Likelihood and Asymptotic Theory for Statistical Inference

Nancy Reid

020 7679 1863
reid@utstat.utoronto.ca
n.reid@ucl.ac.uk

http://www.utstat.toronto.edu/reid/ltccF12.html

LTCC London Taught Course Centre
for PhD students in the mathematical sciences

Approximate Outline

1. Asymptotic theory for likelihood; likelihood root, maximum likelihood estimate, score function; pivotal quantities, exact and approximate ancillary; Laplace approximations for Bayesian inference

2. Higher order approximations for non-Bayesian inference; marginal, conditional and adjusted log-likelihoods; sample space differentiation and approximate ancillary; examples

3. Likelihood inference for complex data structure: time series, spatial models, space-time models, extremes; composite likelihood – definition, summary statistics, asymptotic theory; examples

4. Semi-parametric likelihoods for point process data; empirical likelihood; nonparametric models

`http://www.utstat.toronto.edu/reid/ltcc`

Assessment: Problems assigned weeks 1 to 4; due weeks 2 to 5; discussion on week 5.

Firefox File Edit View History Bookmarks Tools Window Help

Nancy Reid, Toronto

www.utstat.toronto.edu/reid/ltccF12.html

Most Visited · Department of S... · Text Forecasts –... · Welcome to Univ... · TD Canada Trust · Piece of Mind | ...

Text Forecasts – Environment ... · Manage Logbook – The Canadi... · Inbox (24,914) – nancy.reid1@g...

Nancy Reid

Information

....Curriculum Vitae

....Contact Details

Research

Research overview

....Recent papers

....Recent talks

Teaching

....Current

....Previous

Miscellaneous

....doing a project

....FAQ

LTCC Advanced Course: Likelihood Inference November/December, 2012

Course Outline

Running list of references and background reading

- Davison, A.C. (2003) *Statistical Models* (SM)
  - Introduction to likelihood -- Ch 4
  - Likelihood and stochastic models -- Ch 6
- Barndorff-Nielsen, O.E. and Cox, D.R. (1994) *Inference and Asymp*
  - Examples of likelihood functions -- Ch 2.2
- Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics* (CH)
  - Intro to likelihood -- Ch 2.1 (i),(ii)
- Cox, D.R. (2006). *Principles of Statistical Inference* (Cox)
  - Intro to likelihood -- Ch 2.1
- Brazzale, A.R., Davison, A.C. and Reid, N. (2007). *Applied Asympt*
  - Intro to pivots -- Ch. 2
- Reid, N. (2010). Likelihood Inference. in Wiley Interdisciplinary Revi
- Reid, N. (2000). Likelihood in *JASA*

Week 1

- Slides
- Handout (with exercises)

The likelihood function

- Parametric model: $f(y;\theta), \quad y \in \mathcal{Y}, \theta \in \Theta \subset \mathbb{R}^d$
- Likelihood function

$$L(\theta; y) = f(y;\theta), \text{ or } L(\theta; y) = c(y)f(y;\theta), \text{ or } L(\theta; y) \propto f(y;\theta)$$

- typically, $y = (y_1, \ldots, y_n) \qquad x_1, \ldots, x_n \qquad i = 1, \ldots, n$
- $f(y;\theta)$ or $f(y \mid x;\theta)$ is joint density
- under independence $L(\theta; y) \propto \prod f(y_i \mid x_i;\theta)$
- log-likelihood $\ell(\theta; y) = \log L(\theta; y) = \sum \log f(y_i \mid x_i;\theta)$
- θ could have dimension $d > n$ (e.g. genetics), or $d \uparrow n$, or
- θ could have infinite dimension e.g.
- regular model $d < n$ and d fixed as n increases

Examples

- $y_i \sim N(\mu, \sigma^2)$:

$$L(\theta; y) = \prod_{i=1}^{n} \sigma^{-n} \exp\{-\frac{1}{2\sigma^2}\Sigma(y_i - \mu)^2\}$$

- $E(y_i) = x_i^T\beta$:

$$L(\theta; y) = \prod_{i=1}^{n} \sigma^{-n} \exp\{-\frac{1}{2\sigma^2}\Sigma(y_i - x_i^T\beta)^2\}$$

- $E(y_i) = m(x_i), \quad m(x) = \Sigma_{j=1}^{J}\phi_j B_j(x)$:

$$L(\theta; y) = \prod_{i=1}^{n} \sigma^{-n} \exp\{-\frac{1}{2\sigma^2}\Sigma(y_i - \Sigma_{j=1}^{J}\phi_j B_j(x_i))^2\}$$

... examples

- $y_i = \mu + \rho(y_{i-1} - \mu) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$:

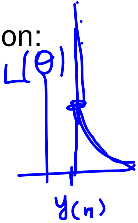
$$L(\theta; y) = \prod_{i=1}^{n} f(y_i \mid y_{i-1}; \theta) f_0(y_0; \theta)$$

- $y_1, \ldots, y_n$ are the times of jumps of a non-homogeneous Poisson process with rate function $\lambda(\cdot)$:

$$\ell\{\lambda(\cdot); y\} = \sum_{i=1}^{n} \log\{\lambda(y_i)\} - \int_0^\tau \lambda(u) du, \quad 0 < y_1 < \cdots < y_n < \tau$$

- $y_1, \ldots, y_n$ i.i.d. observations from a $U(0, \theta)$ distribution:

$$L(\theta; y) = \prod_{i=1}^{n} \theta^{-n}, \quad 0 < y_{(1)} < \cdots < y_{(n)} < \theta$$



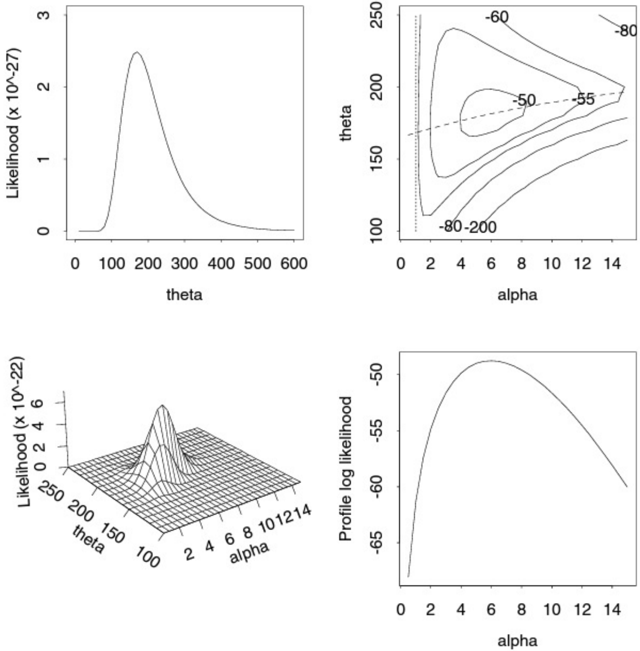


Figure 4.1 Likelihoods for the spring failure data at stress 950 N/mm^2. The upper left panel is the likelihood for the exponential model, and below it is a perspective plot of the likelihood for the Weibull model. The upper right panel shows contours of the log likelihood for the Weibull model; the exponential likelihood is obtained by setting $\alpha = 1$. that is, slicing L along the vertical dotted line. The lower right panel shows the profile log likelihood for α, which corresponds to the log likelihood values along the dashed line in the panel above, plotted against α.

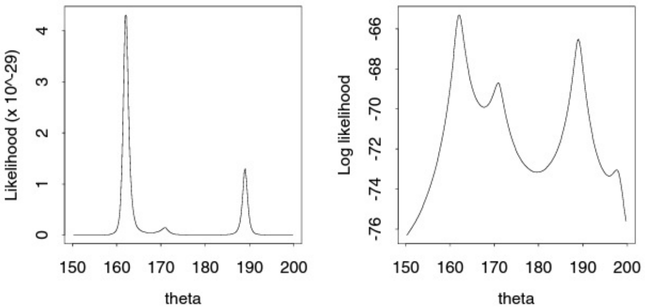


Figure 4.2 Cauchy likelihood and log likelihood for the spring failure data at stress 950N/mm^2.

SM p. 96

Data: times of failure of a spring under stress
225, 171, 198, 189, 189, 135, 162, 135, 117, 162

Principle

"The probability model and the choice of [parameter] serve to translate a subject-matter question into a mathematical and statistical one"

Cox, 2006, p.3

Non-computable likelihoods

- Ising model:

$$f(y;\theta) = \exp\Big(\sum_{(i,j)\in E} \theta_{ij} y_i y_j\Big)\frac{1}{Z(\theta)}$$

- $y_i = \pm 1$; binary property of a node i in a graph with n nodes
- θ_{ij} measures strength of interaction between nodes i and j
- E is the set of edges between nodes
- partition function $Z(\theta) = \sum_y \exp(\sum_{(i,j)\in E} \theta_{ij} y_i y_j)$

Ravikumar et al. (2010). High-dimensional Ising model selection... Ann. Statist. p.1287

... complicated likelihoods

- example: clustered binary data
- latent variable:
  $z_{ir} = x_{ir}'\beta + b_i + \epsilon_{ir}, \quad b_i \sim N(0, \sigma_b^2), \quad \epsilon_{ir} \sim N(0, 1)$
- $r = 1, \ldots, n_i$: observations in a cluster/family/school...
  $i = 1, \ldots, n$ clusters
- random effect b_i introduces correlation between observations in a cluster
- observations: $y_{ir} = 1$ if $z_{ir} > 0$, else 0
- $Pr(y_{ir} = 1 \mid b_i) = \Phi(x_{ir}'\beta + b_i) = p_i \quad \Phi(z) = \int^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$
- likelihood $\theta = (\beta, \sigma_b)$
  $L(\theta; y) = \prod_{i=1}^n \log \int_{-\infty}^{\infty} \prod_{r=1}^{n_i} p_i^{y_{ir}} (1 - p_i)^{1-y_{ir}} \phi(b_i, \sigma_b^2) db_i$
- more general: $z_{ir} = x_{ir}'\beta + w_{ir}'b_i + \epsilon_{ir}$

Widely used

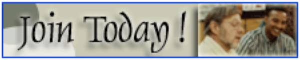
A Short Proof that Phylogenetic Tree Reconstruction by Maximum Li
http://csdl2.computer.org/persagen/DLAbsToc.jsp?resourcePath=/dl/trans/tb
Le Collège français d...
Mark Up Your Docu...
Canada411
Welcome to Universit...
TD Canada Trust
Tech-
JAMA -- Search Result
A Short Proof that Phylogenetic ...
IEEE computer society
Enter Sear
CS Searc
Home
Digital Library
Site Map
Store
Contact Us
Press Room
Shopping Cart
Help
Login
digital library
DIGITAL LIBRARY HOME
BROWSE BY TITLE
BROWSE BY SUBJECT
SEARCH
LIBRARY/INSTITUTION RESOURCES
RESOURCES
SUBSCRIPTION
ABOUT THE DIGITAL LIBRARY
Past Issues >> Table of Contents >> Abstract
IEEE/ACM TRANSACTIONS ON
COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
January-March 2006 (Vol. 3, No. 1) pp. 92-94
A Short Proof that Phylogenetic Tree Reconstructi
Maximum Likelihood Is Hard
Sebastien Roch
Full Article Text:
HTML
PDF
BUY ARTICLE $19

Estimating Genotypic Correlations and Their Standard Errors Using Multivariate Restricted Maxi

file:///Users/nancy/talks/waterloo/cropscience.html

Le Collège français d... Mark Up Your Docu... Canada411 Welcome to Universit... TD Canada Trust Tech–

CROP SCIENCE



HOME | HELP | FEEDBACK | SUBSCRIPTIONS | ARCHIVE | SEARCH | TABLE OF CONTENTS

QUICK SEARCH: [adva

Author: Keyword(s):

Go

Year: Vol: Page:

Published online 1 February 2006
Published in Crop Sci 46:642-654 (2006)
DOI: 10.2135/cropsci2005.0191

CROP BREEDING, GENETICS & CYTOLOGY

Estimating Genotypic Correlations and Their Standard Errors U Multivariate Restricted Maximum Likelihood Estimation with S Proc MIXED

James B. Holland*

USDA-ARS Plant Science Research Unit, Dep. of Crop Science, Box 7620, North Carolina State University, Raleigh, NC 2'

* Corresponding author (James_Holland@ncsu.edu)

Plant breeders traditionally have estimated genotypic and phenotypic correlations between traits using the moments on the basis of a multivariate analysis of variance (MANOVA). Drawbacks of using the methoc moments to estimate variance and covariance components include the possibility of obtaining estimates o

The Review of Financial Studies

Maximum Likelihood Estimation of Latent Affine Processes

David S. Bates
University of Iowa

This article develops a direct filtration-based maximum likelihood methodology for estimating the parameters and realizations of latent affine processes. Filtration is conducted in the transform space of characteristic functions, using a version of Bayes' rule for recursively updating the joint characteristic function of latent variables and the data conditional upon past data. An application to daily stock market returns over 1953–1996 reveals substantial divergences from estimates based on the Efficient Methods of Moments (EMM) methodology; in particular, more substantial and time-varying jump risk. The implications for pricing stock index options are examined.

IEEE Transactions on Information Theory

Single-Symbol Maximum Likelihood Decodable Linear STBCs

Md. Zafar Ali Khan, *Member, IEEE,* and B. Sundar Rajan, *Senior Member, IEEE*

***Abstract*—Space–time block codes (STBCs) from orthogonal designs (ODs) and coordinate interleaved orthogonal designs (CIOD) have been attracting wider attention due to their amenability for fast (single-symbol) maximum-likelihood (ML) decoding, and full-rate with full-rank over quasi-static fading channels. However, these codes are instances of single-symbol decodable codes and it is natural to ask, if there exist codes other than STBCs form ODs and CIODs that allow single-symbol decoding? In this paper, the above question is answered in the affirmative by characterizing all linear STBCs, that allow single-symbol ML decoding (not necessarily full-diversity) over quasi-static fading channels-calling them single-symbol decodable designs (SDD). The class SDD includes ODs and CIODs as proper subclasses. Further, among the SDD, a class of those that offer full-diversity, called Full-rank SDD (FSDD) are characterized and classified. We then concentrate on square designs and derive the maximal rate for square FSDDs using a constructional proof. It follows that 1) except for $N = 2$, square complex ODs are not maximal rate and 2) a rate one square FSDD exist only for two and four transmit antennas. For nonsquare designs, generalized coordinate-interleaved orthogonal designs (a superset of CIODs) are presented and analyzed. Finally, for rapid-fading channels an equivalent matrix channel representation is developed, which allows the results of quasi-static fading channels to be applied to rapid-fading channels. Using this representation we show that for rapid-fading channels the rate of single-symbol decodable STBCs are independent of the number of transmit antennas and inversely proportional to the block-length of the code. Significantly, the CIOD for two transmit antennas is the only STBC that is single-symbol decodable over both quasi-static and rapid-fading channels.**

Index Terms—**Diversity, fast ML decoding, multiple-input and** [illegible]

difference between coded modulation [used for single-input single-output (SISO), single-iutput multiple-output (SIMO)] and space–time codes is that in coded modulation the coding is in time only while in space–time codes the coding is in both space and time and hence the name. STC can be thought of as a signal design problem at the transmitter to realize the capacity benefits of MIMO systems [1], [2], though, several developments toward STC were presented in [3]–[7] which combine transmit and receive diversity, much prior to the results on capacity. Formally, a thorough treatment of STCs was first presented in [8] in the form of trellis codes [space–time trellis codes (STTC)] along with appropriate design and performance criteria.

The decoding complexity of STTC is exponential in bandwidth efficiency and required diversity order. Starting from Alamouti [12], several authors have studied space–time block codes (STBCs) obtained from orthogonal designs (ODs) and their variations that offer fast decoding (single-symbol decoding or double-symbol decoding) over quasi-static fading channels [9]–[27]. But the STBCs from ODs are a class of codes that are amenable to single-symbol decoding. Due to the importance of single-symbol decodable codes, need was felt for rigorous characterization of single-symbol decodable linear STBCs.

Following the spirit of [11], by a linear STBC,[1] we mean those covered by the following definition.

Molecular Biology and Evolution

Accuracy of Coalescent Likelihood Estimates: Do We Need More Sites, More Sequences, or More Loci?

Joseph Felsenstein

Department of Genome Sciences and Department of Biology, University of Washington, Seattle

A computer simulation study has been made of the accuracy of estimates of $\Theta = 4N_e\mu$ from a sample from a single isolated population of finite size. The accuracies turn out to be well predicted by a formula developed by Fu and Li, who used optimistic assumptions. Their formulas are restated in terms of accuracy, defined here as the reciprocal of the squared coefficient of variation. This should be proportional to sample size when the entities sampled provide independent information. Using these formulas for accuracy, the sampling strategy for estimation of Θ can be investigated. Two models for cost have been used, a cost-per-base model and a cost-per-read model. The former would lead us to prefer to have a very large number of loci, each one base long. The latter, which is more realistic, causes us to prefer to have one read per locus and an optimum sample size which declines as costs of sampling organisms increase. For realistic values, the optimum sample size is 8 or fewer individuals. This is quite close to the results obtained by Pluzhnikov and Donnelly for a cost-per-base model, evaluating other estimators of Θ. It can be understood by considering that the resources spent collecting larger samples prevent us from considering more loci. An examination of the efficiency of Watterson's estimator of Θ was also made, and it was found to be reasonably efficient when the number of mutants per generation in the sequence in the whole population is less than 2.5.

Introduction

The availability of molecular sequencing at prices that even population biologists can afford has brought into existence new methods of estimation of population parameters. Sequence samples from populations enable one to make an estimate of the coalescent tree of genes connecting these sequences. I have argued (Felsenstein 1992*a*) that these enable a substantial increase in the accuracy of estimation of population parameters like $\Theta = 4N_e\mu$, the product of effective population size, and the neutral mutation rate per site. (This is usually expressed as θ, the neutral mutation rate per locus but is perhaps better thought of in terms of the neutral mutation rate per site.)

Fu and Li (1993) analyzed my claim further. They developed some approximations to the accuracy of maximum likelihood estimation of Θ. I will show below that these are remarkably good approximations, better than one might have expected. My argument had assumed that an infinite number of sites could be examined and that the coalescent

Fu (1994) developed a method which makes a UPGMA estimate of the coalescent tree and constructs a best linear unbiased estimate conditional on that being the correct tree. In his simulations using the infinite-sites model, his BLUE method achieved variances nearly as low as the Fu and Li lower bound. It is not obvious from this whether it would perform as well with data from an actual finite-sites DNA sequence model of evolution, where the tree is bound to be harder to infer. Nevertheless, the good behavior of BLUE suggests that a full likelihood method based on summing over all coalescent trees might do almost as well as the Fu-Li lower bound.

In the present paper, the results of a computer simulation of coalescent likelihood estimates of Θ will be described, demonstrating that one of Fu and Li's optimistic approximation formulas does do a good job of calculating the accuracy of maximum likelihood estimates of Θ. Formulas based on it can then to be used to inves-

Physical Review D

PHYSICAL REVIEW D **73,** 015013 (2006)

Multidimensional mSUGRA likelihood maps

B. C. Allanach

DAMTP, CMS, Wilberforce Road, Cambridge, CB3 0WA, United Kingdom

C. G. Lester

Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, United Kingdom
(Received 18 November 2005; published 25 January 2006)

We calculate the likelihood map in the full 7-dimensional parameter space of the minimal
symmetric standard model assuming universal boundary conditions on the supersymmetry breakin
Simultaneous variations of m_0, A_0, $M_{1/2}$, $\tan\beta$, m_t, m_b and $\alpha_s(M_Z)$ are applied using a Marko
Monte Carlo algorithm. We use measurements of $b \to s\gamma$, $(g-2)_\mu$ and $\Omega_{DM}h^2$ in order to const
model. We present likelihood distributions for some of the sparticle masses, for the branching
$B_s^0 \to \mu^+\mu^-$ and for $m_{\tilde{\tau}} - m_{\chi_1^0}$. An upper limit of 2×10^{-8} on this branching ratio might be achi
the Tevatron, and would rule out 29% of the currently allowed likelihood. If one allows for non-t
neutralino components of dark matter, this fraction becomes 35%. The mass ordering allows the im
cascade decay $\tilde{q}_L \to \chi_2^0 \to \tilde{l}_R \to \chi_1^0$ with a likelihood of $24 \pm 4\%$. The stop-coannihilation re
highly disfavored, whereas the light Higgs region is marginally disfavored.

DOI: 10.1103/PhysRevD.73.015013 PACS numbers: 14.80.Ly,

US Patent Office

US007058142B2

(12) **United States Patent**
Coene et al.

(10) **Patent No.: US 7,058,14**
(45) **Date of Patent: Jun. 6**

(54) **GENERATION OF AMPLITUDE LEVELS FOR A PARTIAL RESPONSE MAXIMUM LIKELIHOOD (PRML) BIT DETECTOR**

(75) Inventors: **Willem M.J. Coene**, Eindhoven (NL); **Renatus J. Van Der Vleuten**, Eindhoven (NL)

(73) Assignee: **Koninklijke Philips Electronics N.V.**, Eindhoven (NL)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **10/403,544**

(22) Filed: **Mar. 31, 2003**

(56) **References Cited**

U.S. PATENT DOCUMENTS

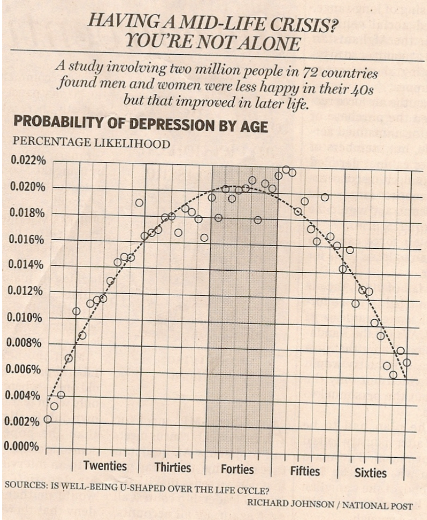
5,113,400	A	5/1992	Gould et al. ....
5,588,011	A	12/1996	Riggle ....
5,666,370	A	9/1997	Ganesan et al. ....
5,764,608	A	6/1998	Satomura ....
5,774,470	A	6/1998	Nishiya et al. ....
6,092,230	A	7/2000	Wood et al. ....
6,278,748	B1	8/2001	Fu et al. ....
6,288,992	B1	9/2001	Okumura et al. ....

Primary Examiner—Pankaj Kumar
(74) *Attorney, Agent, or Firm*—Michael E. Belk

(57) **ABSTRACT**

An apparatus for deriving amplitude values from information signal, which amplitude values can be reference levels for the states of a finite state machin

In the News



National Post, Toronto, Jan 30 2008

Likelihood inference

- direct use of likelihood function
- note that only relative values are well-defined

- define relative likelihood

$$RL(\theta) = \frac{L(\theta)}{\sup_{\theta'} L(\theta')} = \frac{L(\theta)}{L(\hat{\theta})}$$

Royall (199?)

$1 \geq RL(\theta) > \frac{1}{3}$, θ strongly supported,

$\frac{1}{3} \geq RL(\theta) > \frac{1}{10}$, θ supported,

$\frac{1}{10} \geq RL(\theta) > \frac{1}{100}$, θ weakly supported,

$\frac{1}{100} \geq RL(\theta) > \frac{1}{1000}$, θ poorly supported,

$\frac{1}{1000} \geq RL(\theta) > 0$, θ very poorly supported.

SM (4.11)

Derived quantities; $f(y;\theta)$

observed likelihood	$L(\theta; y) = c(y) f(y;\theta)$
log-likelihood	$\ell(\theta; y) = \log L(\theta; y) = \log f(y;\theta) + a(y)$
score	$U(\theta) = \partial \ell(\theta; y)/\partial\theta$
observed information	$j(\theta) = -\partial^2 \ell(\theta; y)/\partial\theta\partial\theta^T$
expected information	$i(\theta) = \mathrm{E}_\theta U(\theta) U(\theta)^T$ called $i_1(\theta)$ in CH

... derived quantities; $f(\underline{y}; \theta)$

observed likelihood	$L(\theta; y) \propto \prod_{i=1}^{n} f(y_i; \theta)$
log-likelihood	$\ell(\theta; y) = \sum_{i=1}^{n} \log f(y; \theta) + a(y)$
score	$U(\theta) = \partial \ell(\theta; y)/\partial\theta = O_p(\sqrt{n})$
maximum likelihood estimate	$\hat{\theta} = \hat{\theta}(y) = \arg\sup_\theta \ell(\theta; y)$
Fisher information	$j(\hat{\theta}) = -\partial^2 \ell(\hat{\theta}; y)/\partial\theta\partial\theta^T$
expected information	$i(\theta) = \mathrm{E}_\theta U(\theta) U(\theta)^T = O(n)$

Bartlett identities

$$E_\theta(j(\theta)) = i(\theta)$$

$$\| \qquad\qquad \|$$

$$E_\theta\left\{-\frac{\partial^2 \ell(\theta; y)}{\partial\theta\,\partial\theta^T}\right\} = E_\theta\{U(\theta)\,U(\theta)^T\}$$

2nd Bartlett ident.

Limiting distributions

- $U(\theta) = \sum_{i=1}^{n} U_i(\theta)$ $\quad U_i = \log f(y_i; \theta)$
- $E\{U(\theta)\} =$ θ $\quad y_1, \ldots, y_n$ iid
- $var\{U(\theta)\} =$ $i(\theta) = n i_1(\theta)$
- $U(\theta)/\sqrt{n} \xrightarrow{\mathcal{L}} N\{0, i_1(\theta)\}$ ← by CLT

... limiting distributions

- $U(\theta)/\sqrt{n} \xrightarrow{\mathcal{L}} N\{0, i_1(\theta)\}$
- $U(\hat{\theta}) = 0 = U(\theta) + (\hat{\theta} - \theta)U'(\theta) + R_n$
- $(\hat{\theta} - \theta) = \{U(\theta)/i(\theta)\}\{1 + o_p(1)\}$
- $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} N\{0, i_1^{-1}(\theta)\}$

$u = \ell'(\theta)$

$\ell'(\hat{\theta}) = 0$

$\hat{\theta} = \arg\max_{\theta} \ell(\theta)$

... limiting distributions

- $\sqrt{n}(\hat{\theta}) \xrightarrow{\mathcal{L}} N\{\theta, i_1^{-1}(\theta)\}$

$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} N(0, i_1^{-1}(\theta))$

- $\ell(\theta) = \ell(\hat{\theta}) + (\theta - \hat{\theta})\ell'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 \ell''(\hat{\theta}) + R_n$

- $2\{\ell(\hat{\theta}) - \ell(\theta)\} = (\hat{\theta} - \theta)^2 i(\theta)\{1 + o_p(1)\}$

- $2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{\mathcal{L}} \chi_d^2$

$\hat{\theta} - \theta \xrightarrow{\mathcal{L}} N(0, i^{-1}(\theta))$

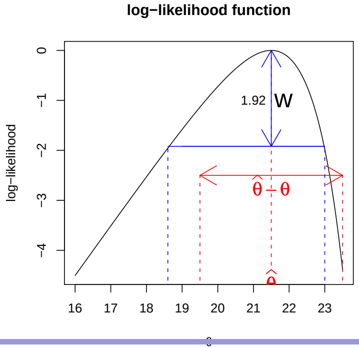
... limiting distributions

- $\sqrt{n}(\hat{\theta}) \xrightarrow{\mathcal{L}} N\{\theta, i_1^{-1}(\theta)\}$
- $\ell(\theta) = \ell(\hat{\theta}) + (\theta - \hat{\theta})\ell'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2\ell''(\hat{\theta}) + R_n$
- $2\{\ell(\hat{\theta}) - \ell(\theta)\} = (\hat{\theta} - \theta)^2 i(\theta)\{1 + o_p(1)\}$
- $2\{\ell(\hat{\theta}) - \ell(\theta)\} \xrightarrow{\mathcal{L}} \chi^2_d$

Inference from limiting distributions

- $\hat{\theta} \dot{\sim} N_d\{\theta, j^{-1}(\hat{\theta})\}$ $\quad j(\hat{\theta}) = -\ell''(\hat{\theta}; y)$
- "θ is estimated to be 21.5 (95% CI 19.5 – 23.5)"
- $_{19.5}21.5_{23.5}$ $\quad \hat{\theta} \pm 2\hat{\sigma}$

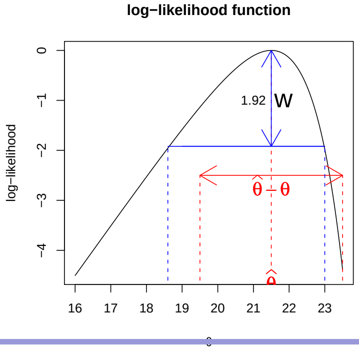
- $w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \dot{\sim} \chi^2_d$
- "likelihood based CI for θ with confidence level 95% is (18.6, 23.0)" $\quad _{18.6}21.5_{23.0}$



Inference from limiting distributions

- $\hat{\theta} \dot{\sim} N_d\{\theta, j^{-1}(\hat{\theta})\}$ $\quad j(\hat{\theta}) = -\ell''(\hat{\theta}; y)$
- "θ is estimated to be 21.5 (95% CI 19.5 – 23.5)"
- $_{19.5}21.5_{23.5}$ $\quad \hat{\theta} \pm 2\hat{\sigma}$

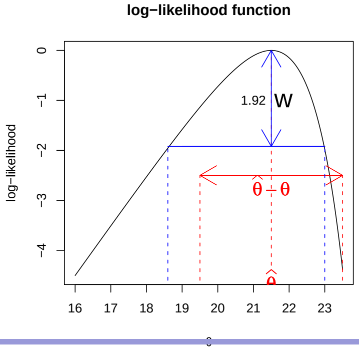
- $w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \dot{\sim} \chi^2_d$
- "likelihood based CI for θ with confidence level 95% is $(18.6, 23.0)$" $\quad _{18.6}21.5_{23.0}$



Inference from limiting distributions

- $\hat{\theta} \,\dot{\sim}\, N_d\{\theta, j^{-1}(\hat{\theta})\}$ $\qquad j(\hat{\theta}) = -\ell''(\hat{\theta}; y)$
- "θ is estimated to be 21.5 (95% CI 19.5 – 23.5)"
- ${}_{19.5}\,21.5\,{}_{23.5}$ $\qquad \hat{\theta} \pm 2\hat{\sigma}$

- $w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} \,\dot{\sim}\, \chi^2_d$
- "likelihood based CI for θ with confidence level 95% is $(18.6, 23.0)$" $\qquad {}_{18.6}\,21.5\,{}_{23.0}$



... inference from limiting distributions

- pivotal quantities and *p*-value functions; θ scalar
- $$r_u(\theta) = U(\theta) j^{-1/2}(\hat\theta) \overset{\cdot}{\sim} N(0,1)$$
- $$\Pr\{U(\theta) j^{-1/2}(\hat\theta) \le u(\theta) j^{-1/2}(\hat\theta)\} \doteq \Phi\{u(\theta) j^{-1/2}(\hat\theta)\}$$
- under sampling from the model $f(y;\theta) = f(y_1, \dots, y_n; \theta)$
- $$p_u(\theta) = \Phi\{u(\theta) j^{-1/2}(\hat\theta)\}$$

  p-value function (of θ, for fixed data)
- shorthand

$$= \quad \Phi\{r_u(\theta)\}, \text{and}$$
$$= \quad \Phi\{r_e(\theta)\},$$
$$= \quad \Phi\{r(\theta)\}$$

are all *p*-value functions for θ, based on limiting dist'ns

... inference from limiting distributions

- pivotal quantities and p-value functions; θ scalar
-
$$r_u(\theta) = U(\theta) j^{-1/2}(\hat\theta) \overset{.}{\sim} N(0,1)$$
-
$$\Pr\{U(\theta) j^{-1/2}(\hat\theta) \le u(\theta) j^{-1/2}(\hat\theta)\} \doteq \Phi\{u(\theta) j^{-1/2}(\hat\theta)\}$$
- under sampling from the model $f(y;\theta) = f(y_1, \ldots, y_n; \theta)$
-
$$p_u(\theta) = \Phi\{u(\theta) j^{-1/2}(\hat\theta)\}$$
p-value function (of θ, for fixed data)
- shorthand

$$= \quad \Phi\{r_u(\theta)\}, \text{and}$$
$$= \quad \Phi\{r_e(\theta)\},$$
$$= \quad \Phi\{r(\theta)\}$$

are all p-value functions for θ, based on limiting dist'ns

$$\frac{1}{\sqrt{n}} U(\theta) \xrightarrow{d} N(0, i(\theta))$$
$$U \cdot i^{-1/2} \overset{.}{\sim} N(0,1)$$

... inference from limiting distributions

- pivotal quantities and *p*-value functions; θ scalar
- $$r_u(\theta) = U(\theta)j^{-1/2}(\hat\theta) \mathrel{\dot\sim} N(0,1)$$
- $$\Pr\{U(\theta)j^{-1/2}(\hat\theta) \le u(\theta)j^{-1/2}(\hat\theta)\} \doteq \Phi\{u(\theta)j^{-1/2}(\hat\theta)\}$$
- under sampling from the model $f(y;\theta) = f(y_1,\ldots,y_n;\theta)$
- $$p_u(\theta) = \Phi\{u(\theta)j^{-1/2}(\hat\theta)\}$$

  p-value function (of θ, for fixed data)
- shorthand

$$\begin{aligned} &= \Phi\{r_u(\theta)\}, \text{and} \\ &= \Phi\{r_e(\theta)\}, \\ &= \Phi\{r(\theta)\} \end{aligned}$$

$(\hat\theta - \theta) j^{1/2}(\hat\theta) = r_e$

$\pm\sqrt{2\{\ell(\hat\theta) - \ell(\theta)\}} = r$

are all *p*-value functions for θ, based on limiting dist'ns

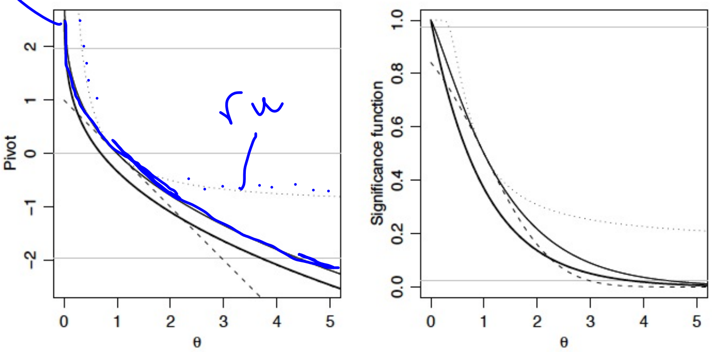


Figure 2.2: Approximate pivots and P-values based on an exponential sample of size $n = 1$. Left: likelihood root $r(\theta)$ (solid), score pivot $s(\theta)$ (dots), Wald pivot $t(\theta)$ (dashes), modified likelihood root $r^*(\theta)$ (heavy), and exact pivot $\theta \sum y_j$ (dot-dash). The modified likelihood root is indistinguishable from the exact pivot. The horizontal lines are at $0, \pm 1.96$. Right: corresponding significance functions, with horizontal lines at 0.025 and 0.975.

Example

- $f(y_i; \theta) = \theta^{y_i} e^{-\theta} / y_i!$
- $\ell(\theta) =$
- $\ell'(\theta) =$
- $\ell''(\theta) =$
- $r_e(\theta) = (s - n\theta)/\sqrt{s}$
- $\Pr(S \leq s) \neq 1 - \Pr(S \geq s)$
- upper and lower p-value functions: $\Pr(S < s)$, $\Pr(S \leq s)$
- mid p-value function: $\Pr(S < s) + 0.5\Pr(S = s)$

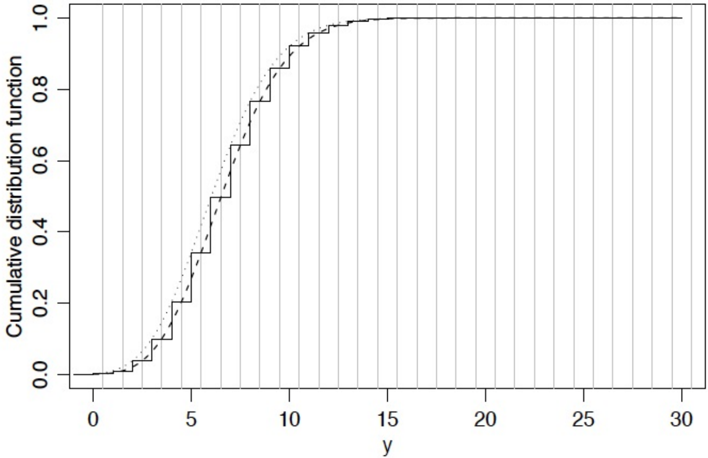


Figure 3.2: Cumulative distribution function for Poisson distribution with parameter 6.7 (solid), with approximations $\Phi\{r^*(y)\}$ (dashes) and $\Phi\{r^*(y + 1/2)\}$ (dots). The vertical lines are at $0.5, 1.5, 2.5, \ldots$

Aside

- for inference re θ, given y, plot $p(\theta)$ vs θ
- for p-value for $H_0 : \theta = \theta_0$, compute $p(\theta_0)$
- for checking whether, e.g. $\Phi\{r_e(\theta)\}$ is a good approximation,
  - compare $p(\theta) = \Phi\{r_e(\theta)\}$ to $p_{\text{exact}}(\theta)$, as a function of θ, fixed y
  - or compare $p(\theta_0)$ to $p_{\text{exact}}(\theta_0)$ as a function of y
- if $p_{\text{exact}}(\theta)$ not available, simulate

Nuisance parameters

- $\theta = (\psi, \lambda) = (\psi_1, \ldots, \psi_q, \lambda_1, \ldots, \lambda_{d-q})$
- $U(\theta) = \begin{pmatrix} U_\psi(\theta) \\ U_\lambda(\theta) \end{pmatrix}, \qquad U_\lambda(\psi, \hat\lambda_\psi) = 0$
- $i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix} \quad j(\theta) = \begin{pmatrix} j_{\psi\psi} & j_{\psi\lambda} \\ j_{\lambda\psi} & j_{\lambda\lambda} \end{pmatrix}$
- $i^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix} \quad j^{-1}(\theta) = \begin{pmatrix} j^{\psi\psi} & j^{\psi\lambda} \\ j^{\lambda\psi} & j^{\lambda\lambda} \end{pmatrix}.$
- $i^{\psi\psi}(\theta) = \{i_{\psi\psi}(\theta) - i_{\psi\lambda}(\theta) i_{\lambda\lambda}^{-1}(\theta) i_{\lambda\psi}(\theta)\}^{-1},$
- $\ell_{\mathrm{P}}(\psi) = \ell(\psi, \hat\lambda_\psi), \qquad j_{\mathrm{P}}(\psi) = -\ell''_{\mathrm{P}}(\psi)$

Nuisance parameters

- $\theta = (\psi, \lambda) = (\psi_1, \ldots, \psi_q, \lambda_1, \ldots, \lambda_{d-q})$
- $U(\theta) = \begin{pmatrix} U_\psi(\theta) \\ U_\lambda(\theta) \end{pmatrix}$, $\quad U_\lambda(\psi, \hat{\lambda}_\psi) = 0$

constrained mle of nuisance p.

- $i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix} \quad j(\theta) = \begin{pmatrix} j_{\psi\psi} & j_{\psi\lambda} \\ j_{\lambda\psi} & j_{\lambda\lambda} \end{pmatrix}$
- $i^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix} \quad j^{-1}(\theta) = \begin{pmatrix} j^{\psi\psi} & j^{\psi\lambda} \\ j^{\lambda\psi} & j^{\lambda\lambda} \end{pmatrix}.$
- $i^{\psi\psi}(\theta) = \{i_{\psi\psi}(\theta) - i_{\psi\lambda}(\theta) i_{\lambda\lambda}^{-1}(\theta) i_{\lambda\psi}(\theta)\}^{-1}$,
- $\ell_{\mathrm{P}}(\psi) = \ell(\psi, \hat{\lambda}_\psi)$, $\quad j_{\mathrm{P}}(\psi) = -\ell''_{\mathrm{P}}(\psi)$

Nuisance parameters

- $\theta = (\psi, \lambda) = (\psi_1, \dots, \psi_q, \lambda_1, \dots, \lambda_{d-q})$
- $U(\theta) = \begin{pmatrix} U_\psi(\theta) \\ U_\lambda(\theta) \end{pmatrix}, \qquad U_\lambda(\psi, \hat\lambda_\psi) = 0$
- $i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix} \quad j(\theta) = \begin{pmatrix} j_{\psi\psi} & j_{\psi\lambda} \\ j_{\lambda\psi} & j_{\lambda\lambda} \end{pmatrix}$
- $i^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix} \quad j^{-1}(\theta) = \begin{pmatrix} j^{\psi\psi} & j^{\psi\lambda} \\ j^{\lambda\psi} & j^{\lambda\lambda} \end{pmatrix}.$
- $i^{\psi\psi}(\theta) = \{i_{\psi\psi}(\theta) - i_{\psi\lambda}(\theta) i_{\lambda\lambda}^{-1}(\theta) i_{\lambda\psi}(\theta)\}^{-1},$
- $\ell_{\mathrm{P}}(\psi) = \ell(\psi, \hat\lambda_\psi), \qquad j_{\mathrm{P}}(\psi) = -\ell_{\mathrm{P}}''(\psi)$

Inference from limiting distributions, nuisance parameters

$(\hat\theta - \theta) \dot\sim N(0, i^{-1}(\theta))$

$$
\begin{aligned}
w_u(\psi) = U_\psi(\psi,\hat\lambda_\psi)^T\{i^{\psi\psi}(\psi,\hat\lambda_\psi)\}U_\psi(\psi,\hat\lambda_\psi) &\quad \dot\sim \quad \chi^2_q \\
w_e(\psi) = (\hat\psi - \psi)\{i^{\psi\psi}(\hat\psi,\hat\lambda)\}^{-1}(\hat\psi - \psi) &\quad \dot\sim \quad \chi^2_q \\
w(\psi) = 2\{\ell(\hat\psi,\hat\lambda) - \ell(\psi,\hat\lambda_\psi)\} = 2\{\ell_P(\hat\psi) - \ell_P(\psi)\} &\quad \dot\sim \quad \chi^2_q;
\end{aligned}
$$

Approximate Pivots

$$
\begin{aligned}
\text{score} \quad r_u(\psi) &= \ell_P'(\psi) j_P(\hat\psi)^{1/2} \dot\sim N(0,1), \quad (1/2 \to -1/2) \\
\text{mle} \quad r_e(\psi) &= (\hat\psi - \psi) j_P(\hat\psi)^{1/2} \dot\sim N(0,1), \\
\text{loglik} \quad r(\psi) &= \text{sign}(\hat\psi - \psi)\sqrt{2\{\ell_P(\hat\psi) - \ell_P(\psi)\}} \dot\sim N(0,1)
\end{aligned}
$$

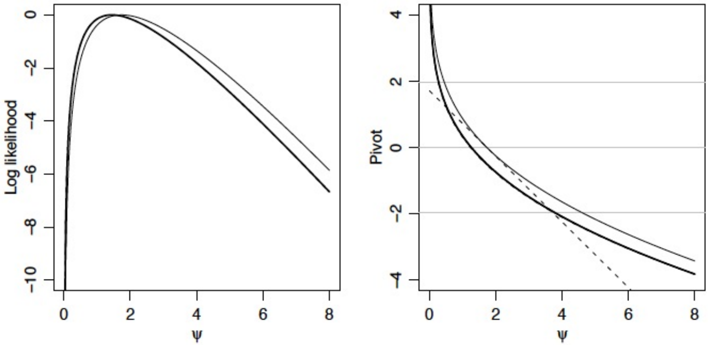


Figure 2.3: Inference for shape parameter ψ of gamma sample of size $n = 5$. Left: profile log likelihood $\ell_{\rm p}$ (solid) and the log likelihood from the conditional density of u given v (heavy). Right: likelihood root $r(\psi)$ (solid), Wald pivot $t(\psi)$ (dashes), modified likelihood root $r^*(\psi)$ (heavy), and exact pivot overlying $r^*(\psi)$. The horizontal lines are at $0, \pm 1.96$.

Properties of likelihood functions and likelihood inference

- the likelihood depends only on the minimal sufficient statistic
- recall:
  $L(\theta; y) = m_1(s; \theta) m_2(y) \iff s$ is minimal sufficient
- equivalently $\dfrac{L(\theta; y)}{L(\theta_0; y)}$ depends only on s
- " the likelihood map is sufficient" Fraser & Naderi, 2006; Barndorff-Nielsen, et al, 1976

  i.e $y \to \bar{L}_0(\cdot; y)$, or $y \to \bar{L}(\cdot; y)$ normed

... properties

- maximum likelihood estimates are equivariant: $\hat{h}(\theta) = h(\hat{\theta})$ for one-to-one $h(\cdot)$
- question: which of w_e, w_u, w are invariant under reparametrization of the full parameter: $\varphi(\theta)$?
- question: which of r_e, r_u, r are invariant under interest-respecting reparameterizations $(\psi, \lambda) \to \{\psi, \eta(\psi, \lambda)\}$?

- consistency of maximum likelihood estimate?
- equivalence of maximum likelihood estimate and root of score equation?

- observed vs. expected information

$u(\theta)\, i^{1/2}(\theta) \quad u(\theta)\, i^{-1/2}(\hat{\theta}) \quad u(\theta)\, j^{-1/2}(\theta) \quad u(\theta)\, j^{-1/2}(\hat{\theta})$

Asymptotics for Bayesian inference

- $$\pi(\theta \mid y) = \frac{\exp\{\ell(\theta; x)\}\pi(\theta)}{\int \exp\{\ell(\theta; x)\}\pi(\theta)d\theta}$$

$y = (y_1, \ldots, y_n)$ iid $f(y; \theta)$

- expand numerator and denominator about $\hat{\theta}$, assuming $\ell'(\hat{\theta}) = 0$

$\pi(\theta \mid y) \mathrel{\dot\sim} N(\hat{\theta}, j^{-1}(\hat{\theta}))$

- $\pi(\theta \mid y) \doteq N(\hat{\theta}, j^{-1}(\hat{\theta}))$

- expand denominator only about $\hat{\theta}$

- result

$$\pi(\theta \mid y) \doteq \frac{1}{(2\pi)^{d/2}} |j(\hat{\theta})|^{-1/2} \exp\{\ell(\hat{\theta}; y) - \ell(\theta; y)\} \frac{\pi(\theta)}{\pi(\hat{\theta})}$$

Asymptotics for Bayesian inference

- $\pi(\theta \mid y) = \dfrac{\exp\{\ell(\theta; x)\}\pi(\theta)}{\int \exp\{\ell(\theta; x)\}\pi(\theta)d\theta}$

- expand numerator and denominator about $\hat{\theta}$, assuming $\ell'(\hat{\theta}) = 0$

- $\pi(\theta \mid y) \doteq N(\hat{\theta}, j^{-1}(\hat{\theta}))$

- expand denominator only about $\hat{\theta}$

- result

$$\pi(\theta \mid y) \doteq \frac{1}{(2\pi)^{d/2}} |j(\hat{\theta})|^{-1/2} \exp\{\ell(\hat{\theta}; y) - \ell(\theta; y)\} \frac{\pi(\theta)}{\pi(\hat{\theta})}$$

Asymptotics for Bayesian inference

- $\pi(\theta \mid y) = \dfrac{\exp\{\ell(\theta; x)\}\pi(\theta)}{\int \exp\{\ell(\theta; x)\}\pi(\theta)d\theta}$

- expand numerator and denominator about $\hat{\theta}$, assuming $\ell'(\hat{\theta}) = 0$

- $\pi(\theta \mid y) \doteq N(\hat{\theta}, j^{-1}(\hat{\theta})$

- expand denominator only about $\hat{\theta}$

- result

$$\pi(\theta \mid y) \doteq \frac{1}{(2\pi)^{d/2}} |j(\hat{\theta})|^{-1/2} \exp\{\ell(\hat{\theta}; y) - \ell(\theta; y)\} \frac{\pi(\theta)}{\pi(\hat{\theta})}$$

$$\pi(\theta|y) = \frac{e^{\ell(\theta)}\pi(\theta)}{\int e^{\ell(\theta)}\pi(\theta)\,dy}$$

$$= e^{\ell(\theta)}\pi(\theta) \Big/ \int e^{\ell(\hat\theta) + \frac{1}{2}(\theta-\hat\theta)^2 \ell''(\hat\theta) + \cdots} (\pi(\hat\theta) + \cdots)$$

$$= \frac{e^{\ell(\theta)}\pi(\theta)}{e^{\ell(\hat\theta)}\pi(\hat\theta)} \cdot \frac{1}{\int e^{\frac{1}{2}(\theta-\hat\theta)^2 \ell''(\hat\theta)} \cdots d\theta} d\theta$$

$$= e^{\ell(\theta)-\ell(\hat\theta)} \left|j(\hat\theta)\right|^{1/2} \cdot \frac{\pi(\theta)}{\pi(\hat\theta)} \frac{1}{(\sqrt{2\pi})} d$$