Ancillarity and Conditional Inference

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 1, Hilary Term 2008

April 2, 2008

Various forms of the *conditionality principle* say that the distribution used for inference should be conditional on any ancillary, such as the instrument actually used.

Note this is a frequentist concept and plays little role in a Bayesian paradigm.

In the Fisherian paradigm, we should not compare the measurement obtained to anything we could have seen, but did not. Rather we should define a relevant *reference set* of values, for example by conditioning with an ancillary statistic, and use this set for inference calculations.

The relevant reference set may not simply be the original sample space!

Consider an experiment with two instruments available:

One instrument is very precise and produces measurements $\mathcal{N}(\theta, 1)$. The other instrument is older and less accurate; it produces measurements which are $\mathcal{N}(\theta, 100)$.

We wish to check whether a parameter $\theta = 0$, the alternative being that $\theta > 0$.

Toss a fair coin and let $A = i, i = 1, 2$ denote that the instrument i is chosen. Perform then the measurement to obtain X. The joint distribution of (X, A) is determined as

$$f(x, a; \theta) = \phi(x - \theta)1_{\{1\}}(a)/2 + \phi\{(x - \theta)/10\}1_{\{2\}}(a)/2.$$

Suppose we have chosen the first instrument and observe $X = 4$. Is this consistent with the assumption $\theta = 0$?

In a Bayesian paradigm we only consider the value observed trough the likelihood function, which modifies the prior distribution into the posterior.

The likelihood function when observing $X = 4, A = 1$ would be

$$L(\theta \mid X = 4, a = 1) \propto \phi(4 - \theta)$$

which in itself gives very strong evidence against $\theta = 0$.

The p-value is

$$p = P(X > 4; \theta = 0) = \{1 - \Phi(4)\}/2 + \{1 - \Phi(.4)\}/2 = .1723,$$

so there is nothing to worry about?

However, we did in fact use the precise instrument. So, with a standard deviation of 1, a value of $X = 4$ should be very unlikely. Why should it matter that we could have used the other instrument, but didn't?

Should we not rather have considered $A = a$ fixed and condition on the actual instrument used? That is, calculate the p-value as

$$\tilde{p} = P(X > 4 \mid A = 1; \theta = 0) = \{1 - \Phi(4)\} = .00003$$

giving very strong evidence against the hypothesis.

In general, if the MLE $\hat{\theta}$ is not sufficient, it is often possible to find an ancillary statistic A so that $(\hat{\theta}, A)$ is jointly sufficient. Then since

$$f(x; \theta) = h(x)k\{\hat{\theta}(x), a(x); \theta\}$$

we also have

$$f(x \mid A = a; \theta) \propto h(x)k\{\hat{\theta}(x), a; \theta\}.$$

Thus $\hat{\theta}$ *is sufficient when considering the conditional distribution given the ancillary* A.

A statistic $A = a(X)$ is said to be *ancillary* if

(i) The distribution of A does not depend on θ;

(ii) there is a statistic $T = t(X)$ so that $S = (T, A)$ taken together are minimal sufficient.

Intuitively A is then *uninformative* about the unknown parameter.

In the example just given, A is such an ancillary statistic since $\hat{\theta} = X$ can play the role of T as (X, A) clearly is jointly (minimal) sufficient.

The word 'ancillary' both means secondary and auxiliary, each meaning referring to each of the two conditions.

Notion of ancillarity seems fundamental in statistics and is due to Fisher, but its role is less clear than that of sufficiency.

It has several time been attempted to give statistical inference a firm foundation through so-called inference principles, for example:

- The *sufficiency principle* (S) says that if $S = s(X)$ is a sufficient statistic, S carries the same evidence for the parameter θ as does X.

It has several time been attempted to give statistical inference a firm foundation through so-called inference principles, for example:

- The *sufficiency principle* (S) says that if $S = s(X)$ is a sufficient statistic, S carries the same evidence for the parameter θ as does X.
- The *conditionality principle* (C) says that if $A = a(X)$ is ancillary, then the conditional distribution given $A = a(x_{\text{obs}})$, carries the same evidence as the unconditional experiment.

Consider an exponential family, with densities

$$f(x;\theta) = b(x)e^{a(\theta)^\top t(x) - c(\theta)}, \quad x \in \mathcal{X}.$$

If the family is linear, then $T = t(X)$ is boundedly complete and sufficient.

This is a non-trivial result. The proof uses analytic function theory and is outside the scope of this course.

The case of a linear exponential family is essentially the only case where a complete sufficient statistic exists, or at least where this can be proved.

For curved exponential families there is typically no complete sufficient statistic.

It has several time been attempted to give statistical inference a firm foundation through so-called inference principles, for example:

- The *sufficiency principle* (S) says that if $S = s(X)$ is a sufficient statistic, S carries the same evidence for the parameter θ as does X.
- The *conditionality principle* (C) says that if $A = a(X)$ is ancillary, then the conditional distribution given $A = a(x_{\text{obs}})$, carries the same evidence as the unconditional experiment.
- The *likelihood principle* (L) says that all evidence in an experiment is summarized in the likelihood function.

Sometimes it does not matter, whether we condition on A or not:

If $T = t(X)$ is complete and sufficient for θ and the distribution of A does not depend on θ, then T and A are independent.

Here is a nice application of this:

If $(X_1, \ldots, X_n)$ is a sample from the normal distribution $\mathcal{N}(\mu, \sigma^2)$ with *known* variance $\sigma^2 = \sigma_0^2$, it holds that $\hat{\mu} = \bar{X}$ complete and sufficient. Since the distribution of $\sum(X_i - \bar{X})^2$ cannot depend on μ, it follows that $\bar{X}$ and $\sum(X_i - \bar{X})^2$ are independent.

Birnbaum's theorem

Whereas some variant of (S) and (C) are commonly accepted among statisticians, (L) is not.

Birnbaum showed in 1972 that (S) and (C) combined are equivalent to (L)!

Reactions on this result have been different. The theorem depends heavily on the precise formulation of the principles (weak and strong forms) and is therefore not generally accepted as a fact.

Bayesian inference obeys (L) in the strongest form.

Attitudes towards this fact are varied...

The proof is surprisingly simple: Let g be an arbitrary bounded function of a and let $m = \mathbf{E}_\theta\{g(A)\}$. Note m does not depend on θ as the distribution of A did not. Now let

$$h\{t(x)\} = \mathbf{E}_\theta[\{g(A) - m\} \mid T = t(x)]$$

which also does not depend on θ because T was sufficient.

Iterating expectations and using the definition of m yields

$$\begin{aligned}\mathbf{E}_\theta\{h(T)\} &= \mathbf{E}_\theta\mathbf{E}_\theta[g\{A\} - m \mid T] \\ &= \mathbf{E}_\theta\{g(A) - m\} = 0\end{aligned}$$

for all θ. Completeness then implies

$$\mathbf{E}_\theta\{g(A) \mid T = t(x)\} = \mathbf{E}\{g(A)\},$$

thus that A and T are independent.

A statistic $T = t(X)$ is said to be *complete* w.r.t. θ if for all functions h

$$\mathbf{E}_\theta\{h(T)\} = 0 \text{ for all } \theta \implies h(t) = 0 \text{ a.s.}$$

It is *boundedly complete* if the same holds when only bounded functions h are considered.

It would be more precise to say the family of densities of T

$$\mathcal{F}_T = \{f_T(t;\theta), \theta \in \Theta\}$$

is complete, but the shorter usage has become common.

The Lehmann-Scheffé theorem says that *if a sufficient statistic is complete, it is also minimal sufficient.*

Nuisance parameters and their treatment

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 2, Hilary Term 2008

April 2, 2008

One instrument produces measurements $\mathcal{N}(\theta, 1)$, the other measurements which are $\mathcal{N}(\theta, 100)$.

We wish to check whether a parameter $\theta = 0$, the alternative being that $\theta > 0$.

Toss a coin *with probability λ of landing heads* and let $A = i, i = 1, 2$ denote that the instrument i is chosen. Perform then the measurement to obtain X. The joint distribution of (X, A) is determined as

$$f(x, a; \theta, \lambda) = \phi(x - \theta)1_{\{1\}}(a)\lambda + \phi\{(x - \theta)/10\}1_{\{2\}}(a)(1 - \lambda).$$

Suppose we have chosen the first instrument and observe $X = 4$. Is this consistent with the assumption $\theta = 0$?

A statistic $A = a(X)$ is said to be *ancillary* if

(i) The distribution of A does not depend on θ;

(ii) there is a statistic $T = t(X)$ so that $S = (T, A)$ taken together are minimal sufficient.

If the MLE $\hat{\theta}$ is not sufficient, it is often possible to find an ancillary statistic A so that $(\hat{\theta}, A)$ is jointly sufficient. Then we also have

$$f(x \mid A = a; \theta) \propto h(x)k\{\hat{\theta}(x), a; \theta\}.$$

Thus *$\hat{\theta}$ is sufficient when considering the conditional distribution given the ancillary A.*

The parameter λ is *nuisance parameter* in the sense that we are not interested in its value, but its value modifies the distribution of our observations.

If we now redo the exercise from the case where λ is known, we have the additional problem that the p-value

$$\begin{aligned} p &= P(X > 4; \theta = 0) \\ &= \{1 - \Phi(4)\}\lambda + \{1 - \Phi(.4)\}(1 - \lambda) \\ &= .00003\lambda + .34458(1 - \lambda) \end{aligned}$$

unfortunately *depends on the unknown λ.*

- The *sufficiency principle* (S) says that if $S = s(X)$ is a sufficient statistic, S carries the same evidence for the parameter θ as does X.
- The *conditionality principle* (C) says that if $A = a(X)$ is ancillary, then the conditional distribution given $A = a(x_{\text{obs}})$, carries the same evidence as the unconditional experiment.
- The *likelihood principle* (L) says that all evidence in an experiment is summarized in the likelihood function.

Birnbaum's theorem says that (S) and (C) combined are equivalent to (L)!

Bayesian inference obeys (L) in the strongest form.

However, *the probability of choosing the instrument seems irrelevant once we know which instrument was in fact used.*

Thus, again we would rather consider $A = a$ fixed and condition on the actual instrument used. That is, also here *calculate the p-value as*

$$\tilde{p} = P(X > 4 \mid A = 1; \theta = 0) = \{1 - \Phi(4)\} = .00003$$

giving very strong evidence against the hypothesis. Note that *λ does not enter in this conditional calculation.*

Motivated by this example, we consider more generally a family of distributions $f(x; \theta), \theta \in \Theta$ where θ is partitioned into $\theta = (\psi, \lambda)$. We also assume that ψ is the *parameter of interest* and λ a *nuisance parameter.*

A statistic $T = t(X)$ is said to be (boundedly) *complete* w.r.t. θ if for all functions h

$$\mathbf{E}_\theta\{h(T)\} = 0 \text{ for all } \theta \implies h(t) = 0 \text{ a.s.}$$

In a linear exponential family, the canonical statistic $T = t(X)$ is boundedly complete and sufficient.

The Lehmann-Scheffé theorem: *if a sufficient statistic is complete, it is also minimal sufficient.*

Basu's theorem: *If $T = t(X)$ is (boundedly) complete and sufficient for θ and the distribution of A does not depend on θ, then T and A are independent.*

Suppose that there is a minimal sufficient statistic $T = t(X)$ partitioned as $T = (S, C) = (s(X), c(X))$ where:

C1: the distribution of C depends on λ but not on ψ;

C2: the conditional distribution of S given $C = c$ depends on ψ but not λ, for all c;

C3: the parameters vary independently, i.e. $\Theta = \Psi \times \Lambda$.

Then the likelihood function factorizes as

$$L(\theta \mid x) \propto f(s, c; \theta) = f(s \mid c; \psi)f(c; \lambda)$$

and we say that C is *ancillary for ψ*, S is *conditionally sufficient for ψ* given C, and C is *marginally sufficient for λ*.

We also say that C is a *cut* for λ.

When C is a cut, the likelihood factorizes as

$$L(\theta \mid x) \propto f(s, c; \theta) = f(s \mid c; \psi) f(c; \lambda) = L_1(\psi \mid s, c) L_2(\lambda \mid c).$$

Since ψ and λ vary independently, we may then maximize L by maximizing each of these factors separately. In other words, the maximum likelihood estimator $\hat{\theta}$ of the parameter θ satisfies

$$\hat{\theta} = (\hat{\psi}, \hat{\lambda}), \quad \text{where } \hat{\psi} = \arg\max_{\psi} L_1(\psi \mid s, c), \ \hat{\lambda} = \arg\max_{\lambda} L_2(\lambda \mid c).$$

Hence we get the *same estimate whether we use the joint distribution $f_{(S,C)}$ for θ, or $f_{S|C}$ for ψ and f_C for λ.*

Note that the equation above may indicate a *simple way of maximizing the likelihood function.*

Consider the hypothesis that the parameter of interest ψ has a specific value, i.e. $H_0 : \psi = \psi_0$. This is a *composite* hypothesis and we wish to find a test of size α so the rejection region R satisfies

$$P(X \in R; \psi_0, \lambda) = \alpha \text{ for all values of } \lambda \in \Lambda.$$

A test is said to be *similar* if this condition holds.

One way of constructing a similar test is to find a statistic C which is *sufficient for λ for fixed $\psi = \psi_0$*. This would in particular be the case if C is a cut. Now look for a set $R(c)$ such that

$$P(X \in R(c) \mid C = c; \psi_0, \lambda) = P(X \in R(c) \mid C = c; \psi_0) = \alpha,$$

where we have used the sufficiency of C to remove λ.

A widely accepted *conditionality principle* says that when C is a cut for a nuisance parameter λ, *inference about ψ should be based on the conditional distribution of S given C.*

In the simple example given, this corresponds to conditioning on the instrument actually used when making inference about θ.

A possibly less well accepted principle says that when C is a cut for λ, *inference about λ should be based on the marginal distribution of C.*

Thus when making inference about the probability λ of choosing the first instrument, we should ignore the fact that the instrument was used, but only consider that it was chosen.

If we define R as $x \in R \iff x \in R(c(x))$ we then get

$$\begin{aligned} P(X \in R; \psi_0, \lambda) &= \mathbf{E}_{(\psi_0,\lambda)}\{P(X \in R \mid C; \psi_0)\} \\ &= \mathbf{E}_{(\psi_0,\lambda)}\{P(X \in R(C) \mid C; \psi_0)\} \\ &= \mathbf{E}_{(\psi_0,\lambda)}(\alpha) = \alpha. \end{aligned}$$

We have thus succeeded in constructing a similar test by this conditioning operation.

A test of this kind is said to have *Neyman structure*. An important result is that if C is complete and sufficient for λ for $\psi = \psi_0$, then *any similar rejection region R has Neyman structure.*

Another example

Consider a sample $X = (X_1, \ldots, X_n)$ from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Since $(\bar{X}, S^2 = \sum_i (X_i - \bar{X}_i)^2)$ is minimal sufficient, the likelihood function becomes

$$L(\mu, \sigma^2 \mid x) \propto f(\bar{x}; \mu, \sigma^2) f(s^2; \sigma^2),$$

where we have used the independence of $\bar{X}$ and S^2 and the fact that S^2 follows a $\sigma^2 \chi^2$-distribution not depending on μ.

Here the situation is less clear cut. It could make sense to think of $\bar{x}$ as being sufficient for μ (which it is if σ^2 is fixed) and S^2 as ancillary for μ and sufficient for σ^2, but *it does not fit into the theory developed* as the distribution of $\bar{X}$ depends on (μ, σ^2).

This is shown as follows. Assume R is a similar rejection region, i.e.

$$P(X \in R; \psi_0, \lambda) = \alpha \text{ for all } \lambda.$$

Then define $h(C) = P(X \in R \mid C; \psi_0) - \alpha$. We get

$$\begin{aligned} \mathbf{E}_{(\psi_0,\lambda)}\{h(C)\} &= \mathbf{E}_{(\psi_0,\lambda)}\{P(X \in R \mid C; \psi_0) - \alpha\} \\ &= \mathbf{E}_{(\psi_0,\lambda)}\{P(X \in R \mid C; \psi_0, \lambda) - \alpha\} \\ &= P(X \in R; \psi_0, \lambda) - \alpha = 0. \end{aligned}$$

Completeness yields $h(C) = 0$ and $P(X \in R \mid C; \psi_0) - \alpha$.

As a consequence of this result it is common, although not universally accepted, to *condition on the statistic sufficient under the hypothesis when testing composite hypothesis,* i.e. to construct tests with Neyman structure.

Since Bayesian inference obeys the likelihood principle only the factorization itself matters:

$$L(\theta \mid x) \propto L_1(\psi \mid s, c) L_2(\lambda \mid c).$$

Still, this fact is not unimportant. Assume that the prior density satisfies

$$\pi(\psi, \lambda) = \eta(\psi) \rho(\lambda),$$

in other words that the parameters ψ and λ are prior independent. Then the posterior density satisfies

$$\pi^*(\psi, \lambda) = \pi(\psi, \lambda \mid x) \propto \eta(\psi) \rho(\lambda) L_1(\psi \mid s, c) L_2(\lambda \mid c) \propto \eta^*(\psi) \rho^*(\lambda).$$

Hence *if C is a cut for λ and ψ and λ are prior independent, they are posterior independent.*

Newton–Raphson Iteration and the Method of Scoring

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 3, Hilary Term 2008

January 28, 2008

Recall that the (expected) Fisher information is

$$I(\theta) = \mathbf{E}\{J(\theta)\}$$

and that for large i.i.d. samples it holds approximately that $\hat{\theta} \sim \mathcal{N}(\theta, I(\theta)^{-1})$.

But it is also approximately true, to be elaborated later, under the same assumptions that

$$\sqrt{J(\hat{\theta})}(\hat{\theta} - \theta) \sim \mathcal{N}(0, 1),$$

so we could write $\hat{\theta} \sim \mathcal{N}(\theta, J(\hat{\theta})^{-1})$.

In fact, the observed information is in many ways preferable to the expected information. Indeed, as $\hat{\theta}$ is approximately sufficient, $J(\hat{\theta})$ is approximately ancillary.

Recall that, under suitable regularity conditions, the maximum likelihood estimate is the solution to the score equation

$$s(\theta) = s(x; \theta) = \frac{\partial}{\partial\theta} l(\theta) = \frac{\partial}{\partial\theta} \log L(\theta; x) = 0,$$

where $S(\theta) = s(X; \theta)$ is the *score statistic*.

Generally the solution to this equation must be calculated by iterative methods. One of the most common methods is the *Newton–Raphson method* and is based on successive approximations to the solution, using Taylor's theorem to approximate the equation.

Formally the iteration becomes

- Choose an initial value θ and calculate $S(\theta)$ and $J(\theta)$;
- `While` $|S(\theta)| > \epsilon$ `Repeat`
  1. $\theta \leftarrow \theta + J(\theta)^{-1} S(\theta)$
  2. Calculate $S(\theta)$ and $J(\theta)$ `go to 1`
- Return θ;

Other criteria for terminating the iteration can be used. To get a criterion which is insensitive to scaling of the variables, one can instead use the criterion $J(\theta)^{-1} S(\theta)^2 > \epsilon$.

Note that, as a by-product of this algorithm, the final value of $J(\theta)$ is the observed information which can be used to assess the uncertainty of $\hat{\theta}$.

Thus, we take an initial value θ_0 and write

$$0 = S(\theta_0) - J(\theta_0)(\theta - \theta_0),$$

ignoring the remainder term. Here

$$J(\theta) = J(\theta; X) = -\frac{\partial}{\partial\theta} S(\theta) = -\frac{\partial^2}{\partial\theta^2} l(\theta).$$

Solving this equation for θ then yields a new value θ_1

$$\theta_1 = \theta_0 + J(\theta_0)^{-1} S(\theta_0)$$

and we keep repeating this procedure as long as $|S(\theta_j)| > \epsilon$, i.e.

$$\theta_{k+1} = \theta_k + J(\theta_k)^{-1} S(\theta_0).$$

If θ_0 is chosen sufficiently near $\hat{\theta}$ convergence is very fast.

It can be computationally expensive to evaluate $J(\theta)$ a large number of times. This is sometimes remedied by only changing J every 10 iterations or similar.

Another problem with the Newton–Raphson method is its lack of stability. When the initial value θ_0 is far from θ it might wildly oscillate and not converge at all. This is sometimes remedied by making smaller steps as

$$\theta \leftarrow \theta + \gamma J(\theta)^{-1} S(\theta)$$

where $0 < \gamma < 1$ is a constant. An alternative (or additional) method of stabilization is to let

$$\theta \leftarrow \theta + \gamma\{J(\theta) + S(\theta)^2\}^{-1} S(\theta)$$

as this avoids taking large steps when $S(\theta)$ is large.

Clearly, $\hat{\theta}$ is a fixed point of this iteration as $S(\hat{\theta}) = 0$ and, conversely, any fixpoint is a solution to the likelihood equation.

If $\hat{\theta}$ is a local maximum for the likelihood function, we must have

$$J(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2} l(\hat{\theta}) > 0.$$

The quantity $J(\hat{\theta})$ determines the sharpness of the peak in the likelihood function around its maximum. It is also known as the *observed information*.

Occasionally we also use this term for $J(\theta)$ where θ is arbitrary but strictly speaking this can be quite inadequate as $J(\theta)$ may well be negative (although positive in expectation).

The iteration has a tendency to be unstable for many reasons, one of them being that $J(\theta)$ may be negative unless θ already is very close to the MLE $\hat{\theta}$. In addition, $J(\theta)$ might sometimes be hard to calculate.

R. A. Fisher introduced the *method of scoring* which simply replaces the observed second derivative with its expectation to yield the iteration

$$\theta \leftarrow \theta + I(\theta)^{-1} S(\theta).$$

In many cases, $I(\theta)$ is easier to calculate and $I(\theta)$ is always positive. This generally stabilizes the algorithm, but here it can also be necessary to iterate as

$$\theta \leftarrow \theta + \gamma\{I(\theta) + S(\theta)^2\}^{-1} S(\theta).$$

In the case of n independent and identically distributed observations we have $I(\theta) = nI_1(\theta)$ so

$$\theta \leftarrow \theta + I_1(\theta)^{-1} S(\theta)/n$$

where $I_1(\theta)$ is the Fisher information in a single observation.

In a linear canonical one-parameter exponential family

$$f(x;\theta) = b(x)e^{\theta t(x) - c(\theta)}$$

we get

$$J(\theta) = \frac{\partial^2}{\partial\theta^2}\{c(\theta) - \theta t(X)\} = c''(\theta) = I(\theta).$$

so *for canonical exponential families the method of scoring and the method of Newton–Raphson coincide.*

If we let $v(\theta) = c''(\theta) = I(\theta) = \mathbf{V}(t(X))$ the iteration becomes

$$\theta \leftarrow \theta + v(\theta)^{-1} S(\theta)/n.$$

The lack of stability of the Newton–Raphson algorithm is not getting better in the multiparameter case. On the contrary there are not only problems with negativity, but the matrix can be singular and not invertible or it can have both positive and negative eigenvalues.

Recall that a symmetric matrix A is *positive definite* if all its eigenvalues are positive or, equivalently, if $x^\top Ax > 0$ for all $x \neq 0$. Sylvester's theorem says that *A is positive definite if and only if $\det(A_R) > 0$ for all submatrices A_R of the form $\{a_{rs}\}_{r,s=1\ldots,R}$.*

The identity of Newton–Raphson and the method of scoring *only holds for the canonical parameter.* If $\theta = g(\mu)$

$$\begin{aligned} J(\mu) &= \frac{\partial^2}{\partial\mu^2}\{c(g(\mu)) - g(\mu)t(X)\} \\ &= \frac{\partial}{\partial\mu}\left[g'(\mu)\tau\{g(\mu)\} - g'(\mu)t(X)\right] \\ &= v\{g(\mu)\}\{g'(\mu)\}^2 + g''(\mu)\left[\tau\{g(\mu)\} - t(X)\right] \end{aligned}$$

where we have let $\tau(\theta) = c'(\theta) = \mathbf{E}_\theta\{t(X)\}$ and $v(\theta) = c''(\theta) = \mathbf{V}_\theta\{t(X)\}$.

The method of scoring is simpler because the last term has expectation equal to 0:

$$I(\mu) = \mathbf{E}\{J(\mu)\} = v\{g(\mu)\}\{g'(\mu)\}^2.$$

It is therefore also here advisable to replace $J(\theta)$ with its expectation, the Fisher information matrix, i.e. iterate as

$$\theta \leftarrow \theta + I(\theta)^{-1} S(\theta)$$

where now $I(\theta)$ is the Fisher information matrix which is always positive definite if the model is not over-parametrized.

Also in the multi-parameter case it can be advisable to stabilize additionally, i.e. by iterating as

$$\theta \leftarrow \theta + \gamma\{I(\theta) + S(\theta)S(\theta)^\top\}^{-1} S(\theta)$$

or

$$\theta \leftarrow \theta + \gamma\{I(\theta) + S(\theta)^\top S(\theta)E\}^{-1} S(\theta),$$

where E is the identity matrix.

The considerations on the previous overheads readily generalize to the multi-parameter case. The approximation to the score equation becomes

$$0 = S(\theta_0) - J(\theta_0)(\theta - \theta_0)$$

where

$$S(\theta)_r = \frac{\partial}{\partial\theta_r} l(\theta), \quad J(\theta)_{rs} = -\frac{\partial^2}{\partial\theta_r\partial\theta_s} l(\theta),$$

i.e. $S(\theta)$ is the *gradient* and $-J(\theta)$ the *Hessian* of $l(\theta)$.

The iterative step can still be written as

$$\theta \leftarrow \theta + J(\theta)^{-1} S(\theta)$$

where we just have to remember that the score statistic S is a *vector* and the Hessian $-J$ a *matrix*.

In a multi-parameter curved exponential family with densities

$$f(x;\beta) = b(x)e^{\theta(\beta)^\top t(x) - c\{\theta(\beta)\}}$$

where β is d-dimensional, we get

$$\begin{aligned} J(\beta) &= \frac{\partial^2}{\partial\beta\partial\beta^\top}\left[c\{\theta(\beta)\} - \theta(\beta)^\top t(X)\right] \\ &= \frac{\partial}{\partial\beta}\left[\left(\frac{\partial\theta}{\partial\beta}\right)^\top \tau\{\theta(\beta)\} - \left(\frac{\partial\theta}{\partial\beta}\right)^\top t(X)\right] \\ &= \frac{\partial^2\theta}{\partial\beta\partial\beta^\top}\left[\tau\{\theta(\beta)\} - t(X)\right] + \left(\frac{\partial\theta}{\partial\beta}\right)^\top v\{\theta(\beta)\}\left(\frac{\partial\theta}{\partial\beta}\right), \end{aligned}$$

where the first term has expectation zero so

$$I(\beta) = \mathbf{E}\{J(\theta)\} = \left(\frac{\partial\theta}{\partial\beta}\right)^\top v\{\theta(\beta)\}\left(\frac{\partial\theta}{\partial\beta}\right).$$

More on nuisance parameters

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 4, Hilary Term 2008

February 1, 2008

This example shows that we have to be very careful when nuisance parameters are present and straight likelihood considerations can lead us astray:

We wish to establish the precision of a new instrument which measures with normal errors. We are therefore taking repeated measurements of individuals $(X_{i1}, X_{i2}), i = 1, \ldots, n$ which are all independent with

$$X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2).$$

Now consider

$$U_i = (X_{i1} + X_{i2})/2, \quad V_i = (X_{i1} - X_{i2})/2.$$

These are again independent and normally distributed as

$$U_i \sim \mathcal{N}(\mu_i, \tau^2), \quad V_i \sim \mathcal{N}(0, \tau^2),$$

where $\tau^2 = \sigma^2/2$.

Suppose that there is a minimal sufficient statistic $T = t(X)$ partitioned as $T = (S, C) = (s(X), c(X))$ where:

- C1: the distribution of C depends on λ but not on ψ;
- C2: the conditional distribution of S given $C = c$ depends on ψ but not λ, for all c;
- C3: the parameters vary independently, i.e. $\Theta = \Psi \times \Lambda$.

Then the likelihood function factorizes as

$$L(\theta \,|\, x) \propto f(s, c; \theta) = f(s \,|\, c; \psi) f(c; \lambda)$$

and we say that C is *ancillary for* ψ, S is *conditionally sufficient for* ψ given C, and C is *marginally sufficient for* λ.

We also say that C is a *cut* for λ and would then

- base inference about λ on the marginal distribution of C;

Clearly, we might as well consider (U_i, V_i) as the original data. Also, the pair (U, W) is minimal sufficient, where $U = (U_1, \ldots, U_n)$ and $W = \sum_i V_i^2$, hence the likelihood function becomes

$$\begin{aligned} L(\mu, \tau^2) &\propto (\tau^2)^{-n/2} e^{-\frac{1}{2\tau^2}\sum_i (u_i - \mu_i)^2} (\tau^2)^{-n/2} e^{-\frac{1}{2\tau^2}\sum_i v_i^2} \\ &= e^{-\frac{1}{2\tau^2}\sum_i (u_i - \mu_i)^2} (\tau^2)^{-n} e^{-\frac{w}{2\tau^2}}. \end{aligned}$$

Thus the maximum likelihood estimator is

$$\hat{\mu}_i = U_i, i = 1, \ldots, n; \quad \hat{\tau}^2 = W/2n.$$

But $W \sim \tau^2 \chi^2(n)$, so for large n, $\hat{\tau}^2 \approx n\tau^2/(2n) = \tau^2/2$!! So *the additional parameters μ_i are a serious nuisance if τ^2 is the parameter of interest*.

Suppose that there is a minimal sufficient statistic $T = t(X)$ partitioned as $T = (S, C) = (s(X), c(X))$ where:

- C1: the distribution of C depends on λ but not on ψ;
- C2: the conditional distribution of S given $C = c$ depends on ψ but not λ, for all c;
- C3: the parameters vary independently, i.e. $\Theta = \Psi \times \Lambda$.

Then the likelihood function factorizes as

$$L(\theta \,|\, x) \propto f(s, c; \theta) = f(s \,|\, c; \psi) f(c; \lambda)$$

and we say that C is *ancillary for* ψ, S is *conditionally sufficient for* ψ given C, and C is *marginally sufficient for* λ.

We also say that C is a *cut* for λ and would then

- base inference about λ on the marginal distribution of C;
- base inference about ψ on the conditional distribution of S given $C = c$.

The previous example shows that straight likelihood considerations may not lead to meaningful results when only a part of the parameter is considered.

There are a number of suggestions for modifying the likelihood function to extract the evidence in the sample concerning a parameter of interest ψ when $\theta = (\psi, \lambda)$. Such modifications are generally known as *pseudo-likelihood* functions.

Examples include: *conditional* likelihood, *marginal* likelihood, *profile* likelihood, *integrated* likelihood, and others, for example local, partial, restricted, residual, penalized, etc. The many names bear witness that straight likelihood considerations may not always be satisfactory.

Consider a sample $X = (X_1, \ldots, X_n)$ from a normal distribution $\mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Recall that $(U, V) = (\bar{X}, S^2 = \sum_i (X_i - \bar{X}_i)^2)$ is minimal sufficient and the likelihood function is

$$L(\mu, \sigma^2 \,|\, x) \propto f(u; \mu, \sigma^2) f(v; \sigma^2).$$

If we do straight maximum likelihood estimation, we have

$$\hat{\mu} = U = \bar{X}, \quad \hat{\sigma}^2 = V/n.$$

However, most statisticians agree that it is sensible to use $\tilde{\sigma}^2 = V/(n-1)$ as the estimator of σ^2. Is this reasonable and is there a general rationale for this?

Note that the *common unbiasedness argument does not work* as $\tilde{\sigma}$ is *not* ubiased for the standard deviation σ, or $\tilde{\sigma}^{-1}$ is *not* unbiased for the precision σ^{-2}.

Suppose we can write the joint density of a sufficient statistic $T = (U, V)$ as

$$f(u; \lambda, \psi) f(v \,|\, u; \psi),$$

where ψ is the parameter of interest. Then, for fixed ψ, U is sufficient for λ. Inference for ψ can now be based on the *conditional likelihood function*

$$L(\psi; v \,|\, u) = f(v \,|\, u; \psi),$$

as the conditional distribution does not involve λ.

The critical issue is whether (useful) information about ψ is lost by ignoring the factor $f(u; \lambda, \psi)$.

In the normal example with many nuisance parameters, $U = (U_i, i = 1, \ldots, n)$ is sufficient for the nuisance parameter $\lambda = (\mu_i, i = 1, \ldots, n)$ and for $\psi = \tau^2$

$$L(\tau^2; w \mid u) = f(w \mid u; \tau^2) = f(w; \tau^2) = (\tau^2)^{-n/2} e^{-\frac{w}{2\tau^2}}.$$

This gives the conditional MLE $\hat{\tau}^2_{|u} = W/n$ which is more sensible.

It may be argued that $U_i \sim \mathcal{N}(\mu_i, \tau^2)$ cannot possibly have useful information about τ^2. Or at least that the information it may have is not useful.

Although the profile likelihood generally can be very useful, it does not help in the the normal example with many nuisance parameters with $\lambda = (\mu_i, i = 1, \ldots, n)$ and $\psi = \tau^2$ we get

$$\hat{L}(\tau^2; w) = f(u; \hat{\mu}, \tau^2) f(w; \tau^2) = (\tau^2)^{-n} e^{-\frac{w}{2\tau^2}},$$

hence also peaks in the wrong place, at $\hat{\tau}^2 = W/(2n)$.

We shall later return to various attempts at modifying the profile likelihood.

This uses conditioning the other way around. Suppose we can write the joint density of a sufficient statistic $T = (U, V)$ as

$$f(u \mid v; \lambda, \psi) f(v; \psi),$$

where ψ is the parameter of interest. Then the nuisance parameter λ can be eliminated by marginalization as it does not enter in the marginal distribution of V. Inference for ψ can now be based on the *marginal likelihood function*

$$L(\psi; v) = f(v; \psi).$$

The issue is also here whether (useful) information about ψ is lost by ignoring the factor $f(u \mid v; \lambda, \psi)$.

Another way of removing nuisance parameters from the likelihood is to use integration. This method is essentially Bayesian and demands the specification of a prior distribution $\pi(\lambda \mid \psi)$ of the nuisance parameter for fixed ψ.

The *integrated likelihood function* is then defined as

$$\bar{L}(\psi) = \int L(\psi, \lambda) \pi(\lambda \mid \psi) \, d\lambda.$$

The integrated likelihood has the *same fundamental relation to the marginal prior and posterior distributions as the ordinary likelihood.*

For if $\pi(\psi)$ is the prior on ψ, the full posterior distribution is determined as

$$\pi^*(\psi, \lambda) \propto \pi(\psi) \pi(\lambda \mid \psi) L(\psi, \lambda)$$

and thus, by integration

$$\pi^*(\psi) \propto \int \pi^*(\psi, \lambda) \, d\lambda = \pi(\psi) \bar{L}(\psi).$$

In the normal example with many nuisance parameters with $\lambda = (\mu_i, i = 1, \ldots, n)$ and $\psi = \tau^2$ we get

$$L(\tau^2; w) = f(w; \tau^2) = (\tau^2)^{-n/2} e^{-\frac{w}{2\tau^2}},$$

which in this case is identical to the conditional likelihood function considered earlier and hence $\hat{\tau}^2_w = W/n$.

Marginal likelihood is in this case also known as *residual likelihood* because it is based on the residuals

$$V_i = X_i - \hat{\mu}_i = X_{i1} - \frac{X_{i1} + X_{i2}}{2} = \frac{X_{i1} - X_{i2}}{2}.$$

The corresponding estimates are then known as *REML* estimates.

In the normal example with many nuisance parameters, we may for example consider μ_i independent and normally distributed as $\mu_i \sim \mathcal{N}(\alpha, \omega^2)$, where (α, ω^2) represent prior knowledge about the population from which μ_i's are taken.

The integrated likelihood for τ^2 can then be calculated as

$$\bar{L}(\tau^2) = f(w; \tau^2) \int \prod_i f(u_i; \mu_i) \pi(\mu_i; \alpha, \omega^2) \, d\mu_i.$$

The integral can be recognized as the marginal distribution of U where now U_i are independent and identically distributed as $\mathcal{N}(\alpha, \tau^2 + \omega^2)$.

Marginal and conditional likelihood changes the problem either by ignoring some of the data (by marginalization) or by ignoring their variability (by conditioning).

Profile likelihood attempts to stick to the original data distribution and likelihood function, but eliminates the nuisance parameters by maximization.

The *profile likelihood function* $\hat{L}(\psi)$ for ψ is defined as

$$\hat{L}(\psi) = \sup_\lambda L(\psi, \lambda) = L\{\psi, \hat{\lambda}(\psi)\},$$

where ψ is the parameter of interest and $\hat{\lambda}(\psi)$ is the MLE of λ when ψ is considered fixed.

Thus

$$\begin{aligned} \bar{L}(\tau^2) &\propto f(w; \tau^2)(\tau^2 + \omega^2)^{-n/2} e^{-\frac{1}{2(\tau^2+\omega^2)} \sum_i (u_i - \alpha)^2} \\ &\propto (\tau^2)^{-n/2} e^{-\frac{w}{2\tau^2}} (\tau^2 + \omega^2)^{-n/2} e^{-\frac{q_\alpha(u)}{2(\tau^2+\omega^2)}} \end{aligned}$$

where

$$Q_\alpha(U) = \sum_i (U_i - \alpha)^2.$$

In this calculation, ω^2 and α are *known and fixed.* If these are 'correct', in the sense that μ_i are in fact behaving as if they were i.i.d. $\mathcal{N}(\alpha, \omega^2)$, then the integrated likelihood will peak around the correct value, else the peak will be shifted to an incorrect position. So the influence of the prior *prevails*.

Empirical Bayes or, equivalently(!), *MLE in the random effects model,* would also estimate α and ω^2 and get it right, as would *Hierarchical Bayes*, assigning a prior on (α, ω^2).

Generalized linear models

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 5, Hilary Term 2008

February 1, 2008

A generalized linear model is based on a family the form

$$f(y;\theta,\phi) = b(y,\phi)e^{\{y\theta - c(\theta)\}/d(\phi)}. \quad (1)$$

For ϕ fixed and θ varying over all possible values, this is a one-dimensional exponential family with canonical statistic $t(y) = y$, canonical parameter $\theta^* = \theta/d(\phi)$, and cumulant generating function

$$\kappa\{\theta^*\} = \kappa\{\theta/d(\phi)\} = c(\theta)/d(\phi) = \log \int b(y,\phi)e^{y\theta^*}\, dy, \quad (2)$$

so

$$\mathbf{E}(Y) = \frac{\partial}{\partial\theta^*}\kappa\{\theta/d(\phi)\} = d(\phi)\frac{\partial}{\partial\theta}\kappa\{\theta/d(\phi)\} = c'(\theta)$$

and

$$\mathbf{V}(Y) = \frac{\partial^2}{\partial\theta^{*2}}\kappa\{\theta/d(\phi)\} = d(\phi)^2\frac{\partial^2}{\partial\theta^2}\kappa\{\theta/d(\phi)\} = c''(\theta)d(\phi).$$

An exponential families with the canonical statistic $t(y) = y$ is also known as a *natural exponential family* (NEF), but terminology varies among authors so beware. Clearly, one can either consider the family (1) as an exponential family with canonical statistic $t(y) = y$ and canonical parameter $\theta^* = \theta/d(\phi)$, or let $t^*(y) = y/d(\phi)$ with parameter θ.

For varying ϕ, the situation is generally much more complex. Sometimes it is an exponential family, sometimes not. Sometimes it is not possible to have $d(\phi)$ varying independently of θ at all, e.g. in the Poisson case.

When $d(\phi)$ is varying, it is a strong restriction on the function $b(y,\phi)$ to assume that the cumulant generating function (2) has the form $\kappa(\theta/\phi) = c(\theta)/d(\phi)$.

Since $\mathbf{V}(Y) = d(\phi)c''(\theta)$ we have $c''(\theta) > 0$ and hence the function $c'(\theta)$ is strictly increasing in θ. We can therefore parametrize the family with its mean μ and define $\theta(\mu)$ by the relation

$$\mu = \mathbf{E}(Y) = c'(\theta), \quad \theta(\mu) = c'^{-1}(\mu)$$

and define the *variance function*

$$v(\mu) = c''\{\theta(\mu)\}$$

so now

$$\mathbf{V}(Y) = d(\phi)v(\mu)$$

and we can readily think of ϕ as a *dispersion parameter*.

An important fact is that *the variance functions identifies the family* in the sense that two families of densities which both have the form (1) and have the same variance function $v(\mu)$, must be identical.

Common variance functions for standard families are

Normal	Poisson	Binomial	Gamma	Inverse Gaussian
1	μ	$\mu(1-\mu)$	μ^2	μ^3

Not all functions $v(\mu)$ can occur as variance functions.

For example, a function of the form $v(\mu) = \mu^\alpha$ is a variance function for a NEF if $\alpha \leq 0$ or $1 \leq \alpha < \infty$, but not if $0 < \alpha < 1$.

Generalized linear models describe independent samples of the form $Y = (Y_1, \ldots, Y_n)$ where each Y_i is a one-dimensional *response* to *covariates* $x_i = (x_{i1}, \ldots, x_{ip})$ having distribution of the form (1), with expectations μ_i and dispersions $d_i(\phi)$.

For simplicity we assume $d_i(\phi) = \phi$ although $d_i(\phi) = \phi/w_i$ with w_i being a known *weight* may be appropriate in some cases. *Formally we assume ϕ known for the moment.*

The *saturated model* makes no further restriction on the parameters μ_i and the maximum likelihood estimator under this model is therefore given as

$$\hat{\mu} = Y,$$

provided the base exponential family is regular.

More generally, we restrict the vector of expectations $\mu = (\mu_1, \ldots, \mu_n)^\top$ through a *linear predictor* $\eta_i = x_i\beta$ written in matrix form as

$$\eta = X\beta$$

where x_i are the rows of X and $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a vector of unknown parameters, and a *link function* g relating the linear predictor to the mean as

$$\eta_i = g(\mu_i).$$

Here care should be taken in the choice of link function, as the parameter space for β must be restricted so that this equation makes sense.

A special role is played by the link function $g(\mu) = \theta(\mu) = c'^{-1}(\mu)$ which is known as *canonical link.*

If we consider the likelihood function we get

$$l(\beta) = \log L(\beta) = \sum_i \{y_i\theta_i - c(\theta_i)\}/\phi = \{y^\top\theta - \sum_i c(\theta_i)\}/\phi,$$

where now

$$\theta_i = \theta(\mu_i) = \theta\{g^{-1}(\eta_i)\} = \theta\{g^{-1}(x_i\beta)\}.$$

If g is the canonical link function, we have $g(\mu) = \theta(\mu)$ and hence $\theta_i = x_i\beta$. This then yields

$$l(\beta) = \{y^\top X\beta - \sum_i c(x_i\beta)\}/\phi = \{(X^\top y)^\top\beta - \sum_i c(x_i\beta)\}/\phi$$

and hence the family of joint distributions is a linear and canonical exponential family with canonical statistic $t(y) = X^\top y$ and β/ϕ as the canonical parameter.

Thus, the likelihood equation for a fixed ϕ again equates the expectation of the sufficient statistic to the observed value. Interpreting vector functions componentwise this has the simple form

$$X^\top \mu(\beta) = X^\top y$$

or equivalently

$$X^\top \{y - \mu(\beta)\} = 0$$

expressing that the residual $y - \mu(\beta)$ is orthogonal to all columns of X.

From general theory of exponential families it is known that *there is at most one solution $\hat{\beta}$ to this equation,* despite the fact that the equation typically is non-linear in β, as $\mu(\beta) = g^{-1}(X\beta)$.

Under reasonable assumption on the behaviour of the covariates x_i, D can be shown to be asymptotically distributed as a χ^2-distribution with degrees of freedom $n - p$ where X is assumed to have full rank p.

In the situation, where the dispersion parameter ϕ is considered *unknown* it is therefore customary to use the estimator

$$\tilde{\phi} = \frac{D_1(\hat{\mu}, Y)}{n - p}.$$

Note that this is *not* a maximum likelihood estimator, and there are good reasons for not using the MLE:

For a general link function, the score statistic can be written in the form

$$S(\beta) = Z^\top W\{y - \mu(\beta)\}/\phi$$

where Z is a matrix with elements

$$Z(\beta)_{ij} = \frac{\partial \eta_i}{\partial \beta_j}$$

and $W(\beta)$ is a diagonal matrix with diagonal elements equal to $W_{ii} = 1/v\{\mu_i(\beta)\}$.
Fisher's method of scoring leads to a *iterative weighted least squares regression* procedure (IRLS) for solving these, which now can be used for all generalized linear models, only the calculation of the matrix Z and the weights W being special to the model considered, depending in a simple way on the link and variance functions. Details are omitted here.

Firstly, the problem of finding the MLE of ϕ could be computationally very difficult in general, and the computational problem very different for different variance functions.

Secondly, there would be a problem with the nuisance parameter β distorting the estimate, in particular if the dimension p of β is large.

The estimate for ϕ used is thus based on 'approximate marginal likelihood', estimating ϕ on the basis of the approximate χ^2-distribution for the deviance. The MLE of μ is the same for all values of ϕ and is therefore appropriate as is.

The goodness of fit of a specific generalized linear model is assessed in the usual way using the *deviance*

$$\begin{aligned} D(\hat{\mu}; y) &= -2\{l(\hat{\mu}; y) - l(y; y)\} \\ &= -2\{l_1(\hat{\mu}; y) - l_1(y; y)\}/\phi = D_1(\hat{\mu}; y)/\phi, \end{aligned}$$

where $l(y; y)$ is the maximized log-likelihood in the saturated model and $l(\hat{\mu}; y)$ is the maximized log-likelihood in the model considered.

The symbol l_1 is used for the log-likelihood in the case $\phi = 1$ and similarly for D_1.

The Multivariate Gaussian Distribution

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 6, Hilary Term 2008

February 1, 2008

A d-dimensional random vector $X = (X_1, \ldots, X_d)$ is has a *multivariate Gaussian distribution* or *normal* distribution on $\mathcal{R}^d$ if there is a vector $\xi \in \mathcal{R}^d$ and a $d \times d$ matrix Σ such that

$$\lambda^\top X \sim \mathcal{N}(\lambda^\top \xi, \lambda^\top \Sigma \lambda) \quad \text{for all } \lambda \in R^d. \tag{1}$$

We then write $X \sim \mathcal{N}_d(\xi, \Sigma)$.

Taking $\lambda = e_i$ or $\lambda = e_i + e_j$ where e_i is the unit vector with i-th coordinate 1 and the remaining equal to zero yields:

$$X_i \sim \mathcal{N}(\xi_i, \sigma_{ii}), \quad \mathrm{Cov}(X_i, X_j) = \sigma_{ij}.$$

Hence ξ is the *mean vector* and Σ the *covariance matrix* of the distribution.

The definition (1) makes sense if and only if $\lambda^\top \Sigma \lambda \geq 0$, i.e. if Σ is *positive semidefinite*. Note that we have allowed distributions with variance zero.

The multivariate moment generating function of X can be calculated using the relation (1) as

$$m_d(\lambda) = E\{e^{\lambda^\top X}\} = e^{\lambda^\top \xi + \lambda^\top \Sigma \lambda / 2}$$

where we have used that the univariate moment generating function for $\mathcal{N}(\mu, \sigma^2)$ is

$$m_1(t) = e^{t\mu + \sigma^2 t^2/2}$$

and let $t = 1$, $\mu = \lambda^\top \xi$, and $\sigma^2 = \lambda^\top \Sigma \lambda$.

In particular this means that *a multivariate Gaussian distribution is determined by its mean vector and covariance matrix.*

Assume $X^\top = (X_1, X_2, X_3)$ with X_i independent and $X_i \sim \mathcal{N}(\xi_i, \sigma_i^2)$. Then

$$\lambda^\top X = \lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 \sim \mathcal{N}(\mu, \tau^2)$$

with

$$\mu = \lambda^\top \xi = \lambda_1 \xi_1 + \lambda_2 \xi_2 + \lambda_3 \xi_3, \quad \tau^2 = \lambda_1^2 \sigma_1^2 + \lambda_2^2 \sigma_2^2 + \lambda_3^2 \sigma_3^2.$$

Hence $X \sim \mathcal{N}_3(\xi, \Sigma)$ with $\xi^\top = (\xi_1, \xi_2, \xi_3)$ and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}.$$

If Σ is *positive definite*, i.e. if $\lambda^\top \Sigma \lambda > 0$ for $\lambda \neq 0$, the distribution has density on $\mathcal{R}^d$

$$f(x \,|\, \xi, \Sigma) = (2\pi)^{-d/2} (\det K)^{1/2} e^{-(x-\xi)^\top K (x-\xi)/2}, \tag{2}$$

where $K = \Sigma^{-1}$ is the *concentration matrix* of the distribution. We then also say that Σ is *regular*.

If $X_1, \ldots, X_d$ are independent and $X_i \sim \mathcal{N}(\xi_i, \sigma_i^2)$ their joint density has the form (2) with $\Sigma = \mathrm{diag}(\sigma_i^2)$ and $K = \Sigma^{-1} = \mathrm{diag}(1/\sigma_i^2)$.

Hence *vectors of independent Gaussians are multivariate Gaussian.*

In the bivariate case it is traditional to write

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1 \sigma_2 \rho \\ \sigma_1 \sigma_2 \rho & \sigma_2^2 \end{pmatrix},$$

with ρ being the *correlation* between X_1 and X_2. Then

$$\det(\Sigma) = \sigma_1^2 \sigma_2^2 (1 - \rho^2) = \det(K)^{-1}$$

and

$$K = \frac{1}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)} \begin{pmatrix} \sigma_2^2 & -\sigma_1 \sigma_2 \rho \\ -\sigma_1 \sigma_2 \rho & \sigma_1^2 \end{pmatrix}.$$

Thus the density becomes

$$f(x \,|\, \xi, \Sigma) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{(1 - \rho^2)}} \times e^{-\frac{1}{2(1-\rho^2)} \left\{ \frac{(x_1 - \xi_1)^2}{\sigma_1^2} - 2\rho \frac{(x_1 - \xi_1)(x_2 - \xi_2)}{\sigma_1 \sigma_2} + \frac{(x_2 - \xi_2)^2}{\sigma_2^2} \right\}}.$$

The contours of this density are ellipses and the corresponding density is bell-shaped with maximum in (ξ_1, ξ_2).

The marginal distributions of a vector X can all be Gaussian without the joint being multivariate Gaussian:

For example, let $X_1 \sim \mathcal{N}(0, 1)$, and define X_2 as

$$X_2 = \begin{cases} X_1 & \text{if } |X_1| > c \\ -X_1 & \text{otherwise.} \end{cases}$$

Then, using the symmetry of the univariate Gausssian distribution, X_2 is also distributed as $\mathcal{N}(0, 1)$.

However, the joint distribution is not Gaussian unless $c = 0$ since, for example, $Y = X_1 + X_2$ satisfies

$$P(Y = 0) = P(X_2 = -X_1) = P(|X_1| \leq c) = \Phi(c) - \Phi(-c).$$

Note that for $c = 0$, the correlation ρ between X_1 and X_2 is 1 whereas for $c = \infty$, $\rho = -1$.
It follows that there is a value of c so that X_1 and X_2 are uncorrelated, and still not jointly Gaussian.

Adding two independent Gaussians yields a Gaussian:
If $X_1 \sim \mathcal{N}_d(\xi_1, \Sigma_1)$ and $X_2 \sim \mathcal{N}_d(\xi_2, \Sigma_2)$ and $X_1 \perp\!\!\!\perp X_2$

$$X_1 + X_2 \sim \mathcal{N}_d(\xi_1 + \xi_2, \Sigma_1 + \Sigma_2).$$

To see this, just note that

$$\lambda^\top (X_1 + X_2) = \lambda^\top X_1 + \lambda^\top X_2$$

and use the univariate addition property.

Linear transformations preserve multivariate normality:
If A is an $r \times d$ matrix, $b \in \mathcal{R}^r$ and $X \sim \mathcal{N}_d(\xi, \Sigma)$, then

$$Y = AX + b \sim \mathcal{N}_r(A\xi + b, A\Sigma A^\top).$$

Again, just write

$$\gamma^\top Y = \gamma^\top (AX + b) = (A^\top \gamma)^\top X + \gamma^\top b$$

and use the corresponding univariate result.

Partition X into into X_1 and X_2, where $X_1 \in \mathcal{R}^r$ and $X_2 \in \mathcal{R}^s$ with $r + s = d$.
Partition mean vector, concentration and covariance matrix accordingly as

$$\xi = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}, \quad K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

so that Σ_{11} is $r \times r$ and so on. *Then, if $X \sim \mathcal{N}_d(\xi, \Sigma)$*

$$X_2 \sim \mathcal{N}_s(\xi_2, \Sigma_{22}).$$

This follows simply from the previous fact using the matrix

$$A = (0_{sr}\ I_s).$$

where 0_{sr} is an $s \times r$ matrix of zeros and I_s is the $s \times s$ identity matrix.

If Σ_{22} is regular, it further holds that

$$X_1 \,|\, X_2 = x_2 \sim \mathcal{N}_r(\xi_{1|2}, \Sigma_{1|2}),$$

where

$$\xi_{1|2} = \xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \xi_2) \quad \text{and} \quad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

In particular, *if $\Sigma_{12} = 0$ if and only if X_1 and X_2 are independent.*

From the matrix identities

$$K_{11}^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Sigma_{1\,|\,2} \tag{3}$$

and

$$K_{11}^{-1}K_{12} = -\Sigma_{12}\Sigma_{22}^{-1}, \tag{4}$$

it follows that then the conditional expectation and concentrations also can be calculated as

$$\xi_{1|2} = \xi_1 - K_{11}^{-1}K_{12}(x_2 - \xi_2) \quad \text{and} \quad K_{1|2} = K_{11}.$$

Note that the *marginal covariance is simply expressed in terms of Σ* where as the *conditional concentration is simply expressed in terms of K*.

Consider $\mathcal{N}_3(0, \Sigma)$ with covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

The concentration matrix is

$$K = \Sigma^{-1} = \begin{pmatrix} 3 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

The marginal distribution of (X_2, X_3) has covariance and concentration matrix

$$\Sigma_{23} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad (\Sigma_{23})^{-1} = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

The conditional distribution of (X_1, X_2) given X_3 has concentration and covariance matrix

$$K_{12} = \begin{pmatrix} 3 & -1 \\ -1 & 1 \end{pmatrix}, \quad \Sigma_{12|3} = (K_{12})^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix}.$$

Similarly, $\mathbf{V}(X_1 \,|\, X_2, X_3) = 1/k_{11} = 1/3$, etc.

Laplace's Method of Integration

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 7, Hilary Term 2008

February 8, 2008

Consider the Gamma function

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t}\, dt$$

and recall that for integers λ we have

$$\Gamma(\lambda + 1) = \lambda!$$

We get

$$\Gamma(\lambda + 1) = \int_0^\infty t^\lambda e^{-t}\, dt.$$

Substituting $y = t/\lambda$ and letting $g(y) = y - \log y$ we get

$$\Gamma(\lambda + 1) = \lambda \int_0^\infty (\lambda y)^\lambda e^{-\lambda y}\, dy = \lambda^{\lambda+1} \int_0^\infty e^{-\lambda g(y)}\, dy.$$

Consider an integral of form

$$I = \int_a^b e^{-\lambda g(y)} h(y)\, dy$$

where

1. λ is large;
2. $g(y)$ is a smooth function which has a local minimum at y^* in the interior of the interval (a, b);
3. $h(y)$ is smooth.

The integral can be the moment generating function of the distribution of $g(Y)$ when Y has density h, it could be a posterior expectation of $h(Y)$, or just an integral.

When λ is large, the contribution to this integral is essentially entirely originating from a neigbourhood around y^*.

To use Laplace's method we differentiate twice and get

$$g'(y) = 1 - 1/y, \quad g''(y) = 1/y^2$$

so that $y^* = 1$, $g(y^*) = 1$ and $g''(y^*) = 1$. Laplace's method now yields

$$\begin{aligned}\Gamma(\lambda + 1) &= \lambda^{\lambda+1} e^{-\lambda g(y*)} \sqrt{\frac{2\pi}{\lambda g''(y^*)}} \left\{1 + O\left(\frac{1}{\lambda}\right)\right\} \\ &= \lambda^{\lambda+1/2} e^{-\lambda} \sqrt{2\pi} \left\{1 + O\left(\frac{1}{\lambda}\right)\right\}\end{aligned}$$

which is known as *Stirling's formula*.

We formalize this by Taylor expansion of the function g around y^*:

$$g(y) = g(y^*) + g'(y^*)(y - y^*) + g''(y^*)(y - y^*)^2/2 + \cdots$$

Since y^* is a local minimum, we have $g'(y^*) = 0$, $g''(y^*) > 0$, and thus

$$g(y) - g(y^*) = g''(y^*)(y - y^*)^2/2 + \cdots$$

If we further approximate $h(y)$ linearly around y^* we get

$$\begin{aligned}I &= \int_a^b e^{-\lambda g(y)} h(y)\, dy \\ &\approx e^{-\lambda g(y*)} h(y^*) \int_{-\infty}^{\infty} e^{-\lambda g''(y^*)(y-y^*)^2/2}\, dy \\ &\quad + e^{-\lambda g(y*)} h'(y^*) \int_{-\infty}^{\infty} (y - y^*) e^{-\lambda g''(y^*)(y-y^*)^2/2}\, dy \\ &= e^{-\lambda g(y*)} h(y^*) \sqrt{\frac{2\pi}{\lambda g''(y^*)}} + 0.\end{aligned}$$

By expanding the function g further, the error of approximation can be improved for a constant function h so that

$$\begin{aligned}\tilde{I} &= \int_a^b e^{-\lambda g(y)}\, dy \\ &= e^{-\lambda g(y*)} \sqrt{\frac{2\pi}{\lambda g''(y^*)}} \left\{1 + \frac{5\rho_3^* - 3\rho_4^*}{24\lambda} + O\left(\frac{1}{\lambda^2}\right)\right\},\end{aligned}$$

where

$$\rho_3^* = \frac{g^{(3)}(y^*)}{\{g''(y^*)\}^{3/2}}, \quad \rho_4^* = \frac{g^{(4)}(y^*)}{\{g''(y^*)\}^2}.$$

We have exploited that we know the integral and expectation of a Gaussian density with concentration $g''(y^*)\lambda$. The approximation is typically very accurate and satisfies

$$\begin{aligned}I &= \int_a^b e^{-\lambda g(y)} h(y)\, dy \\ &= e^{-\lambda g(y*)} h(y^*) \sqrt{\frac{2\pi}{\lambda g''(y^*)}} \left\{1 + O\left(\frac{1}{\lambda}\right)\right\} = A\left\{1 + O\left(\frac{1}{\lambda}\right)\right\}\end{aligned}$$

meaning that the relative error

$$\frac{I - A}{A}$$

is $O(\lambda^{-1})$ and thus remains bounded for $\lambda \to \infty$, *even when multiplied with* λ.

In this fashion we can also get *Stirling's improved formula* as

$$\Gamma(\lambda + 1) = \lambda^{\lambda+1/2} e^{-\lambda} \sqrt{2\pi} \left\{1 + \frac{1}{12\lambda} + O\left(\frac{1}{\lambda^2}\right)\right\}$$

which is remarkably accurate, even for rather small values of λ, as this table of $\log \Gamma(\lambda + 1)$ shows:

λ	Exact	Stirling	Improved
2	0.6931472	0.6518048	0.6926268
4	3.1780538	3.1572615	3.1778807
8	10.6046029	10.5941899	10.6045527
16	30.6718601	30.6666508	30.6718456
32	205.1681995	205.1668957	205.1681970

Alternatively, if the variation of h around y^* is not neglible, or a more accurate approximation is desired, one can incorporate h in g as

$$\tilde{g}_\lambda(y) = g(y) - \frac{1}{\lambda}\log h(y)$$

and get the approximation

$$\begin{aligned} I &= \int_a^b e^{-\lambda g(y)} h(y)\, dy \\ &= \int_a^b e^{-\lambda \tilde{g}_\lambda(y)}\, dy \\ &= e^{-\lambda \tilde{g}_\lambda(\tilde{y}_\lambda)} \sqrt{\frac{2\pi}{\lambda \tilde{g}_\lambda''(\tilde{y}_\lambda)}} \left\{ 1 + \frac{5\tilde{\rho}_3 - 3\tilde{\rho}_4}{24\lambda} + O\left(\frac{1}{\lambda^2}\right) \right\}, \end{aligned}$$

where now $\tilde{y}_\lambda$ maximizes $\tilde{g}_\lambda(y)$, and other quantities are similarly defined.

The multivariate case is completely analogous. Here we again write

$$g(y) = g(y^*) + \frac{\partial g(y^*)}{\partial y}(y-y^*) + (y-y^*)^\top \frac{\partial^2 g(y^*)}{\partial y \partial y^\top}(y-y^*)/2 + \cdots$$

and exploit that the vector of partial derivatives $\frac{\partial g(y^*)}{\partial y}$ must vanish, whereby

$$\begin{aligned} I &= \int_B e^{-\lambda g(y)} h(y)\, dy \\ &= e^{-\lambda g(y^*)} h(y^*) \int_{\mathcal{R}^d} e^{-\lambda (y-y^*)^\top \frac{\partial^2 g(y^*)}{\partial y \partial y^\top}(y-y^*)/2 + \cdots}\, dy \\ &= e^{-\lambda g(y^*)} h(y^*) (2\pi/\lambda)^{d/2} \left| \frac{\partial^2 g(y^*)}{\partial y \partial y^\top} \right|^{-1/2} \left\{ 1 + O\left(\frac{1}{\lambda}\right) \right\}. \end{aligned}$$

Bayesian Asymptotics

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 8, Hilary Term 2008

February 8, 2008

For large λ we have the approximation

$$I = \int_a^b e^{-\lambda g(y)} h(y)\, dy = e^{-\lambda g(y*)} h(y^*) \sqrt{\frac{2\pi}{\lambda g''(y^*)}} \left\{1 + O\left(\frac{1}{\lambda}\right)\right\}$$

A more accurate approximation is

$$I = e^{-\lambda \tilde{g}_\lambda(\tilde{y}_\lambda)} \sqrt{\frac{2\pi}{\lambda \tilde{g}''_\lambda(\tilde{y}_\lambda)}} \left\{1 + \frac{5\tilde{\rho}_3 - 3\tilde{\rho}_4}{24\lambda} + O\left(\frac{1}{\lambda^2}\right)\right\},$$

where $\tilde{y}_\lambda$ maximizes $\tilde{g}_\lambda(y)$ and

$$\tilde{\rho}_3 = \frac{g^{(3)}(\tilde{y}_\lambda)}{\{g''(\tilde{y}_\lambda)\}^{3/2}}, \quad \tilde{\rho}_4 = \frac{g^{(4)}(\tilde{y}_\lambda)}{\{g''(\tilde{y}_\lambda)\}^2}.$$

In the multivariate case we have

$$\begin{aligned} I &= \int_B e^{-\lambda g(y)} h(y)\, dy \\ &= e^{-\lambda g(y^*)} h(y^*) \int_{\mathcal{R}^d} e^{-\lambda (y-y^*)^\top \frac{\partial^2 g(y^*)}{\partial y \partial y^\top}(y-y^*)/2 + \ldots}\, dy \\ &= e^{-\lambda g(y^*)} h(y^*) (2\pi/\lambda)^{d/2} \left|\frac{\partial^2 g(y^*)}{\partial y \partial y^\top}\right|^{-1/2} \left\{1 + O\left(\frac{1}{\lambda}\right)\right\} \end{aligned}$$

and additional accuracy up to $O(\lambda^{-2})$ can be obtained using derivatives of third and fourth order as in the univariate case.

We consider a standard asymptotic setup, involving $X_1, \ldots, X_n, \ldots$ random variables which, conditional on a d-dimensional parameter θ are independent and identically distributed with density $f(x \,|\, \theta)$, and $\pi(\theta)$ is the prior distribution of the parameter θ.

The posterior density is determined as

$$\pi^*(\theta) = f(\theta \,|\, x) \propto e^{l(\theta)} \pi(\theta),$$

where $l(\theta) = \log L(\theta)$ is the log-likelihood function. Letting

$$\bar{l}_n(\theta) = l(\theta)/n = \frac{1}{n} \sum_1^n \log f(X_i \,|\, \theta),$$

the law of large numbers yields that for $n \to \infty$,

$$\bar{l}_n(\theta) \to \mathbf{E}_\theta\{\log f(X \,|\, \theta)\} = -H(\theta),$$

where $H(\theta)$ is the *entropy* of the density $f(\cdot \,|\, \theta)$.

Thus the variation in the posterior density

$$\pi^*(\theta) \propto e^{n\bar{l}_n(\theta)} \pi(\theta)$$

will for sufficiently large n be dominated by the contribution from the likelihoood funtion. Expanding $l(\theta)$ around the maximum likelihood estimate $\hat{\theta}$ yields

$$\pi^*(\theta) \propto e^{n\bar{l}_n(\hat{\theta})} \pi(\hat{\theta}) e^{-(\theta-\hat{\theta})^\top j_n(\hat{\theta})(\theta-\hat{\theta})/2} \propto e^{-(\theta-\hat{\theta})^\top j_n(\hat{\theta})(\theta-\hat{\theta})/2}$$

where $j_n(\hat{\theta}) = nj(\hat{\theta})$ is the observed information matrix, so, approximately for large n, the posterior distribution of θ is

$$\theta \sim \mathcal{N}_d\{\hat{\theta}, j_n(\hat{\theta})^{-1}) = \mathcal{N}_d(\hat{\theta}, j(\hat{\theta})^{-1}/n\}.$$

Note this expression makes perfect sense, as $\hat{\theta}$ is not random in the posterior distribution.

A more accurate approximation is obtained by expanding around the posterior mode θ^*_π to get

$$\pi^*(\theta) \propto e^{-(\theta-\theta^*_\pi)^\top j_n(\theta^*_\pi)(\theta-\theta^*_\pi)/2}$$

yielding, approximately for large n, the posterior distribution of θ as

$$\theta \sim \mathcal{N}_d\{\theta^*_\pi, j_n(\theta^*_\pi)^{-1}) = \mathcal{N}_d(\hat{\theta}, j(\theta^*_\pi)^{-1}/n\}.$$

Note both differences and similarities to the analogous frequentist results

$$\hat{\theta} \sim \mathcal{N}_d\{\theta, i_n(\theta)^{-1}\} \quad \hat{\theta} \sim \mathcal{N}_d\{\theta, i_n(\hat{\theta})^{-1}\}, \quad \hat{\theta} \sim \mathcal{N}_d\{\theta, j_n(\hat{\theta})^{-1}\},$$

where the two latter needs appropriate interpretation to make perfect sense.

We can obtain an accurate approximation of the posterior distribution by applying Laplace's method to the normalization constant:

$$\begin{aligned} \pi^*(\theta) &= \frac{\exp\{l(\theta)\}\pi(\theta)}{\int_\Theta \exp\{l(\theta)\}\pi(\theta)\, d\theta} \\ &= (2\pi)^{-d/2} \exp\{l(\theta) - l(\hat{\theta})\} \frac{\pi(\theta)}{\pi(\hat{\theta})} \left|nj(\hat{\theta})\right|^{1/2} \{1 + O(n^{-1})\} \\ &= (2\pi/n)^{-d/2} \exp\{l(\theta) - l(\hat{\theta})\} \frac{\pi(\theta)}{\pi(\hat{\theta})} \left|j(\hat{\theta})\right|^{1/2} \{1 + O(n^{-1})\}. \end{aligned}$$

Note in particular the expression for the normalization constant

$$\int_\Theta f(x \,|\, \theta)\pi(\theta)\, d\theta = (2\pi/n)^{d/2} L(\hat{\theta})\pi(\hat{\theta}) \left|j(\hat{\theta})\right|^{-1/2} \{1 + O(n^{-1})\}.$$

Recall that for competing models M_1 and M_2 with parameters $\theta_1 \in \Theta_1 \in \mathcal{R}^{d_1}$ and $\theta_2 \in \Theta_2 \in \mathcal{R}^{d_2}$ and prior distributions π_1, π_2, the *Bayes factor* B in favour of M_1 over M_2 is

$$B = \frac{f(x_1, \ldots, x_n \,|\, M_1)}{f(x_1, \ldots, x_n \,|\, M_2)} = \frac{\int_{\Theta_1} f(x \,|\, \theta_1, M_1)\pi_1(\theta_1)\, d\theta_1}{\int_{\Theta_2} f(x \,|\, \theta_2, M_2)\pi_2(\theta_2)\, d\theta_2}.$$

Using the approximate expression obtained for the normalization constants, we get

$$B = (2\pi)^{(d_1-d_2)/2} n^{(d_2-d_1)/2} \frac{L(\hat{\theta}_1)}{L(\hat{\theta}_2)} \frac{\left|j_2(\hat{\theta}_2)\right|^{1/2}}{\left|j_1(\hat{\theta}_1)\right|^{1/2}} \{1 + O(n^{-1})\}.$$

To study the asymptotic behaviour of the Bayes factor we take logarithms and collect terms of similar order to get

$$\begin{aligned}\log B &= n\{\bar{l}_n(\hat{\theta}_1) - \bar{l}_n(\hat{\theta}_2)\} + \frac{d_2 - d_1}{2}\log n \\ &\quad -\frac{1}{2}\log\left\{\left|j_1(\hat{\theta}_2)\right| / \left|j_1(\hat{\theta}_1)\right|\right\} - \frac{d_2 - d_1}{2}\log(2\pi) + O(n^{-1}).\end{aligned}$$

The dominating terms are those on the first line, as all other terms are of smaller order for $n \to \infty$. Ignoring the latter we get

$$\log B \approx \{l(\hat{\theta}_1) - l(\hat{\theta}_2)\} - \frac{d_1 - d_2}{2}\log n.$$

The right-hand side is the *Bayesian Information Criterion* (BIC). It reflects that, for large n, the Bayes factor will favour the model with highest maximized likelihood (the first term), but will also penalize the model having the largest number of parameters.

Model comparison and selection

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lectures 9 and 10, Hilary Term 2008

March 2, 2008

Consider two alternative models $M_1 = \{f(x;\theta), \theta \in \Theta_1\}$ and $M_2 = \{f(x;\theta), \theta \in \Theta_2\}$ for a sample $(X = x) = (X_1 = x_1, \ldots, X_n = x_n)$.

We can apparently address the question of which of these are more adequate by considering the likelihood ratio

$$\Lambda = \frac{\sup_{\Theta_1} L(\theta)}{\sup_{\Theta_2} L(\theta)} = \frac{L(\hat\theta_1)}{L(\hat\theta_2)}.$$

Note that the quantities $L(\hat\theta_i)$ can be considered as the *profile likelihood* $\hat L_i$ of the 'model label' i, considering θ as a nuisance parameter.

If the models are *nested* in the sense that

$$\Theta_1 \subseteq \Theta_2$$

the likelihood ratio

$$\Lambda = \frac{\sup_{\Theta_1} L(\theta)}{\sup_{\Theta_2} L(\theta)} = \frac{L(\hat\theta_1)}{L(\hat\theta_2)}$$

will always be less than or equal to 1, so will always prefer the larger model as a description for the data.

There are many reasons this is not adequate, hence Λ as above is rarely used as a measure of relative accuracy of two models.

If the models are nested, one may in principle consider the *p-value*

$$p = P\{-2\log\Lambda \geq -2\log\lambda_{\text{obs}}; M_1\} \qquad (1)$$

i.e. the probability that the ratio Λ is less that the observed value, assuming the simpler model is true.

If the p-value is very small, corresponding to Λ_1 being unusually small, this will be taken as evidence against M_1, and so M_2 is favoured.

In contrast, if p is moderate, M_1 would be favoured over M_2 as the simpler explanation of the data.

This approach has several problems, including:

- it does not make clear sense unless M_2 has been established as adequate
- it does not make sense if the models M_i are not nested
- when many models M_i are considered, it is hard to control the probability of favouring an incorrect model by chance.

The *Bayes factor* B in favour of M_1 over M_2 is

$$B = \frac{f(x \mid M_1)}{f(x \mid M_2)} = \frac{\int_{\Theta_1} f(x \mid \theta, M_1)\pi_1(\theta)\,d\theta}{\int_{\Theta_2} f(x \mid \theta, M_2)\pi_2(\theta)\,d\theta} = \frac{\bar L_1}{\bar L_2},$$

where $\bar L_i$ are the *integrated likelihoods* for the models M_i.

When the integrated likelihood is approximated with using Laplace's method, we get the *Bayesian Information Criterion*

$$\bar L_i \approx \text{constant} + \text{BIC}_i = l(\hat\theta_i) - \frac{d_i}{2}\log n.$$

The prior distributions π_i do not enter in the expression for BIC which may or may not be seen as an advantage.

Models with a *high* value of BIC would be preferred over models with a low value of BIC.

One can get a more accurate approximation of the Bayes factor by adding terms

$$-\frac{1}{2}\log\left\{\left|j_i(\hat\theta_2)\right|\right\} + \frac{d_i}{2}\log(2\pi)$$

but this correction is not increasing with n, so it is most commonly ignored.

For the comparison of two models we get

$$\begin{aligned}\Delta\text{BIC} &= l(\hat\theta_1) - l(\hat\theta_2) + \frac{d_1 - d_2}{2}\log n \\ &= -\log\Lambda + \frac{d_1 - d_2}{2}\log n.\end{aligned}$$

Thus, in comparison with straight maximized likelihood, the simpler model gets preference by entertaining a lower penalty.

In the nested case, if $d_1 < d_2$ and the true value of the parameter $\theta_0 \in M_1 \subseteq M_2$, the deviance $-2\log\Lambda$ would under suitable regularity conditions be approximately $\chi^2(d_2 - d_1)$ and the penalty term will thus dominate for large values of n, so the simpler model will be correctly chosen.

In this sense, *BIC will asymptotically choose the simplest model which is correct.*

This classic criterion has been developed to choose between different subsets of variables in linear regression.

Consider the problem of predicting an n-dimensional vector Y with expectation μ from explanatory variables X. The total mean square prediction error would be

$$\mathbf{E}(||Y - \hat{Y}||^2) = \mathbf{E}\{||\mu - \hat{\mu}||^2\} + \mathbf{E}\{||Y - \mathbf{E}(Y)||^2\},$$

where $||v||^2 = \sum_i v_i^2$is the squared error norm.

The second term in this expression is the intrinsic random error and we can do nothing about it. The first term is the *squared prediction risk*

$$R = \mathbf{E}\{||\mu - \hat{\mu}||^2\}$$

and we would wish to choose a model for $\mu(X)$ which makes this risk small.

If it holds that $\mu = X\beta$ and we use a linear model of the form

$$\mu_S(X) = X(S)\beta_S$$

where S is a subset of d elements of the covariates so

$$x_i(S) = (x_{ij}, j \in S)$$

we thus have the prediction risk

$$R = \mathbf{E}\{||X\beta - X(S)\hat{\beta}_S||^2\} = d\sigma^2 + B(S)$$

where $B(S)$ is a bias term

$$B(S) = ||\mu - \mu_S(X)||^2 = ||X\beta - X(S)\beta_S||^2$$

with $B(S) = 0$ if the true distribution satisfies $\beta_j = 0$ for $j \notin S$.

The corresponding residual sum of squares has expectation

$$\mathbf{E}(\text{RSS}) = \mathbf{E}\{||Y - X(S)\hat{\beta}||^2\} = (n - d)\sigma^2 + B(S).$$

Thus, if we add $(2d - n)\sigma^2$ to both sides this equation, we get an unbiased estimate of the prediction risk from the residual sum of squares

$$\hat{R}(S) = \text{RSS} + (2d - n)\sigma^2.$$

Mallows C_p uses now an unbiased estimate of σ^2, typically based on the residual sum of squares for the model with all the variables included, to estimate the risk so that

$$C_p = \frac{\text{RSS}}{\hat{\sigma}^2} + 2d - n.$$

Choosing a model S can now be based on this criterion. Note that this also penalizes models with many parameters.

Akaike's Information Criterion (AIC) is based on exactly the same idea as C_p, but it is more general and is not restricted to regression models.

Akaike suggests assessing the prediction error by the *Kullback-Leibler distance* to the true distribution g:

$$D(g, \theta) = \int g(x) \log f(x, \theta)\, dx - \int g(x) \log g(x)\, dx = S(g, \theta) + H(g).$$

The AIC is an approximately unbiased estimate of $-2nS(g, \hat{\theta})$ which can be shown to reduce to

$$\text{AIC}_i = l(\hat{\theta}_i) - d_i$$

so

$$\Delta\text{AIC} = -\log\Lambda + (d_1 - d_2).$$

AIC gives typically lower penalty for complexity than BIC.

Maximum likelihood asymptotics

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 11, Hilary Term 2008

February 25, 2008

In terms of the original variable S we get

$$f_{S_n}(s) = \frac{e^{-x^2/2}}{\sigma\sqrt{2\pi n}}\left\{1 + \frac{\rho_3 H_3(x)}{6\sqrt{n}} + \frac{3\rho_4 H_4(x) + \rho_3^2 H_6(x)}{72n}\right\} + O(n^{-3/2}),$$

where $x = (s - n\mu)/(\sigma\sqrt{n})$.

Since $H_3(0) = 0$ this is particularly accurate when s is close to $n\mu$, as the first correction term then disappears.

If we wish a similar accuracy for other values of s we use the idea of *tilting* the distribution by shifting the log-density with a linear term, as we shall see next.

Let $X_1, \ldots, X_n$ be independent and identically distributed with density f and moment generating function $M(t) = \mathbf{E}e^{tX}$. The *cumulant generating function* of X is

$$K(t) = \log M(t) = \sum_{r=1}^{\infty} \frac{\kappa_r}{r!} t^r,$$

and the coefficient

$$\kappa_r = \frac{\partial^r K(0)}{\partial r\, t}$$

is the *cumulant of order r*. *The first two cumulants are the mean and variance*

$$\kappa_1 = \mu = \mathbf{E}(X), \quad \kappa_2 = \sigma^2 = \mathbf{V}(X).$$

Associate an exponential family of densities with the originial density f as

$$f(x;\gamma) = f(x)e^{x\gamma - K(\gamma)},$$

where K is the cumulant generating function of f. Clearly, $f(x;0) = f(x)$. We say that $f(x;\gamma)$ is obtained by *tilting* f by γ.

If X_i have density $f(x;\gamma)$, the sum S_n has density

$$f_{S_n}(s;\gamma) = f_{S_n}(s)e^{s\gamma - nK(\gamma)},$$

implying that

$$f_{S_n}(s) = e^{nK(\gamma) - s\gamma} f_{S_n}(s;\gamma).$$

Since this equation holds for all γ *we can now choose γ freely to suit our purpose*.

If X and Y are independent random variable, their cumulants satisfy

$$\kappa_r(aX + bY) = a^r\kappa_r(X) + b^r\kappa^r(Y).$$

The standardized *standardized cumulants*

$$\rho_r = \kappa_r/\kappa_2^{r/2}, r = 3, 4, \ldots$$

are thus invariant under translations and scaling

$$\rho_r(aX + b) = \rho_r$$

and therefore determine the shape of the density.

In the normal distribution, $\kappa_r = 0$ for $r > 2$ and cumulants ρ_r for $r > 2$ therefore indicate departures from normality.

The third cumulant ρ_3 is is known as the *skewness*, and the fourth cumulant ρ_4 as the *kurtosis* of the distribution.

If we use an Edgeworth expansion to approximate $f_{S_n}(s;\gamma)$ we can thus choose γ so that the expectation $\mathbf{E}_\gamma(S_n) = s$.

Since the mean of S_n in the tilted distribution is $nK'(\gamma)$ we should choose $nK'(\hat{\gamma}) = s$. As the variance of S_n in the tilted distribution is $nK''(\gamma)$, the resulting approximation is then

$$f_{S_n}(s) \approx e^{nK(\hat{\gamma}) - s\hat{\gamma}} \frac{1}{\{2\pi nK''(\hat{\gamma})\}^{-1/2}},$$

which can be extremely accurate.

Note that the Edgeworth approximation uses a normal approximation around the *mean* of the distribution whereas Laplace's method uses its *mode*. The tilting technique can be useful in both cases.

F. Y. Edgeworth (1845-1926), Professor of Political Economy at Oxford, showed that the density of

$$S_n^* = \frac{\sum_1^n X_i - n\mu}{\sigma\sqrt{n}}$$

could be approximated as

$$f_{S_n^*}(x) \approx \phi(x)\left\{1 + \frac{\rho_3 H_3(x)}{6\sqrt{n}} + \frac{3\rho_4 H_4(x) + \rho_3^2 H_6(x)}{72n}\right\} + O(n^{-3/2})$$

where ϕ is standard normal density and the omitted terms are $O(n^{-3/2})$ and H_r are *Hermite polynomials*

$$H_r(x) = (-1)^r\phi^{(r)}(x)/\phi(x).$$

For example, $H_3(x) = x^3 - 3x$, $H_4(x) = x^4 - 6x^2 + 3$, $H_6(x) = x^16 - 15x^4 + 45x^2 - 15$.

Use this approximation for an exponential family with parameter θ, then $K(t) = K(\theta + t) - K(t)$ and thus $\hat{\gamma} = \hat{\theta} - \theta$, where $\hat{\theta}$ is the MLE, yielding

$$f_{S_n}(s;\theta) \approx e^{n\{K(\hat{\theta}) - K(\theta)\} - s(\hat{\theta} - \theta)} \frac{1}{\{2\pi nK''(\hat{\theta})\}^{-1/2}} \propto \frac{L(\theta)}{L(\hat{\theta})}|j(\hat{\theta})|^{-1/2}.$$

Since $nK'(\hat{\theta}) = s$ we have

$$\frac{\partial\hat{\theta}}{\partial s} = \frac{1}{nK''(\hat{\theta})} = \frac{1}{nj(\hat{\theta})}$$

so a change of variables leads to the following approximate formula for the density of the MLE

$$f(\hat{\theta};\theta) \approx\propto \frac{L(\theta)}{L(\hat{\theta})}|j(\hat{\theta})|^{1/2}.$$

Similar methods can be used to show that, in wide generality, if A is ancillary so that $(\hat{\theta}, A)$ is minimal sufficient, then approximately, and quite often exactly,

$$f(\hat{\theta} \mid A = a; \theta) \approx\propto \frac{L(\theta)}{L(\hat{\theta})} |j(\hat{\theta})|^{1/2},$$

which is known as *Barndorff–Nielsen's formula*. Note that normalization constant may depend on θ and a.

Note similarity to the approximate Bayesian posterior:

$$\pi^*(\theta) \approx\propto \frac{L(\theta)}{L(\hat{\theta})} \left|j(\hat{\theta})\right|^{1/2}$$

where we have ignored the contribution $\pi(\theta)/\pi(\hat{\theta})$ from the prior. Only the interpretations are different!

Multivariate Gaussian Analysis

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 12, Hilary Term 2008

February 25, 2008

A square matrix A has *trace*

$$\mathrm{tr}(A) = \sum_i a_{ii}.$$

The trace has a number of properties:

1. $\mathrm{tr}(\gamma A + \mu B) = \gamma\,\mathrm{tr}(A) + \mu\,\mathrm{tr}(B)$ for γ, μ being scalars;
2. $\mathrm{tr}(A) = \mathrm{tr}(A^\top)$;
3. $\mathrm{tr}(AB) = \mathrm{tr}(BA)$
4. $\mathrm{tr}(A) = \sum_i \lambda_i$ where λ_i are the *eigenvalues* of A.

For a positive definite covariance matrix Σ, the multivariate Gaussian distribution has density on $\mathcal{R}^d$

$$f(x \,|\, \xi, \Sigma) = (2\pi)^{-d/2}(\det K)^{1/2} e^{-(x-\xi)^\top K(x-\xi)/2}, \qquad (1)$$

where $K = \Sigma^{-1}$ is the *concentration matrix* of the distribution. If $X_1 \sim \mathcal{N}_d(\xi_1, \Sigma_1)$ and $X_2 \sim \mathcal{N}_d(\xi_2, \Sigma_2)$ and $X_1 \perp\!\!\!\perp X_2$

$$X_1 + X_2 \sim \mathcal{N}_d(\xi_1 + \xi_2, \Sigma_1 + \Sigma_2).$$

If A is an $r \times d$ matrix, $b \in \mathcal{R}^r$ and $X \sim \mathcal{N}_d(\xi, \Sigma)$, then

$$Y = AX + b \sim \mathcal{N}_r(A\xi + b, A\Sigma A^\top).$$

For symmetric matrices the last statement follows from taking an orthogonal matrix O so that $OAO^\top = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ and using

$$\mathrm{tr}(OAO^\top) = \mathrm{tr}(AO^\top O) = \mathrm{tr}(A).$$

The trace is thus *orthogonally invariant*, as is the determinant:

$$\det(OAO^\top) = \det(O)\det(A)\det(O^\top) = 1\det(A)1 = \det(A).$$

There is an important trick that we shall use again and again: For $\lambda \in \mathcal{R}^d$

$$\lambda^\top A\lambda = \mathrm{tr}(\lambda^\top A\lambda) = \mathrm{tr}(A\lambda\lambda^\top)$$

since $\lambda^\top A\lambda$ is a scalar.

Partition X into X_1 and X_2, where $X_1 \in \mathcal{R}^r$ and $X_2 \in \mathcal{R}^s$ with $r + s = d$ and partition mean vector, concentration and covariance matrix accordingly.

Then, if $X \sim \mathcal{N}_d(\xi, \Sigma)$

$$X_2 \sim \mathcal{N}_s(\xi_2, \Sigma_{22}).$$

If Σ_{22} *is regular, it further holds that*

$$X_1 \,|\, X_2 = x_2 \sim \mathcal{N}_r(\xi_{1|2}, \Sigma_{1|2}),$$

where

$$\xi_{1|2} = \xi_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \xi_2) \quad \text{and} \quad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

In particular, *if* $\Sigma_{12} = 0$ *if and only if* X_1 *and* X_2 *are independent.*

Consider first the case where $\xi = 0$ and a sample $X_1 = x_1, \ldots, X_n = x_n$ from a multivariate Gaussian distribution $\mathcal{N}_d(0, \Sigma)$ with Σ regular. Using (1), we get the likelihood function

$$\begin{aligned} L(K) &= (2\pi)^{-nd/2}(\det K)^{n/2} e^{-\sum_{\nu=1}^n x_\nu^\top K x^\nu/2} \\ &\propto (\det K)^{n/2} e^{-\sum_{\nu=1}^n \mathrm{tr}\{Kx_\nu x_\nu^\top\}/2} \\ &= (\det K)^{n/2} e^{-\mathrm{tr}\{K\sum_{\nu=1}^n x_\nu x_\nu^\top\}/2} \\ &= (\det K)^{n/2} e^{-\mathrm{tr}(KW)/2}. \end{aligned} \qquad (4)$$

where

$$W = \sum_{\nu=1}^n x_\nu x_\nu^\top$$

is the matrix of *sums of squares and products*.

From the matrix identities

$$K_{11}^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Sigma_{1\,|\,2} \qquad (2)$$

and

$$K_{11}^{-1}K_{12} = -\Sigma_{12}\Sigma_{22}^{-1}, \qquad (3)$$

it follows that then the conditional expectation and concentrations also can be calculated as

$$\xi_{1|2} = \xi_1 - K_{11}^{-1}K_{12}(x_2 - \xi_2) \quad \text{and} \quad K_{1|2} = K_{11}.$$

Note that the *marginal covariance is simply expressed in terms of* Σ where as the *conditional concentration is simply expressed in terms of* K.

Writing the trace out

$$\mathrm{tr}(KW) = \sum_i \sum_j k_{ij} W_{ji}$$

emphasizes that it is linear in both K and W and we can recognize this as a linear and canonical exponential family with K as the canonical parameter and $-W/2$ as the canonical sufficient statistic. Thus, the likelihood equation becomes

$$\mathbf{E}(-W/2) == -n\Sigma/2 = -W/2$$

since $\mathbf{E}(W) = n\Sigma$. Solving, we get

$$\hat{K}^{-1} = \hat{\Sigma} = W/n$$

in analogy with the univariate case.

Rewriting the likelihood function as

$$\log L(K) = \frac{n}{2}\log(\det K) - \mathrm{tr}(KW)/2$$

we can of course also differentiate to find the maximum, leading to

$$\frac{\partial}{\partial k_{ij}}\log(\det K) = w_{ij}/n,$$

which in combination with the previous result yields

$$\frac{\partial}{\partial K}\log(\det K) = K^{-1}.$$

This can also be derived directly by writing out the determinant, and it holds for any non-singular square matrix!

If W_1 and W_2 are independent with $W_i \sim \mathcal{W}_d(n_i, \Sigma)$, then

$$W_1 + W_2 \sim \mathcal{W}_d(n_1 + n_2, \Sigma).$$

If A is an $r \times d$ matrix and $W \sim \mathcal{W}_d(n, \Sigma)$, then

$$AWA^\top \sim \mathcal{W}_r(n, A\Sigma A^\top).$$

For $r = 1$ we get that when $W \sim \mathcal{W}_d(n, \Sigma)$ and $\lambda \in R^d$,

$$\lambda^\top W \lambda \sim \sigma_\lambda^2 \chi^2(n),$$

where $\sigma_\lambda^2 = \lambda^\top \Sigma \lambda$.

The Wishart distribution is the sampling distribution of the matrix of sums of squares and products. More precisely:

A random $d \times d$ matrix W has a *d-dimensional Wishart distribution* with parameter Σ and n *degrees of freedom* if

$$W \stackrel{\mathcal{D}}{=} \sum_{i=1}^{n} X_\nu X_\nu^\top$$

where $X_\nu \sim \mathcal{N}_d(0, \Sigma)$. We then write

$$W \sim \mathcal{W}_d(n, \Sigma).$$

The Wishart is the multivariate analogue to the χ^2:

$$\mathcal{W}_1(n, \sigma^2) = \sigma^2 \chi^2(n).$$

If $W \sim \mathcal{W}_d(n, \Sigma)$ its mean is $\mathbf{E}(W) = n\Sigma$.

If $W \sim \mathcal{W}_d(n, \Sigma)$, where Σ is regular, then *W is regular with probability one if and only if $n \geq d$*.

When $n \geq d$ the Wishart distribution has density

$$\begin{aligned} & f_d(w \mid n, \Sigma) \\ &= c(d, n)^{-1} (\det \Sigma)^{-n/2} (\det w)^{(n-d-1)/2} e^{-\mathrm{tr}(\Sigma^{-1} w)/2} \end{aligned}$$

for w positive definite, and 0 otherwise.

The Wishart constant $c(d, n)$ is

$$c(d, n) = 2^{nd/2} (2\pi)^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\{(n + 1 - i)/2\}.$$

Wilks' Λ and Hotelling's T^2.

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 13, Hilary Term 2008

March 2, 2008

If X and Y are independent , $X \sim \Gamma(\alpha_x, \gamma)$, and $Y \sim \Gamma(\alpha_y, \gamma)$, then the ratio $X/(X+Y)$ follows a Beta distribution:

$$B = \frac{X}{X+Y} \sim \mathcal{B}(\alpha_x, \alpha_y).$$

A multivariate analogue of this result involves the Wishart distribution and asserts.

If $W_1 \sim \mathcal{W}_d(f_1, \Sigma)$ and $W_2 \sim \mathcal{W}_d(f_2, \Sigma)$ with $f_1 \geq d$, then the distribution of

$$\Lambda = \frac{\det(W_1)}{\det(W_1 + W_2)}$$

does not depend on Σ and is denoted by $\Lambda(d, f_1, f_2)$. The distribution is known as *Wilks' distribution*.

To see that the distribution of Λ does not depend on Σ, we choose a matrix A such that $A\Sigma A^\top = I_d$. Then

$$\tilde{W}_i = AW_iA^\top \sim \mathcal{W}_d(f_i, I_d)$$

and

$$\tilde{\Lambda} = \frac{\det(\tilde{W}_1)}{\det(\tilde{W}_1 + \tilde{W}_2)} = \frac{\det(A)\det(W_1)\det(A^\top)}{\det(A)\det(W_1+W_2)\det(A^\top)} = \Lambda.$$

Clearly, the distribution of $\tilde{\Lambda}$ does not depend on Σ and as $\tilde{\Lambda} = \Lambda$ this also holds for the latter.

Wilks' distribution is closely related to the Beta distribution. *It holds that*

$$\Lambda \stackrel{\mathcal{D}}{=} \prod_{i=1}^{d} B_i$$

where B_i are independent and follow Beta distributions with

$$B_i \sim \mathcal{B}\{(f_1 + 1 - i)/2, f_2/2)\}.$$

Indeed the distribution of

$$(W_1 + W_2)^{-1}W_1$$

is also known as *the multivariate Beta distribution*.

We first need a useful result about determinants of block matrices.

If A is a $d \times d$ symmetric matrix partitioned into blocks of dimension $r \times r$, $r \times s$, and $s \times s$ as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

it holds that

$$\det A = \det(A_{11} - A_{12}A_{22}^{-1}A_{21})\det(A_{22}). \qquad (1)$$

Here the entire expression should be considered equal to 0 if A_{22} is not invertible and $\det(A_{22}) = 0$.

This follows from a simple calculation

$$\begin{aligned} \det(A) &= \det\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \det\begin{pmatrix} I_{r\times r} & 0_{r\times s} \\ -A_{22}^{-1}A_{21} & I_{s\times s} \end{pmatrix} \\ &= \det\begin{pmatrix} A_{11} - A_{12}A_{22}^{-1}A_{21} & A_{12} \\ 0_{s\times r} & A_{22} \end{pmatrix} \\ &= \det(A_{11} - A_{12}A_{22}^{-1}A_{21})\det(A_{22}). \end{aligned}$$

Consider a partitioning of W and Σ into blocks as

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where Σ_{11} is an $r \times r$ matrix, Σ_{22} is $s \times s$, etc.

If $W \sim \mathcal{W}_d(f, \Sigma)$ and $\Sigma_{12} = \Sigma_{21} = 0$ then

$$\frac{\det(W)}{\det(W_{11})\det(W_{22})} \sim \Lambda(r, f-s, s) = \Lambda(s, f-r, r).$$

To see this is true we first use the matrix identity (1) to write

$$\frac{\det(W)}{\det(W_{11})\det(W_{22})} = \frac{\det(W_{1|2})}{\det(W_{11})} = \frac{\det(W_{1|2})}{\det(W_{1|2} + W_{12}W_{22}^{-1}W_{21})},$$

where $W_{1|2} = W_{11} - W_{12}W_{22}^{-1}W_{21}$.

Next we need to use that if $\Sigma_{12} = 0$ and thus $\Sigma_{1|2} = \Sigma_{11}$, it further holds that $W_{1|2}$ and $W_{12}W_{22}^{-1}W_{21}$ are independent and both Wishart distributed as

$$W_{1|2} \sim \mathcal{W}_r(f-s, \Sigma_{11}), \quad W_{12}W_{22}^{-1}W_{21} \sim \mathcal{W}_r(s, \Sigma_{11}).$$

We abstain from giving further details.

Wilks' distribution occurs as the likelihood ratio test for independence. Consider $X_1, \ldots, X_n \sim \mathcal{N}_d(0, \Sigma)$. The likelihood function is

$$L(K) = (\det K)^{n/2} e^{-\operatorname{tr}(KW)/2}.$$

As this is maximized by

$$\hat{K} = nW^{-1}$$

we have

$$L(\hat{K}) = (\det W)^{-n/2} e^{-nd/2}.$$

If $\Sigma_{12} = 0$ we similarly have

$$L(\hat{K}_{11}, \hat{K}_{22}) = (\det W_{11})^{-n/2} e^{-nr/2} (\det W_{22})^{-n/2} e^{-ns/2}.$$

Hence the likelihood ratio statistic is

$$\frac{L(\hat{K}_{11}, \hat{K}_{22})}{L(\hat{K})} = \left\{ \frac{\det(W)}{\det(W_{11}) \det(W_{22})} \right\}^{n/2} = \Lambda^{n/2}.$$

To see this we exploit the matrix identity (1) and calculate a determinant in two different ways. We may without loss of generality let $\mu = 0$. We have

$$\det \begin{pmatrix} W & -Y/\sqrt{c} \\ Y/\sqrt{c} & 1 \end{pmatrix} = \det(W + YY^\top/c) \cdot 1,$$

But we also have

$$\begin{aligned} \det \begin{pmatrix} W & -Y/\sqrt{c} \\ Y/\sqrt{c} & 1 \end{pmatrix} &= \det(1 + Y^\top W^{-1} Y/c) \det W \\ &= (1 + Y^\top W^{-1} Y/c) \det W \\ &= (1 + T^2/f) \det W. \end{aligned}$$

Let $Y \sim \mathcal{N}_d(\mu, c\Sigma)$ and $W \sim \mathcal{W}_d(f, \Sigma)$ with $f \geq d$, and $Y \perp\!\!\!\perp W$. Then

$$T^2 = f(Y - \mu)^\top W^{-1} (Y - \mu)/c$$

is known as *Hotelling's* T^2. This is the multivariate analogue of Student's t (or rather t^2).

It is equivalent to the likelihood ratio statistic for testing $\mu = 0$ from a sample $X_1, \ldots, X_n$ where then $Y = \bar{X}$, $W = \sum_i (X_i - \bar{X})$, $f = n - 1$, and $c = 1/n$.

It holds that

$$\frac{1}{1 + T^2/f} \sim \Lambda(d, f, 1) = \Lambda(1, f - d + 1, d).$$

Hence

$$\frac{1}{1 + T^2/f} = \frac{1}{1 + Y^\top W^{-1} Y/c} = \frac{\det W}{\det(W + YY^\top/c)}.$$

The result now follows by noting that $Y \sim \mathcal{N}_d(0, c\Sigma)$ implies $YY^\top/c \sim \mathcal{W}_d(1, \Sigma)$. Since

$$\Lambda(d, f, 1) = \Lambda(1, f - d + 1, d)$$

and the latter is a Beta distribution, *it also holds that*

$$\frac{f - d + 1}{fd} T^2 \sim F(d, f + 1 - d)$$

where F denotes Fisher's F-distribution.

Inverse Wishart Distribution and Conjugate Bayesian Analysis

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 14, Hilary Term 2008

March 2, 2008

Recall that the Wishart density has the form

$$f_d(w \mid f, \Sigma) \propto (\det w)^{(f-d-1)/2} e^{-\operatorname{tr}(\Sigma^{-1}w)/2}.$$

Since the likelihood function for Σ is

$$L(K) = (\det K)^{f/2} e^{-\operatorname{tr}(KW)/2},$$

a conjugate family of distributions for K is given by

$$\pi(K; a, \Psi) \propto (\det K)^{a/2-1} e^{-\operatorname{tr}(K\Psi)/2},$$

which thus specifies a Wishart distribution for the concentration matrix.

If $W_1 \sim \mathcal{W}_d(f_1, \Sigma)$ and $W_2 \sim \mathcal{W}_d(f_2, \Sigma)$ with $f_1 \geq d$, then the distribution of

$$\Lambda = \frac{\det(W_1)}{\det(W_1 + W_2)}$$

is Wilks' distribution and denoted by $\Lambda(d, f_1, f_2)$. It holds that

$$\Lambda \stackrel{\mathcal{D}}{=} \prod_{i=1}^{d} B_i$$

where B_i are independent and follow Beta distributions with

$$B_i \sim \mathcal{B}\{(f_1 + 1 - i)/2, f_2/2)\}.$$

We then say that Σ *follows an inverse Wishart distribution* if $K = \Sigma^{-1}$ follows a Wishart distribution, formally expressed as

$$\Sigma \sim \mathcal{IW}_d(\delta, \Psi) \iff K = \Sigma^{-1} \sim \mathcal{W}_d(\delta + d - 1, \Psi^{-1}),$$

i.e. if the density of K has the form

$$f(K \mid \delta, \Psi) \propto (\det K)^{\delta/2-1} e^{-\operatorname{tr}(\Psi K)/2}.$$

We repeat the expression for the standard Wishart density:

$$f_d(w \mid f, \Sigma) \propto (\det w)^{(f-d-1)/2} e^{-\operatorname{tr}(\Sigma^{-1}w)/2}.$$

It follows that the family of inverse Wishart distributions is a conjugate family for Σ.

Wilks' distribution occurs as the likelihood ratio test for independence. Consider $W \sim \mathcal{W}_d(f, \Sigma)$ and the hypothesis that $\Sigma_{12} = 0$ for a fixed block partitioning of Σ into $r \times r$, $r \times s$ and $s \times s$ matrices. The likelihood ratio statistic then becomes

$$\frac{L(\hat{K}_{11}, \hat{K}_{22})}{L(\hat{K})} = \left\{ \frac{\det(W)}{\det(W_{11}) \det(W_{22})} \right\}^{n/2} = U^{n/2},$$

where

$$U \sim \Lambda(r, f - s, s) = \Lambda(s, f - r, r).$$

It follows that

$$\Lambda(d, f_1, f_2) = \Lambda(f_2, f_1 + f_2 - d, d).$$

If the prior distribution of Σ is $\mathcal{IW}_d(\delta, \Psi)$ and $W \mid \Sigma \sim \mathcal{W}_d(f, \Sigma)$, we get for the posterior density of K that

$$\begin{aligned} f(K \mid \delta, \Psi, W) &\propto (\det K)^{f/2} e^{-\operatorname{tr}(KW)/2} \\ &\quad \times (\det K)^{\delta/2-1} e^{-\operatorname{tr}(\Psi K)/2} \\ &= (\det K)^{(f+\delta)/2-1} e^{-\operatorname{tr}\{(\Psi+W)K\}/2}, \end{aligned}$$

and hence the posterior distribution is simply $\mathcal{IW}_d(\delta + f, \Psi + W) = \mathcal{IW}_d(\delta^*, \Psi^*)$.

We can thus interpret the parameter δ as a prior equivalent sample size and Ψ as the value of a matrix of sums and squares and products from a previous sample.

This is the equivalent of Student's t-distribution. Let $Y \sim \mathcal{N}_d(\mu, c\Sigma)$, $W \sim \mathcal{W}_d(f, \Sigma)$ with $f \geq d$, and $Y \perp\!\!\!\perp W$.

$$T^2 = f(Y - \mu)^\top W^{-1}(Y - \mu)/c$$

is known as Hotelling's T^2.

It holds that

$$\frac{1}{1 + T^2/f} \sim \Lambda(d, f, 1) = \Lambda(1, f - d + 1, d)$$

and

$$\frac{f - d + 1}{fd} T^2 \sim F(d, f + 1 - d)$$

where F denotes Fisher's F-distribution.

We need the full form of the Wishart density for K, as constants may become important and recall that

$$\begin{aligned} & f_d(K \mid \delta, \Psi) \\ &= q(d, \delta)^{-1} (\det \Psi)^{(\delta+d-1)/2} (\det K)^{\delta/2-1} e^{-\operatorname{tr}(\Psi K)/2} \end{aligned}$$

The constant $q(d, \delta)$ is

$$q(d, \delta) = 2^{(\delta+d-1)d/2} (2\pi)^{d(d-1)/4} \prod_{i=1}^{d} \Gamma\{(\delta + d - i)/2\}.$$

Consider now alternative models M_1 with Σ arbitrary and M_2 with Σ of block diagonal form:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}.$$

If the associated prior distributions are for M_1 that $\Sigma \sim \mathcal{IW}_d(\delta, I_d)$ and for M_2 that $\Sigma_{11} \sim \mathcal{IW}_r(\delta, I_r)$, $\Sigma_{22} \sim \mathcal{IW}_s(\delta, I_s)$, we can now calculate the Bayes factor.

Sequential Bayesian Updating

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lectures 15 and 16, Hilary Term 2008

March 6, 2008

The previous considerations take on a particular dynamic form when also the parameter or *state* θ is changing with time. More precisely, we consider a Markovian model for the *state dynamics* of the form

$$f(\theta_0) = \pi(\theta_0), \quad f(\theta_{i+1} \,|\, \boldsymbol{\theta_i}) = f(\theta_{i+1} \,|\, \theta_i)$$

where the evolving states $\theta_0, \theta_1, \ldots$ are not directly observed, but information about them are available through sequential *observations* $X_i = x_i$, where

$$f(x_i \,|\, \boldsymbol{\theta_i}, \mathbf{x_{i-1}}) = f(x_i \,|\, \theta_i)$$

so the joint density of states and observations is

$$f(\mathbf{x_n}, \boldsymbol{\theta_n}) = \pi(\theta_0) \prod_{i=1}^{n} f(\theta_{i+1} \,|\, \theta_i) f(x_i \,|\, \theta_i).$$

We consider data arriving sequentially $X_1, \ldots, X_n, \ldots$ and wish to update inference on an unknown parameter θ online.

In a Bayesian setting, we have a prior distribution $\pi(\theta)$ and at time n we have a density for data conditional on θ as

$$f(x_1, \ldots, x_n \,|\, \theta) = f(x_1 \,|\, \theta) f(x_2 \,|\, x_1, \theta) \cdots f(x_n \,|\, \mathbf{x_{n-1}}, \theta)$$

where we have let $\mathbf{x_i} = (x_1, \ldots, x_i)$. Note that we are not assuming $X_1, \ldots, X_n, \ldots$ to be independent conditionally on θ.

At time n, we may have updated our distribution of θ to its posterior

$$\pi_n(\theta) = f(\theta \,|\, \mathbf{x_n}) \propto \pi(\theta) f(\mathbf{x_n} \,|\, \theta).$$

This type of model is common in robotics, speech recognition, target tracking, and steering/control, for example of large ships, airplanes, and space ships.

The natural tasks associated with inference about the evolving state θ_i are known as

- *Filtering:* Find $f(\theta_n \,|\, \mathbf{x_n})$. What is the current state?

If we obtain a new observation $X_{n+1} = x_{n+1}$ we may either start afresh and write

$$\pi_{n+1}(\theta) = f(\theta \,|\, \mathbf{x_{n+1}}) \propto \pi(\theta) f(\mathbf{x_{n+1}} \,|\, \theta)$$

or we could claim that just before time $n+1$, our knowledge of θ is summarized in the distribution $\pi_n(\theta)$ so we just use this as a prior distribution for the new piece of information and update as

$$\tilde{\pi}_{n+1}(\theta) \propto \pi_n(\theta) f(x_{n+1} \,|\, \mathbf{x_n}, \theta).$$

Indeed, *these updates are identical* since

$$\begin{aligned} \tilde{\pi}_{n+1}(\theta) &\propto \pi_n(\theta) f(x_{n+1} \,|\, \mathbf{x_n}, \theta) \\ &\propto \pi(\theta) f(\mathbf{x_n} \,|\, \theta) f(x_{n+1} \,|\, \mathbf{x_n}, \theta) \\ &= \pi(\theta) f(\mathbf{x_{n+1}} \,|\, \theta) \propto \pi_{n+1}(\theta). \end{aligned}$$

This type of model is common in robotics, speech recognition, target tracking, and steering/control, for example of large ships, airplanes, and space ships.

The natural tasks associated with inference about the evolving state θ_i are known as

- *Filtering:* Find $f(\theta_n \,|\, \mathbf{x_n})$. What is the current state?
- *Prediction:* Find $f(\theta_{n+1} \,|\, \mathbf{x_n})$. What is the next state?

We may summarize these facts by replacing the usual expression for a Bayesian updating scheme

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

with

$$\textit{revised} \propto \textit{current} \times \textit{new likelihood}$$

represented by the formula

$$\pi_{n+1}(\theta) \propto \pi_n(\theta) \times L_{n+1}(\theta) = \pi_n(\theta) f(x_{n+1} \,|\, \mathbf{x_n}, \theta).$$

In this dynamic perspective we notice that at time n we only need to keep a representation of π_n and otherwise can ignore the past.

The current π_n contains all information needed to revise knowledge when confronted with new information $L_{n+1}(\theta)$.

We sometimes refer to this way of updating as *recursive*.

This type of model is common in robotics, speech recognition, target tracking, and steering/control, for example of large ships, airplanes, and space ships.

The natural tasks associated with inference about the evolving state θ_i are known as

- *Filtering:* Find $f(\theta_n \,|\, \mathbf{x_n})$. What is the current state?
- *Prediction:* Find $f(\theta_{n+1} \,|\, \mathbf{x_n})$. What is the next state?
- *Smoothing:* Find $f(\theta_j \,|\, \mathbf{x_n}), j < n$. What was the past state at time j?

If the filter distribution $f(\theta_n \,|\, \mathbf{x_n})$ is available we may calculate the *predictive distribution* as

$$f(\theta_{n+1} \,|\, \mathbf{x_n}) = \int_{\theta_n} f(\theta_{n+1} \,|\, \theta_n) f(\theta_n \,|\, \mathbf{x_n})\, d\theta_n \qquad (1)$$

which uses the current filter distribution and the dynamic model. When a new observation $X_{n+1} = x_{n+1}$ is obtained, we can use

$$\text{revised} \propto \text{current} \times \text{new likelihood}$$

to update the filter distribution as

$$f(\theta_{n+1} \,|\, \mathbf{x_{n+1}}) \propto f(\theta_{n+1} \,|\, \mathbf{x_n}) f(x_{n+1} \,|\, \theta_{n+1}), \qquad (2)$$

i.e. the *updated filter distribution is found by combining the current predictive with the incoming likelihood.* The predictive distributions can now be updated to yield a general recursive scheme of *predict-observe-filter-predict-observe-filter*...

The filtering relations become particularly simple, since the conditional distributions all are normal, and we are only concerned with expectations and variances.

We repeat Thiele's argument as an instance of the general theory developed.

Suppose at time n we have the filter distribution of θ_n as $\mathcal{N}(\mu_n, \omega_n^2)$. Then the predictive distribution of θ_{n+1} is

$$\theta_{n+1} \,|\, \mathbf{x_n} \sim \mathcal{N}(\mu_n, \omega_n^2 + \sigma^2).$$

We can think of μ_n as our current 'best measurement' of θ_{n+1}, with this variance.

The contribution from the observation is a measurement of θ_{n+1} with a value of x_{n+1} and a variance τ^2. The best way of combining these estimates is to take a weighted average with the inverse variances as weights.

When we have more time, we may similarly look retrospectively and try to reconstruct the movements of θ. This calculation is slightly more subtle than filtering. We first get

$$\begin{aligned} f(\theta_{j-1} \,|\, \mathbf{x_n}) &= \int_{\theta_j} f(\theta_{j-1} \,|\, \theta_j, \mathbf{x_n}) f(\theta_j \,|\, \mathbf{x_n})\, d\theta_j \\ &= \int_{\theta_j} f(\theta_{j-1} \,|\, \theta_j, \mathbf{x_{j-1}}) f(\theta_j \,|\, \mathbf{x_n})\, d\theta_j, \end{aligned}$$

where we have used that

$$\begin{aligned} f(\theta_{j-1} \,|\, \theta_j, \mathbf{x_n}) &\propto f(\theta_{j-1} \,|\, \theta_j, \mathbf{x_{j-1}}) f(x_j, \ldots, x_n \,|\, \theta_j, \theta_{j-1}) \\ &= f(\theta_{j-1} \,|\, \theta_j, \mathbf{x_{j-1}}) f(x_j, \ldots, x_n \,|\, \theta_j) \end{aligned}$$

so

$$f(\theta_{j-1} \,|\, \theta_j, \mathbf{x_n}) = f(\theta_{j-1} \,|\, \theta_j, \mathbf{x_{j-1}}).$$

It follows that our new filter distribution has expectation

$$\mu_{n+1} = \frac{\mu_n/(\omega_n^2 + \sigma^2) + x_{n+1}/\tau^2}{(\omega_n^2 + \sigma^2)^{-1} + \tau^{-2}} = \frac{\tau^2 \mu_n + (\sigma_2 + \omega_n^2) x_{n+1}}{\tau^2 + \sigma^2 + \omega_n^2}$$

and variance

$$\omega_{n+1}^2 = \frac{1}{(\omega_n^2 + \sigma^2)^{-1} + \tau^{-2}} = \frac{\tau^2(\sigma^2 + \omega_n^2)}{\tau^2 + \sigma^2 + \omega_n^2}.$$

Clearly this result could also have been obtained from expanding the sum of squares in the expression for the filter distribution (2)

$$f(\theta_{n+1} \,|\, \mathbf{x_n}) \propto \exp\left\{ -\frac{(\theta_{n+1} - \mu_n)^2}{2(\sigma^2 + \omega_n^2)} + \frac{(\theta_{n+1} - x_{n+1})^2}{2\tau^2} \right\}.$$

Since further

$$f(\theta_{j-1} \,|\, \theta_j, \mathbf{x_{j-1}}) \propto f(\theta_j \,|\, \theta_{j-1}) f(\theta_{j-1} \,|\, \mathbf{x_{j-1}})$$

we thus get

$$\begin{aligned} f(\theta_{j-1} \,|\, \mathbf{x_n}) &\propto \int_{\theta_j} f(\theta_j \,|\, \theta_{j-1}) f(\theta_{j-1} \,|\, \mathbf{x_{j-1}}) f(\theta_j \,|\, \mathbf{x_n})\, d\theta_j \\ &\propto f(\theta_{j-1} \,|\, \mathbf{x_{j-1}}) \int_{\theta_j} f(\theta_j \,|\, \theta_{j-1}) f(\theta_j \,|\, \mathbf{x_n})\, d\theta_j, \end{aligned}$$

Which is the basic *smoothing recursion*:

$$f(\theta_{j-1} \,|\, \mathbf{x_n}) \propto f(\theta_{j-1} \,|\, \mathbf{x_{j-1}}) \int_{\theta_j} f(\theta_j \,|\, \theta_{j-1}) f(\theta_j \,|\, \mathbf{x_n})\, d\theta_j. \qquad (3)$$

It demands that we have stored a representation of the filter distributions $f(\theta_{j-1} \,|\, \mathbf{x_{j-1}})$ as well as the dynamic state model.

We may elaborate the expression for μ_{n+1} and write it as a correction of μ_n or of x_{n+1} as

$$\mu_{n+1} = \mu_n + \frac{\sigma^2 + \omega_n^2}{\tau^2 + \sigma^2 + \omega_n^2}(x_{n+1} - \mu_n)$$

or

$$\mu_{n+1} = x_{n+1} - \frac{\tau^2}{\tau^2 + \sigma^2 + \omega_n^2}(x_{n+1} - \mu_n)$$

showing how at each stage n the filtered value is obtained by modifying the observed and predicted values when the prediction is not on target.

The Kalman filter readily generalizes to the multivariate case and more complex models for the state evolution and observation equation. We abstain from further details.

This special case of the previous is traditionally attributed to Kalman from a result in 1960, but was in fact developed in full detail by the Danish statistician T.N. Thiele in 1880.

It is based on the Markovian state model

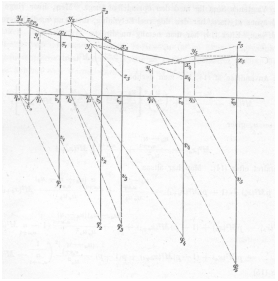
$$\theta_{i+1} \,|\, \theta_i \sim \mathcal{N}(\theta_i, \sigma_{i+1}^2), \quad \theta_0 = 0$$

and the simple observational model

$$X_i \,|\, \theta_i \sim \mathcal{N}(\theta_i, \tau_i^2), i = 1, \ldots$$

where typically $\sigma_i^2 = (t_i - t_{i-1})\sigma^2$ and $\tau_i^2 = \tau^2$ with t_i denoting the time of the ith observation. For simplicity we shall assume $t_i = i$ and $w_i = 1$ in the following.

This geometric construction of the Kalman filter and smoother is taken from Thiele (1880).



One of the most recent developments in modern statistics is using Monte Carlo methods for representing the predictive and filtered distributions.

We assume that we at time n have represented the filter distribution (2) by a sample

$$f(\theta_n \mid \mathbf{x_n}) \sim \{\theta_n^1, \ldots, \theta^M\}$$

so that we would approximate any integral w.r.t. this density as

$$\int h(\theta_n) f(\theta_n \mid \mathbf{x_n})\, d\theta_n \approx \sum_{i=1}^{M} h(\theta_n^i).$$

The values $\{\theta_n^1, \ldots, \theta_n^M\}$ are generally referred to as *particles*.

More generally, we may have the particles associated with weights

$$f(\theta_n \mid \mathbf{x_n}) \sim \{(\theta_n^1, w_n^1), \ldots, (\theta^M, w_n^M)\}$$

with $\sum_{i=1}^{M} w_n^i = 1$, so that the integral is approximated by

$$\int h(\theta_n) f(\theta_n \mid \mathbf{x_n})\, d\theta_n \approx \sum_{i=1}^{M} h(\theta_n^i) w_n^i. \tag{4}$$

Typically, w_i will reflect that we have been sampling from a *proposal distribution* $g(\theta_n)$ rather than the *target distribution* $f(\theta_n \mid \mathbf{x_n})$ so the weights are calculated as

$$w_n^i = f(\theta_n^i \mid \mathbf{x_n}) / g(\theta_n^i).$$

When filtering to obtain particles representing the next stage of the filtering distribution we *move* each particle a random amount by drawing θ_{n+1}^i at random from a proposal distribution $g_{n+1}(\theta \mid \theta_n^i, \mathbf{x_{n+1}})$ and subsequently *reweight* the particle as

$$w_{n+1}^i \propto w_n^i \frac{f(\theta_{n+1}^i \mid \theta_n^i) f(x_{n+1} \mid \theta_{n+1}^i)}{g_{n+1}(\theta_{n+1}^i \mid \theta_n^i, \mathbf{x_{n+1}})}$$

the numerator being proportional to $f(\theta_{n+1}^i \mid \theta_n^i, x_{n+1})$.

There are many possible proposal distributions but a common choice is a normal distribution with an approximately correct mean and slightly enlarged variance.

The approximate inverse variance of the integral (4) is for the constant function $h \equiv 1$ equal to

$$\tilde{M}_n = \frac{1}{\sum_i (w_n^i)^2}$$

which is known as *effective number of particles*. It is maximized for $w^i \equiv 1/M$ which represents weights obtained when sampling from the correct distribution.

As the filtering evolves, it may happen that some weights become very small, reflecting bad particles, which are placed in areas of small probability. This leads to the effective number of particles becoming small.

To get rid of these, M new particles are *resampled* with replacement, the probability for choosing particle i at each sampling being equal to w^i so that bad particles have high probability of not being included. This creates now a new set of particles which now all have weight $1/M$.
However, some particles will now be repeated in the sample and when this has been done many times, there may be only few particles left.

Various schemes then exist for *replenishing* and sampling new particles.

This can also be done routinely at each filtering, for example by first sampling two new particles for every existing one and subsequently resampling as above to retain exactly M particles.