High-order Asymptotic Expansions for Likelihood-based Statistics With Application to Testing for Signal Presence in Particle Physics Experiments

Basitha Hewa* & Igor Volobouev† & Alex Trindade*

(†) Dept. of Physics & Astronomy, Texas Tech University

(*) Dept. of Mathematics & Statistics, Texas Tech University

September 2020
- **Classical Setup:** Test a one-sided hypothesis for a single parameter via Likelihood Ratio, Score, and Wald tests.
- **Goal:** Accurate p-values.
- **Classical Solution:** Use asymptotics or simulation!
- **Non-Classical Setup:** Required Type I error rate is $\alpha \sim 10^{-7}$.
- **Solution 1 (asymptotic):** Derive **high-order** Edgeworth approximations to p-values.
- **Solution 2 (asymptotic):** Derive **high-order** saddlepoint approximations to p-values *(with new twists)*.
- Compare accuracies on simulated data *(bump-hunting expmts)*.
- Practical implementation: going beyond the toy problem...
Motivation: Discovery of Higgs Boson (The God Particle, Nobel Prize 2013)

![Graph showing CMS Preliminary data for Higgs Boson](image)
Model 1: Overall density is a mixture of signal $s(x)$ and background $b(x)$ densities:

$$p(x|\alpha) = \alpha s(x) + (1 - \alpha) b(x)$$

(1)

Signal fraction $\alpha$ based on IID sample $x_1, \ldots, x_n$ is estimated by maximizing the log-likelihood

$$\ell(\alpha) = \sum_{i=1}^{n} \log p(x_i|\alpha).$$

(2)

Leading to the MLE

$$\hat{\alpha} := \arg \max_{\alpha \in \mathbb{R}} \ell(\alpha)$$

(3)

Goal: produce accurate tests of $\mathcal{H}_0 : \alpha = 0$ vs. $\mathcal{H}_1 : \alpha > 0$.

Only unknown parameter is $\alpha \in \mathbb{R}$. 

Model 2: Sample size is not \textit{a priori} known, so treat data $x_1, \ldots, x_N$ as arising from a \textit{Poisson process} with intensity function:

$$\Lambda(x|\lambda) = \lambda s(x) + \mu b(x)$$ (4)

Signal fraction $\lambda$ is estimated by maximizing the log-likelihood

$$\ell(\lambda) = -(\lambda + \mu) + \sum_{i=1}^{N} \log \Lambda(x_i|\lambda)$$ (5)

Leading to the MLE

$$\hat{\lambda} := \arg \max_{\lambda \in \mathbb{R}} \ell(\lambda)$$ (6)

Goal: produce accurate tests of $\mathcal{H}_0 : \lambda = 0$ vs. $\mathcal{H}_1 : \lambda > 0$.

Only unknown parameter is $\lambda \in \mathbb{R}$.
Background is either standard Uniform or Exponential on $[0, 1]$: 

$$b(x) = \begin{cases} 
1, & \text{if } x \in [0, 1] \\
0, & \text{if } x \not\in [0, 1] 
\end{cases}, \quad b(x) = \begin{cases} 
e^{-x}/(1 - e^{-1}), & \text{if } x \in [0, 1] \\
0, & \text{if } x \not\in [0, 1] 
\end{cases}$$

Signal is truncated Gaussian on $[0, 1]$: 

$$s(x) = \begin{cases} 
e^{-\frac{(x-\mu)^2}{2\sigma^2}} / \int_0^1 e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy, & \text{if } x \in [0, 1] \\
0, & \text{if } x \not\in [0, 1] 
\end{cases}$$

Whenever specific settings of the signal are needed, we use 

$$\mu = 0.5, \quad \text{and} \quad \sigma = 0.1.$$ 

Toy problem! Everything known except mix proportion ($\alpha$ or $\lambda$). . .
For $\alpha$ and $\hat{\alpha}$ (similar statements hold for $\lambda$ and $\hat{\lambda}$ with $n \mapsto \mu$):

- $\ell_i(\alpha) = \partial^i \ell / \partial \alpha^i$, the $i$-th derivative of $\ell(\alpha)$
- $J(\alpha) = -\ell_2(\alpha)$
- **Expected information** number: $I(\alpha) = \mathbb{E}[J(\alpha)]$
- **Observed information** number: $J(\hat{\alpha}) = -\ell_2(\hat{\alpha})$

Assume usual regularity conditions for consistency and asymptotic normality of $\hat{\alpha}$ are satisfied:

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\alpha)^{-1}), \quad \mathcal{I}(\alpha) = \lim_{n \to \infty} \frac{1}{n} I(\alpha)$$

$$\implies \hat{\alpha} \sim \mathcal{N}(\alpha, \sigma_{\hat{\alpha}}^2(\alpha)), \quad \sigma_{\hat{\alpha}}^2(\alpha) = I(\alpha)^{-1}$$

(By not restricting $\alpha \in [0, 1]$ we avoid “exotic” asymptotics at the boundaries...
In lack of a UMP test, we have the following:

**Table:** Promising statistics for tests on $\alpha$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>$T_{LR}$</td>
<td>$2[\ell(\hat{\alpha}) - \ell(0)]$</td>
</tr>
<tr>
<td>Wald (Expected)</td>
<td>$T_W$</td>
<td>$\hat{\alpha}^2 I(0)$</td>
</tr>
<tr>
<td>Wald (Observed)</td>
<td>$T_{W2}$</td>
<td>$\hat{\alpha}^2 J(\hat{\alpha})$</td>
</tr>
<tr>
<td>Score</td>
<td>$T_S$</td>
<td>$\ell_1(0)^2 / I(0)$</td>
</tr>
<tr>
<td>Wald-type 3</td>
<td>$T_{W3}$</td>
<td>$\hat{\alpha}^2 / \sigma_3^2$</td>
</tr>
<tr>
<td>Wald-type 4</td>
<td>$T_{W4}$</td>
<td>$\hat{\alpha}^2 / \sigma_4^2$</td>
</tr>
</tbody>
</table>

The Wald-type 3 & 4 statistics are variants of $T_{W2}$ (used by physicists) that use shortcuts for computing $J(\alpha)$ so as to avoid differentiating $\ell(\alpha)$. 
For one-sided testing use *signed* version of any of the statistics (say $T$) in the Table:

$$R = \text{sgn}(\hat{\alpha}) \sqrt{T}.$$ 

Under $\mathcal{H}_0$, to first order $R \sim Z$, where $Z \sim \mathcal{N}(0, 1)$, whence

$$p\text{-value} = P(Z > r), \quad r = \text{sgn}(\hat{\alpha}) \sqrt{t}$$

In general, $R_n \sim Z$ to $k$-th order, means that

approx error $= R_n - Z = O_p(n^{-k/2})$

$$\Rightarrow P(R_n \leq r) = \Phi(r) + \frac{a_{1,n}}{n^{1/2}} + \frac{a_{2,n}}{n^1} + \frac{a_{3,n}}{n^{3/2}} + \cdots + \frac{a_{k-1,n}}{n^{(k-1)/2}} + O(n^{-k/2})$$
Tools for Higher-Order Asymptotic Theory (e.g., Severini, 2000)

- Taylor expansions of $\ell(\alpha)$ near true value of $\alpha$
- Joint cumulants for the derivatives of $\ell(\alpha)$ under $H_0$; in our case can express everything as a function of:
  \[ V_k = \mathbb{E}\ell_k(0), \quad k = 1, 2, \ldots \]

- Edgeworth-type series: approx pdf for $X \approx \mathcal{N}(\kappa_1, \kappa_2)$ via the Gram-Charlier series
  \[ f(x) = \frac{\phi(z)}{\sqrt{\kappa_2}} \left[ 1 + \sum_{k=3}^{\infty} \beta'_k H_k(z) \right], \quad z = \frac{x - \kappa_1}{\sqrt{\kappa_2}} \]

- $H_j(z)$ are the Hermite polynomials.
- Coefficients $\beta'_j$ are chosen to match cumulants $\kappa_j$ of $X$ (by inversion of its CGF $K(s) = \log \mathbb{E}\exp\{sX\}$).
CDF of $X$ obtained by integrating $f(x)$, grouping together terms in powers of $n^{-1/2}$, resulting in the Edgeworth expansion.

For a “typical” likelihood-based statistic we obtain

$$F(x) = \Phi(z) - \phi(z) \left[ \sum_{k=2}^{11} \beta_k H_k(z) + O(n^{-5/2}) \right], \quad z = \frac{x - \kappa_1}{\sqrt{\kappa_2}},$$

<table>
<thead>
<tr>
<th>Statistic</th>
<th>$R$</th>
<th>$R \neq R_{LR}$</th>
<th>$R = R_{LR}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>$R \neq R_{LR}$</td>
<td>$\frac{\kappa_3}{6}$</td>
<td>$\frac{\kappa_4}{24}$</td>
<td>$\kappa_5$</td>
</tr>
<tr>
<td>$R = R_{LR}$</td>
<td>$\frac{\kappa_3}{6}$</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
The challenge now is to express (approximate) the $\kappa_j$ (which are unknown) in terms of the $V_k$ (which can be computed)!!!

Has to be done case-by-case for each statistic $R$.

Start from suitable Taylor expansions in probability for $\hat{\alpha}$, the maximizer of $\ell(\alpha)$, and use some tricks...

Required **A LOT OF BOOKKEEPING** (20th century).

In 21st century this can be replaced with careful programming of a symbolic algebra system (Maple/Mathematica).

Above challenge has been worked out to 3rd order for classical statistics (LR, Wald, Score), by assuming $X \approx N(0, 1)$, so we:

- assumed $X \approx N(\kappa_1, \kappa_2)$ (gives greater accuracy), and
- worked out 5th order expansions for all statistics in Table 1.
• Represent the log-likelihood derivatives by

\[ \frac{d^k \ell(\alpha)}{d\alpha^k} \bigg|_{\alpha=0} = nV_k + \sqrt{n}Z_k, \quad \text{recall} \quad V_k := \mathbb{E}\ell_k(0) \]

and \( Z_k \) is an \( O_p(1) \) random variable with zero mean.

• Construct high order Taylor expansion for \( \ell(\alpha) \) at \( \alpha = 0 \), and solve for \( \hat{\alpha} \) in terms of \( V_k \) and \( Z_k \):

\[ \sqrt{n}\hat{\alpha} = \sum_{k=0}^{k_{\text{max}}} a_k(Z, V) n^{-k/2} + O_p(n^{-(k_{\text{max}}+1)/2}) \]

• The multivariate polynomials \( a_k(Z, V) \) are functions of \((Z_1, Z_2, \ldots)\) and \((V_1, V_2, \ldots)\).
• These polynomials are complicated but only need to be derived once (e.g., using symbolic computing).
• They do not depend on the model or statistic!
Example: For Score Statistic $R_S = Z_1 / \sqrt{-V_2}$

- Very simple, and holds for all orders of accuracy!!!
- All other statistics are more complicated...
- Makes it possible to analytically derive all cumulants.
- Cumulants depend only on following (dimensionless & location-scale invariant) expressions:

\[
\gamma = \frac{V_3}{2(-V_2)^{3/2}}, \quad \rho = -\frac{V_4}{6V_2^2}, \quad \xi = \frac{V_5}{24(-V_2)^{5/2}}, \quad \zeta = \frac{V_6}{120V_2^3}
\]

**Table:** Approximations to the first 6 cumulants of $R_S$ for the two models under consideration. The error in these approximations is $O(n^{-5/2})$ (Mixture model) or $O(\mu^{-5/2})$ (Poisson model).

<table>
<thead>
<tr>
<th>Model</th>
<th>$\hat{\kappa}_1$</th>
<th>$\hat{\kappa}_2$</th>
<th>$\hat{\kappa}_3$</th>
<th>$\hat{\kappa}_4$</th>
<th>$\hat{\kappa}_5$</th>
<th>$\hat{\kappa}_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixture</td>
<td>0</td>
<td>1</td>
<td>$\gamma / \sqrt{n}$</td>
<td>$(\rho - 3)/n$</td>
<td>$(\xi - 10\gamma)/n^{3/2}$</td>
<td>$(30 + \zeta - 10\gamma^2 - 15\rho)/n^2$</td>
</tr>
<tr>
<td>Poisson</td>
<td>0</td>
<td>1</td>
<td>$\gamma / \sqrt{\mu}$</td>
<td>$\rho / \mu$</td>
<td>$\xi / \mu^{3/2}$</td>
<td>$\zeta / \mu^2$</td>
</tr>
</tbody>
</table>
But Wait: why do we need higher-order asymptotics?

- When $n$ is small (not necessarily the case in these experiments).
- When **Type I error rate** ($q_o$) is very small..., how small?
- In “signal-hunting” particle physics experiments the gold standard is $5\sigma$:
  \[ q_0 = P(Z > 5) = 2.87 \times 10^{-7} \]
- This puts us way out in the tail of the $N(0, 1)$...
- (And is the reason why simulation is undesirable; to get 100 values exceeding $q_0$ requires $\sim 10^9$ runs!)
Quantifying Deviations From Normality

- 5th order Edgeworth-approx: $F_R(r) - F_R^{\text{edge}}(r) = O(n^{-5/2})$.
- Consider normal approx error

$$\Delta R(r) = r - \tilde{r}, \quad \tilde{r} = \Phi^{-1}(F_R^{\text{edge}}(r))$$

- Implies:

$$\tilde{r} = r \text{ to an accuracy of } O(n^{-5/2}) \text{ under } \mathcal{H}_0$$

- If $R$ is exactly $\mathcal{N}(0, 1)$:

$$\Delta R(r) = 0$$

- If $R$ differs greatly from $\mathcal{N}(0, 1)$:

large values of $\Delta R(r)$
Example: $\Delta R(r)$ for Poisson Model With Exp. Background (under $\mathcal{H}_0$)

- **LR Statistic**
- **Score Statistic**
- **Wald Statistic**
- **Poisson Model Exponential Background ($\mu = 20$)**

Legend:
- 2nd order Edgeworth
- 3rd order Edgeworth
- 4th order Edgeworth
- 5th order Edgeworth
- SPA
- Simulation
Example: P-values at $r = 5$ for Mixture Model With Unif. Background (under $\mathcal{H}_0$)

![LR Statistic Graph]

![Score Statistic Graph]

![Wald Statistic Graph]

![Mixture Model Uniform Background Graph]
SPA is an efficient “automatic” procedure to perform the inversion:

$$K(s) = \sum_{j=1}^{\infty} \frac{s^j}{j!} \kappa_j \quad \Rightarrow \quad F(x) = P(X \leq x)$$

(k + 1)-th order SPA for the CDF of $\tilde{X}_n$ (Daniels, 1987):

$$\hat{F}_{n,k}(x) = \Phi(\hat{w}\sqrt{n}) - \phi(\hat{w}\sqrt{n}) \left[ \frac{c_0}{n^{1/2}} + \frac{c_1}{n^{3/2}} + \cdots + \frac{c_k}{n^{k+1/2}} \right]$$

The (asymptotic) truncation error of $\hat{F}_{n,k}(x)$ is:

$$\frac{\hat{F}_{n,k}(x)}{F(x)} = 1 + \mathcal{O}(n^{-k-3/2}) \quad \iff \quad F(x) - \hat{F}_{n,k}(x) = \mathcal{O}(n^{-k-3/2})$$
Saddlepoint vs. Edgeworth Approximation (SPA vs. Edge)

- Since we have \( \{\hat{\kappa}_1, \ldots, \hat{\kappa}_6\} \) for \( R \) (Table 3), SPA with \( n = 1 \) is an alternative to the Edgeworth approximation of p-values.
- Starting with \( \hat{K}_m(s) = \sum_{j=1}^{m} \hat{\kappa}_j s^j / j! \), we note from Figs 1 & 2 that 5th order Edge and SPA give same \( F(r) \)...

**Theorem**

Let \( \hat{G}_{1,k}(x) \) be estimated \( \hat{F}_{1,k}(x) \) by using \( \hat{K}_m(s) = \sum_{j=1}^{m} \hat{\kappa}_j s^j / j! \) with \( \hat{\kappa}_j = \kappa_j + O_p(n^{-\alpha}) \). Then:

\[
\frac{\hat{G}_{1,k}(x)}{F(x)} = 1 + O_p(n^{-\min\{\alpha,(m-1)/2,k+3/2\}})
\]

- In our case, \( m = 6 \) and \( \alpha = 5/2 \), so if we take \( k = 1 \), we get:

\[
\frac{\hat{G}_{1,1}(x)}{F(x)} = 1 + O_p(n^{-5/2})
\]
Thus with $\hat{K}_6(s)$, both Edge and SPA give same 5th order estimated $F(r)$... provided CGF is **convex**!

**Edge:** doesn’t care about convexity, but requires **new** painstaking analytical computations as $m$ changes...

**SPA:** remains essentially the **same** as $m$ changes, but CGF must be convex...

**Idea:** convexify CGF by doubling number of cumulants:

$\{\hat{\kappa}_1, \ldots, \hat{\kappa}_6\} =$ approx cumulants on hand (rest are $O(n^{-5/2})$)

$\{\kappa_7, \ldots, \kappa_{12}\} =$ solve for these by minimizing

$$\sum_{j=7}^{12} \kappa_j^2$$

subject to convexity (**quadratic programming**)
Example: P-values at $r = 5$ for Mixture Model With Unif. Background (under $\mathcal{H}_0$)

- **CGFs: Wald, Mixture, Uniform, n = 20**
  - $K(s)$
  - Asymptotic $N(0,1)$
  - Deg. 6 approx
  - Deg. 12 approx (convex)

- **Delta(R): Wald, Mixture, Uniform (n=20)**
  - Edgeworth 5th order
  - Simulation

- **CGFs: Score, Mixture, Uniform, n = 20**
  - $K(s)$
  - Asymptotic $N(0,1)$
  - Deg. 6 approx
  - Deg. 12 approx (convex)

- **Delta(R): Score, Mixture, Uniform (n=20)**
  - SPA 6 cumulants
  - SPA 12 cums convex
  - Edgeworth 5th order
  - Simulation
SPA vs. Edge: if both use same $\hat{K}_6(s)$, and it’s convex, then

$$F_{R}^{\text{edge}}(r) \approx F_{R}^{\text{spa}}(r).$$

If CGF not convex, then SPA can easily be “fixed”, whereas Edge may give results of dubious quality...

SPA CDF: guaranteed to be positive; Edge can be negative...

Score statistic has a very simple asymptotic expansion, which makes it (relatively) easy to derive any number of (estimated) cumulants!

Application of SPA to these instances is immediate, whereas Edge requires substantial analytical effort!!!

(Question: Is it possible to combine these good properties of Score statistic with efficiency of LR statistic into a new statistic?)
**Doable:** $s(x) \& b(x) \mapsto b(x|\phi)$.

- extend everything we have done to the nuisance parameter setting (multivariate Edge/SPA).

**Problem:** $s(x) \mapsto s(x|\theta)$ means $\theta$ is not identifiable under $\mathcal{H}_0$:
- classical inference for treating nuisance parameters then breaks down...
- Davies (Biometrika, 1987): appropriate p-value is an excursion probability
  \[
  p\text{-value} = P\left(\max_{\theta \in \Theta} R(\theta) > c\right)
  \]

**Theory of Random Fields (TRF):** emerged as only analytical solution so far (large-scale searches in neuroimaging, astrophysics, etc.)

- $R(\theta)$ is viewed as Gaussian random field over manifold $\Theta \subset \mathbb{R}^d$
- $\phi$ has been profiled out of $R(\theta, \phi) : \phi \mapsto \hat{\phi}$
- provides closed-form approximaton when $c$ is large...
Excursion set of field above level $c$:

$$\mathcal{A}_c = \{ \theta \in \Theta : R(\theta) > c \}$$

Euler characteristic of excursion set:

$$\phi(\mathcal{A}_c) = \text{geometric property of field}$$

Fundamental result in TRF:

$$\mathbb{E}[\phi(\mathcal{A}_c)] = \sum_{i=0}^{d} a_i f_i(c)$$

- $a_i$: positive constants (to be determined by Monte Carlo)
- $f_i(\cdot)$: known “universal” functions

For large $c$: (Taylor et al., *Annals of Probability*, 2005)

$$p\text{-value} = P(\max_{\theta \in \Theta} R(\theta) > c) \approx \mathbb{E}[\phi(\mathcal{A}_c)] \equiv p_{\text{global}}$$
Suppose $\theta \neq 0$ & $\phi \neq 0$

- **Solution 1 (straightforward):** treat all parameters via TRF in conjunction with Edge/SPA $O(n^{-5/2})$ normalized versions of LR statistic

$$ r \mapsto \tilde{r} = \Phi^{-1}(\hat{F}_{RLR}(r)) $$

- **Solution 2 (exotic):** adjust global significance of test statistic, leading to (conservative) estimate of $p_{global}$ in context of TRF...

$$ p_{global} = P(R_{LR}(\hat{\theta}) > r(\hat{\theta})) $$

- $r(\theta)$ is observed (local) value of $R_{LR}(\theta)$ computed from sample,
- $\hat{\theta} = \text{arg max}_{\theta \in \Theta} r(\theta)$. 

(Volobouev, I. & Trindade, A., JINST, 2018)
Algorithm for Solution 2: details...

- Normal approx error for each observed (local) \( r \equiv r(\theta) \) as before:
  \[
  \Delta R(r(\theta)) = r(\theta) - \tilde{r}(\theta)
  \]

- Locate:
  \[
  \theta^* = \arg \max_{\theta \in \Theta} \Delta R(r(\theta))
  \]

  - Search can use same grid as TRF search for \( \hat{\theta} = \arg \max r(\theta) \).

- Calculate **global significance** of signal \( p_{\text{global}} \) via TRF, and express it in terms of the **global** \( r \):
  \[
  r_{\text{global}} = \Phi^{-1}(1 - p_{\text{global}})
  \]

- Adjust global \( r \):
  \[
  r_{\text{global}}^{\text{adj}} = r_{\text{global}} - \Delta R(r(\theta^*))
  \]

- Global (adjusted) \( p \)-value is then:
  \[
  p_{\text{global}}^{\text{adj}} = 1 - \Phi(r_{\text{global}}^{\text{adj}})
  \]


THE END!