Almost-perfect Signal Detection

Anthony Davison

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Simple example

- Define set of cuts to select possible signal.
- Expect background b ($\pm$ error) from uninteresting sources.
- Observe n events
  1. For n rather larger than b, quantify **significance** of deviation (strong significance required, i.e. "5σ").
  2. For $n \approx b$, establish **upper** limit on possible excess from interesting new source.
- **Background** and **calibration efficiencies** (intensity of signal) should be considered as nuisance parameters.
- Typically there are many **search channels** (different configurations in which the particle could be produced and decay).

At Banff the following model was proposed:

$$\begin{aligned} N_i &\sim \text{Poisson}(\lambda_{1i}\,\psi + \lambda_{2i}), \\ Y_i &\sim \text{Poisson}(\lambda_{2i}\,t_i), \\ Z_i &\sim \text{Poisson}(\lambda_{1i}\,u_i), \end{aligned}$$

- $i = 1, \ldots, c$, where c is the number of search channels,
- Y and Z represent auxiliary independent experiments meant to measure background and intensity respectively,
- t_i and u_i are known positive constants,
- different channels are assumed independent,
- Then the parameters are
  - ψ, the **interest parameter** (**signal of interest**),
  - $\lambda = (\lambda_{11}, \lambda_{21}, \ldots, \lambda_{1c}, \lambda_{2c})$, the **nuisance parameter**.

Likelihood

- The proposed Poisson model has log likelihood

$$\begin{aligned}\ell(\psi,\lambda) &= \log f(n,y,z;\psi,\lambda)\\ &\equiv \sum_{i=1}^{c}\{n_i\log(\lambda_{1i}\,\psi+\lambda_{2i})+z_i\log\lambda_{1i}\\ &\quad + y_i\log\lambda_{2i}-(\psi+u_i)\,\lambda_{1i}-(1+t_i)\,\lambda_{2i}\}\end{aligned}$$

- A $(3c, 2c+1)$ curved exponential family.
- A full exponential family with a single channel ($c=1$).
- For physicists, $\psi \geq 0$.
- But for the use of the model $\psi > -\lambda_{2i}/\lambda_{1i}$, for each i, so can have $\psi < 0$, at least mathematically.
- Aim to use **higher order likelihood inference** for ψ.
- Details for this example in Davison and Sartori (2008, *Statistical Science*).

Higher Order Likelihood Inference

First order likelihood theory

- Parametric model $f(y;\theta)$ with (notional?) sample size n, log likelihood $\ell(\theta)$, observed information $j(\theta) = -\partial^2\ell(\theta)/\partial\theta^2$, MLE $\widehat{\theta}$.
- If θ scalar, first order inference based on limiting $N(0,1)$ laws of

likelihood root, $\quad r(\theta) = \text{sign}(\widehat{\theta} - \theta)\left[2\left\{\ell(\widehat{\theta}) - \ell(\theta)\right\}\right]^{1/2};$

score statistic, $\quad s(\theta) = j(\widehat{\theta})^{-1/2}\partial\ell(\theta)/\partial\theta;$

Wald statistic, $\quad t(\theta) = j(\widehat{\theta})^{1/2}(\widehat{\theta} - \theta)$:

as $n \to \infty$ have, for instance,

$$\mathrm{P}\left\{r(\theta) \le r_{\text{obs}}\right\} = \Phi(r_{\text{obs}}) + O(n^{-1/2}),$$

which yields tests and confidence sets for θ based on an observed r_{obs}.

Higher order likelihood theory

- For continuous responses, third order inference based on limiting $N(0,1)$ distribution of **modified likelihood root**, also called **modified directed deviance**,

$$r^*(\theta) = r(\theta) + \frac{1}{r(\theta)} \log\left\{\frac{q(\theta)}{r(\theta)}\right\},$$

  where $q(\theta)$ depends on model, can be $s(\theta)$, $t(\theta)$, or similar.

- Inference uses **significance function**

$$\Phi\{r^*(\theta)\} \dot\sim U(0,1), \quad \text{for true } \theta.$$

- Level $1-2\alpha$ confidence interval contains those θ for which

$$\alpha \le \Phi\{r^*(\theta)\} \le 1-\alpha.$$

- When $\theta = (\psi, \lambda)$, with ψ scalar

$$r(\psi) = \mathsf{sign}(\widehat{\psi} - \psi)\left[2\left\{\ell(\widehat{\theta}) - \ell(\widehat{\theta}_\psi)\right\}\right]^{1/2}$$

  where $\widehat{\theta}_\psi$ is the MLE of θ for fixed ψ.

- The function $q(\psi)$ is

$$q(\psi) = \frac{|\ \varphi(\widehat{\theta}) - \varphi(\widehat{\theta}_\psi) \ \ \varphi_\lambda(\widehat{\theta}_\psi)\ |}{|\ \varphi_\theta(\widehat{\theta})\ |}\left\{\frac{|\ j(\widehat{\theta})\ |}{|\ j_{\lambda\lambda}(\widehat{\theta}_\psi)\ |}\right\}^{1/2}$$

  where φ (which has the dimension of θ) is the canonical parameter of a local exponential family approximation to the model and where, for example, φ_θ denotes the matrix $\partial\varphi/\partial\theta^{\mathrm{T}}$.

- Parameterization-invariant.
- Computation almost as easy as first order asymptotics.
- Error $O(n^{-3/2})$ in continuous response models.
- Gives continuous approximation to discrete response models.
- Relative (not absolute) error, so highly accurate in tails.
- **Bayesian version** with prior π uses

$$q_B(\psi) = \ell_{\rm p}'(\psi) j_{\rm p}(\widehat{\psi})^{-1/2} \left\{ \frac{\left| j_{\lambda\lambda}(\widehat{\theta}_\psi) \right|}{\left| j_{\lambda\lambda}(\widehat{\theta}) \right|} \right\}^{1/2} \frac{\pi(\widehat{\theta})}{\pi(\widehat{\theta}_\psi)},$$

where $\ell_{\rm p}(\psi) = \ell(\widehat{\theta}_\psi)$ is the **profile log likelihood** for ψ, and $j_{\rm p} = -\ell_{\rm p}''$ is the corresponding information. Again, easy computation and high accuracy.

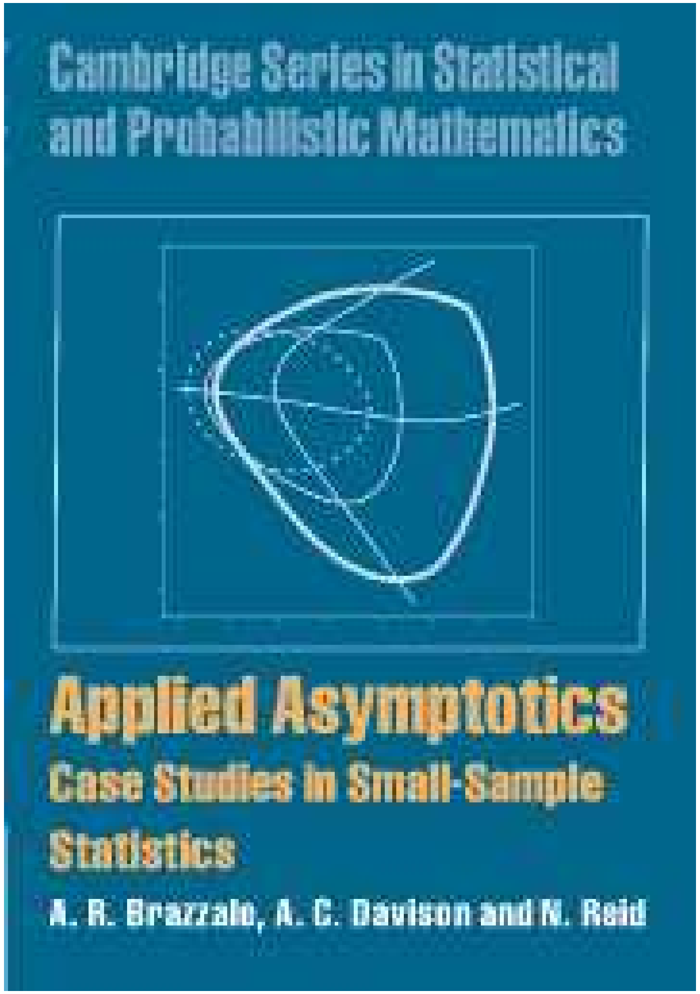
Bayesian approach

- Aim for prior that is noninformative for ψ in presence of nuisance parameter ξ.
- Tibshirani (1989, *Biometrika*) shows that this prior is proportional to

$$|i_{\psi\psi}(\psi,\xi)|^{1/2} g(\xi)\,\mathrm{d}\psi\mathrm{d}\xi,$$

  when ξ is orthogonal to ψ and
  - $i_{\psi\psi}(\psi,\xi)$ denotes the (ψ,ψ) element of the Fisher information matrix,
  - $g(\xi)$ is an arbitrary positive function satisfying mild regularity conditions.
- This gives a Jeffreys prior for ψ that is also a **matching prior**: it gives $(1-\alpha)$ one-sided Bayesian posterior confidence intervals that contain ψ with frequentist probability $(1-\alpha)+\mathcal{O}(n^{-1})$.
- Calculations are (miraculously) explicit for the Poisson model.

Publicité



Brazzale, Davison, Reid (2007) *Applied Asymptotics*, Cambridge University Press.

Application to Signal Detection

Single channel ($c = 1$): Example

- The data are

n	y	z	t	u
1	8	14	27	80

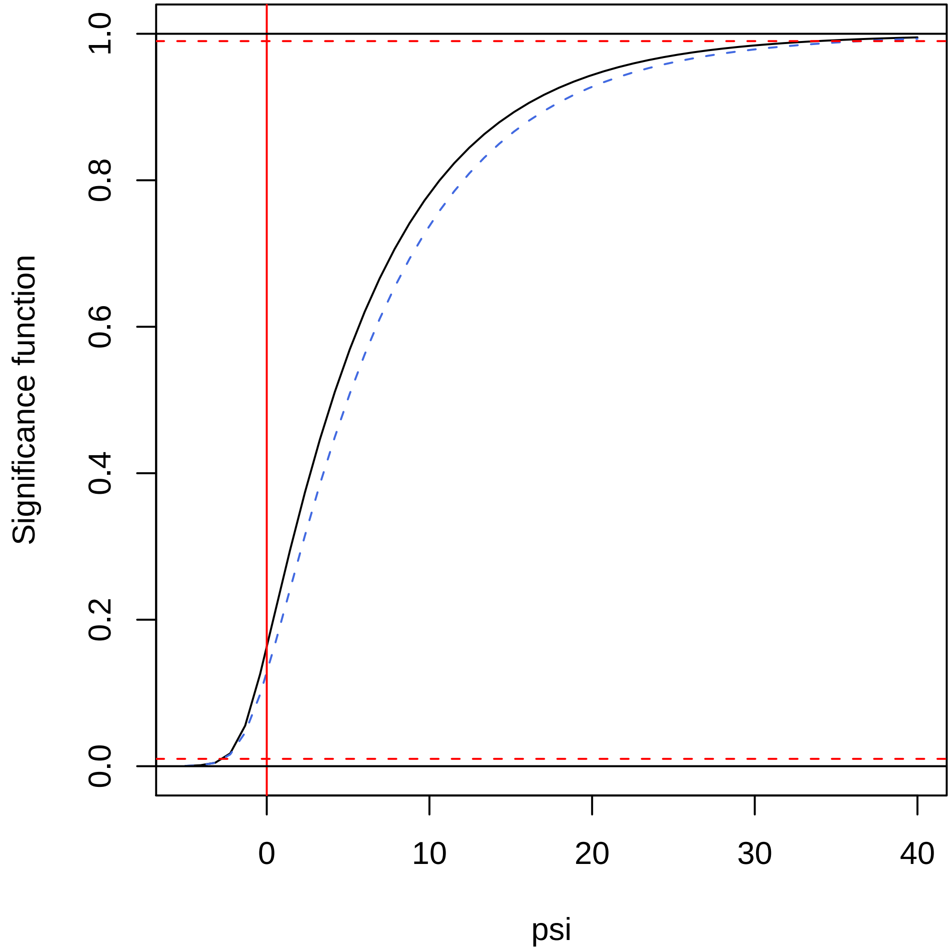
- The MLE of ψ is 4.02.
- The p-value for testing $\psi = 0$ versus $\psi > 0$ is 0.16276 for r and 0.12714 for r^*, both indicating almost no evidence of signal.
- The 0.99 lower and upper limits for ψ obtained from r are

$$-2.64 \quad (\mapsto 0), \qquad 33.84.$$

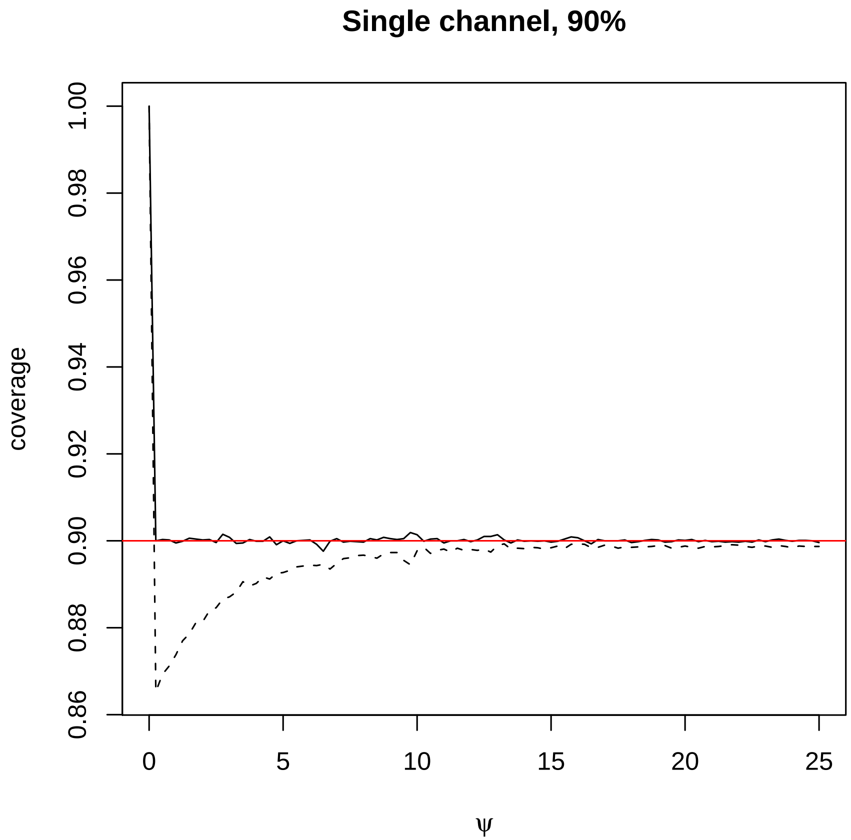
- The analogous limits obtained from r^* are

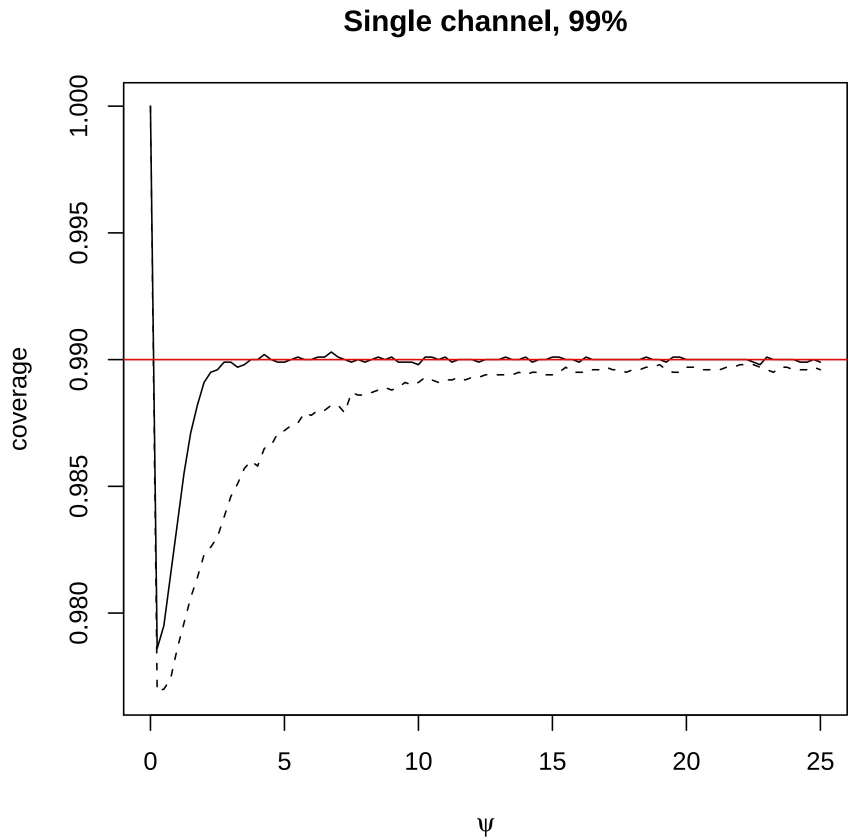
$$-2.60 \quad (\mapsto 0), \qquad 36.52.$$

Single channel: An example



One channel: Coverage of confidence limits





Target coverage (red), coverage of r^* (black), coverage of r_B^* (dashes), as a function of interest parameter ψ.

Multichannel simulation, $c = 10$

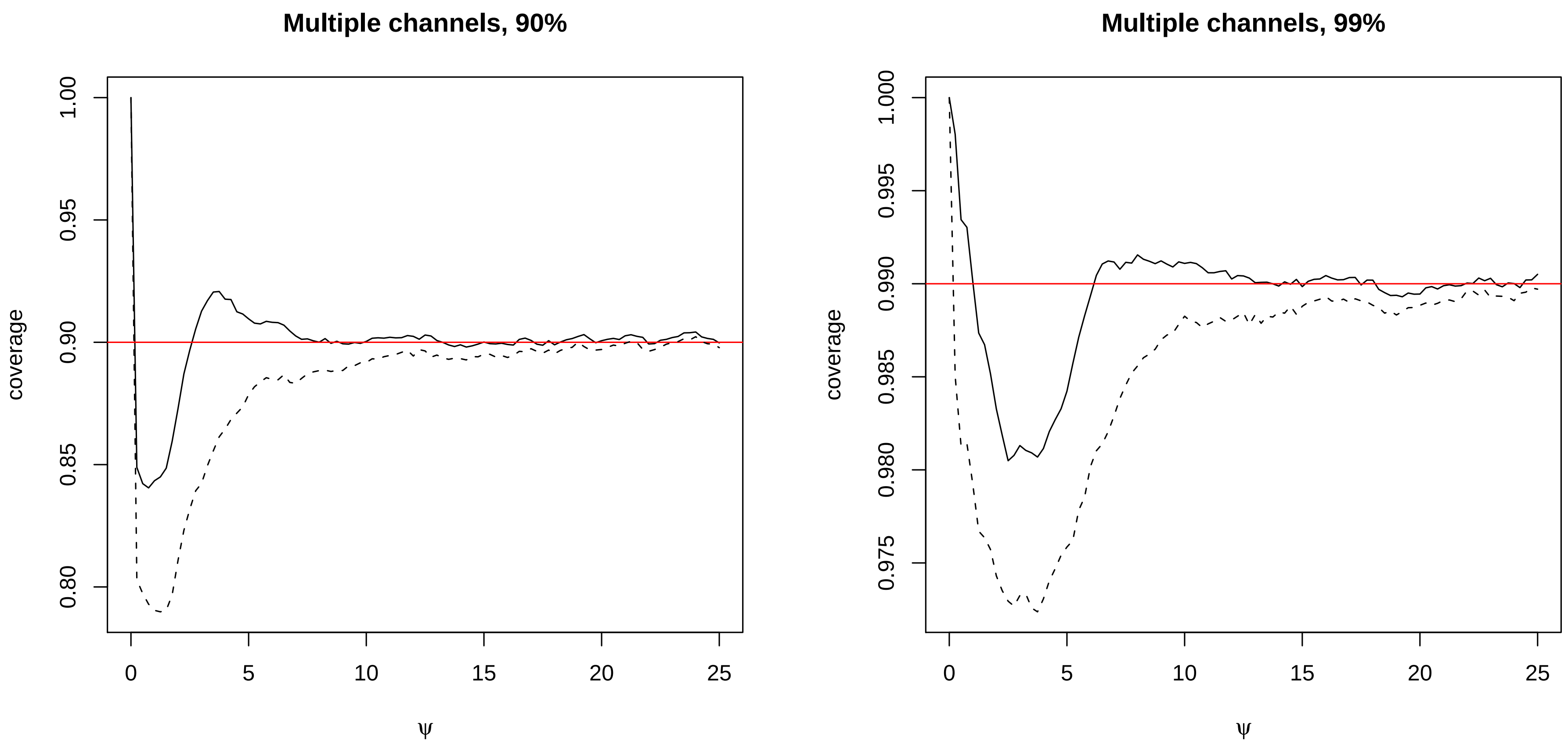
TABLE 3

Empirical coverage probabilities in a multiple-channel simulation with 10,000 replications, $\psi = 2$, $\beta = (0.20, 0.30, 0.40, \ldots, 1.10)$, $\gamma = (0.20, 0.25, 0.30, \ldots, 0.65)$, $t = (15, 17, 19, \ldots, 33)$ and $u = (50, 55, 60, \ldots, 95)$

Probability	$\boldsymbol{r}$	$\boldsymbol{r^*}$	$\boldsymbol{r_B^*}$
0.0100	0.0099	0.0101	0.0109
0.0250	0.0244	0.0255	0.0273
0.0500	0.0493	0.0519	0.0542
0.1000	0.0967	0.1012	0.1035
0.5000	**0.4869**	0.5043	0.5027
0.9000	**0.8900**	0.9013	0.8942
0.9500	**0.9421**	0.9499	**0.9427**
0.9750	**0.9687**	0.9759	**0.9689**
0.9900	**0.9875**	0.9913	**0.9864**

Figures in bold differ from the nominal level by more than simulation error.

Multichannel simulation, $c = 10$



Target coverage (red), coverage of r^* (black), coverage of r^*_B (dashes), as a function of interest parameter ψ.

Discussion

- Modified likelihood root can yield highly accurate inference in this toy problem.
- It's pretty good even with many nuisance parameters.
- Some (boundary) cases give problems with $q(\psi)$: then we use $r(\psi)$.
- Noninformative Bayesian solution provides (slightly) worse confidence intervals, but is quite feasible, could instead use an informative prior.

- Modified likelihood root can yield highly accurate inference in this toy problem.
- It's pretty good even with many nuisance parameters.
- Some (boundary) cases give problems with $q(\psi)$: then we use $r(\psi)$.
- Noninformative Bayesian solution provides (slightly) worse confidence intervals, but is quite feasible, could instead use an informative prior.

Now for neutrinos . . .

Neutrino ordering

Fun with neutrinos

- Discussion based on Heavens and Sellentin (2018, arXiv): *Objective Bayesian analysis of neutrino masses and hierarchy*, suggested by Louis.
- Neutrino masses: $0 \le \mu_L \le \mu_M \le \mu_H$ meV
- Two hierarchies: normal and inverted
- Measurements $y_1 = 75\text{meV}^2$, $y_2 = 2524\text{meV}^2$, and probability models
  - **normal hierarchy**:

$$\begin{aligned} y_1 &\sim \mathcal{N}(\mu_M^2 - \mu_L^2, 1.8^2), \\ y_2 &\sim \mathcal{N}\{\mu_H^2 - (\mu_M^2 + \mu_L^2)/2, 40^2\}, \end{aligned}$$

  - **inverted hierarchy**:

$$\begin{aligned} y_1 &\sim \mathcal{N}(\mu_H^2 - \mu_M^2, 1.8^2), \\ y_2 &\sim \mathcal{N}\{(\mu_H^2 + \mu_M^2)/2 - \mu_L^2, 40^2\} \end{aligned}$$

- 95% credible region: $\mathrm{P}(\mu_L + \mu_M + \mu_H \le 120\text{meV}) = 0.95$, interpreted as penalized likelihood corresponding to $y_3 = 0$ observed from half-normal distribution

$$|\mathcal{N}\{\mu_L + \mu_M + \mu_H, (120/1.96)^2\}|.$$

- Possible typo: replace $y_2 = 2524$ by $y_2 = 2514$ in inverted case?

First idea

- Normal hierarchy indicated if

$$\mu_H^2 - \mu_M^2 > \mu_M^2 - \mu_L^2 \quad \Leftrightarrow \quad \psi = \mu_H^2 + \mu_L^2 - 2\mu_M^2 > 0,$$

  so try and use higher order methods to test this.

- As a warm-up, try testing non-nested hypotheses:
  - Cox (1961, *Tests of Separate Families of Hypotheses*, 4th Berkeley Symposium),
  - Cox (1962, *Further results on tests of separate families of hypotheses*, JRSSB),
  - massive literature in econometrics,

- Boils down to use of likelihood ratios for the models.

Separate families test with neutrino data

- Natural test statistic is difference of maximised log likelihoods,

$$T = 2\left(\widehat{\ell}_{\rm Normal} - \widehat{\ell}_{\rm Inverted}\right),$$

  with observed value $t_{\rm obs} = 1.748$.
- $T \dot{\sim}$ normal with unknown mean and variance.
- We estimate the significance probabilities for testing the normal and inverted models,

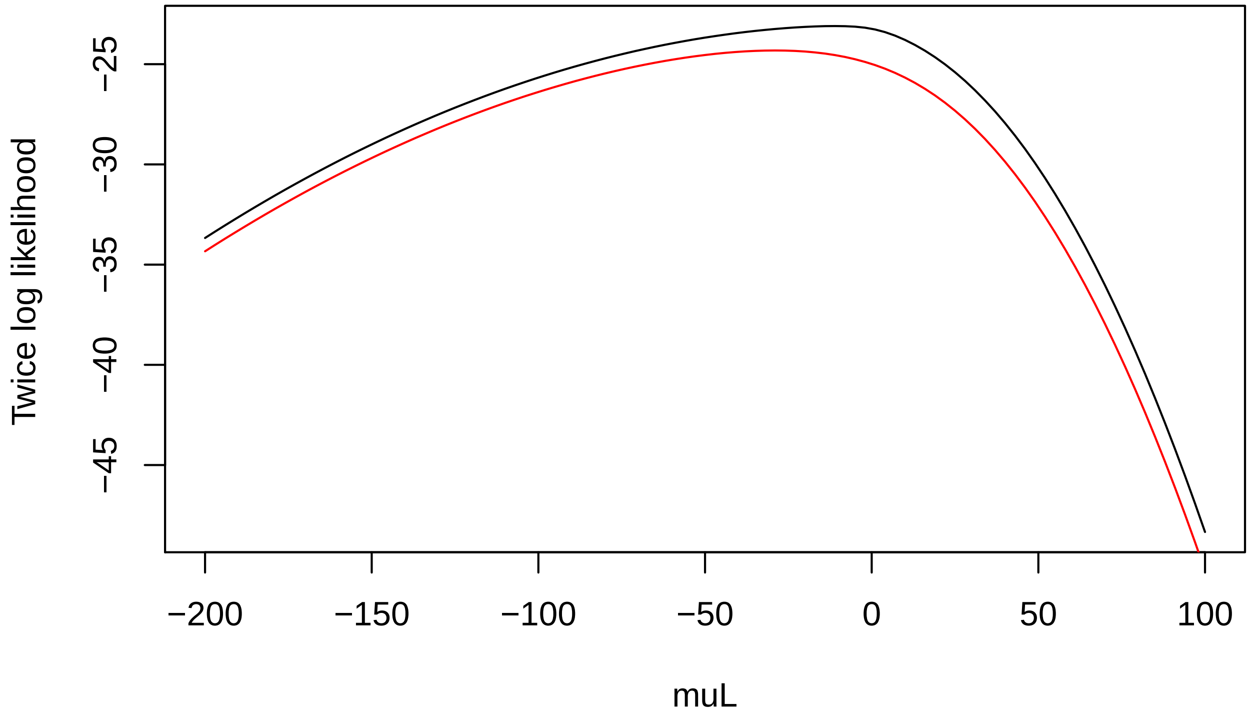
$$p_{\rm Normal} = {\rm P}_{\rm Normal}(T \le t_{\rm obs}), \quad p_{\rm Inverted} = {\rm P}_{\rm Inverted}(T \ge t_{\rm obs})$$

  by simulating data from the best-fitting normal and inverted models, and get

$$p_{\rm Normal} \approx 0.44, \quad p_{\rm Inverted} \approx 0.55,$$

  so the data do not distinguish the models.
- Equivalent to (two) bootstrap hypothesis tests (Davison and Hinkley, 1997, *Bootstrap Methods and their Application*, Chapter 4).

Profile log likelihood for μ_L



$2\ell_{\mathrm{p}}(\mu_L)$ for normal hierarchy (black) and inverted hierarchy (red)

Oops . . .

- On reflection the result is obvious: there are three parameters to match the means of y_1 and y_2, and this can be done (almost) perfectly:
  - normal model: $\mu_L = 0.238$, $\mu_M = 8.665$, $\mu_H = 50.611$, $\mathrm{E}(y_1) = 75.018$, $\mathrm{E}(y_2) = 2523.90$,
  - inverted model: $\mu_L = 0.009$, $\mu_M = 49.856$, $\mu_H = 50.603$, $\mathrm{E}(y_1) = 75.000$, $\mathrm{E}(y_2) = 2523.14$,

  so the only difference between the fits is due to the penalty, which is larger for the inverted model, for which $\mu_L + \mu_M + \mu_H \approx 100$, whereas for the normal model, $\mu_L + \mu_M + \mu_H \approx 59$.
- The same argument applies to Bayesian analyses: the log odds depends only on the extent that the credible region for the sum of masses favours the normal model, and the effects of a prior.
- Check: replace 120meV by stronger penalty $\mathrm{P}(\mu_L + \mu_M + \mu_H \leq 12\text{meV}) = 0.95$, which disfavours the inverted hierarchy even more, and now get

$$p_{\text{Normal}} \approx 0.72, \quad p_{\text{Inverted}} \approx 0.024,$$

  so we would reject the inverted model at the 5% level, as expected.
- Conclusion: more (but different!) data are needed to distinguish the hierarchies.

Thanks!

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE