{k-1} = x$, HM *proposes* a new value Y_k drawn from this transition density $q(x, .)$. This value Y_k is *accepted* as the new value X_k with probability

$$p(x, y) := \min\Big(1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\Big);$$

otherwise, X_k is taken as the previous value X_{k-1}. One can check that this does indeed define a Markov chain, which satisfies (the continuous form of) (DB) and has invariant (= equilibrium) distribution π. Here $q(x, y) := p(|x - y|)$, for some transition density p of a symmetric random walk (the choice is usually not critical, so can be made for convenience).

What is critical in applying MCMC in practice is the rate of convergence. We have to run the chain for a long enough 'burn-in' period for it to be 'approximately in equilibrium'.

VII. BAYESIAN STATISTICS

1. Classical statistics and its limitations.

Broadly speaking, statistics splits into two main streams:
(i) classical, or frequentist, and
(ii) Bayesian.

Much of classical statistics is devoted to the following general areas: Estimation of parameters (I), Hypothesis testing (II). Again, this is not exhaustive: the main remaining area is Non-parametric statistics (VI).

Estimation of parameters itself splits, into
(ia). Point estimation [e.g., maximum-likelihood estimates],
(ib). Interval estimation [e.g., confidence intervals].
Both these are open to interpretational objections. A point estimate is a single number, which will almost certainly be wrong [i.e., will differ from the value of the parameter it estimates]. How wrong? And what should we do about this?

A confidence interval is more informative, because it includes an error estimate. For instance, its mid-point can be regarded as a point estimate, and half its length as an error estimate – leading to conclusions of the form

$$\theta = 3.76 \pm 0.003 \qquad (*)$$

– with *confidence* 95% [or 99 %, or whatever]. What does this mean? It is not a probability statement:
either θ lies between 3.73 and 3.79 [when (*) is true, so holds with probability 100 %]
or it doesn't [when (*) is false, so holds with probability 0 %].
Problem: We don't know which!
Interpretation. If a large number of statisticians independently replicated the analysis leading to (*), then about 95 % of them would succeed in producing confidence intervals covering the unknown parameter θ. But
(a) We wouldn't know *which* 95 %,
(b) This is of doubtful relevance anyway. The large number of independent replications will usually never take place in practice. So confidence statements like (*) lack, in practice, a direct interpretation. [They are 'what happens to probability statements in classical statistics when we put the numbers in'.]

A further problem is that small changes in our data can lead to abrupt discontinuities in our conclusions. In borderline situations, θ 'just within'

the confidence interval and 'just outside' represent diametrically opposite outcomes, but the data may be very close. Small changes in input *should* only lead to small changes in output, rather than abrupt changes.

Hypothesis testing is open to similar objections. It is usual to have a null hypothesis, H_0, representing our present theory (the 'default option'), and an alternative hypothesis, H_1, representing some proposed alternative theory. At the end of the investigation, we have to choose between two alternatives. We may be wrong: we may
reject H_0 when it is true, and choose H_1 [Type I error, probability α, the significance level], or
reject H_1 when it is true, and choose H_0 [Type II error, probability β].
We then have a trade-off between α and β. It is not always clear how to do this sensibly, still less optimally [it is customary to choose $\alpha = 0.05$ or 0.01, and then try to minimise β, but this is merely conventional]. Again, problems present themselves:
(i) We won't know whether our choice between H_0 and H_1 was correct;
(ii) Small changes in the data can lead to abrupt changes between choosing H_0 and choosing H_1.

Thus both the main branches of classical parametric statistics lead to abruptly discontinuous conclusions and present interpretational difficulties. One justification for Bayesian statistics is that it avoids these. There are many others: we shall argue for Bayesian statistics below on its merits.

2. Prior knowledge and how to update it.

The difficulties identified above arise because in classical statistics we rely *entirely* on the data, that is, on the sample we obtained. The mathematics involved in classical statistics amounts to comparing the sample we actually obtained with the large (usually, infinite) class of hypothetical samples we might have obtained but didn't. These include the samples that we would obtain if we repeated our sampling independently – or that other statisticians would obtain if they independently replicated our work. This is where the term 'frequentist' for classical statistics originates: e.g., in 95 % confidence intervals, independently replicated confidence intervals would cover the parameter θ with frequency 0.95.

The other aspect of classical statistics crucial for our purposes is that it ignores everything before sampling. This is often unreasonable. For instance, we may know a good deal about the situation under study, based on prior

experience. Such situations are typical in, e.g., industrial quality control: suppose we are employed by a rope manufacturer, and are testing the breaking strain of ropes in a current batch. We may have to hand large amounts of data obtained from tests on previous batches from the same production line. In hypothesis testing, such prior knowledge is tacitly assumed, because we need it to be able to formulate H_0 and H_1 sensibly. But we may not be willing to enter the 'accept or reject' framework of hypothesis testing [which some statisticians believe is inappropriate and damaging]: how then can we use prior knowledge? In the estimation framework also, we may know a lot about θ before sampling [as in the rope example above]: indeed, if we do *not* have some prior knowledge of the situation to be studied, we would in practice not have enough prior interest in it to be willing to invest the time, trouble and money to study it statistically.

Bayesian statistics addresses these aspects by providing a framework in which

1. The statistician knows something before sampling: he has some *prior knowledge*.

2. He then draws a sample, and analyses the *data* to extract some relevant information.

3. He then *updates his prior information* with his *data (or sample) information*, to obtain *posterior information*

(prior: before (sampling); posterior: after (sampling)).

This verbal description of the Bayesian approach is attractive, because it resembles how we learn. Life involves (indeed, largely consists of) a constant, ongoing process of acquiring new information and using it to update our previous ('prior') information/beliefs/attitudes/policies.

To implement the Bayesian approach, we need some mathematics. The formulae below derive from the work of the English clergyman

Thomas BAYES (1702-1761): *An essay towards solving a problem in the doctrine of chances* (1763, posth.).

Recall that if A, B are events of positive probability,

$$P(A) > 0, \qquad P(B) > 0,$$

the *conditional probability* of A *given* (or *knowing*) B is

$$P(A|B) := P(A \cap B)/P(B).$$

Symmetrically,

$$P(B|A) := P(B \cap A)/P(A) = P(A \cap B)/P(A).$$

Combining,

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A),$$

or

$P(B|A) = P(A|B)P(B)/P(A)$ (BAYES' FORMULA, or BAYES' THEOREM).

Interpretation.

1. Think of A as a 'cause', B as an 'effect'. We naturally first think of P(effect B|cause A). We can use Bayes' formula to get from this to P(cause A|effect B) (think of B as an effect we can see, A as an effect we can't see).

2. Suppose we are interested in event B. We begin with an initial, *prior* probability $P(B)$ for its occurrence. This represents how probable we initially consider B to be [this depends on us: we will have to estimate $P(B)$!]. Suppose we then observe that event A occurs. This gives us new information, which affects how probable we should now consider B to be, *after* observing A [or, to use the technical term, *a posteriori*]. Bayes' theorem tells us how to do this updating: we multiply by the ratio $P(A|B)/P(A)$:

$$P(B|A) = P(B).P(A|B)/P(A) :$$

posterior probability of B = prior probability of B $\times$ updating ratio.

We first observe some extreme cases.

Independence. If A, B are independent, $P(A \cap B) = P(A).P(B)$, so

$$P(B|A) = P(A \cap B)/P(A) = P(A).P(B)/P(A) = P(B),$$

and similarly $P(A|B) = P(A)$: updating ratio = 1, posterior probability = prior probability – conditioning on something independent has no effect.

Inclusion.

1. $A \subset B$: here, $P(A \cap B) = P(A)$, $P(A|B) = P(A \cap B)/P(B) = P(A)/P(B)$;

updating ratio $P(A|B)/P(A) = 1/P(B)$, posterior probability = 1.

2. $B \subset A$: here, $P(A \cap B) = P(B), P(A|B) = P(A \cap B)/P(B) = P(B)/P(B) = 1$;

updating ratio $P(A|B)/P(A) = 1/P(A)$, posterior probability = $P(B)/P(A)$.

Partitions. The event B partitions the sample space Ω (the space of all possible outcomes) into two disjoint events B, B^c whose union is Ω. Then A is the disjoint union of $A \cap B$ and $A \cap B^c$, so

$$P(A) = P(A \cap B) + P(A \cap B^c) = P(A|B)P(B) + P(A|B^c)P(B^c),$$

by definition of conditional probability. Similarly, if $B_1, B_2, \cdots, B_n$ form a partition (are disjoint events with union Ω), A is the disjoint union of events $A \cap B_1, \cdots, A \cap B_n$. So by the additivity property of probability,

$$P(A) = \Sigma_{r=1}^n P(A \cap B_r) = \Sigma_{r=1}^n P(A|B_r)P(B_r) \quad \text{(FORMULA OF TOTAL PROBABILITY)},$$

using the definition of conditional probability again.

Such expressions are often used for the denominator in Bayes' formula:

$$P(B_r|A) = P(B_r)P(A|B_r)/P(A) = P(B_r)P(A|B_r)/\Sigma_k P(B_k)P(A|B_k).$$

3. Prior and posterior densities.

Suppose now we are studying a parameter θ. Suppose we have data x [x may be a single number, i.e. a scalar, or a vector $x = (x_1, \cdots, x_n)$ of numbers; we shall simply write x in both cases]. Recall that x is an observed value of a random variable, X say. In the *density case*, this random variable has a (probability) *density* (function), $f(x)$ say, a non-negative function that integrates to 1:

$$f(x) \geq 0, \qquad \int f(x)dx = 1$$

(here and below, integrals with limits unspecified are over everything).
Interpretation. $P(X \in A) = \int_A f(x)dx$ for all subsets A of the real line $\mathbf{R}$ [actually, we need to restrict to suitable – 'measurable' – sets A, but it suffices for our purposes to consider intervals or half-lines. For instance, if $A = (-\infty, x]$,

$$F(x) := P(X \in (-\infty, x]) = P(X \leq x) = \int_{-\infty}^x f(y)dy \quad \forall x \in \mathbf{R};$$

as x varies, $F(x)$ gives the (probability) distribution (function) of X.]
In brief: the density $f(x)$ describes the *uncertainty* in the data x.

The distinctive feature of Bayesian statistics is that it treats *parameters* θ in the same way as *data* x. Our initial (prior) uncertainty about θ should also be described by a density $f(\theta)$:

$$f(\theta) \geq 0, \qquad \int_{-\infty}^{\infty} f(\theta)d\theta = 1,$$

$$P(\theta \in A) = \int_A f(\theta)d\theta \qquad \forall A \subset \mathbf{R},$$

where the probability on the left is a *prior probability*. The analogue for densities of Bayes' formula

$$P(B|A) = P(B)P(A|B)/P(A)$$

now becomes

$$f(\theta|x) = f(\theta)f(x|\theta)/f(x). \qquad (*)$$

The density on the left is the *posterior density* of θ *given* the data x; it describes our uncertainty about θ knowing x.

Now densities integrate to 1:

$$\int f(\theta|x)d\theta = 1,$$

so $\int [f(\theta)f(x|\theta)/f(x)]d\theta = 1$:

$$\int f(\theta)f(x|\theta)d\theta = f(x).$$

Combining,

$$f(\theta|x) = f(\theta)f(x|\theta)/\int f(\theta)f(x|\theta)d\theta.$$

In the *discrete case*, θ and/or x may take discrete values $\theta_1, \theta_2, \cdots, x_1, x_2, \cdots$ only, with probabilities $f(\theta_1), f(\theta_2), \cdots, f(x_1), f(x_2), \cdots$. The above formulae still apply, but with *integrals replaced by sums*:

$$P(X \in A) = \Sigma_{x\in A} f(x), \qquad P(\theta \in B) = \Sigma_{\theta \in B} f(\theta),$$

$$f(x) = \Sigma_\theta f(\theta f(x|\theta),$$

$$f(\theta|x) = f(\theta)f(x|\theta)/\Sigma_\theta f(\theta)f(x|\theta).$$

In the formula $f(\theta|x) = f(\theta)f(x|\theta)/f(x)$, it is θ, the parameter under study, which is the main focus of interest. Consequently, the denominator $f(x)$ – whose role is simply to ensure that the posterior density $f(\theta|x)$ integrates to 1 (i.e., really is a density) – can be omitted (or understood from context). This replaces the *equation* above by a *proportionality statement*:

$$f(\theta|x) \propto f(\theta)f(x|\theta)$$

(here $\propto$, read as 'is proportional to', relates to the variability in θ, which is where the action is). Now $f(x|\theta)$ can be viewed in two ways:

(i) for fixed θ as a function of x. It is then the density of x when θ is the true parameter value,
(ii) for fixed/known/given data values x as a function of θ. It is then called the *likelihood* of θ (Fisher), familiar from IS, Ch. I, Ch. II, etc.

The formula above now reads, in words:

$$\textbf{posterior} \propto \textbf{prior} \times \textbf{likelihood.}$$

This is the essence of Bayesian statistics. It shows how Bayes' theorem (of which this formula is a version) may be used to *update* the *prior* information on θ before sampling by using the information in the *data* x – which is contained in the *likelihood* factor $f(x|\theta)$ by which one multiplies – to give the *posterior* information on θ after sampling. Thus posterior information combines two sources: prior information and data/sample/likelihood information.

4. Examples.

Example 1. Bernoulli trials with Beta prior ([O'H], Ex. 1.4, p.5).

Here θ represents the probability of a head on tossing a biased coin. On the basis of prior information, θ is assumed to have a prior density proportional to $\theta^{p-1}(1-\theta)^{q-1}$ $(0 \le \theta \le 1)$ for $p, q > 0$:

$$f(\theta) \propto \theta^{p-1}(1-\theta)^{q-1} \qquad (0 \le \theta \le 1).$$

Writing

$$B(p,q) := \int_0^1 \theta^{p-1}(1-\theta)^{q-1} d\theta$$

(the *Beta function)*,

$$f(\theta) = \theta^{p-1}(1-\theta)^{q-1}/B(p,q).$$

[We quote the *Eulerian integral* for the Beta function: for

$$\Gamma(p) := \int_0^\infty e^{-x}x^{p-1}dx \quad (p>0), \quad B(p,q) = \Gamma(p)\Gamma(q)/\Gamma(p+q) \quad (p,q>0).]$$

Note that, as p, q vary, the shape of $f(\theta)$ varies – e.g, the graph is u-shaped if $0 < p, q < 1$, n-shaped if $p, q > 1$. Here p, q are called *hyperparameters* - they are parameters describing the parameter θ.

Suppose now we toss the biased coin n times (independently), observing x heads. Then x is our data. It has a discrete distribution, the binomial $B(n,\theta)$, described by

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad (x = 0, 1, \cdots, n).$$

We apply Bayes' theorem to update our prior information on θ – our prior values of p, q – by our data x. Now

$$f(x) = \int f(\theta)f(x|\theta)d\theta = \int \frac{\theta^{p-1}(1-\theta)^{q-1}}{B(p,q)}.\binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta$$

$$= \binom{n}{x}.\frac{1}{B(p,q)}.\int_0^1 \theta^{p+x-1}(1-\theta)^{q+n-x-1}d\theta = \binom{n}{x}.\frac{B(p+x,q+n-x)}{B(p,q)}.$$

So Bayes' theorem gives

$$f(\theta|x) = f(\theta)f(x|\theta)/f(x) = \binom{n}{x}.\frac{1}{B(p,q)}.\theta^{p+x-1}(1-\theta)^{q+n-x-1}/\binom{n}{x}.\frac{B(p+x,q+n-x)}{B(p,q)}$$

or

$$f(\theta|x) = \frac{\theta^{p+x-1}(1-\theta)^{q+n-x-1}}{B(p+x,q+n-x)}.$$

The posterior density of θ is thus another Beta density, $B(p+x, q+n-x)$.
Summarising:
• prior $B(p,q)$ is updated by data x heads in n tosses to posterior
$B(p+x, q+n-x)$.
Graphs. To graph the three functions of θ – prior, likelihood and posterior – first find their maxima.
Likelihood: $f(x|\theta)$ has a maximum where $\log f(x|\theta)$ has a maximum, i.e. where
$x\log\theta + (n-x)\log(1-\theta)$ has a maximum, i.e. where

$$\frac{x}{\theta} - \frac{n-x}{1-\theta} = 0: \qquad x - x\theta = n\theta - x\theta: \qquad \theta = x/n.$$

Prior: similarly, $f(\theta)$ has a maximum where $\log f(\theta)$ does, i.e. where

$$\frac{p-1}{\theta} - \frac{q-1}{1-\theta} = 0: \quad p - p\theta - 1 + \theta = q\theta - \theta: \quad \theta = (p-1)/(p+q-2).$$