Improved Inference for Signal Discovery Under Exceptionally Low False Positive Error Rates

(to appear in *Journal of Instrumentation*)

Igor Volobouev & Alex Trindade

Dept. of Physics & Astronomy, Texas Tech University

Dept. of Mathematics & Statistics, Texas Tech University

December 2018

Model is signal/background density mixture without nuisance parameters:

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

Signal fraction α is estimated by maximizing the log-likelihood:

$$\ell(\alpha) = \sum_{i=1}^{n} \log p(x_i|\alpha). \tag{2}$$

$$\hat{\alpha} := \arg\max_{\alpha \in \mathbb{R}} \ell(\alpha) \tag{3}$$

- Goal: produce accurate tests of $\mathcal{H}_0 : \alpha = 0$ vs. $\mathcal{H}_1 : \alpha > 0$.
- Only unknown parameter is α; a toy problem...
- But we'll make it more realistic at the end.

Let $b(x)$ follow a uniform distribution on $[0,1]$, and $s(x)$ a truncated Gaussian on $[0,1]$:

$$b(x) = \begin{cases} 1, & \text{if } x \in [0,1] \\ 0, & \text{if } x \notin [0,1] \end{cases}, \tag{4}$$

$$s(x) = \begin{cases} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Big/ \int_0^1 e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy , & \text{if } x \in [0,1] \\ 0, & \text{if } x \notin [0,1] \end{cases}. \tag{5}$$

This will be the model used in simulations, and whenever specific settings of the signal are needed, we use

$$\mu = 0.5, \qquad \text{and} \qquad \sigma = 0.1. \tag{6}$$

- $\ell_i(\alpha) = \partial^i \ell / \partial \alpha^i$, 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:

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

$$\Longrightarrow \hat{\alpha} \dot{\sim} \mathcal{N}(\alpha, \sigma^2_{\hat{\alpha}}(\alpha)), \qquad \sigma^2_{\hat{\alpha}}(\alpha) = I(\alpha)^{-1}$$

(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 α.

Method	Statistic	Value
Likelihood Ratio	T_{LR}	$2[\ell(\hat{\alpha}) - \ell(0)]$
Wald (Expected)	T_{W}	$\hat{\alpha}^2 I(0)$
Wald (Observed)	T_{W2}	$\hat{\alpha}^2 J(\hat{\alpha})$
Score	T_{S}	$\ell_1(0)^2 / I(0)$
Wald-type 3	T_{W3}	$\hat{\alpha}^2 / \sigma_3^2$
Wald-type 4	T_{W4}	$\hat{\alpha}^2 / \sigma_4^2$

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
$$p\text{-value} \;= P(Z > r), \qquad r = \text{sgn}(\hat{\alpha})\sqrt{t}$$
- In general, $R_n \sim Z$ to k-th order, means that
$$\text{approx error} = R_n - Z = O_p(n^{-k/2})$$

$$\Rightarrow P(R_n \leq r) = \Phi(r) + \frac{a_{1,n}}{n^{1/2}} + \frac{a_{2,n}}{n^1} + \frac{a_{3,n}}{n^{3/2}} + \cdots + \frac{a_{k-1,n}}{n^{(k-1)/2}} + O(n^{-k/2})$$

- Taylor expansions of $\ell(\alpha)$ near true value of α
- Joint cumulants for the derivatives of $\ell(\alpha)$ under $\mathcal{H}_0$:

$$n\nu_{ijkl} = (i, j, k, l)\text{-th joint cumulant of } \{\ell_1(0), \ldots, \ell_4(0)\}$$

- Edgeworth-type series: construct an approximate pdf for R (which is approx $\mathcal{N}(0,1)$) via the Gram-Charlier expansion:

$$f_R(z) = \phi(z)\left(1 + \sum_{j=1}^{\infty} \beta_j H_j(z)\right),$$

- $H_j(z)$ are the Hermite polynomials.
- Coefficients β_j are chosen to match the cumulants κ_j of R_n.

Integrate Gram-Charlier expansion and collect terms in powers of $n^{-1/2}$:

$$\begin{aligned} F_R(z) &= \Phi(z) - \phi(z)\Bigg[\kappa_1 + \frac{1}{6}\kappa_3 H_2(z) \\ &\quad + \frac{1}{2}(\kappa_1^2 + \kappa_2 - 1)z + \left(\frac{1}{6}\kappa_1\kappa_3 + \frac{1}{24}\kappa_4\right) H_3(z) \\ &\quad + \frac{1}{72}\kappa_3^2 H_5(z) + O(n^{-3/2})\Bigg]. \end{aligned} \tag{7}$$

(There are some technical assumptions on the cumulant behavior of R...)

- The above expression for $F_R(z)$ holds for any statistic R which is approx $\mathcal{N}(0,1)$.
- The challenge is to be able to express (approximate) the κ_j (which are unknown) in terms of the ν_{ijkl} (which can be computed)!!!
- Has to be done case-by-case for each statistic R: start from suitable Taylor expansions in probability, 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).
- The relationships in the above challenge have been worked out (to 3rd order) for the classical statistics (LR, Wald, Score), so we:
  - corrected some typos in the existing expressions (Severini, 2000),
  - worked out 4th order expansions for the classical statistics, and
  - worked out 3rd order expansions for the non-classical statistics.

- $\mathbb{E}$ denotes "expectation under the null": $\mathbb{E}[q] := \int q(x)b(x)dx$.
- $\mathbb{E}_s$ denotes "expectation under the signal": $\mathbb{E}_s[q] := \int q(x)s(x)dx$.
- Defining $V_i := \mathbb{E}\ell_i(0)$, the Edgeworth expansions for all the statistics in Table 1 depend only on the following (dimensionless & location-scale invariant) expressions:

$$\gamma := \frac{V_3}{2\left(-V_2\right)^{3/2}} = \frac{\mathbb{E}_s\left[\frac{s^2}{b^2}\right] - 3\mathbb{E}_s\left[\frac{s}{b}\right] + 2}{\left(\mathbb{E}_s\left[\frac{s}{b}\right] - 1\right)^{3/2}}, \tag{8}$$

$$\rho := -\frac{V_4}{6V_2^2} = \frac{\mathbb{E}_s\left[\frac{s^3}{b^3}\right] - 4\mathbb{E}_s\left[\frac{s^2}{b^2}\right] + 6\mathbb{E}_s\left[\frac{s}{b}\right] - 3}{\left(\mathbb{E}_s\left[\frac{s}{b}\right] - 1\right)^2}. \tag{9}$$

Plugging the (thus) approximated $\hat{\kappa}_j$ into $F_R(z)$ in (7), gives, e.g.

Signed Wald statistic:

$$P(R_W \leq z) = \Phi(z) - \phi(z)\left[n^{-1/2}\left(\frac{1}{6}\gamma H_2(z)\right) + n^{-1}\left(\frac{1}{2}(\rho - \gamma^2 - 1)z + \frac{1}{24}(\rho - 3)H_3(z) + \frac{\gamma^2}{72}H_5(z)\right) + \mathcal{O}(n^{-3/2})\right].$$

Signed likelihood ratio statistic:

$$P(R_{LR} \leq z) = \Phi(z) - \phi(z)\left[n^{-1/2}\left(-\frac{\gamma}{6}\right) + n^{-1}\left(\frac{1}{12}(3\rho - 2\gamma^2)z\right) + \mathcal{O}(n^{-3/2})\right].$$

- When n is small (typically not the case in these experiments).
- When Type I error rate (q_1) is very small..., how small?
- In "signal-hunting" particle physics experiments the gold standard is 5σ:

$$q_0 = P(Z > 5) = 2.87 \times 10^{-7}$$

- This puts us way out in the tail of the $\mathcal{N}(0, 1)$...

- Consider normal approx error

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

- With Edgeworth-approx $F_R(\cdot)$:

$$\tilde{r} = r \text{ to an accuracy of } O(n^{-3/2}) \text{ under } \mathcal{H}_0$$

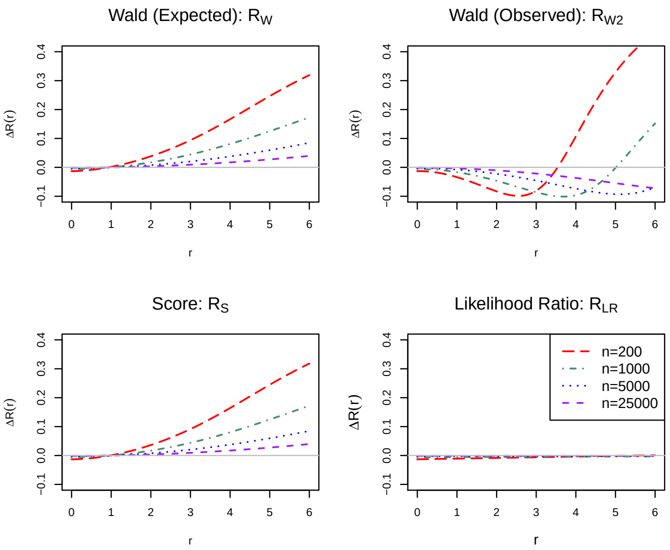
- If R is exactly $\mathcal{N}(0,1)$:

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

- Large values of $\Delta R(r)$:

Edgeworth-approx had a large effect in "normalizing" R

Ex: $\Delta R(r)$ for Flat Background With Gaussian Signal



- $\mu = 0.5$ and $\sigma = 0.1$, implies $\gamma \approx 1.1094$ and $\rho \approx 2.6893$.
- Wald statistic at $r = 5$ for $n = 200$:

$$\Delta R_{\mathrm{W}}(r) = 0.25$$

  - means p-value is wrong by factor of $P(Z > 5)/P(Z > 5.25) \approx 3.8$
  - higher signal significance will be claimed than supported by data.
- LR statistic at $r = 4$ for $n = 200$:

$$\Delta R_{\mathrm{LR}}(r) = -0.005$$

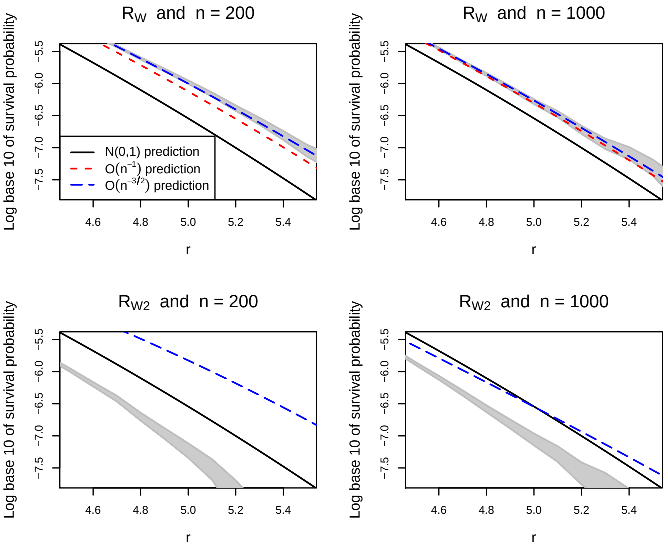
  - means p-value is wrong by factor of $P(Z > 4)/P(Z > 3.995) \approx 0.98$
  - reported signal significance will be about right.

- $m = 10^9$ Monte Carlo replicates for $n = 200, 1000, 5000, 25000$.
- Compared distributions of R with Edgeworth-predictions.
- For 3 classical stats (R_{W}, R_{LR}, R_{S}): distributional shape parameters (mean, standard deviation, skewness, kurtosis) are in good agreement with $\mathcal{O}(n^{-3/2})$ predictions for $n \geq 1000$.
- The agreement worsens for $n = 200$...
- Statistically significant disagreements with $\mathcal{N}(0, 1)$ predictions are in red values in next Table.

Simulations: Predicted Mean (Equals 0 if Exactly $\mathcal{N}(0,1)$)

	n	$O(n^{-3/2})$ Prediction	Simulated Value	Simulation Uncertainty
R_W	200	0	-9.0×10^{-5}	3.2×10^{-5}
	1000	0	-0.9×10^{-5}	
	5000	0	-1.6×10^{-5}	
	25000	0	4.8×10^{-5}	
R_{W2}	200	-39.222×10^{-3}	-40.724×10^{-3}	0.032×10^{-3}
	1000	-17.541×10^{-3}	-17.670×10^{-3}	
	5000	-7.844×10^{-3}	-7.872×10^{-3}	
	25000	-3.508×10^{-3}	-3.461×10^{-3}	
R_{W3}	200	-39.222×10^{-3}	-40.817×10^{-3}	0.032×10^{-3}
	1000	-17.541×10^{-3}	-17.678×10^{-3}	
	5000	-7.844×10^{-3}	-7.872×10^{-3}	
	25000	-3.508×10^{-3}	-3.461×10^{-3}	
R_{W4}	200	-50.617×10^{-3}	-48.508×10^{-3}	0.032×10^{-3}
	1000	-19.820×10^{-3}	-19.106×10^{-3}	
	5000	-8.300×10^{-3}	-8.152×10^{-3}	
	25000	-3.599×10^{-3}	-3.516×10^{-3}	
R_{LR}	200	-13.074×10^{-3}	-13.199×10^{-3}	0.032×10^{-3}
	1000	-5.847×10^{-3}	-5.860×10^{-3}	
	5000	-2.615×10^{-3}	-2.631×10^{-3}	
	25000	-1.169×10^{-3}	-1.121×10^{-3}	
R_S	200	0	2.7×10^{-5}	3.2×10^{-5}
	1000	0	-1.6×10^{-5}	
	5000	0	1.3×10^{-5}	
	25000	0	0.1×10^{-5}	

Simulations: Predicted Survival Probabilities (R_W & R_{W2})



Simulations: Predicted Survival Probabilities (R_{LR} & R_S)

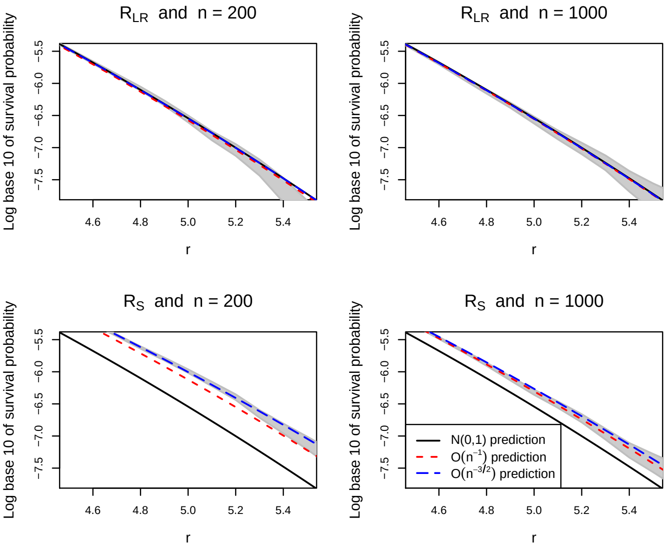


Table: $O(n^{-3/2})$ Edgeworth predictions for the $(1 - q_0)$-th quantiles of the statistics in Table 1 ($\pm$ std. err.), compared to their corresponding values computed based on 10^9 simulations. (Predictions deviating by more than twice the std. err. from their simulated values are bolded.)

Statistic	Method	$n = 200$	$n = 1000$	$n = 5000$	$n = 25000$
R_{W}	Predicted	5.267 $\pm$ 0.012	5.131 $\pm$ 0.012	5.061 $\pm$ 0.012	5.028 $\pm$ 0.012
	Simulated	5.286	5.123	5.046	5.028
R_{W2}	Predicted	**5.392 $\pm$ 0.013**	**4.998 $\pm$ 0.013**	**4.908 $\pm$ 0.011**	4.946 $\pm$ 0.011
	Simulated	4.736	4.790	4.872	4.943
R_{W3}	Predicted	**5.394 $\pm$ 0.013**	**5.000 $\pm$ 0.013**	**4.908 $\pm$ 0.011**	4.946 $\pm$ 0.011
	Simulated	4.744	4.791	4.872	4.944
R_{W4}	Predicted	**5.292 $\pm$ 0.014**	**4.945 $\pm$ 0.013**	**4.918 $\pm$ 0.011**	4.956 $\pm$ 0.011
	Simulated	4.755	4.822	4.893	4.954
R_{LR}	Predicted	4.999 $\pm$ 0.011	4.996 $\pm$ 0.011	4.998 $\pm$ 0.011	4.999 $\pm$ 0.011
	Simulated	4.990	4.988	4.983	4.998
R_{S}	Predicted	5.265 $\pm$ 0.012	5.131 $\pm$ 0.012	5.061 $\pm$ 0.012	5.028 $\pm$ 0.012
	Simulated	5.259	5.121	5.050	5.026

Table: Type I error probabilities: rejection is based on the $O(n^{-3/2})$ Edgeworth-predicted quantiles from previous Table. (Values that deviate by more than twice the simulation uncertainty of 0.17×10^{-7} from the nominal value are bolded.)

n	R_{W}	R_{W2}	R_{W3}	R_{W4}	R_{LR}	R_{S}
200	3.03×10^{-7}	$\mathbf{0.07 \times 10^{-7}}$	$\mathbf{0.07 \times 10^{-7}}$	$\mathbf{0.13 \times 10^{-7}}$	2.68×10^{-7}	2.73×1
1000	2.71×10^{-7}	$\mathbf{0.78 \times 10^{-7}}$	$\mathbf{0.78 \times 10^{-7}}$	$\mathbf{1.35 \times 10^{-7}}$	2.77×10^{-7}	2.72×1
5000	2.65×10^{-7}	$\mathbf{2.44 \times 10^{-7}}$	$\mathbf{2.44 \times 10^{-7}}$	2.56×10^{-7}	2.72×10^{-7}	2.77×1
25000	2.87×10^{-7}	2.82×10^{-7}	2.82×10^{-7}	2.83×10^{-7}	2.83×10^{-7}	2.86×1

- Since all statistics have same asymptotics as MLE (to 1st order), when truth is $\alpha = \alpha_1$, we have

$$R_n \dot{\sim} N(\alpha_1, \sigma^2_{\hat{\alpha}}(\alpha_1))$$

- Let c_n be the Edgeworth-predicted quantiles from Table such that

$$P_{\alpha=0}(R_n > c_n) = q_0, \qquad \Rightarrow c_n \to 5\sigma_{\hat{\alpha}}(0)$$

- Then the power is:

$$1 - \beta \quad = \quad P_{\alpha=\alpha_1}(R_n > c_n) = P\left(Z > \frac{c_n - \alpha_1}{\sigma_{\hat{\alpha}}(\alpha_1)}\right)$$

- Thus choosing $\alpha_1 = 5\sigma_{\hat{\alpha}}(0)$ we should have as $n \to \infty$:

$$1 - \beta \longrightarrow P\left(Z > 0\right) = 0.5$$

- (Keeps difficulty in finding signal approx constant...)

Table: Values of the Cramer-Rao uncertainty for $\mathcal{H}_0$, $\sigma_{\hat{\alpha}}$, and corresponding values of $\alpha = \alpha_1$ used as the actual model signal fraction under $\mathcal{H}_1$.

n	$\sigma_{\hat{\alpha}}$	$\alpha_1 = 5\sigma_{\hat{\alpha}}$
200	5.2401×10^{-2}	0.26200
1000	2.3434×10^{-2}	0.11717
5000	1.0480×10^{-2}	0.05240
25000	0.4687×10^{-2}	0.02343

Table: Type II error probabilities (determined empirically), using the predicted & simulated quantiles from Table. The smallest predicted value at each n is in bold. (Simulation uncertainty $\approx 2 \times 10^{-5}$.)

R	Method	Sample Size (n) 200	1000	5000	25000
R_W	Predicted	0.59592	0.55021	0.52433	0.51130
	Simulated	0.60241	0.54701	0.51848	0.51143
R_{W2}	Predicted	0.80891	0.63069	0.53250	0.51192
	Simulated	0.58540	0.54601	0.51775	0.51074
R_{W3}	Predicted	0.80695	0.63082	0.53257	0.51193
	Simulated	0.58535	0.54601	0.51775	0.51074
R_{W4}	Predicted	0.77782	0.59609	0.52830	0.51156
	Simulated	0.58515	0.54577	0.51778	0.51074
R_{LR}	Predicted	**0.58873**	**0.54827**	**0.52385**	**0.51119**
	Simulated	0.58533	0.54501	0.51796	0.51074
R_S	Predicted	0.59510	0.55011	0.52432	0.51129
	Simulated	0.59296	0.54672	0.52024	0.51059

- Often remarked in math-stat books...
- Mykland (1999): proves that k-th cumulant of R_{LR} vanishes to $O(n^{-k/2})$ for all $k \geq 3$
  - $\kappa_3 = 0$ to $O(n^{-3/2})$ (but $\kappa_j \neq 0$ for $j \leq 2$)
  - $\kappa_4 = 0$ to $O(n^{-2})$ (but $\kappa_j \neq 0$ for $j \leq 3$)
  - etc.
- Mykland speculates: this fact "...would seem to be the main asymptotic property governing the accuracy behavior..." of R_{LR}.
- Why? Because the high-order cumulants are precisely the coefficients of the highest degree $H_k(\cdot)$ in the Edgeworth exp...

- Doable: $s(x)$ & $b(x) \mapsto b(x|\phi)$.
  - extend everything we have done in nuisance parameter setting (multivariate Edgeworth exp.).
- Problem: $s(x) \mapsto s(x|\boldsymbol{\theta})$ means $\boldsymbol{\theta}$ 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**

$$\text{p-value} = P(\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} R(\boldsymbol{\theta}) > c)$$

  - Theory of Random Fields (TRF): emerged as only **analytical** solution so far (large-scale searches in neuroimaging, astrophysics, etc.)
    - $R(\boldsymbol{\theta})$ is viewed as Gaussian random field over manifold $\boldsymbol{\Theta} \subset \mathbb{R}^d$
    - ϕ has been profiled out of $R(\boldsymbol{\theta}, \phi) : \phi \mapsto \hat{\phi}$
    - provides closed-form approximaton when c is large...

- Excursion set of field above level c:

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

- **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)

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

(THIS IS THE MAIN CONTRIBUTION OF THIS PAPER!)

Suppose $\boldsymbol{\theta} \neq 0$ & $\boldsymbol{\phi} \neq 0$

- Solution 1 (straightforward): treat **all** parameters via TRF in conjuction with Edgeworth $O(n^{-3/2})$ normalized versions of LR statistic

$$r \mapsto \tilde{r} = \Phi^{-1}(F_R(r))$$

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

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

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

- 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_{global} via TRF, and express it in terms of the **global** r:

$$r_{global} = \Phi^{-1}(1 - p_{global})$$

- Adjust global r:

$$r^{adj}_{global} = r_{global} - \Delta R(r(\theta^*))$$

- Global (adjusted) p-value is then:

$$p^{adj}_{global} = 1 - \Phi(r^{adj}_{global})$$

- Suppose $x_1, \ldots, x_N$ iid $\sim p(x|\alpha)$ and $N \sim \mathcal{P}(\nu(\alpha))$, then have two cases:
  - if $\nu(\alpha)$ **independent** of α so that $\nu(\alpha) \equiv \nu$, inference on α is as before

$$\ell(\alpha, \nu) = \sum_{i=1}^{n} \log p(x_i|\alpha) - \nu + n \log \nu = \ell(\alpha) + \ell(\nu)$$

  - **otherwise**, log-likelihood is not separable, but under a simplified regime where $\alpha = \nu_s/(\nu_s + \nu_b)$ with ν_b known, we have

$$\ell(\nu_s) = \sum_{i=1}^{n} \log p(x_i|\alpha) - (\nu_s + \nu_b) + n \log(\nu_s + \nu_b)$$

  and can now re-do all calcs for the new parameter ν_s...

THE END!