Almost-perfect Signal Detection

Anthony Davison
Simple example

- Define set of cuts to select possible signal.
- Expect background $b$ (± 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:

\[ N_i \sim \text{Poisson}(\lambda_{1i} \psi + \lambda_{2i}), \]
\[ Y_i \sim \text{Poisson}(\lambda_{2i} t_i), \]
\[ Z_i \sim \text{Poisson}(\lambda_{1i} u_i), \]

\( i = 1, \ldots, c, \) 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

- \( \psi \), the interest parameter (signal of interest),
- \( \lambda = (\lambda_{11}, \lambda_{21}, \ldots, \lambda_{1c}, \lambda_{2c}) \), the nuisance parameter.
The proposed Poisson model has log likelihood

\[
\ell(\psi, \lambda) = \log f(n, y, z; \psi, \lambda) \\
\equiv \sum_{i=1}^{c} \{ n_i \log(\lambda_1^i \psi + \lambda_2^i) + z_i \log \lambda_1^i \\
+ y_i \log \lambda_2^i - (\psi + u_i) \lambda_1^i - (1 + t_i) \lambda_2^i \}\]

- 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 \(\psi > -\lambda_2^i/\lambda_1^i\), for each \(i\), so can have \(\psi < 0\), at least mathematically.
- Aim to use **higher order likelihood inference** for \(\psi\).
- Details for this example in Davison and Sartori (2008, *Statistical Science*).
Higher Order Likelihood Inference

Detecting a Signal
  Higher Order Likelihood
  Inference
  First order
  Higher order
  Nuisance parameters
  Properties
  Bayes
  Application to Signal Detection
  Neutrino ordering
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 $\hat{\theta}$.

- If $\theta$ scalar, first order inference based on limiting $\mathcal{N}(0,1)$ laws of
  
  - **likelihood root**, $r(\theta) = \text{sign}(\hat{\theta} - \theta) \left[ 2 \left\{ \ell(\hat{\theta}) - \ell(\theta) \right\} \right]^{1/2}$;
  
  - **score statistic**, $s(\theta) = j(\hat{\theta})^{-1/2} \partial \ell(\theta) / \partial \theta$;
  
  - **Wald statistic**, $t(\theta) = j(\hat{\theta})^{1/2} (\hat{\theta} - \theta)$:

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

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

  which yields tests and confidence sets for $\theta$ based on an observed $r_{\text{obs}}$. 
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

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

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

$$\alpha \leq \Phi\{r^*(\theta)\} \leq 1 - \alpha.$$
With nuisance parameters?

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

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

where $\hat{\theta}_\psi$ is the MLE of $\theta$ for fixed $\psi$.

□ The function $q(\psi)$ is

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

where $\varphi$ (which has the dimension of $\theta$) is the canonical parameter of a local exponential family approximation to the model and where, for example, $\varphi_\theta$ denotes the matrix $\partial \varphi / \partial \theta^T$. 
Properties of higher order approximations

- 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 $\pi$ uses

$$q_B(\psi) = \ell_p'(\psi) j_p(\hat{\psi})^{-1/2} \left\{ \frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|}{|j_{\lambda\lambda}(\hat{\theta})|} \right\}^{1/2} \frac{\pi(\hat{\theta})}{\pi(\hat{\theta}_\psi)},$$

where $\ell_p(\psi) = \ell(\hat{\theta}_\psi)$ is the **profile log likelihood** for $\psi$, and $j_p = -\ell''_p$ is the corresponding information. Again, easy computation and high accuracy.
Aim for prior that is noninformative for $\psi$ in presence of nuisance parameter $\xi$.

Tibshirani (1989, *Biometrika*) shows that this prior is proportional to

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

when $\xi$ is orthogonal to $\psi$ and

- $i_{\psi\psi}(\psi, \xi)$ denotes the $(\psi, \psi)$ element of the Fisher information matrix,
- $g(\xi)$ is an arbitrary positive function satisfying mild regularity conditions.

This gives a Jeffreys prior for $\psi$ that is also a **matching prior**: it gives $(1 - \alpha)$ one-sided Bayesian posterior confidence intervals that contain $\psi$ with frequentist probability $(1 - \alpha) + O(n^{-1})$.

Calculations are (miraculously) explicit for the Poisson model.
Application to Signal Detection
Single channel \((c = 1)\): Example

- The data are

<table>
<thead>
<tr>
<th>n</th>
<th>y</th>
<th>z</th>
<th>t</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
<td>14</td>
<td>27</td>
<td>80</td>
</tr>
</tbody>
</table>

- The MLE of \(\psi\) 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 \(\psi\) obtained from \(r\) are

  \(-2.64 \ (\mapsto 0), \ 33.84.\)

- The analogous limits obtained from \(r^*\) are

  \(-2.60 \ (\mapsto 0), \ 36.52.\)
Single channel: An example
Target coverage (red), coverage of $r^\ast$ (black), coverage of $r^\ast_B$ (dashes), as a function of interest parameter $\psi$. 
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 $u = (50, 55, 60, \ldots, 95)$

<table>
<thead>
<tr>
<th>Probability</th>
<th>$r$</th>
<th>$r^*$</th>
<th>$r^*_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0100</td>
<td>0.0099</td>
<td>0.0101</td>
<td>0.0109</td>
</tr>
<tr>
<td>0.0250</td>
<td>0.0244</td>
<td>0.0255</td>
<td>0.0273</td>
</tr>
<tr>
<td>0.0500</td>
<td>0.0493</td>
<td>0.0519</td>
<td>0.0542</td>
</tr>
<tr>
<td>0.1000</td>
<td>0.0967</td>
<td>0.1012</td>
<td>0.1035</td>
</tr>
<tr>
<td>0.5000</td>
<td><strong>0.4869</strong></td>
<td>0.5043</td>
<td>0.5027</td>
</tr>
<tr>
<td>0.9000</td>
<td><strong>0.8900</strong></td>
<td>0.9013</td>
<td>0.8942</td>
</tr>
<tr>
<td>0.9500</td>
<td><strong>0.9421</strong></td>
<td>0.9499</td>
<td><strong>0.9427</strong></td>
</tr>
<tr>
<td>0.9750</td>
<td><strong>0.9687</strong></td>
<td>0.9759</td>
<td><strong>0.9689</strong></td>
</tr>
<tr>
<td>0.9900</td>
<td><strong>0.9875</strong></td>
<td>0.9913</td>
<td><strong>0.9864</strong></td>
</tr>
</tbody>
</table>

Figures in bold differ from the nominal level by more than simulation error.
Target coverage (red), coverage of $r^*$ (black), coverage of $r^*_B$ (dashes), as a function of interest parameter $\psi$. 
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

- Neutrino masses: $0 \leq \mu_L \leq \mu_M \leq \mu_H$ meV
- Two hierarchies: normal and inverted
- Measurements $y_1 = 75 \text{meV}^2$, $y_2 = 2524 \text{meV}^2$, and probability models
  - **normal hierarchy**:
    
    $y_1 \sim \mathcal{N}(\mu_M^2 - \mu_L^2, 1.8^2)$,
    
    $y_2 \sim \mathcal{N}\{\mu_H^2 - (\mu_M^2 + \mu_L^2)/2, 40^2}\}$,

  - **inverted hierarchy**:
    
    $y_1 \sim \mathcal{N}(\mu_H^2 - \mu_M^2, 1.8^2)$,
    
    $y_2 \sim \mathcal{N}\{(\mu_H^2 + \mu_M^2)/2 - \mu_L^2, 40^2}\}$

- 95% credible region: $P(\mu_L + \mu_M + \mu_H \leq 120 \text{meV}) = 0.95$, 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?
Normal hierarchy indicated if

\[ \mu_H^2 - \mu_M^2 > \mu_M^2 - \mu_L^2 \iff \psi = \mu_H^2 + \mu_L^2 - 2\mu_M^2 > 0, \]

so try and use higher order methods to test this.

As a warm-up, try testing non-nested hypotheses:

- Cox (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,

\[ T = 2 \left( \hat{\ell}_{\text{Normal}} - \hat{\ell}_{\text{Inverted}} \right), \]

with observed value \( t_{\text{obs}} = 1.748 \).

- \( T \sim \) normal with unknown mean and variance.

- We estimate the significance probabilities for testing the normal and inverted models,

\[ p_{\text{Normal}} = P_{\text{Normal}}(T \leq t_{\text{obs}}), \quad p_{\text{Inverted}} = P_{\text{Inverted}}(T \geq t_{\text{obs}}) \]

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

\[ p_{\text{Normal}} \approx 0.44, \quad p_{\text{Inverted}} \approx 0.55, \]

so the data do not distinguish the models.

- Equivalent to (two) bootstrap hypothesis tests (Davison and Hinkley, 1997, \textit{Bootstrap Methods and their Application}, Chapter 4).
Profile log likelihood for $\mu_L$

$2\ell_p(\mu_L)$ for normal hierarchy (black) and inverted hierarchy (red)
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: $\mu_L = 0.238$, $\mu_M = 8.665$, $\mu_H = 50.611$, $E(y_1) = 75.018$, $E(y_2) = 2523.90$,
- inverted model: $\mu_L = 0.009$, $\mu_M = 49.856$, $\mu_H = 50.603$, $E(y_1) = 75.000$, $E(y_2) = 2523.14$,

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 $P(\mu_L + \mu_M + \mu_H \leq 12\text{meV}) = 0.95$, 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.