# Integrated likelihoods in models with stratum nuisance parameters

Riccardo De Bin

Department of Mathematics - University of Oslo

based on a joint work with *Nicola Sartori* (University of Padova) and *Thomas A. Severini* (Northwestern University)

# Outline of the talk

- Introduction
  - global and partial interest
  - pseudo-likelihoods
  - integrated likelihoods for non-Bayesian inference
- Two-index asymptotics
- Integrated likelihoods in two-index asymptotics
- Examples
- Conclusions

# Introduction: global and partial interest

- Consider a parametric statistical model for data $y$, realization of a random variable $Y$, with parameter $\theta$ and model function $p_Y(y;\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^d$;

- the likelihood function is $L(\theta) = L(\theta; y) = \prod_{i=1}^{n} p_Y(y;\theta)$ and the log-likelihood is $\ell(\theta) = \log L(\theta)$;

- it is useful to distinguish between:
  - global interest about the whole $\theta$;
  - partial interest about a $p$-dimensional sub-parameter $\psi$, such that $\theta = (\psi, \lambda)$;
    - $\psi$ is the parameter of interest;
    - $\lambda$ is a nuisance parameter.

## Introduction: pseudo-likelihoods

Inference on $\psi$:

- when possible, based on conditional or marginal likelihood;
  - genuine likelihoods for the parameter of interest;
  - all standard likelihood properties satisfied;
  - often not available outside special families of distributions.

- otherwise, a standard approach consists of
  - finding $\hat{\lambda}_\psi$, the constrained maximum likelihood estimate of $\lambda$ for a given $\psi$;
  - maximizing the profile likelihood,

$$L_P(\psi) = L(\psi, \hat{\lambda}_\psi).$$

## Introduction: pseudo-likelihoods

As the profile-likelihood can perform poorly (more details later), alternative have been proposed:

- modifications of the profile likelihood: $L_M(\psi) = L_P(\psi)M(\psi)$, where $M(\psi)$ is a function useful to reduce the profile score bias (see, e.g., Severini, 2000, Chapter 9);
- integrated likelihoods: see Kalbfleisch & Sprott (1970); Berger et al. (1999); Severini (2000, 2007, 2010)

$$L_I(\psi) = \int_\Lambda p_Y(y; \psi, \lambda) g(\lambda; \psi) d\lambda;$$

  - $g(\lambda; \psi)$ is called weight function;
  - $\Lambda$ is the parametric space of $\lambda$.

## Introduction: modified profile likelihood

The modified profile likelihood (Barndorff-Nielsen, 1983) has form $L_{MP}(\psi) = L_P(\psi)M(\psi)$, with

$$M(\psi) = \left|\ell_{\lambda;\hat{\lambda}}(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda}, a)\right|^{-1} \left|-\ell_{\lambda\lambda}(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda}, a)\right|^{1/2},$$

where $\ell_{\lambda;\hat{\lambda}}$ denotes a mixed derivative and $a$ an ancillary statistic.

- based on the $p^*$ formula (Barndorff-Nielsen, 1980);
- can be derived as an approximation of a conditional or marginal likelihood, when one of the two exists;
- higher order properties.

## Introduction: integrated likelihoods for non-Bayesian inference

Integrated likelihoods:

- always available (unlike conditional or marginal likelihoods);
- based on averaging;
  - no maximization-related problems, see Berger et al. (1999);
- role of the weight function:
  - no necessarily a genuine density function;
  - difference with a random-effect modelization of $\lambda$;
- inference on $\psi$ proceeds by treating $L_I(\psi)$ as a genuine likelihood.

## Introduction: integrated likelihoods for non-Bayesian inference

To be useful for non-Bayesian inference on $\psi$, $L_I(\psi)$ should satisfy, at least approximately, the following properties:

- first two Bartlett identities:
  - $E[\ell_{I\psi}(\psi)] = 0$;
  - $E[\ell_{I\psi\psi} - \ell_{I\psi}(\psi)\ell_{I\psi}(\psi)^\top] = 0$;
- insensitivity to the choice of the weight function:
  - $L_I^{g_1}(\psi)/L_I^{g_2}(\psi) = 1$;
- invariance to interest-respecting reparametrization (see, e.g., Pace & Salvan, 1997, Section 4).

NB: here $\ell_{I\psi}(\psi) = \partial\ell_I(\psi)/\partial\psi$ and $\ell_{I\psi\psi}(\psi) = \partial^2\ell_I(\psi)/\partial\psi\partial\psi$

## Introduction: integrated likelihoods for non-Bayesian inference

In general, these properties are not satisfied (Severini, 2007);

- e.g., $E[\ell_{I\psi}(\psi)] = O(1)$;

Severini (2007) proposed the following procedure:

- model reparameterized by a nuisance parameter $\lambda$ unrelated to $\psi$:
  - weakly unrelated $\hat{\lambda}_\psi = \hat{\lambda} + O(n^{-1})$ for $\hat{\psi} = \psi + O(n^{-1/2})$;
  - strongly unrelated $\hat{\lambda}_\psi = \hat{\lambda} + O(n^{-1/2})$ for $\hat{\psi} = \psi + O(1)$.
- $g(\lambda; \psi)$ independent of $\psi$.

## Introduction: zero-score-expectation parameter

Severini (2007) suggests to use the zero-score-expectation parameter as a nuisance parameter unrelated to $\psi$,

$$\phi \equiv \phi(\psi, \lambda; \hat{\psi}),$$

defined by the solution of the equation

$$E_{(\hat{\psi},\phi)}[\ell_\lambda(\psi,\lambda)] \equiv E_{(\psi_0,\lambda_0)}[\ell_\lambda(\psi,\lambda)]\Big|_{(\psi_0,\lambda_0)=(\hat{\psi},\phi)} = 0;$$

- $\phi$ is data dependent (through $\hat{\psi}$);
- $\phi$ strongly unrelated to $\psi$.

## Introduction: zero-score-expectation parameter

In Severini (2007) it is shown that, by using $\phi$ and $g(\phi)$:

- $E[\ell_{I\psi}(\psi)] = 0 + O(n^{-1})$;
- $E[\ell_{I\psi\psi} - \ell_{I\psi}(\psi)\ell_{I\psi}(\psi)^\top] = 0 + O(n^{-1})$;
- $L_I^{g_1}(\psi)/L_I^{g_2}(\psi) = 1 + O(n^{-1/2})$;
- $L_I(\psi)$ parameterization invariant to order $O(n^{-1/2})$ for fixed $\psi$, to order $O(n^{-1})$ for $\psi = \hat{\psi} + O(n^{-1/2})$.

Moreover:

$$L_I(\psi) = cL_{MP}(\psi)\{1 + O(n^{-1/2})O(|\psi - \hat{\psi}|)\},$$

where $c$ is a generic constant.

## Two-index asymptotics: setting

Consider:

- stratified data $y = [y_i]$, $i = 1, \ldots, q$, where $y_i$ is a realization of a random variable $Y_i$ of dimension $m_i$;
- $Y_1, \ldots, Y_q$ be independent;
- a parametric statistical model in which:
  - $\psi$ is a common parameter of interest;
  - $\lambda = (\lambda_1, \ldots, \lambda_q)$ is a nuisance parameter;
  - the stratum-dependent $\lambda_i$ is called incidental nuisance parameter (Neyman & Scott, 1948).

The full likelihood is $L(\psi, \lambda) = \prod_{i=1}^{q} p_{Y_i}(y_i; \psi, \lambda_i)$, while the log-likelihood $\ell(\psi, \lambda) = \log L(\psi, \lambda) = \sum_{i=1}^{q} \ell^i(\psi, \lambda_i)$.

## Two-index asymptotics: setting

In the following, we consider an asymptotic scenario in which:

- both the within-stratum sample size ($m_i$) and the number of strata ($q$) go to infinity;
- two-index asymptotics (Barndorff-Nielsen, 1996).
- initially, we consider $m_i = m$;
- two-index asymptotics is more relevant to cases in which the number of strata is large relative to the total sample size.

NB: for simplicity of notation, hereafter we consider scalar parameters $\psi, \lambda_1, \ldots, \lambda_q$.

## Two-index asymptotics: gamma with common shape parameter

Let $Y_{ij}$, $j = 1, \ldots, m$, $i = 1, \ldots, q$, independent random variables with Gamma$(\psi, \lambda_i)$ distribution.

The standard approach, based on the profile likelihood

$$\ell_P(\psi) = \psi s + mq\psi \log m\psi - mq\psi - mq \log \Gamma(\psi),$$

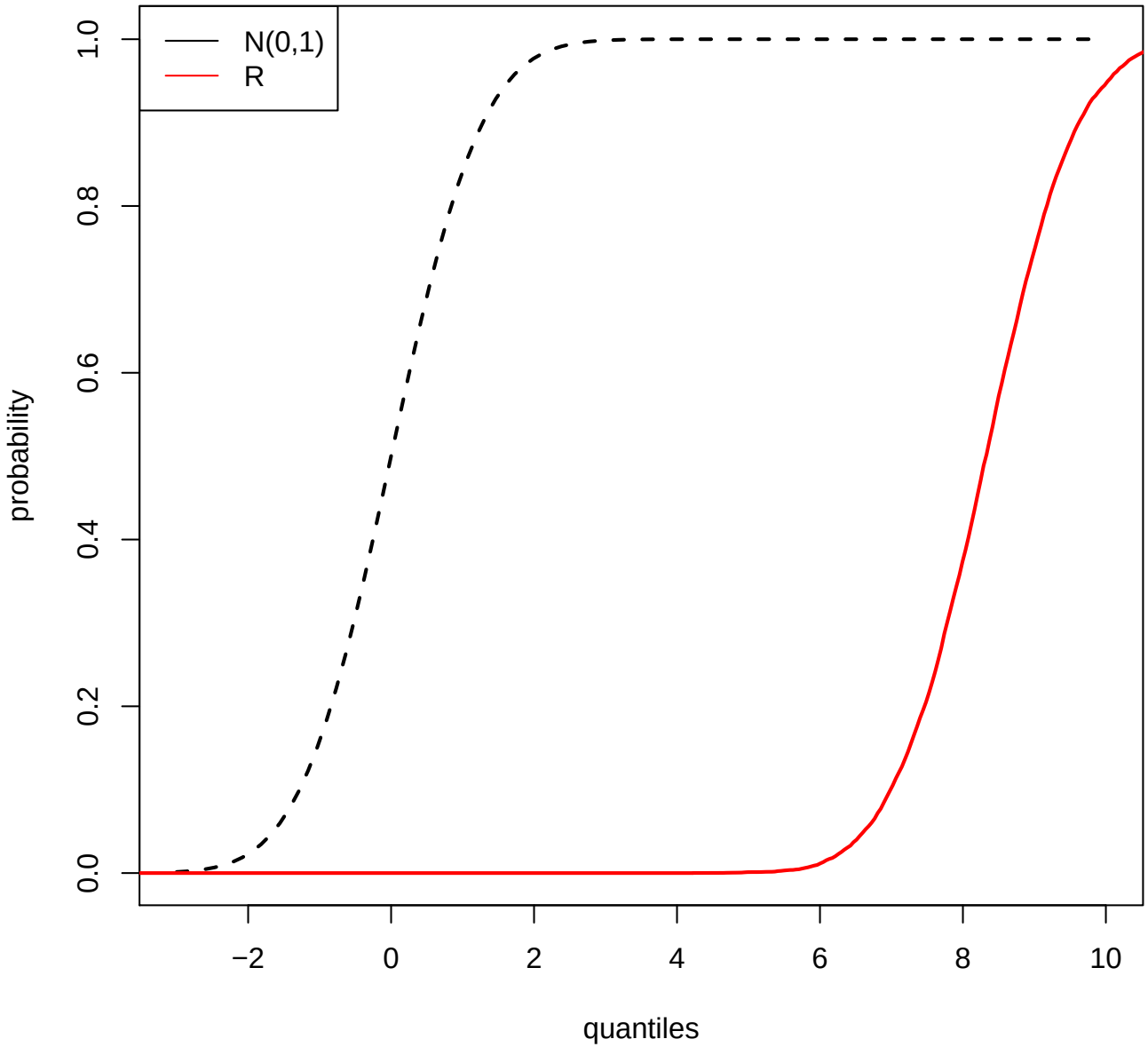
where $s = u - m \sum_{i=1}^{q} \log v_i$, with

- $u = \sum_{i=1}^{q} \sum_{j=1}^{m} \log y_{ij}$;
- $v_i = \sum_{j=1}^{m} y_{ij}$;

which can be used, e.g., to compute

$$R = \mathrm{sgn}(\hat{\psi} - \psi) \sqrt{2\{\ell_P(\hat{\psi}) - \ell_P(\psi)\}}$$

# Two-index asymptotics: gamma with common shape parameter




q=1000;
m=7;

B=9000.

## Two-index asymptotics: modified profile likelihood

We mentioned, as a suitable modification of the profile likelihood, Barndorff-Nielsen (1983)'s $L_{MP}$. With stratified nuisance parameters (Sartori, 2003),

- $E[\ell_{MP\psi}(\psi)] = E[\partial \ell_{MP}(\psi)/\partial\psi] = O(q/m)$;
- $\hat{\psi}_{MP} = \begin{cases} \psi + O(1/\sqrt{mq}) & \text{when } q/m^3 = o(1) \\ \psi + O(1/m^2) & \text{otherwise} \end{cases}$;
- $W_{MP} = 2\{\ell_{MP}(\hat{\psi}_{MP}) - \ell_{MP}(\psi)\} \sim \chi^2_1$ when $q/m^3 = o(1)$;
- $R_{MP} = \text{sgn}(\hat{\psi}_{MP} - \psi)\sqrt{W_{MP}(\psi)} \sim N(0;1)$ if $q/m^3 = o(1)$.

## Two-index asymptotics: modified profile likelihood

**Gamma with common shape parameter example**

In our example it is also possible to compute $\ell_{MP}(\psi)$

$$\ell_{MP}(\psi) = \psi s + q(m\psi - 0.5)\log m\psi - mq\psi - mq\,\log\Gamma(\psi)$$

and the related signed ratio statistic

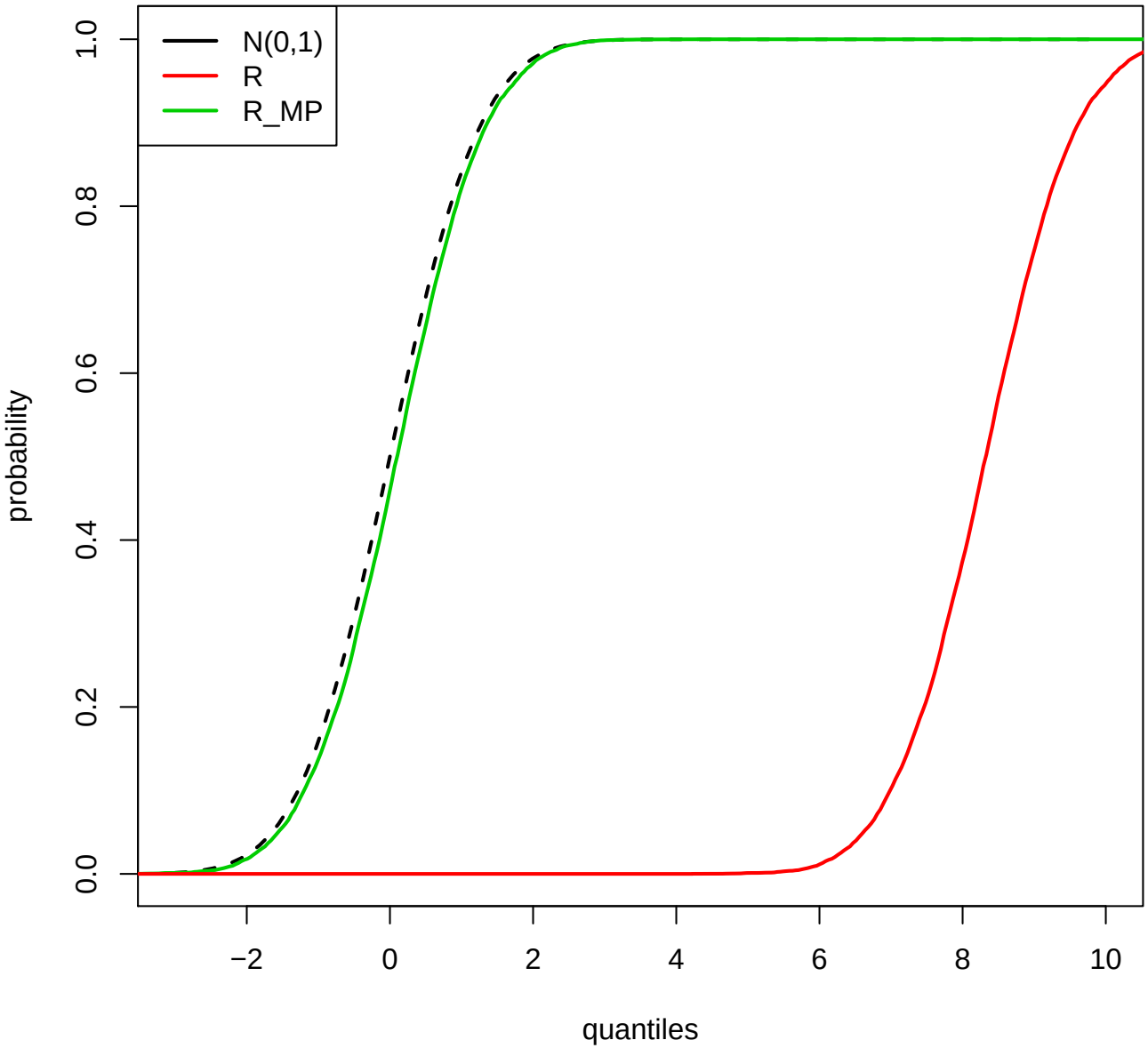
$$R_{MP} = \text{sgn}(\hat{\psi}_{MP} - \psi)\sqrt{\hat{\psi}_{MP}) - \ell_{MP}(\psi)}.$$

**NB:** note that, in the two-index-asymptotics framework,

$$R_{MP} = R^* + O_p(1/\sqrt{mq}),$$

where $R^*$ is the signed modified direct likelihood ratio statistics (Barndorff-Nielsen, 1986). For details, see Sartori et al. (1999).

# Two-index asymptotics: gamma with common shape parameter




q=1000;
m=7;

B=9000.

## Integrated likelihoods in two-index asymptotics: introduction

For a stratified model, an integrated likelihood has form

$$L_I(\psi) = \prod_{i=1}^{q} \int_{\Lambda} p_{Y_i}(y_i; \psi, \lambda_i) g(\lambda_i; \psi) d\lambda_i;$$

and, similarly, the integrated log-likelihood

$$\ell_I(\psi) = \sum_{i=1}^{q} \log \int_{\Lambda} p_{Y_i}(y_i; \psi, \lambda_i) g(\lambda_i; \psi) d\lambda_i;$$

Here we assume:

- each $\lambda_i$ has the same meaning;
- each $\lambda_i$ has the same parametric space $\Lambda$.

## Integrated likelihoods in two-index asymptotics: Laplace approximation

Analytic approximation of $\ell_I(\psi)$ based on stratum-specific Laplace approximations:

$$\ell_I(\psi) = \ell_P(\psi) + \sum_{i=1}^{q} \log g(\hat{\lambda}_{i\psi}; \psi) - \sum_{i=1}^{q} \log\{-\ell_{\lambda_i\lambda_i}(\psi, \hat{\lambda}_{i\psi})\}^{1/2} + O_p(\frac{q}{m});$$

Properties depend on the parameterization and on the choice of $g$:

- in general,
  $E[\ell_{I\psi}(\psi); \psi, \lambda] = O(q)$;
- with $\lambda_i$ unrelated to $\psi$ and $g$ independent of $\psi$,
  $E[\ell_{I\psi}(\psi); \psi, \lambda] = O(\frac{q}{m})$.

IMPORTANT: in the following, we only consider the latter case.

## Integrated likelihoods in two-index asymptotics: likelihood quantities

We can exploit similarities between $\ell_I(\psi)$ and $\ell_{MP}(\psi)$. In particular, it can be shown (De Bin et al., 2015) that:

$$\begin{aligned}
\ell_I(\psi) &= \ell_{MP}(\psi) + O_p(\sqrt{\frac{q}{m^3}}) + O_p(\frac{1}{m}); \\
\frac{1}{\sqrt{mq}}\ell_{I\psi}(\psi) &= \frac{1}{\sqrt{mq}}\ell_{MP\psi}(\psi) + O_p(\sqrt{\frac{q}{m^3}}) + O_p(\frac{1}{m}); \quad (1) \\
\frac{1}{mq}\ell_{I\psi\psi}(\psi) &= \frac{1}{mq}\ell_{MP\psi\psi}(\psi) + O_p(\frac{1}{m}). \quad (2)
\end{aligned}$$

These results are valid for $\psi = \hat{\psi}_{MP} + O_p(1/\sqrt{mq})$, which is true provided that $q/m^3 = o(1)$ (Sartori, 2003).

## Integrated likelihoods in two-index asymptotics: likelihood quantities

Let $\hat{\psi}_I$ denote the maximizer of $\ell_I(\psi)$. Given

$$\sqrt{mq}(\hat{\psi}_I - \psi) = \frac{\frac{1}{\sqrt{mq}}\ell_{I\psi}(\psi)}{-\frac{1}{mq}\ell_{I\psi\psi}(\psi)} + O_p(\frac{1}{\sqrt{mq}});$$

$$\sqrt{mq}(\hat{\psi}_{MP} - \psi) = \frac{\frac{1}{\sqrt{mq}}\ell_{MP\psi}(\psi)}{-\frac{1}{mq}\ell_{MP\psi\psi}(\psi)} + O_p(\frac{1}{\sqrt{mq}});$$

from equation (1) and (2) we obtain

$$\sqrt{mq}(\hat{\psi}_I - \psi) = \sqrt{mq}(\hat{\psi}_{MP} - \psi) + O_p(\sqrt{\frac{q}{m^3}}) + O_p(\frac{1}{m}).$$

From this relationship, given the asymptotic properties of $\hat{\psi}_{MP}$,

- $\hat{\psi}_I = \psi + O_p(1/\sqrt{mq})$ when $q/m^3 = o(1)$;
- $\hat{\psi}_I = \psi + O_p(1/m^2)$ otherwise.

## Integrated likelihoods in two-index asymptotics: likelihood quantities

Moreover, when $q/m^3 = o(1)$, $-\ell_{I\psi\psi}(\hat{\psi}_I)^{1/2}(\hat{\psi}_I - \psi)$ is asymptotically normally distributed with error

$$O_p(\sqrt{\frac{q}{m^3}}) + O_p(\frac{1}{m}) + O_p(\frac{1}{\sqrt{mq}}).$$

From this result (De Bin et al., 2015)

- $W_I(\psi) = 2\{\ell_I(\hat{\psi}_I) - \ell_I(\psi)\}$ is asymptotically distributed as a $\chi^2_1$ with the same error;
- $R_I(\psi) = \text{sgn}(\hat{\psi}_I - \psi)\sqrt{W_I}$ is asymptotically distributed as a $N(0,1)$ with the same error.

## Integrated likelihoods in two-index asymptotics: alternative

Alternatively, we could evaluate the asymptotic procedures under the marginal model

$$p^{\ddagger}(y;\psi) = \prod_{i=1}^{q} \int_{\Lambda} p_i(y_i;\psi,\lambda_i)\pi(\lambda_i;\psi)d\lambda_i;$$

where we assume that $\lambda_1,\ldots,\lambda_q$ are i.i.d. r.v. with an (unknown) density function $\pi(\cdot;\psi)$ (Kiefer & Wolfowitz, 1956; Strasser, 1996).

Using Laplace approximations

$$\ell_I(\psi) = \ell_P(\psi) + \sum_{i=1}^{q} \log g(\hat{\lambda}_{i\psi};\psi) - \sum_{i=1}^{q} \log|-\ell_{\lambda_i\lambda_i}(\psi,\hat{\lambda}_{i\psi})|^{1/2} + O_p(\frac{q}{m})$$

and

$$\ell^{\ddagger}(\psi) = \ell_P(\psi) + \sum_{i=1}^{q} \log \pi(\hat{\lambda}_{i\psi};\psi) - \sum_{i=1}^{q} \log|-\ell_{\lambda_i\lambda_i}(\psi,\hat{\lambda}_{i\psi})|^{1/2} + O_p(\frac{q}{m}).$$

## Integrated likelihoods in two-index asymptotics: alternative

Therefore

$$\ell_I(\psi) = \ell^{\ddagger}(\psi) + \sum_{i=1}^{q} \log \frac{g(\hat{\lambda}_{i\psi};\psi)}{\pi(\hat{\lambda}_{i\psi};\psi)} + O\left(\frac{q}{m}\right)$$

and in term of score

$$\ell_{I\psi}(\psi) = \ell^{\ddagger}_{\psi}(\psi) + \sum_{i=1}^{q} \underbrace{\frac{\partial}{\partial\psi} \log \frac{g(\hat{\lambda}_{i\psi};\psi)}{\pi(\hat{\lambda}_{i\psi};\psi)}}_{D_i(\psi)} + O\left(\frac{q}{m}\right);$$

where $D_i(\psi)$:

- in general is $O_p(1)$ with mean $O(1)$;
- with $\lambda_i$ and $\psi$ orthogonal, if $g(\lambda_i;\psi)$ is independent of $\psi$, it has mean $O(1/m)$.

## Integrated likelihoods in two-index asymptotics: alternative

Denoting with $\hat{\psi}^{\ddagger}$ the maximizer of $l^{\ddagger}(\psi)$,

$$\sqrt{mq}(\hat{\psi}_I - \psi) = \sqrt{mq}(\hat{\psi}^{\ddagger} - \psi) + \bar{D}\sqrt{mq} + O_p\left(\frac{1}{mq}\right) + O_p\left(\frac{\sqrt{q}}{m^{\frac{3}{2}}}\right),$$

where $\bar{D} = (i^{\ddagger})^{-1} \sum_{i=1}^{q} D_i(\psi)$, with $i^{\ddagger}$ denoting the expected information in the marginal model.

Provided that $q/m^3 = o(1)$, the term $(\hat{\psi}_I - \hat{\psi}^{\ddagger})$

- is, in general, $O_p(1/m)$;
- with $\lambda_i$ and $\psi$ orthogonal, if $g(\lambda_i; \psi)$ is independent of $\psi$, it is $O_p(\max\{1/m^2, 1/\sqrt{m^2 q}\})$.

## Integrated likelihoods in two-index asymptotics: alternative

Moreover,

$$\ell_I(\bar{\psi})-\ell_I(\psi) = \{\ell^{\ddagger}(\hat{\psi}^{\ddagger})-\ell^{\ddagger}(\psi)\}+\{\ell^{\ddagger}(\hat{\psi}_I)-\ell^{\ddagger}(\hat{\psi}^{\ddagger})\}+\{H(\hat{\psi}_I)-H(\psi)\}+\{B(\hat{\psi}_I)-B(\psi)\},$$

where $H(\psi) = \sum_{i=1}^{q} H_i(\psi)$,

$$H_i(\psi) = \log \frac{g(\hat{\lambda}_{i\psi};\psi)}{\pi(\hat{\lambda}_{i\psi};\psi)}$$

and $B(\psi)$ is a remainder term of order $O_p(q/m)$.
We can show that

- $\{\ell^{\ddagger}(\hat{\psi}_I) - \ell^{\ddagger}(\hat{\psi}^{\ddagger})\} = O_p(1/m)$;
- $\{H(\hat{\psi}_I) - H(\psi)\} = O_p(1/\sqrt{m})$;
- $\{B(\hat{\psi}_I) - B(\psi)\} = O_p(\sqrt{q/m^3})$.

## Integrated likelihoods in two-index asymptotics: alternative

Hence,

$$R_I = R^{\ddagger} + O_p(1/\sqrt{m}) + O_p(\sqrt{q/m^3}),$$

and the error in normal approximation to the asymptotic distribution of $R_I$ is

$$O_p(1/\sqrt{mq}) + O_p(1/\sqrt{m}) + O_p(\sqrt{q/m^3}).$$

- $O_p(1/\sqrt{mq})$ is based on the overall sample size $mq$;
- $O_p(1/\sqrt{m})$ reflects the effect of the weight function;
- $O_p(\sqrt{q/m^3})$ is due to the error in the Laplace approximation.

## Examples: Gamma common shape parameter

Back to our example, using

- the zeta-score-expectation parameter: $\phi_i = \frac{\hat{\psi}\lambda_i}{\psi}$;
- a weight function independent of $\psi$: $g(\phi_i) = 1$;

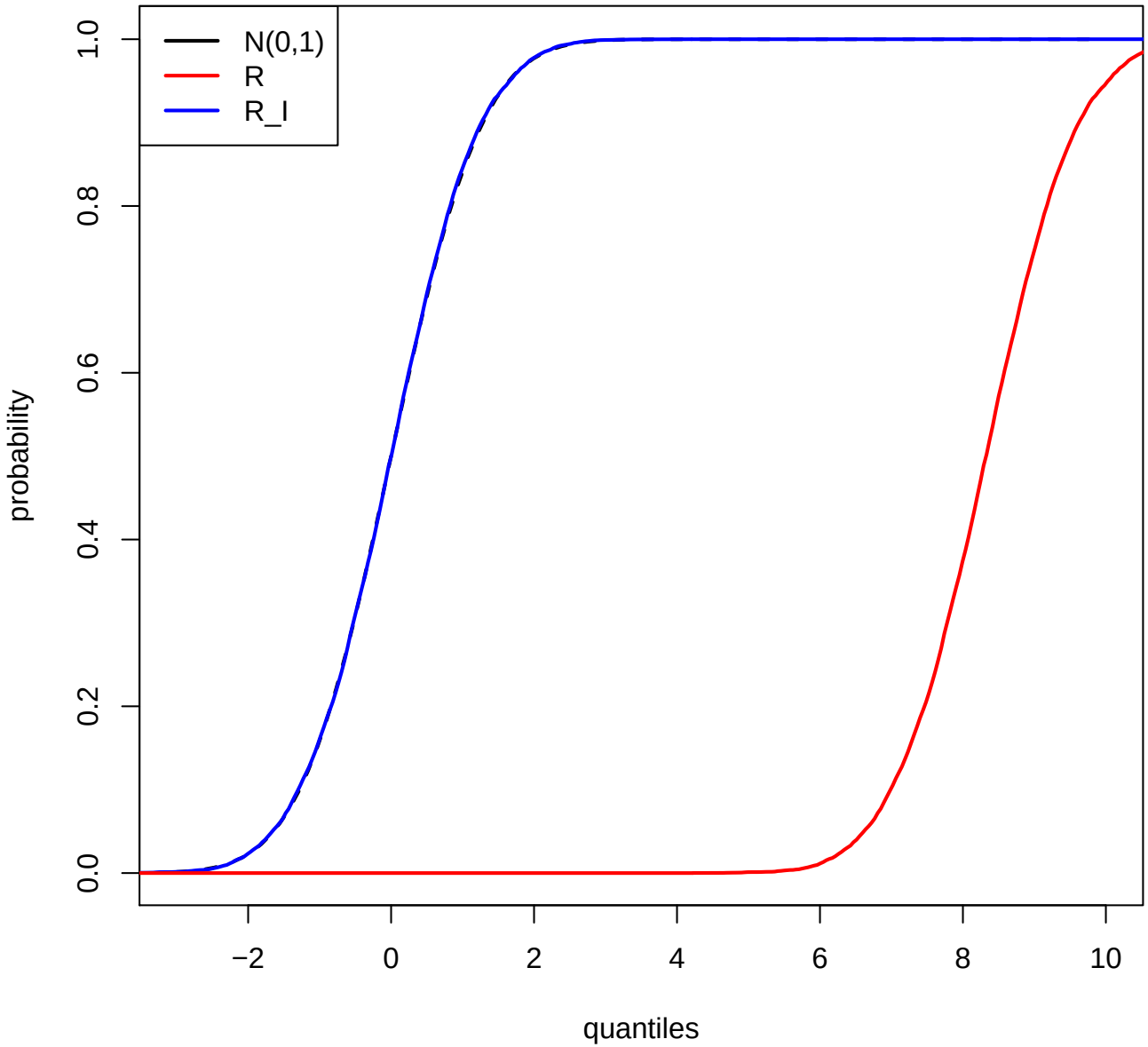
we obtain the integrated likelihood

$$L_I(\psi) = \prod_{i=1}^{q} \int_{R^+} e^{\psi \sum_{j=1}^{m} \log y_{ij} + m\psi \log \frac{\phi_i \psi}{\hat{\psi}} - \frac{\phi_i \psi}{\hat{\psi}} \sum_{j=1}^{q} y_{ij} - m \log \Gamma(\psi)} d\phi_i,$$

and consequently

$$R_I(\psi) = \text{sgn}(\hat{\psi}_I - \psi)\sqrt{2\{\ell_I(\hat{\psi}_I) - \ell_I(\psi)\}}.$$

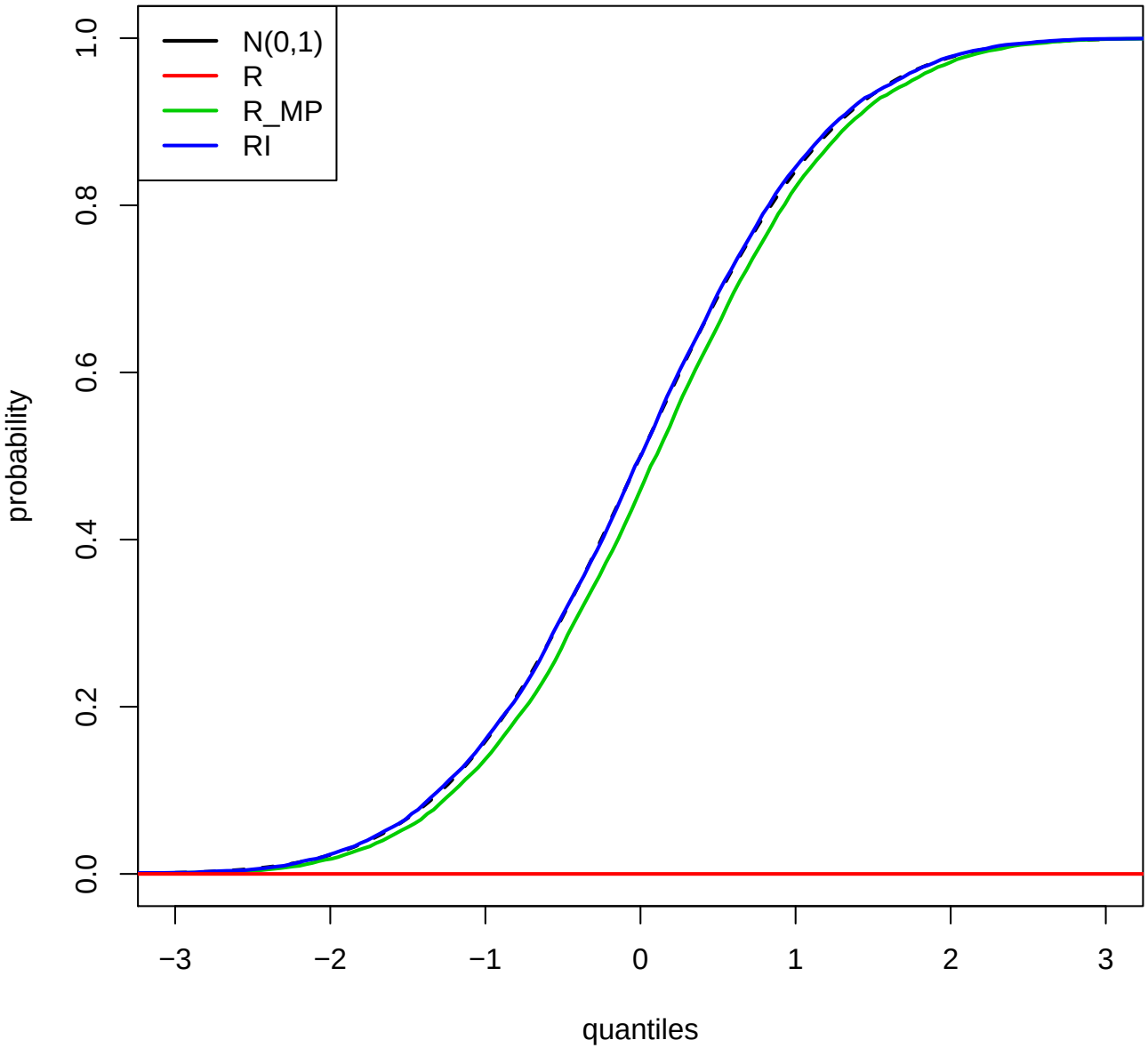
# Examples: Gamma common shape parameter




q=1000;
m=7;

B=9000.

# Examples: Gamma common shape parameter




q=1000;
m=7;

B=9000.

## Examples: Gamma common shape parameter

In this specific case,

$$
\begin{aligned}
L_I(\psi) &= \prod_{i=1}^{q} \int_{R^+} e^{\psi \sum_{j=1}^{m} \log y_{ij} + m\psi \log \frac{\phi_i \psi}{\hat{\psi}} - \frac{\phi_i \psi}{\hat{\psi}} \sum_{j=1}^{q} y_{ij} - m \log \Gamma(\psi)} d\phi_i \\
&= \dots \\
&= \prod_{i=1}^{q} e^{\psi(\sum_{j=1}^{m} \log y_{ij} - m \log \sum_{j=1}^{q} y_{ij}) - m \log \Gamma(\psi) + \log \Gamma(m\psi)} \\
&= L_C(\psi).
\end{aligned}
$$

## Examples: matched binomial

Let $Y_{i1}$ and $Y_{i2}$, $i = 1, \ldots, q$, independent random variables with distribution $Bi(m, p_{i1})$ and $Bi(1, p_{i2})$ respectively.

- stratum-dependent nuisance parameter: $\lambda_i = \log\{p_{i1}/(1 - p_{i1})\}$;
- parameter of interest (common among strata): $\psi = \log\{p_{i2}/(1 - p_{i2})\} - \log\{p_{i1}/(1 - p_{i1})\}$;

Here:

- $L(\psi, \lambda) = \frac{e^{(y_{i1}+y_{i2})\lambda + y_{i2}\psi}}{(1+e^{\lambda})^m (1+e^{\psi+\lambda})}$;
- the conditional likelihood is a noncentral hypergeometric distribution (Davison, 1988, Example 6.1).

## Examples: matched binomial

In this example there is no explicit formula for the zero-score-expectation parameter. Therefore:

- orthogonal parameter based on expected information, by choosing a weight function based on the original parameterization that would act like a uniform one in an orthogonal parameterization (Cox & Reid, 1993);
- zero-score-expectation-parameter, computed numerically.

## Examples: matched binomial

Based on Cox & Reid (1993)'s idea:

- $\partial\xi_i/\partial\lambda_i = me^{\lambda_i}/(1+e^{\lambda_i})^2 + e^{\psi+\lambda_i}/(1+e^{\psi+\lambda_i})^2$;
- $L_I(\psi) = \prod_{i=1}^{q} \int \frac{e^{(y_{i1}+y_{i2})\lambda_i + y_{i2}\psi}}{(1+e_i^{\lambda})^{m+2}(1+e^{\psi+\lambda_i})^3}[e_i^{\lambda}(1+e^{\lambda_i+\psi})^2 + e^{\psi+\lambda_i}(1+e_i^{\lambda})^2]d\lambda_i$.
- after a change of variable $\lambda_i(\omega_i) = \log\{\omega_i/(1-\omega_i)\}$,

$$L_I(\psi) = \prod_{i=1}^{q} e^{\psi y_{i2}}[{}_2F_1(1, y_{i1}+y_{i2}+1, m+2, 1-e^{\psi}) + e^{\psi}\ {}_2F_1(3, y_{i1}+y_{i2}+1, m+2, 1-e^{\psi})]$$

where ${}_2F(a,b,c,z)_1 = [\Gamma(c)/\{\Gamma(b)\Gamma(c-b)\}]\int_0^1 x^{b-1}(1-x)^{c-b-1}(1-zx)^{-a}dx$ (Abramowitz & Stegun, 1964, formula 15.3.1, pag. 558).

## Examples: matched binomial

Based on the zero-score expectation parameterization:

- $\phi$ is the solution of $K_\lambda(\hat{\psi}, \phi) - K_\lambda(\psi, \lambda) = 0,$ (3)
  - $K$ is the cumulant function;
  - the subscript denotes the derivative with respect to $\lambda$;
  - once we fix $\hat{\psi}$, $\psi$ and $\lambda$, it is possible to solve it numerically and get the corresponding $\phi$;
- changing of variable from $\phi$ to $\lambda$ in the integral,

$$L_I(\psi) = \int L(\psi, \phi) d\phi = \int L(\psi, \lambda) \frac{K_{\lambda\lambda}(\psi, \lambda)}{\partial K_{\lambda\lambda}(\hat{\psi}, \phi(\psi, \lambda; \hat{\psi}))} d\lambda.$$

where the Jacobian is obtained by differentiating (3) with respect to $\lambda$.

# Examples: matched binomial

- $q = 300$, $m = 7$

| | Nominal (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1.0 | 2.5 | 5.0 | 10.0 | 25.0 | 50.0 | 75.0 | 90.0 | 95.0 | 97.5 | 99.0 |
| $R$ | 0.0 | 0.0 | 0.1 | 0.5 | 2.6 | 10.4 | 26.3 | 47.1 | 60.4 | 70.9 | 81.5 |
| $R_C$ | 1.0 | 2.3 | 5.0 | 10.2 | 25.0 | 49.7 | 74.4 | 89.5 | 94.6 | 97.3 | 98.8 |
| $R_I$ | 0.6 | 1.7 | 3.6 | 7.8 | 21.0 | 44.8 | 70.6 | 87.4 | 93.5 | 96.9 | 98.5 |
| $R_O$ | 0.5 | 1.6 | 3.3 | 7.3 | 19.8 | 42.9 | 68.6 | 85.8 | 92.5 | 96.0 | 98.2 |
| $R_{MP}$ | 0.6 | 1.8 | 3.7 | 8.0 | 21.1 | 44.8 | 70.2 | 87.1 | 93.2 | 96.5 | 98.3 |

# Examples: matched binomial

- $q = 300$, $m_i = 5, 7, 9$ each replicated 100 times.

| | Nominal (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1.0 | 2.5 | 5.0 | 10.0 | 25.0 | 50.0 | 75.0 | 90.0 | 95.0 | 97.5 | 99.0 |
| $R$ | 0.0 | 0.1 | 0.2 | 0.6 | 2.8 | 9.9 | 25.4 | 46.1 | 59.7 | 70.4 | 80.7 |
| $R_C$ | 1.0 | 2.8 | 5.1 | 10.2 | 25.1 | 49.8 | 75.1 | 90.2 | 94.7 | 97.3 | 99.0 |
| $R_I$ | 0.8 | 2.1 | 4.0 | 8.0 | 20.5 | 44.4 | 70.4 | 87.5 | 93.4 | 96.5 | 98.5 |
| $R_O$ | 0.6 | 1.9 | 3.8 | 7.6 | 19.8 | 43.2 | 69.4 | 86.5 | 92.9 | 96.0 | 98.2 |
| $R_{MP}$ | 0.7 | 2.0 | 4.1 | 8.1 | 20.8 | 44.5 | 70.4 | 87.4 | 93.2 | 96.3 | 98.4 |

- $q = 300$, $m_i = 3$ in 30 strata, $m_i = 7$ in 270 strata.

| | Nominal (%) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1.0 | 2.5 | 5.0 | 10.0 | 25.0 | 50.0 | 75.0 | 90.0 | 95.0 | 97.5 | 99.0 |
| $R$ | 0.0 | 0.0 | 0.2 | 0.6 | 2.2 | 8.4 | 23.0 | 42.9 | 56.3 | 67.8 | 79.0 |
| $R_C$ | 1.0 | 2.4 | 4.8 | 9.4 | 24.7 | 49.2 | 74.7 | 89.8 | 94.5 | 97.5 | 99.0 |
| $R_I$ | 0.7 | 1.5 | 3.3 | 6.8 | 19.6 | 42.5 | 69.1 | 86.8 | 92.8 | 96.4 | 98.5 |
| $R_O$ | 0.7 | 1.5 | 3.3 | 6.8 | 19.5 | 42.0 | 68.3 | 86.2 | 92.5 | 96.0 | 98.3 |
| $R_{MP}$ | 0.7 | 1.6 | 3.5 | 7.2 | 20.3 | 43.3 | 69.6 | 86.8 | 92.8 | 96.2 | 98.4 |

## Remarks: strata sample size

Regarding the within-stratum sample size:

- results hold for $m_i \neq m_j$, $i \neq j$;
- usually it is assumed that $m_i$ are asymptotically balanced:
  - $m_i$ are of order $O(m)$ but not $o(m)$;
- we simply need $\max_{1 \leq i \leq q} m_i/n \to 0$ for $q \to \infty$;
  - $n = m_1 + \cdots + m_q$ is the total number of observations;
- this is true even if some strata have $m_i = o(m)$, provided the number of such strata does not increase with $q$.

## Examples: First order non stationary autoregressive model

Consider the first-order autoregressive model defined by

$$y_{ij} = \lambda_i + \rho y_{ij-1} + \varepsilon_{ij} \, ,$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$, $j = 1, \ldots, m_i$, $i = 1, \ldots, q$.

Consider the non-stationary case. The log-likelihood is

$$\ell(\rho, \sigma^2, \lambda_i) = \sum_{i=1}^{q} \left\{ -\frac{m_i}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^{m_i} (y_{ij} - \lambda_i - \rho y_{ij-1})^2 \right\} ,$$

the parameter of interest is $\psi = (\rho, \sigma^2)$ and the stratum-dependent nuisance parameter $\lambda_i$.

## Examples: First order non stationary autoregressive model

Lancaster (2002) provides an information orthogonal parameter,

$$\xi_i = \lambda_i \exp\{b(\rho)\},$$

where $b_i(\rho) = \frac{1}{m_i} \sum_{j=1}^{m_i-1} \frac{m_i-j}{j} \rho^j$.

Alternatively, we can solve

$$E_{\rho_0,\sigma_0^2,\lambda_{i0}}[\ell_{\lambda_i}(\rho,\sigma^2,\lambda_i)]|_{(\rho_0,\sigma_0^2,\lambda_{i0})=(\hat{\rho},\hat{\sigma}^2,\phi_i)} = 0\,.$$

and obtain the zero-score-expectation parameter

$$\phi_i = \frac{\lambda_i}{1+(\hat{\rho}-\rho)b_i'(\hat{\rho})},$$

where $b_i'(\rho) = (1/m_i)\sum_{j=1}^{m_i-1}(m_i-j)\rho^{j-1}$ and $\hat{\rho}$ is the mle.

## Examples: First order non stationary autoregressive model

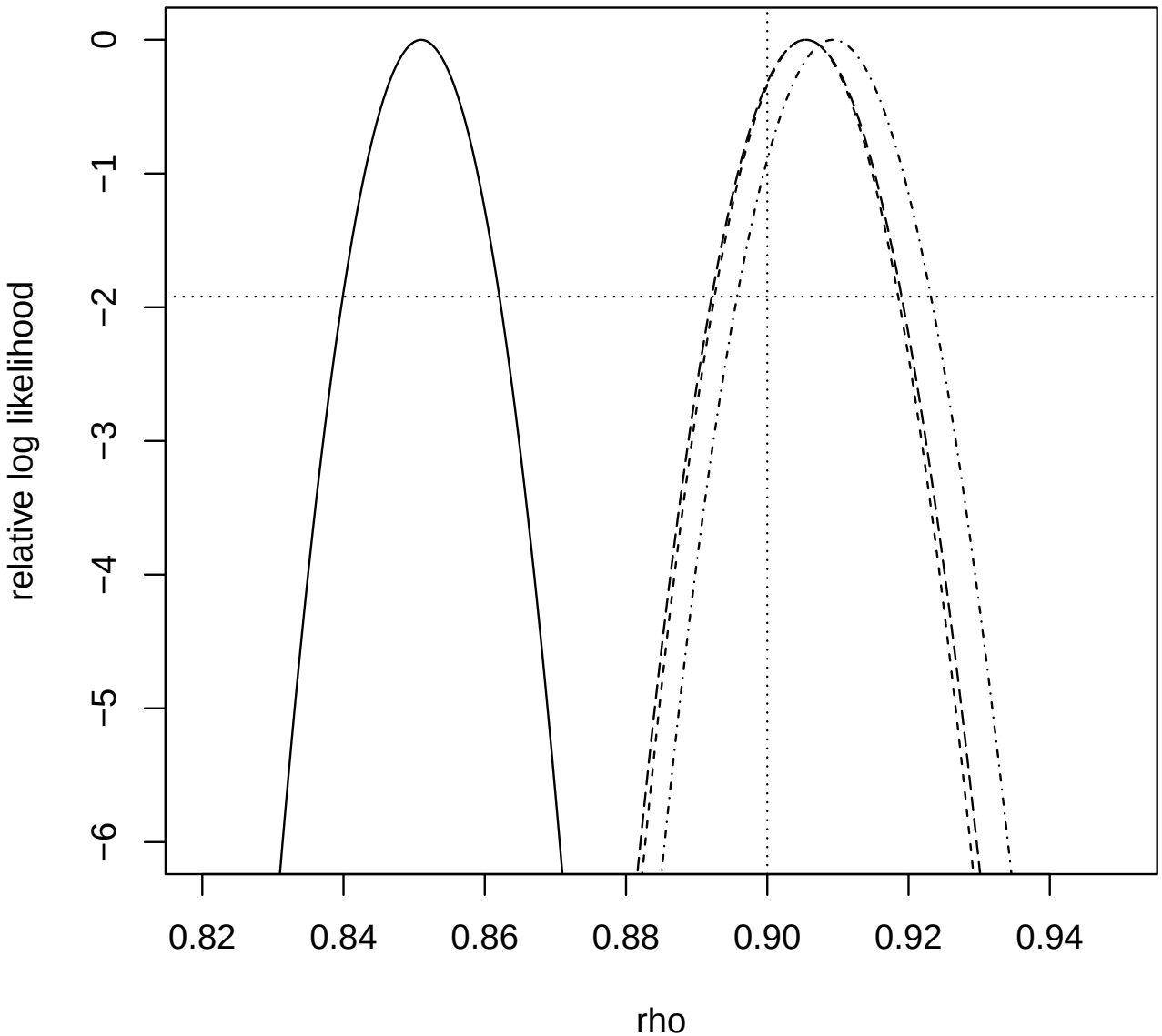
Focusing interest on the parameter $\rho$, we have

$$\begin{aligned}
\ell_P(\rho) &= -\frac{\sum_{i=1}^{q} m_i}{2} \log SS(\rho) \\
\ell_O(\rho) &= -\frac{\sum_{i=1}^{q} (m_i - 1)}{2} \log SS(\rho) + \sum_{i=1}^{q} b_i(\rho) \\
\ell_I(\rho) &= -\frac{\sum_{i=1}^{q} (m_i - 1)}{2} \log SS(\rho) - \sum_{i=1}^{q} \log\{1 + (\hat{\rho} - \rho) b_i'(\hat{\rho})\}
\end{aligned}$$

where

- $SS(\rho) = \sum_{i=1}^{q} \sum_{j=1}^{m_i} \{w_{ij}(\rho) - \bar{w}_i(\rho)\}^2$;
- $w_{ij}(\rho) = y_{ij} - \rho y_{ij-1}$;
- $\bar{w}_i(\rho) = {m_i}^{-1} \sum_{j=1}^{m_i} w_{ij}(\rho)$;
- we used a constant weight function for $\lambda_i$ and $\log \sigma$.

# Examples: First order non stationary autoregressive model




q=500;
m=8;

B=10000.

## Remarks: within stratum dependence

From this example we see:

- the within-stratum observations do not need to be independent;
- within each stratum the asymptotic properties of the likelihood quantities must hold for $m \to \infty$;
- the zero-score-expectation parameterization depends on the data through an estimate (typically mle);
- alternative estimates may lead to more accurate inference.

## Conclusions: final remarks

- We studied the frequentist asymptotic properties of integrated likelihoods quantities in a two-index-asymptotic setting;
- inference based on a properly constructed integrated likelihood is more accurate than that based on the profile likelihood;
- this kind of integrated likelihood has asymptotic properties similar to those of higher order methods such as the modified profile likelihood.

## Conclusions: final remarks

- the integrated likelihood is a tool for inference that can be applied in wide generality;
- the computation of integrals is required;
- it is necessary to find a parameterization in which the nuisance parameter is unrelated to the parameter of interest:
  - information orthogonal parameterization may be not easy to compute or may not even exist (vector parameter of interest);
  - the zero-score-expectation parameterization can always be defined and has an algorithmic form that can be easily implemented.

# References I

Abramowitz, M. & Stegun, I. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, New York.

Barndorff-Nielsen, O. (1980). Conditionality resolutions. *Biometrika* **67**, 293–310.

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343–365.

Barndorff-Nielsen, O. (1986). Inference on full or partial parameters based on the standardized signed log-likelihood ratio. *Biometrika* **73**, 307–322.

Barndorff-Nielsen, O. (1996). Two order asymptotic. In *Frontiers in Pure and Applied Probability II: Proceedings of the Fourth Russian-Finnish Symposium Prob. Th. Math. Statist.*, A. Melnikov, ed. TVP Science, Moscow.

Berger, J., Liseo, B. & Wolpert, R. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science* **14**, 1–22.

# References II

Cox, D. & Reid, N. (1993). A note on the calculation of adjusted profile likelihood. *Journal of the Royal Statistical Society. Series B (Methodological)* **55**, 467–471.

Davison, A. (1988). Approximate conditional inference in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)* **50**, 445–461.

De Bin, R., Sartori, N. & Severini, T. A. (2015). Integrated likelihoods in models with stratum nuisance parameters. *Electronic Journal of Statistics* **9**, 1474–1491.

Kalbfleisch, J. & Sprott, D. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society. Series B (Methodological)* **32**, 175–208.

Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics* **27**, 887–906.

Lancaster, T. (2002). Orthogonal parameters and panel data. *Review of Economic Studies* **69**, 647–666.

# References III

Neyman, J. & Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

Pace, L. & Salvan, A. (1997). *Principles of Statistical Inference: from a neo-Fisherian perspective*. World Scientific, Singapore.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika* **90**, 533–549.

Sartori, N., Bellio, R., Salvan, A. & Pace, L. (1999). The directed modified profile likelihood with many nuisance parameters. *Biometrika* **86**, 735–742.

Severini, T. (2000). *Likelihood Methods in Statistics*. Oxford University Press, Oxford.

Severini, T. (2007). Integrated likelihood functions for non-Bayesian inference. *Biometrika* **94**, 529–542.

Severini, T. (2010). Likelihood ratio statistics based on an integrated likelihood. *Biometrika* **97**, 481–496.

Strasser, H. (1996). Asymptotic efficiency of estimates for models with incidental nuisance parameters. *The Annals of Statistics* **24**, 879–901.