Conditional and Marginal Inference: A (Highly Selective) Review

Anthony Davison
Institute of Mathematics
EPFL

Contents

Standard Likelihood Inference

Data $y_1, \ldots, y_n$ is a realization of random variables $Y_1, \ldots, Y_n$ from a parametric statistical model $f(y; \theta)$, $\theta \in \Theta$.

Log likelihood, observed information, expected information are

$$L(\theta) = f(y; \theta), \quad \ell(\theta) = \log f(y; \theta), \quad J(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}, \quad I(\theta) = \mathrm{E}\left\{J(\theta)\right\}.$$

If $\widehat{\theta}$ is MLE, then 'large sample' results

$$\begin{aligned} \widehat{\theta} &\dot{\sim} N\left\{\theta, I(\theta)^{-1}\right\}, \\ W(\theta) = 2\left\{\ell(\widehat{\theta}) - \ell(\theta)\right\} &\dot{\sim} \chi^2_p, \end{aligned}$$

yield standard first-order inferences on θ: tests, confidence intervals and sets, ...

Motivation

- Standard approach valid for large n. When is n large enough? What if n is small?
- Normal-theory confidence intervals not invariant to reparametrization: invariant intervals desirable.
- Often in practice $\theta = (\psi, \lambda)$, where ψ is interest parameter, and λ, though essential for the model to be appropriate, is not of real concern. Estimation of λ can damage inference for ψ, so would like to produce version of likelihood with λ eliminated.

Interest-preserving reparametrization

$Y = \exp(X)$, where $X \sim N(\mu, \sigma^2)$. Let

$$\psi = \mathrm{E}(Y) = \exp(\mu + \sigma^2/2), \quad \lambda = \mathrm{var}(Y) = \exp(2\mu + \sigma^2)\left\{\exp(\sigma^2) - 1\right\}.$$

Then interval (ψ_-, ψ_+) for ψ should transform to $(\log \psi_-, \log \psi_+)$, if the model is rewritten in terms of $\psi' = \log \psi$ and $\lambda' = \sigma^2$.

Marginal and Conditional Likelihoods

If we can factorize the likelihood as

$$f(y;\psi,\lambda) \propto f(t_1 \mid a;\psi) \times f(t_2 \mid t_1, a;\psi,\lambda),$$

or as

$$f(y;\psi,\lambda) \propto f(t_1 \mid t_2, a;\psi) \times f(t_2 \mid a;\psi,\lambda),$$

then we can use the first terms in the expressions as **marginal likelihood** or as **conditional likelihood** for ψ.

Statistic a is ancillary: see later.

If such a factorization exists, can perform inference for ψ without needing to account for λ.

Neyman–Scott problem

Standard example in which ordinary MLE performs catastrophically.

Let $Y_{ij} \sim N(\mu_i, \sigma^2)$, $i = 1, \ldots, m$, $j = 1, \ldots, k$. Suppose interest parameter is $\psi = \sigma^2$, nuisance parameter is $\lambda = (\mu_1, \ldots, \mu_m)$.

MLEs are

$$\widehat{\mu}_i = \overline{y}_{i\cdot} = k^{-1} \sum_j y_{ij}, \quad \widehat{\sigma}^2 = (mk)^{-1} \sum_{ij} (y_{ij} - \overline{y}_{i\cdot})^2,$$

so

$$\mathrm{E}(\widehat{\sigma}^2) = \sigma^2 \frac{(k-1)}{k}, \quad \mathrm{var}(\widehat{\sigma}^2) = 2\sigma^4 \frac{(k-1)}{mk^2}.$$

Thus if k is fixed and m increases, $\widehat{\sigma}^2$ converges to the wrong value. When $k = 2$, for example, $\widehat{\sigma}^2 \xrightarrow{P} \sigma^2/2$, so $\widehat{\sigma}^2$ is both biased in small samples and inconsistent as $m \to \infty$.

Contents

Conditioning in Statistics

Two basic roles:

- induce relevance of inference to situation under study;
- eliminate nuisance parameters.

Ancillarity is related to first use.

Illustration: : We have two machines to use to estimate θ, with variances 1, 100. We flip a coin to decide which one to use. If the coin comes down heads, say, and we use the first machine, it is irrelevant that we might have used the second. Thus we condition on the coin flip.

Ancillary Statistics

Ancillary statistic A is a function of the minimal sufficient statistic S whose distribution does not depend on θ. Suppose $S = (T, A)$. Then

$$f(y;\theta) = f(y \mid s)f(s;\theta) = f(y \mid s)f(a)f(t \mid a;\theta),$$

so inference should be based on the distribution of T given A.

The value a of A controls precision of estimators.

Example (Regression model): Individuals arrive with (X, Y), and the goal is to estimate features of the conditional distribution of Y given X. Here

$$f(x, y) = f(y \mid x;\psi)f(x)$$

so if ψ appears only on the conditional distribution of Y given X, we base inference on the conditional density.

Location Model

Toy example: useful to fix ideas.

$Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f(y;\theta) = g(y-\theta)$, where g known.

Minimal sufficient statistic for θ is (in general) the order statistics $Y_{(1)} < \cdots < Y_{(n)}$, whose joint density is

$$f(y_{(1)}, \ldots, y_{(n)}; \theta) = n! \prod_{j=1}^{n} g(y_{(j)} - \theta), \quad y_{(1)} < \cdots < y_{(n)}.$$

MLE $\widehat{\theta}$ is unique solution to

$$0 = \sum_{j=1}^{n} \frac{\partial \log g(Y_{(j)} - \widehat{\theta})}{\partial \theta} = \sum_{j=1}^{n} \frac{\partial \log g(A_j)}{\partial \theta},$$

where $A_j = Y_{(j)} - \widehat{\theta}$.

Inference in Location Model

Configuration $A = (A_1, \ldots, A_n)$ is a function of $Y_{(1)}, \ldots, Y_{(n)}$ but its distribution is independent of θ. A is ancillary.

Have transformed

$$(Y_{(1)}, \ldots, Y_{(n)}) \quad \mapsto \quad (A_1, \ldots, A_n, \widehat{\theta}),$$

so conditional inference is based on distribution on $\widehat{\theta}$ given A.

Obviously $Z(\theta) = \widehat{\theta} - \theta$ is a pivot, so if we let $z_\alpha(a)$ denote the α quantile of the conditional distribution of $Z(\theta)$ given $A = a$, a $(1 - 2\alpha)$ conditional confidence interval for θ will have limits

$$\widehat{\theta} - z_{1-\alpha}(a), \quad \widehat{\theta} - z_\alpha(a)$$

that depend on the observed a.

Conditional Density of $\widehat{\theta}$

Conditional density turns out to be

$$f(\widehat{\theta} \mid a; \theta) = \frac{f(\widehat{\theta}, a; \theta)}{f(a)} = \frac{\prod_{j=1}^{n} g(a_j + \widehat{\theta} - \theta)}{\int_{-\infty}^{\infty} \prod_{j=1}^{n} g(a_j + u)\, du}.$$

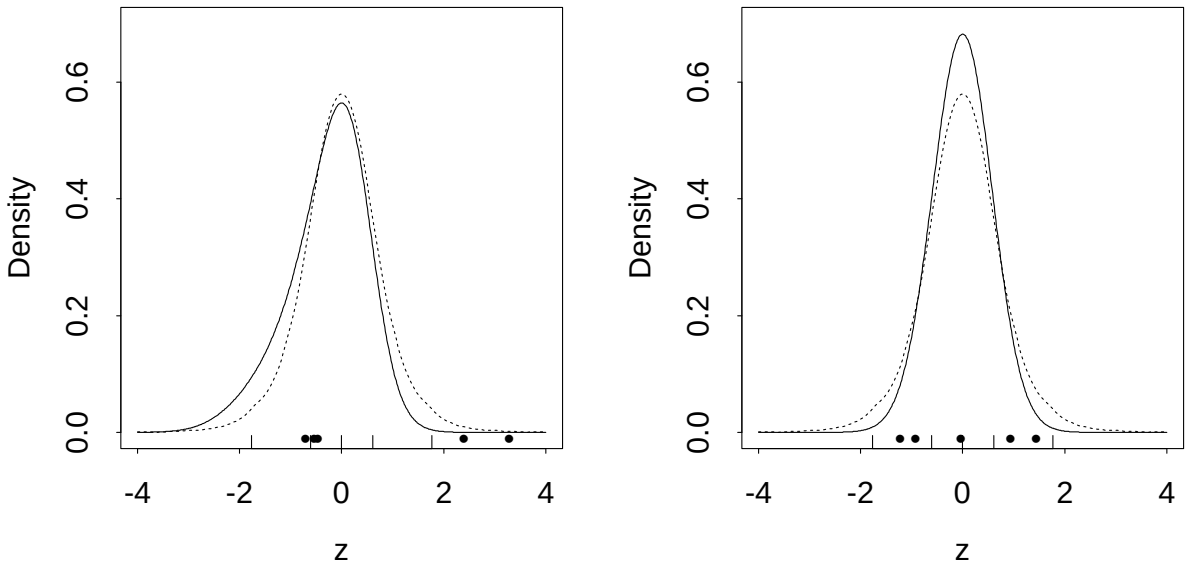
This contains all the information in the data concerning θ, appropriately conditioned.

Changing variables to $Z(\theta) = \widehat{\theta} - \theta$ and integrating gives

$$\mathrm{P}\{Z(\theta) \leq z \mid A = a\} = \frac{\int_{-\infty}^{z} \prod_{j=1}^{n} g(a_j + u)\, du}{\int_{-\infty}^{\infty} \prod_{j=1}^{n} g(a_j + u)\, du},$$

from which we obtain $z_\alpha(a)$ and $z_{1-\alpha}(a)$ by numerical integration.

Example: $n = 5$ observations from t_3



Unconditional 0.025,0 0.975 quantiles ± 1.61

Conditional quantiles $-2.11, 1.03$ (left), $-1.18, 1.13$ (right).

Re-expression of Conditional Density

Write log likelihood as $\ell(\theta; \widehat{\theta}, a)$, and then note that

$$f(\widehat{\theta} \mid a; \theta) = \frac{\exp\left\{\ell(\theta; \widehat{\theta}, a)\right\}}{\int_{-\infty}^{\infty} \exp\left\{\ell(v; \widehat{\theta}, a)\right\}\, dv}.$$

Laplace approximation to the integral gives

$$f(\widehat{\theta} \mid a; \theta) = (2\pi)^{-1/2} |J(\widehat{\theta}; \widehat{\theta}, a)|^{1/2} \exp\left\{\ell(\theta; \widehat{\theta}, a) - \ell(\widehat{\theta}; \widehat{\theta}, a)\right\} \left\{1 + O(n^{-1})\right\}.$$

where $J(\widehat{\theta}; \widehat{\theta}, a)$ is observed information, and renormalizing gives

$$f(\widehat{\theta} \mid a; \theta) = (2\pi)^{-1/2} \overline{c}(a) |J(\widehat{\theta}; \widehat{\theta}, a)|^{1/2} \exp\left\{\ell(\theta; \widehat{\theta}, a) - \ell(\widehat{\theta}; \widehat{\theta}, a)\right\},$$

which expresses the density of interest for inference only in terms of ℓ and related quantities.

Laplace Approximation

Consider

$$I_n = \int_{-\infty}^{\infty} e^{-nh(u)}\, du,$$

where $h(u)$ is a smooth convex function with minimum at $u = \tilde{u}$, where $dh(\tilde{u})/du = 0$ and $d^2h(\tilde{u})/du^2 > 0$.

Let $h_2 = d^2h(\tilde{u})/du^2$, $h_3 = d^3h(\tilde{u})/du^3$, and so forth.

Taylor series gives $h(u) \doteq h(\tilde{u}) + \frac{1}{2}h_2(u-\tilde{u})^2$, so

$$I_n \doteq e^{-nh(\tilde{u})} \int_{-\infty}^{\infty} e^{-nh_2(u-\tilde{u})^2/2}\, du = \left(\frac{2\pi}{nh_2}\right)^{1/2} e^{-nh(\tilde{u})}.$$

More detailed accounting gives

$$I_n = \left(\frac{2\pi}{nh_2}\right)^{1/2} e^{-nh(\tilde{u})} \times \left\{1 + n^{-1}\left(\frac{5h_3^2}{24h_2^3} - \frac{h_4}{8h_2^2}\right) + O\left(n^{-2}\right)\right\}.$$

Leading term is **Laplace approximation** to I_n; call it $\tilde{I}_n$.

Notes:

- Error is relative: $I_n/\tilde{I}_n = 1 + O(n^{-1})$.
- $\tilde{I}_n$ involves only h and its derivatives at $\tilde{u}$, so is easily obtained numerically.
- Asymptotic approximation — adding terms may make it worse.
- Limits not crucial (normal integral over ± 3 is ≈ 1).
- In multivariate case get

$$\tilde{I}_n = \left(\frac{2\pi}{n}\right)^{p/2} |h_2|^{-1/2} e^{-nh(\tilde{u})},$$

where $|h_2|$ is determinant of Hessian matrix of h at $\tilde{u}$.

Another Useful Integral Approximation

Consider

$$J_n(u_0) = \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{u_0} a(u) e^{-ng(u)} \left\{1 + O(n^{-1})\right\} du,$$

where u is scalar, $a(u) > 0$, and in addition to possessing the properties of $h(u)$ above, g is such that $g(\tilde{u}) = 0$.

Two changes of variable give

$$J_n(u_0) = \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{r_0^*} e^{-nr^{*2}/2} \left\{1 + O(n^{-1})\right\} dr^* = \Phi(n^{1/2} r_0^*) + O(n^{-1}),$$

where

$$r_0^* = r_0 + (r_0 n)^{-1} \log\left(\frac{v_0}{r_0}\right), \quad r_0 = \text{sign}(u_0 - \tilde{u})\{2g(u_0)\}^{1/2}, \quad v_0 = \frac{g'(u_0)}{a(u_0)}.$$

Application to $f(\widehat{\theta} \mid a; \theta)$

Aim to find conditional confidence interval (θ_-, θ_+) for θ, by solving

$$\mathrm{P}(\widehat{\theta} \leq t \mid A = a; \theta) = \alpha, 1 - \alpha,$$

with t replaced by observed value of $\widehat{\theta}$.

Previous approximation gives

$$\int_{-\infty}^{t} (2\pi)^{-1/2} \overline{c}(a) |J(\widehat{\theta}; \widehat{\theta}, a)|^{1/2} \exp\left\{\ell(\theta; \widehat{\theta}, a) - \ell(\widehat{\theta}; \widehat{\theta}, a)\right\} d\widehat{\theta} = \Phi\left\{r^*(\theta)\right\} + O(n^{-1})$$

where $r^*(\theta) = r(\theta) + r(\theta)^{-1} \log\{v(\theta)/r(\theta)\}$, with

$$r(\theta) = \mathrm{sign}(t - \theta) \left[2\left\{\ell(t; t, a) - \ell(\theta; t, a)\right\}\right]^{1/2}, \; v(\theta) = \frac{\ell_{;\widehat{\theta}}(t; t, a) - \ell_{;\widehat{\theta}}(\theta; t, a)}{|J(t; t, a)|^{1/2}}.$$

Sample Space Derivative

Note **sample space derivative** of ℓ,

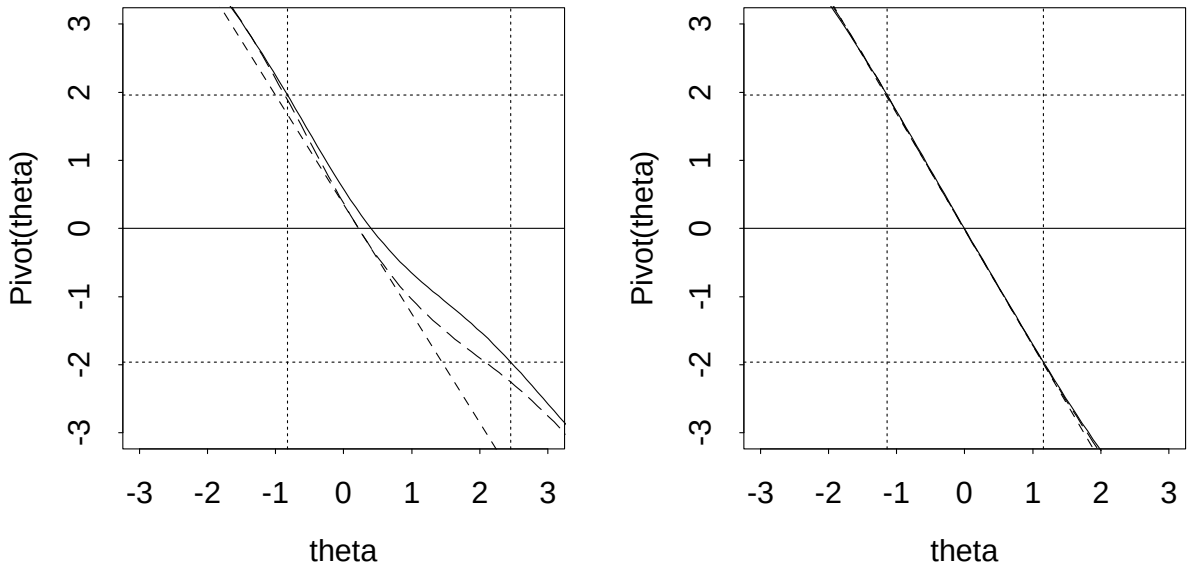
$$\ell_{;\widehat{\theta}}(\theta;\widehat{\theta},a) = \frac{\partial \ell(\theta;\widehat{\theta},a)}{\partial \widehat{\theta}}.$$

Must express ℓ as a function of $\widehat{\theta}$ and a so that it can be differentiated partially with respect to $\widehat{\theta}$, holding a fixed. Usually approximations needed. For the location model, $\ell_{;\widehat{\theta}}(t;t,a) = 0$ for any t, and

$$\ell_{;\widehat{\theta}}(\theta;\widehat{\theta},a) = -\frac{\partial \ell(\theta;\widehat{\theta},a)}{\partial \theta} = -\ell_{\theta;}(\theta;\widehat{\theta},a),$$

say. Thus $v(\theta) = \ell_{\theta;}(\theta;t,a)/|J(t;t,a)|^{1/2}$ is the score statistic.

Example: $n = 5$ observations from t_3



Pivots $r^*(\theta)$ (solid), $r(\theta)$ (large dash), $z(\theta) = J(\widehat{\theta})^{1/2}(\widehat{\theta} - \theta)$ (small dash) as functions of θ.

Contents

Marginal Inference

Recall factorization

$$f(y;\psi,\lambda) \propto f(t_1 \mid a;\psi) \times f(t_2 \mid t_1, a;\psi,\lambda),$$

where we focus on second term and aim to use as marginal likelihood.

Illustration: : Consider $y = X\beta + \sigma\varepsilon$, where the $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} N(0,1)$, β is $p \times 1$. Minimal sufficient statistic is

$$T_1 = S^2 = (n-p)^{-1}(y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta}), \quad T_2 = \widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y.$$

These are independent, with $(n-p)S^2/\sigma^2 \sim \chi^2_{n-p}$. If the interest parameter is σ^2 with nuisance parameter β , then

$$f(y;\sigma^2,\beta) = f(y \mid \widehat{\beta}, s^2) f(\widehat{\beta};\beta,\sigma^2) f(s^2;\sigma^2).$$

Base inference on

$$L_m(\sigma^2) = f(s^2;\sigma^2) = \left(\frac{n-p}{2\sigma^2}\right)^{(n-p)/2} \frac{(s^2)^{(n-p)/2-1}}{\Gamma\left(\frac{n-p}{2}\right)} \exp\left\{-\frac{(n-p)s^2}{2\sigma^2}\right\},$$

where $\sigma^2 > 0$, $s^2 > 0$. Marginal MLE is $\widehat{\sigma}^2_m = s^2$, with expected information $I_m = (n-p)/(2\sigma^4)$.

For previous example,

$$\begin{aligned} S^2 &= \frac{1}{n-p}(y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta}) \\ &= \frac{1}{m(k-1)} \sum_{i=1}^{m}\sum_{j=1}^{k}(y_{ij} - \overline{y}_i)^2 \sim \frac{\sigma^2}{m(k-1)}\chi^2_{m(k-1)}, \end{aligned}$$

and marginal likelihood automatically gives correct inference for σ^2.

Partial Likelihood

Survival time Y has hazard $\xi h(y)$ and density $\xi h(y)\exp\{-\xi H(y)\}$; $H(y) = \int_0^y h(u)\,du$ is the baseline cumulative hazard. Observe that

$$\int_u^\infty h(s)e^{-\gamma H(s)}\,ds = \gamma^{-1}e^{-\gamma H(u)}.$$

Suppose $n = 4$ continuous observations fall as

$$0 < Y_2 < Y_3^+ < Y_1 < Y_4.$$

Minimal sufficient statistic is the set of failure times and censoring indicators $(Y_1, 1), (Y_2, 1), (Y_3, 0), (Y_4, 1)$; 1–1 correspondence with order statistics $(Y_{(1)}, Y_{(2)}, Y_{(3)}, Y_{(4)})$ and inverse ranks $(2, 3^+, 1, 4)$.

Had Y_3 been observed, the joint density of the Y_j would be

$$\xi_2 h(y_2) e^{-\xi_2 H(y_2)} \times \xi_3 h(y_3) e^{-\xi_3 H(y_3)} \times \xi_1 h(y_1) e^{-\xi_1 H(y_1)} \times \xi_4 h(y_4) e^{-\xi_4 H(y_4)},$$

so the probability that $0 < Y_2 < Y_1 < Y_4$ with Y_3 censored somewhere to the right of Y_2 is

$$\xi_2 \xi_1 \xi_4 \int_0^\infty dy_2 \int_{y_2}^\infty dy_1 \int_{y_1}^\infty dy_4 \, h(y_2) h(y_1) h(y_4) e^{-(\xi_2+\xi_3) H(y_2) - \xi_1 H(y_1) - \xi_4 H(y_4)},$$

and this equals the **partial likelihood**

$$\frac{\xi_2}{\xi_1+\xi_2+\xi_3+\xi_4} \times \frac{\xi_1}{\xi_1+\xi_4} = \prod_j \frac{\xi_j}{\sum_{i \in \mathcal{R}_j} \xi_i},$$

where the product is over those j for which Y_j is uncensored and $\mathcal{R}_j$ denotes the risk set of individuals available to fail at the jth failure time.

Restricted Maximum Likelihood (REML)

Normal linear mixed model

$$y = X\beta + Zb + \varepsilon,$$

where $X_{n\times p}$ and $Z_{n\times q}$ are known, $\beta_{p\times 1}$ is unknown parameter vector, and the random vectors $b_{q\times 1}$ and $\varepsilon_{n\times 1}$ are independent with respective $N_q(0, \Omega_b)$ and $N_n(0, \sigma^2 I_n)$ distributions. Suppose $\sigma^2\Upsilon^{-1} = \sigma^2 I_n + Z\Omega_b Z^{\mathrm{T}}$ exists and that Υ depends on parameters ψ but not on β.

Profile log likelihood based on y is

$$\ell_{\mathrm{p}}(\psi, \sigma^2) \equiv \tfrac{1}{2}\log|\Upsilon| - \frac{1}{2\sigma^2}(y - X\widehat{\beta}_\psi)^{\mathrm{T}}\Upsilon(y - X\widehat{\beta}_\psi) - \frac{n}{2}\log\sigma^2,$$

which can be maximized directly to estimate ψ, σ^2.

REML

Construct marginal likelihood for σ^2 and ψ, eliminating β.

In general normal linear model $y \sim N_n(X\beta, \sigma^2\Upsilon^{-1})$, must find analogue of s^2.

Motivation: distribution of s^2 does not depend on $\mu_1, \ldots, \mu_m$, so seek quantity independent of β. Can take any $n - p$ linearly independent residuals from least squares regression of y on X, giving (eventually)

$$\ell_{\mathrm{m}}(\psi, \sigma^2) \equiv \tfrac{1}{2}\log|\Upsilon| - \tfrac{1}{2}\log|X^{\mathrm{T}}\Upsilon X| - \frac{1}{2\sigma^2}(y - X\widehat{\beta}_\psi)^{\mathrm{T}}\Upsilon(y - X\widehat{\beta}_\psi) - \frac{n-p}{2}\log\sigma^2,$$

where Υ and $\widehat{\beta}_\psi$ depend on ψ.

Note addition of $\frac{p}{2}\log\sigma^2 - \frac{1}{2}\log|X^{\mathrm{T}}\Upsilon X|$ to profile log likelihood.

Now maximize with respect to σ^2 and ψ.

Example: Short Time Series

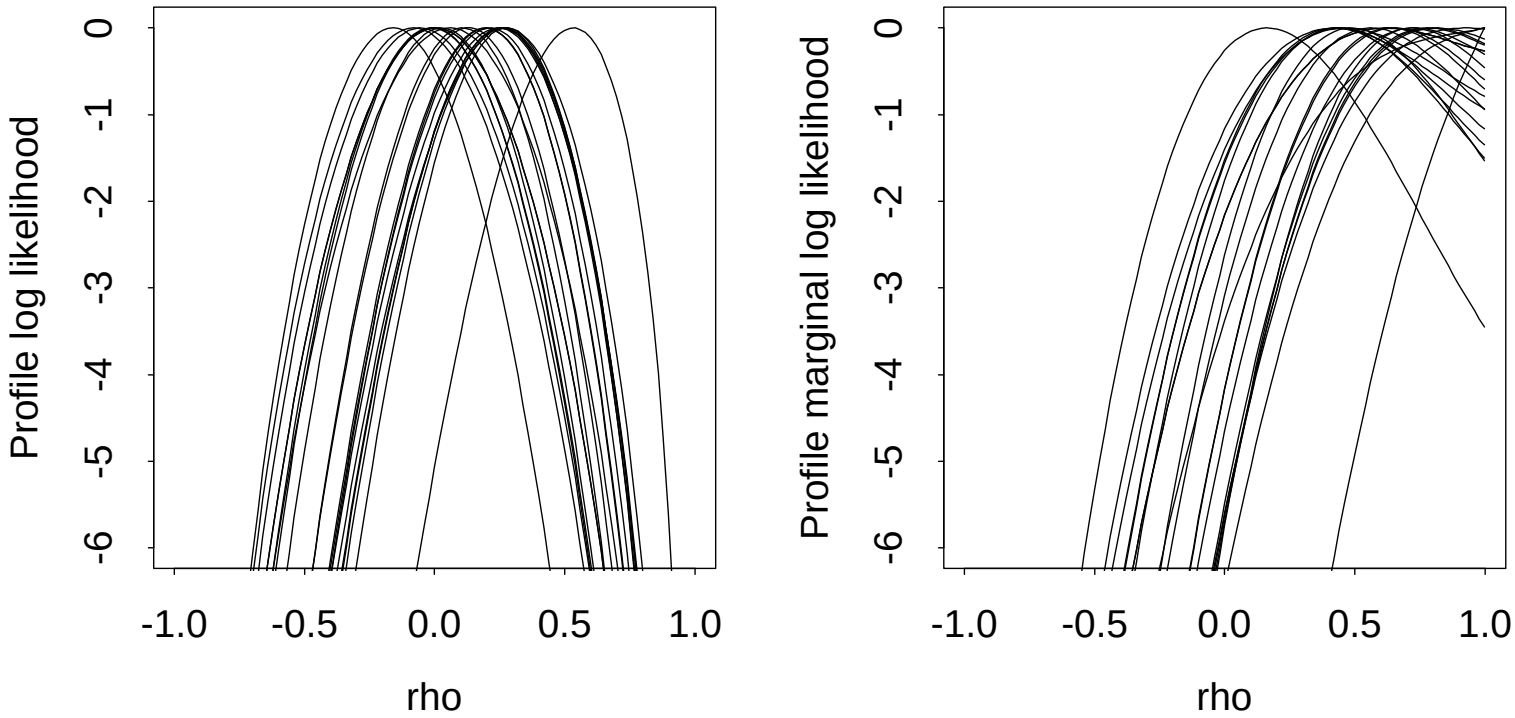
Consider m short time series of length k,

$$y_{ij} = \mu_i + \sigma\varepsilon_{ij}, \quad \varepsilon_{ij} \sim AR(1), i = 1, \ldots, m, j = 1, \ldots, k,$$

with correlation ρ. Here $\psi \equiv \rho$.

Numerical experiment: $m = 10$, $k = 5$, $\rho = 0.7$.

Example: Short Time Series



Regression-Scale Model

Linear model

$$Y = X\beta + \exp(\tau)\varepsilon,$$

where ε contains independent variables with density $e^{-d(u)}$. Log likelihood is

$$\ell(\beta, \tau; y) = -n\tau - \sum_{j=1}^{n} d\left\{e^{-\tau}(y_j - x_j^{\mathrm{T}}\beta)\right\}.$$

MLEs $\widehat{\beta}$ and $\widehat{\tau}$ determined by

$$\sum_{j=1}^{n} x_j d'\left(A_j\right) = 0, \quad n - \sum_{j=1}^{n} A_j d'\left(A_j\right) = 0,$$

where $A_j = e^{-\widehat{\tau}}(Y_j - x_j^{\mathrm{T}}\widehat{\beta})$, for $j = 1, \ldots, n$. Let

$A_0 = (A_1, \ldots, A_{n-p-1})$ be a non-degenerate subset of the A_j.

Construct pivots $U_1 = e^{-\widehat{\tau}}(\widehat{\beta}_1 - \beta_1)$, U_2 is vector with elements $\widehat{\tau} - \tau$ and $U_{-1} = e^{-\widehat{\tau}}(\widehat{\beta}_{-1} - \beta_{-1})$. Then compute

$$f\left(u_1, u_2 \mid a; \beta, \tau\right) = f(\widehat{\beta}, \widehat{\tau} \mid a; \beta, \tau) \left| \frac{\partial(\widehat{\beta}, \widehat{\tau})}{\partial(u_1, u_2)} \right|,$$

and base inference for β_1 on marginal distribution of U_1 given A.

Laplace-type approximation gives

$$\mathrm{P}\left(U_1 \leq u_1^0 \mid A = a\right) = \Phi\left\{r^*(\beta_1^0)\right\}\left\{1 + O(n^{-1})\right\},$$

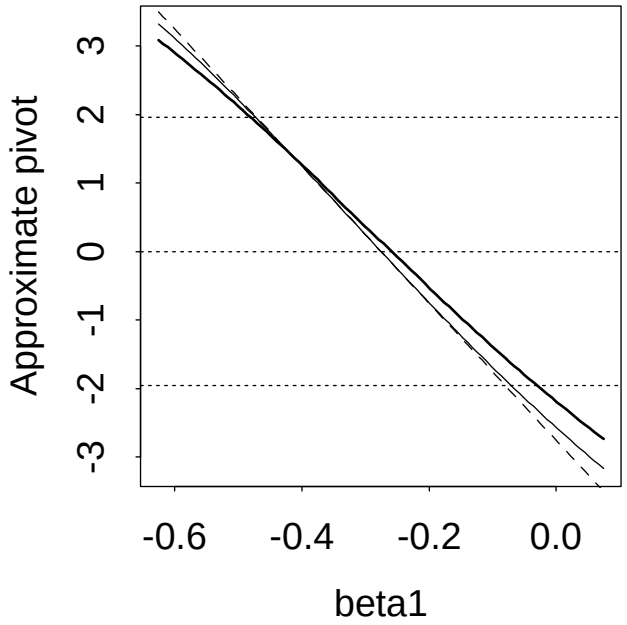
where $r^*(\beta_1^0) = r(\beta_1^0) + r(\beta_1^0)^{-1} \log\{v(\beta_1^0)/r(\beta_1^0)\}$, and $r(\beta_1^0)$ and $v(\beta_1^0)$ are expressed in terms of log likelihood, observed information, etc.

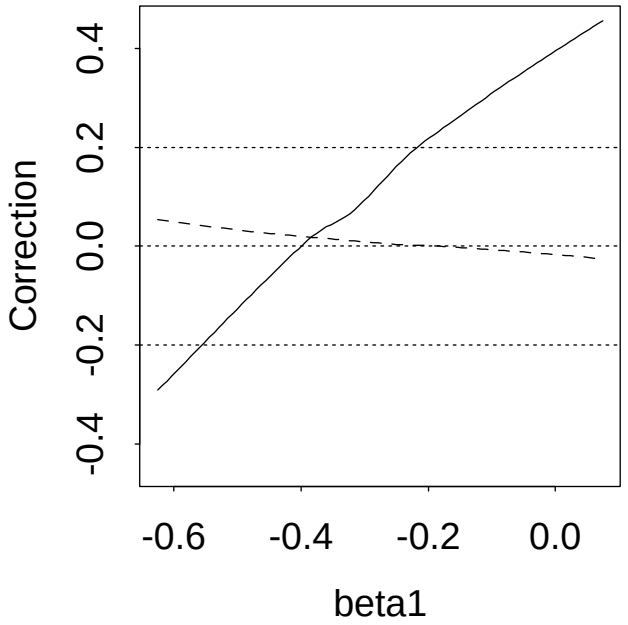
Then seek β_1^+, β_1^- that satisfy

$$\mathrm{P}\left\{ U_1 \leq e^{-\widehat{\tau}}(\widehat{\beta}_1 - \beta_1) \middle| a \right\} \doteq \Phi\left\{r^*(\beta_1)\right\} = \alpha, 1 - \alpha.$$

Nuclear Plant Data

$n = 32$, six covariates; t_5 errors.





Contents

Conditional Likelihood

Exponential family density is

$$f(t_1, t_2; \psi, \lambda) = \exp\left\{t_1\psi + t_2^{\mathrm{T}}\lambda - \kappa(\psi, \lambda)\right\} m(t_1, t_2),$$

where t_1 is scalar and t_2 has dimension $(p-1)\times 1$, and λ is treated as a $(p-1)\times 1$ nuisance parameter. Conditional density for T_1 given T_2 is

$$f(t_1 \mid t_2; \psi) = \frac{f(t_1, t_2; \psi, \lambda)}{f(t_2; \psi, \lambda)},$$

and we approximate both densities by saddlepoint approximation.

Saddlepoint Approximation

$\overline{X}$ is average of independent continuous scalar random variables $X_1, \ldots, X_n$ with cumulant-generating function $K(u) = \log M_X(u)$.
Saddlepoint approximation to the density of $\overline{X}$ at x is

$$f_{\overline{X}}(x) \doteq \left\{ \frac{n}{2\pi K''(\tilde{u})} \right\}^{1/2} \exp\left[n \left\{ K(\tilde{u}) - \tilde{u}x \right\} \right],$$

where $\tilde{u} = \tilde{u}(x)$ satisfies $K'(u) = x$.

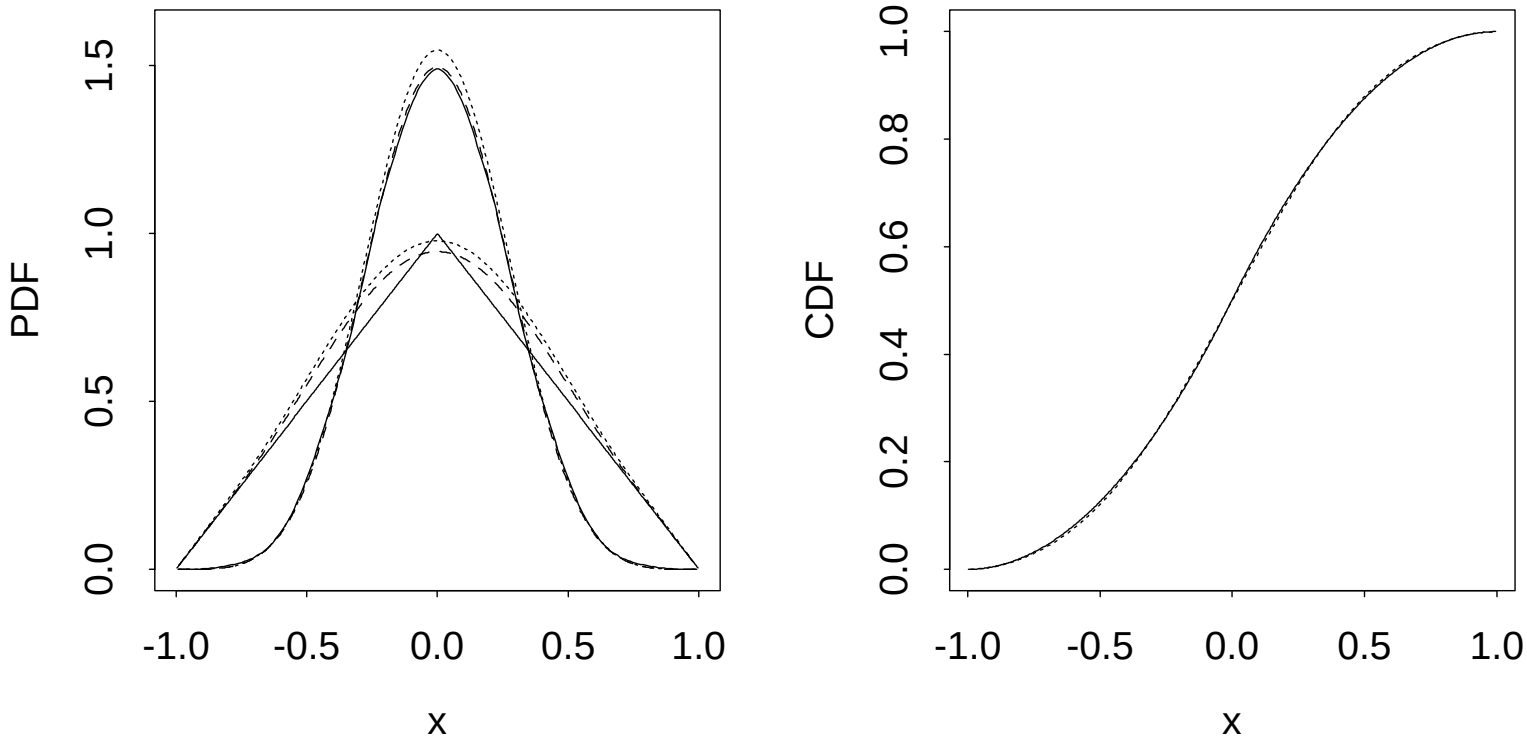
Distribution function approximation

$$\mathrm{P}\left(\overline{X} \leq x\right) \doteq \Phi\left\{ r^*(x) \right\},$$

where $r^*(x) = r(x) + r(x)^{-1} \log\{v(x)/r(x)\}$ and

$$r(x) = \mathrm{sign}(\tilde{u}) \left[2n \left\{ \tilde{u}x - K(\tilde{u}) \right\} \right]^{1/2}, \quad v(x) = \tilde{u} \left\{ nK''(\tilde{u}) \right\}^{1/2}.$$

$U(-1,1)$ Distribution, $n = 2, 5$



Comments

As with Laplace approximation, error is $O(n^{-1})$, and is relative. Approximations often extremely accurate in practice.

Can generalise to conditional densities and distribution functions:

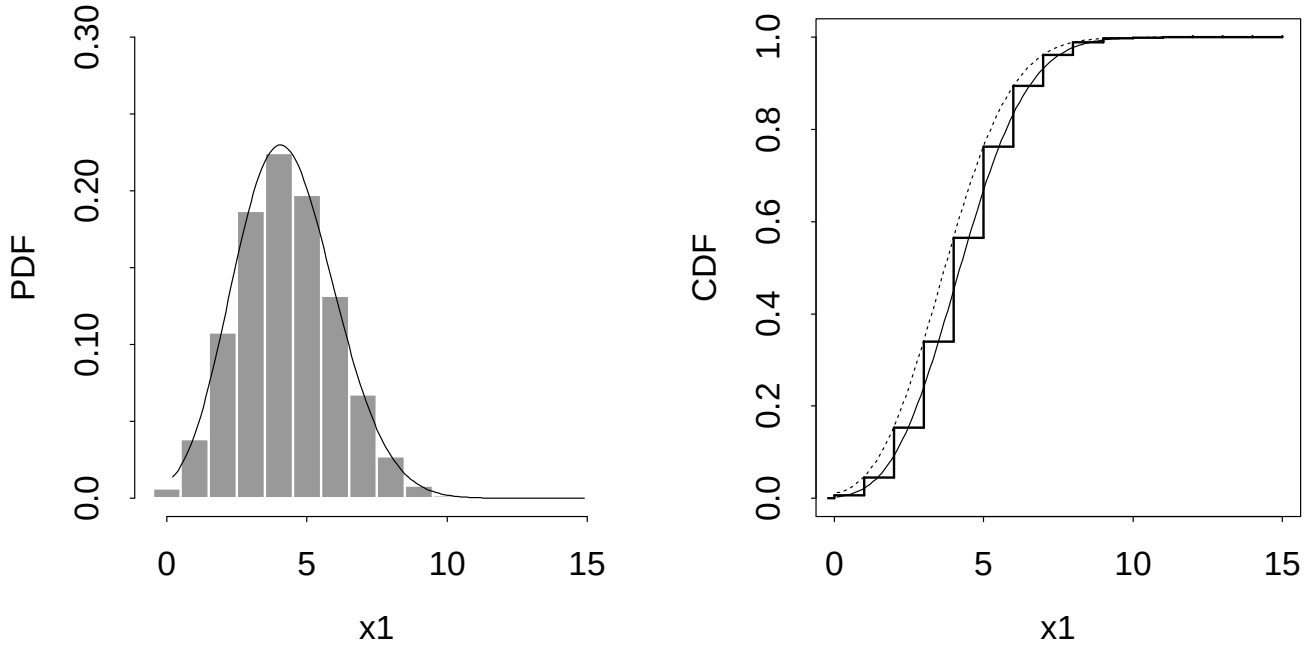
$$\mathrm{P}(\overline{X}_1 \leq x_1 \mid \overline{X}_2 = x_2) \doteq \Phi\left\{r^*(x_1)\right\},$$

where again $r^*(x_1) = r(x_1) + r(x_1)^{-1}\log\{v(x_1)/r(x_1)\}$, but now with

$$\begin{aligned} r(x_1) &= \mathrm{sign}(\tilde{u}_1)\left[2n\left\{K(\tilde{u}_0) - (0, x_2^{\mathrm{T}})\tilde{u}_0\right\} - n\left\{K(\tilde{u}) - \tilde{u}^{\mathrm{T}}x\right\}\right]^{1/2}, \\ v(x_1) &= \tilde{u}_1 n^{1/2}|K''(\tilde{u})|^{1/2}/|K''_{22}(\tilde{u}_0)|^{1/2}. \end{aligned}$$

Approximation to Binomial Distribution

V_1, V_2 are independent Poisson variables; consider V_1 given $V_1 + V_2 = 15$.



Exponential Family

Variables (T_1, T_2) with exponential family density

$$f(t_1, t_2; \psi, \lambda) = \exp\{t_1\psi + t_2^{\mathrm{T}}\lambda - \kappa(\psi, \lambda)\}\, m(t_1, t_2),$$

have joint cumulant-generating function

$$K(u) = K(u_1, u_2) = \kappa(\psi + u_1, \lambda + u_2) - \kappa(\psi, \lambda),$$

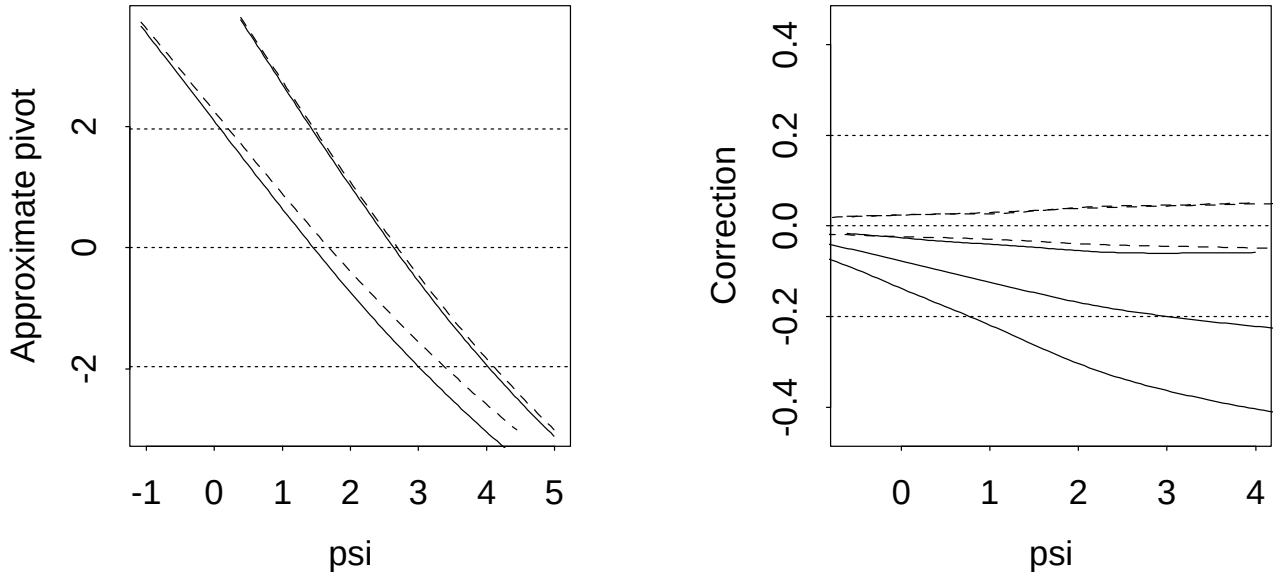
and T_2 has cumulant-generating function $K(0, u_2)$.

Saddlepoint approximation to $\mathrm{P}(T_1 \leq t_1 \mid T_2 = t_2; \psi)$ is of form $\Phi\{r^*(\psi)\}$, where $r^*(\psi) = r(\psi) + r(\psi)^{-1}\log\{r(\psi)/v(\psi)\}$, with

$$r(\psi) = \mathrm{sign}(\widehat{\psi}-\psi)[2\{\ell(\widehat{\theta})-\ell(\widehat{\theta}_\psi)\}]^{1/2},\ v(\psi) = (\widehat{\psi}-\psi)\left\{\frac{|J(\widehat{\theta})|}{|J_{\lambda\lambda}(\widehat{\theta}_\psi)|}\right\}^{1/2}.$$

Nodal Involvement Data

$n = 53$ binary observations with $p = 6$ covariates. Focus on coefficient of one covariate, when all others are included, and when they're not included.



Comments

Can extend ideas to many classes of parametric models, giving highly accurate inference for scalar (and vector) interest parameters.

Similar ideas give modified profile likelihoods (as for REML above).

Monte Carlo simulation plays a role also, particularly using Metropolis–Hastings algorithm for exact sampling conditional on sufficient statistics.

More details and references can be found in *Statistical Models*, Chapter 12 (Davison, 2003).

Books by Severini (2000), Barndorff-Nielsen and Cox (1989, 1994), Jensen (1995), review papers by Reid (1995, Statistical Science; 2003, Annals).