Improved inference for the signal significance

I. Volobouev$^{a,1}$ and A. Trindade$^b$

$^a$Department of Physics & Astronomy, Texas Tech University, Lubbock, TX, 79409-1051, U.S.A.

$^b$Department of Mathematics & Statistics, Texas Tech University, Lubbock, TX, 79409-1042, U.S.A.

E-mail: i.volobouev@ttu.edu

ABSTRACT: We study the properties of several likelihood-based statistics commonly used in testing for the presence of a known signal under a mixture model with known background, but unknown signal fraction. Under the null hypothesis of no signal, all statistics follow a standard normal distribution in large samples, but substantial deviations can occur at practically relevant sample sizes. Approximations for respective $p$-values are derived to various orders of accuracy using the methodology of Edgeworth expansions. Adherence to normality is studied, and the magnitude of deviations is quantified according to resulting $p$-value inflation or deflation. We find that approximations to third-order accuracy are generally sufficient to guarantee $p$-values with nominal false positive error rates in the 5$\sigma$ range ($p$-value $= 2.87 \times 10^{-7}$) for the classic Wald, score, and likelihood ratio (LR) statistics at relatively low sample sizes. Not only does LR have better adherence to normality, but it also consistently outperforms all other statistics in terms of false negative error rates. The reasons for this are shown to be connected with high-order cumulant behavior gleaned from fourth order Edgeworth expansions. Finally, a conservative procedure is suggested for making finite sample adjustments while accounting for the look elsewhere effect via the theory of random fields.

KEYWORDS: Analysis and statistical methods; Data processing methods

ArXiv ePrint: 1609.00752
1 Introduction

In particle physics, a typical search for a new particle or resonance consists in studying invariant masses of decay products and identifying a narrow, localized signal spike on top of a slowly varying background. If the main purpose of the study is determination of signal existence, the mass of the particle enters the analysis as the model nuisance parameter. When it can be expected that the natural width of the particle under search is small, one can usually assume that the shape of the hypothesized signal is Gaussian and that its width is determined by the detector resolution. On the other hand, if the expected width is comparable to or substantially larger than the resolution, it has to be treated as another nuisance parameter in the search.

Reliable determination of the frequentist statistical significance of signal evidence, i.e., the $p$-value, is crucial for substantiating any claim of discovery. The $p$-value is derived by estimating how often the statistical fluctuations in the background alone can produce an apparent signal excess similar to or larger than that observed in the data. However, the parameters of the signal, such as its location and width, are no longer identifiable in a model which represents pure background. As
the Wilks’ theorem [1] does not hold in such situations, the classical nuisance parameter treatment
based on profile likelihood no longer results in a simple asymptotic behavior of the $p$-value [2].

In particle physics, the problem of proper accounting for the false positive signal detection
over a wide search area is referred to as the look elsewhere effect (LEE) [3]. While this effect and
its associated “trial factor” (i.e., the increase in the $p$-value in comparison with the case of fixed
nuisances) can be estimated numerically, reliable direct determination of the $p$-value by computer
simulations is often prohibitively expensive in terms of required CPU time. Fortunately, an accurate
asymptotic treatment of this problem is possible with the aid of techniques developed within the
theory of random fields [4, 5]. The statistic used to test for signal presence is associated with
a random field in the space formed by nuisance parameters. The probability that the field maximum
exceeds some level depends on the postulated geometry of the parameter space and on the field
covariance function. It turns out that, if a sufficiently high level is chosen, the asymptotic behavior of
the $p$-value is determined by a few coefficients only,\footnote{These are the coefficients that enter the asymptotic expansion of the Euler characteristic of the field excursion set.} and full knowledge of the covariance function
is not necessary. These coefficients can be evaluated numerically at a much reduced computational
cost in comparison with the brute-force $p$-value determination. In application to particle physics
searches, this approach has become known as the Gross-Vitells Method (GVM) [6]; see also [7].

Determination of the LEE trial factor via GVM relies upon the crucial assumption that the
pointwise distributions of the fields are known. The relevant theory is developed for centered
Gaussian fields with unit variance and a few related fields, such as $\chi^2$. However, for finite samples,
departures of the signal testing statistic from normality can be substantial. To ensure applicability of
the LEE theory, we must assess how well its assumptions are satisfied and constrain deviations from
these assumptions. With this purpose in mind, we investigate finite sample effects for a number of
statistics that can be used to claim signal discovery.

The specific problem we wish to address is accurate testing for the presence of a signal
fraction, $\alpha$, in the context of a model involving an additive combination of signal and background
distributions. In what follows, it will be assumed that both the signal and the background are
described by known probability density functions, $s(x)$ and $b(x)$ respectively, for some univariate or
multivariate $x$, and that the data are fitted by maximum likelihood to the (mixture) model density:

$$p(x|\alpha) = \alpha s(x) + (1 - \alpha) b(x).$$  \hfill (1.1)

Here, $\alpha$ is the parameter of interest, to be inferred from available data. For searches with unknown
signal location and/or width, this model corresponds to a single point in the nuisance parameter
space. If the observed sample points $x_1, \ldots, x_n$ consist of independent and identically distributed
realizations from model (1.1), the log-likelihood is

$$\ell(\alpha) = \sum_{i=1}^n \log p(x_i|\alpha).$$  \hfill (1.2)

The maximum likelihood estimator (MLE) $\hat{\alpha}$ is obtained by maximizing $\ell(\alpha)$ over all $\alpha \in \mathbb{R}$. It is
easily seen that $\ell(\alpha)$ is strictly concave under continuous background and signal models, and thus
$\hat{\alpha}$ is straightforward to locate numerically. The overall goal is to produce accurate tests of

$$\mathcal{H}_0 : \alpha = 0 \quad \text{vs.} \quad \mathcal{H}_1 : \alpha > 0.$$  \hfill (1.3)
In the statistics literature, (1.1) is known as a \textit{mixture model} [8]. Asymptotically (i.e., for $n \to \infty$), $\hat{\alpha}$ is consistent, normal, and efficient under mild regularity conditions.\footnote{It should be pointed out that in this study we do not impose the $\alpha \in [0, 1]$ constraint which is sometimes adopted in high energy physics applications [9]. In practice, this constraint would not lead to a noticeable improvement in the result interpretability. At the same time, it would introduce an unnecessary complication in various derivations by breaking the likelihood regularity conditions at the boundaries and by adding a point mass at 0 to the $\hat{\alpha}$ distribution.} In practice, however, we must maintain confidence in the statements made for finite $n$. Deviations from normality can be assessed analytically through so-called “higher-order” statistical inference techniques [10]. In this manuscript, we develop higher-order approximations for model (1.1) with the purpose to both provide such an assessment and improve frequentist coverage of the relevant tests. As outlined in section 7, adjustments made to the local significance of the test statistic can also be translated into a subsequent conservative estimate of the global $p$-value.

If $b(x)$ has nuisance parameters, the likelihood is normally profiled over these parameters. Although we do not specifically tackle it here, application of the techniques detailed in [10] is also possible in this situation, leading to suitably modified versions of the asymptotic expansions we derive in the current paper.\footnote{However, the required calculations are substantially more challenging.} Another complication we do not address here is the possibility for the sample size $N$ to be random, since in a typical particle physics experiment $N$ is Poisson-distributed with mean $\nu(\alpha)$, say. The analysis we present can be thought of as being either: (a) conditional on a fixed value of $N = n$, or (b) a version of the random $N$ scenario that assumes independence between the number of data points observed ($\nu$) and the signal fraction ($\alpha$). Both of these scenarios lead to the (same) likelihood function for $\alpha$ given in (1.2).

The remainder of this paper is organized as follows. Section 2 introduces the statistics that will be used to test (1.3) and provides an overview of the higher-order inference techniques employed. The improved $p$-value approximations to signal significance are developed in detail in section 3. After describing an example that poignantly illustrates the need for $p$-value adjustment in section 4, we confirm the validity of the proposed approximations via simulation experiments presented in section 5. Finally, section 6 undertakes a study of type I and type II errors. We end the paper with a discussion which propels the proposed methodology beyond its “toy problem” status. Specifically, we recommend that higher-order Edgeworth-approximated $p$-values be used in concert with GVM to mitigate the LEE.

\section{Higher-order inference techniques}

To set up the notation for what follows, we denote by $\ell_i(\alpha) = \partial^i \ell / \partial \alpha^i$ the $i$-th derivative of $\ell(\alpha)$. Now define $J(\alpha) = -\ell_2(\alpha)$, and the \textit{expected information} number as $I(\alpha) = \mathbb{E}[J(\alpha)]$. Also implicit in the notation is the argument at which a particular derivative is calculated, e.g., $J(\hat{\alpha}) = -\ell_2(\alpha)|_{\alpha = \hat{\alpha}}$, which is the usual definition of the \textit{observed information} number. We assume that the usual regularity conditions for consistency and asymptotic normality of $\hat{\alpha}$ are satisfied, e.g., ([10], chapter 3). Thus, we place no restrictions on the parameter space, so that $\alpha \in \mathbb{R}$.

Parametric statistical inference then seeks a statistic $T \equiv T(x)$, a function of the data vector $x$, which is used to formulate a rejection rule for the null hypothesis $H_0$, thereby providing, in some quantifiable sense, an optimal testing procedure (see, e.g., [11] for a detailed exposition relating
to the notion of a uniformly most powerful, or UMP, test). The three classical test statistics most commonly used in this context are the Likelihood Ratio (LR), Wald, and Score. The LR statistic, $T_{\text{LR}}$, is given by $2[\ell(\hat{\alpha}) - \ell(\alpha_0)]$, where $\alpha_0$ denotes the true or hypothesized value of $\alpha$ (for testing the hypothesis in (1.3), $\alpha_0 = 0$). The Wald statistic is given by $(\hat{\alpha} - \alpha_0)^2/\sigma_\alpha^2$, where $\sigma_\alpha^2$ is some estimate of the $\hat{\alpha}$ variance. The Score statistic, $T_S$, is given by $\ell_1(\alpha_0)/I(\alpha_0)$. It is well known that in lack of a UMP test, these statistics, and LR in particular, are generally near-optimal [12]. Since the asymptotic (large sample) variance of the MLE $\hat{\alpha}$ is $\sigma_\alpha^2 = I(\alpha_0)^{-1}$, the Wald statistics can use any of the consistent estimators $I(\alpha_0)^{-1}$, $I(\hat{\alpha})^{-1}$, $J(\alpha_0)^{-1}$, or $J(\hat{\alpha})^{-1}$, when standardizing it. The expected and observed versions denoted, respectively, by $T_W$ and $T_{W2}$, are the ones in most common usage. Finally, we describe two other noteworthy Wald-type statistics that use a shortcut technique enjoying wide acceptance in particle physics after its popularization by [13, 14]. Letting $\alpha_{\text{min}} < \alpha_{\text{max}}$ denote the two solutions in question, define $\sigma^- = \hat{\alpha} - \alpha_{\text{min}}$ and $\sigma^+ = \alpha_{\text{max}} - \hat{\alpha}$. Two additional approaches for estimating $\sigma_\alpha^2$ are then $\sigma_3 = (\sigma^+ + \sigma^-)/2$, and $\sigma_4 = \sigma^-$, leading to $T_{W3}$ and $T_{W4}$, respectively. A summary description of these six statistics, with $\alpha_0$ set to 0, is given in table 1.

**Table 1.** Definition of the most common versions of the Likelihood Ratio, Wald, and Score statistics for testing the hypothesis in (1.3). The symbols used to specify the value are defined in the text.

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

Under $H_0$, the LR, Wald, and Score statistics are asymptotically distributed as $\chi^2_1$, to first order.\(^4\) We can write this concisely as $T \sim \chi^2_1$, valid to $O(n^{-1/2})$. The statement that $T \sim \chi^2_k$ to $k$-th order, or $O(n^{-k/2})$, informally means that for finite $n$ a corrected quantity $(1 + O(n^{-k/2})/k + O(n^{-k/2}))$ can be found which is distributed as $\chi^2_k$, for $k = 1, 2, 3, \text{etc.}$ In many situations, it is possible to construct higher-order approximations ($k \geq 2$) explicitly. These are arranged in powers of $n^{-1/2}$, and give us control over the differences between the finite $n$ distribution of a statistic and its limiting behavior.

For testing the one-sided alternative $H_1$, the signed version of any of the statistics in table 1 (say $T$) can be used, by defining

$$R = \text{sgn}(\hat{\alpha})\sqrt{T}. \quad (2.1)$$

In this case, and under $H_0$, $R$ is asymptotically distributed as a standard normal, $N(0, 1)$, to first

---

\(^4\)This fact remains true for Wald and Score if any of the following versions of “information number” are used in the definition of the statistic: $I(0)$, $I(\hat{\alpha})$, $J(0)$, or $J(\hat{\alpha})$. **
order, whence the corresponding $p$-value is

$$P(R > r), \quad \text{where} \quad r = \text{sgn}(\hat{a})\sqrt{t},$$

(2.2)

and $t$ is the observed value of $T$ calculated from the sample on hand. In the ensuing discussion we reserve the symbols $\Phi(\cdot)$ and $\phi(\cdot)$ for the cumulative distribution and probability density functions, respectively, of a $N(0, 1)$ distribution.

Tools for developing higher-order asymptotic theory include Taylor series expansions of $\ell(\cdot)$ near $\alpha = 0$, joint cumulants for the derivatives of $\ell(\cdot)$, and Edgeworth approximations to distributions. To any finite order, relationships between cumulants and moments can be determined routinely via a symbolic algebra system (starting from the Taylor expansion of the cumulant generating function). To obtain Edgeworth approximations to the probabilities of a given statistic, its cumulant behavior. The appropriate version of (2.5) for the case in which

$$P(R > r), \quad \text{where} \quad r = \text{sgn}(\hat{a})\sqrt{t},$$

(2.2)

and $t$ is the observed value of $T$ calculated from the sample on hand. In the ensuing discussion we reserve the symbols $\Phi(\cdot)$ and $\phi(\cdot)$ for the cumulative distribution and probability density functions, respectively, of a $N(0, 1)$ distribution.

Tools for developing higher-order asymptotic theory include Taylor series expansions of $\ell(\cdot)$ near $\alpha = 0$, joint cumulants for the derivatives of $\ell(\cdot)$, and Edgeworth approximations to distributions. To any finite order, relationships between cumulants and moments can be determined routinely via a symbolic algebra system (starting from the Taylor expansion of the cumulant generating function). To obtain Edgeworth approximations to the probabilities of a given statistic, its cumulant behavior. The appropriate version of (2.5) for the case in which

$$P(R > r), \quad \text{where} \quad r = \text{sgn}(\hat{a})\sqrt{t},$$

(2.2)

and $t$ is the observed value of $T$ calculated from the sample on hand. In the ensuing discussion we reserve the symbols $\Phi(\cdot)$ and $\phi(\cdot)$ for the cumulative distribution and probability density functions, respectively, of a $N(0, 1)$ distribution.

Tools for developing higher-order asymptotic theory include Taylor series expansions of $\ell(\cdot)$ near $\alpha = 0$, joint cumulants for the derivatives of $\ell(\cdot)$, and Edgeworth approximations to distributions. To any finite order, relationships between cumulants and moments can be determined routinely via a symbolic algebra system (starting from the Taylor expansion of the cumulant generating function). To obtain Edgeworth approximations to the probabilities of a given statistic, its cumulant behavior. The appropriate version of (2.5) for the case in which

$$P(R > r), \quad \text{where} \quad r = \text{sgn}(\hat{a})\sqrt{t},$$

(2.2)

and $t$ is the observed value of $T$ calculated from the sample on hand. In the ensuing discussion we reserve the symbols $\Phi(\cdot)$ and $\phi(\cdot)$ for the cumulative distribution and probability density functions, respectively, of a $N(0, 1)$ distribution.

Tools for developing higher-order asymptotic theory include Taylor series expansions of $\ell(\cdot)$ near $\alpha = 0$, joint cumulants for the derivatives of $\ell(\cdot)$, and Edgeworth approximations to distributions. To any finite order, relationships between cumulants and moments can be determined routinely via a symbolic algebra system (starting from the Taylor expansion of the cumulant generating function). To obtain Edgeworth approximations to the probabilities of a given statistic, its cumulant behavior. The appropriate version of (2.5) for the case in which

$$P(R > r), \quad \text{where} \quad r = \text{sgn}(\hat{a})\sqrt{t},$$

(2.2)

and $t$ is the observed value of $T$ calculated from the sample on hand. In the ensuing discussion we reserve the symbols $\Phi(\cdot)$ and $\phi(\cdot)$ for the cumulative distribution and probability density functions, respectively, of a $N(0, 1)$ distribution.

Tools for developing higher-order asymptotic theory include Taylor series expansions of $\ell(\cdot)$ near $\alpha = 0$, joint cumulants for the derivatives of $\ell(\cdot)$, and Edgeworth approximations to distributions. To any finite order, relationships between cumulants and moments can be determined routinely via a symbolic algebra system (starting from the Taylor expansion of the cumulant generating function). To obtain Edgeworth approximations to the probabilities of a given statistic, its cumulant behavior. The appropriate version of (2.5) for the case in which

$$P(R > r), \quad \text{where} \quad r = \text{sgn}(\hat{a})\sqrt{t},$$

(2.2)

and $t$ is the observed value of $T$ calculated from the sample on hand. In the ensuing discussion we reserve the symbols $\Phi(\cdot)$ and $\phi(\cdot)$ for the cumulative distribution and probability density functions, respectively, of a $N(0, 1)$ distribution.

Tools for developing higher-order asymptotic theory include Taylor series expansions of $\ell(\cdot)$ near $\alpha = 0$, joint cumulants for the derivatives of $\ell(\cdot)$, and Edgeworth approximations to distributions. To any finite order, relationships between cumulants and moments can be determined routinely via a symbolic algebra system (starting from the Taylor expansion of the cumulant generating function). To obtain Edgeworth approximations to the probabilities of a given statistic, its cumulant behavior. The appropriate version of (2.5) for the case in which

$$P(R > r), \quad \text{where} \quad r = \text{sgn}(\hat{a})\sqrt{t},$$

(2.2)
The fundamental strategy in seeking improved inference in the finite \( n \) situation is that of a more accurate approximations to tail probabilities of \( R \) in the form:

\[
P(R \leq r) = \Phi(r) + \text{correction} + \text{error},
\]

where the “correction” is a function of both \( n \) and \( r \), and the “error” is \( O(n^{-1}) \) or smaller. These are accomplished primarily through Edgeworth expansions. Throughout the paper we refer to the resulting tail probabilities as Edgeworth “approximations” or “predictions”.

This strategy applies also to \( T = R^2 \); the only change being that \( \Phi(\cdot) \mapsto 2\Phi(\cdot) - 1 \), corresponding to a \( \chi^2_1 \) tail probability. The underlying theory was gradually developed over the past century, with some of the most important contributions emerging from the pioneering work of [15]. A more recent and updated treatment of the methodology is given by [10]. In the next section we detail the explicit calculations involved.

3 Approximated \( p \)-values and normalizing transformations

To concisely compute the approximate \( p \)-values for the statistics under study, we define two versions of the expectation operator: \( \mathbb{E} \) and \( \mathbb{E}_s \) will denote expectation under \( H_0 \) and under the signal, respectively:

\[
\mathbb{E}[q] := \int q(x)b(x)dx, \quad \mathbb{E}_s[q] := \int q(x)s(x)dx.
\]

Additionally, with the notation \( V_i := \mathbb{E}[\ell_i(0)] \), we find that the following (dimensionless and location-scale invariant) quantities play a key role in the expressions below:

\[
\begin{align*}
\gamma &= \frac{V_3}{2V_2^{3/2}} = \frac{\mathbb{E}_s\left[\frac{x^3}{b^3}\right] - 3\mathbb{E}_s\left[\frac{x}{b}\right] + 2}{(\mathbb{E}_s\left[\frac{x}{b}\right] - 1)^{3/2}}, \\
\rho &= -\frac{V_4}{6V_2^2} = \frac{\mathbb{E}_s\left[\frac{x^4}{b^4}\right] - 4\mathbb{E}_s\left[\frac{x^2}{b^2}\right] + 6\mathbb{E}_s\left[\frac{x}{b}\right] - 3}{(\mathbb{E}_s\left[\frac{x}{b}\right] - 1)^2}.
\end{align*}
\]

We now give explicit higher-order expansions for the signed versions of the statistics from table 1. The bulk of the work involves computing joint cumulants for the derivatives of \( \ell(\cdot) \); these forming the basis for approximating the cumulants that are then substituted into (2.5). Specifically, and following ([10], chapter 5), in what follows we denote by \( n\nu_{ijkl} \) the \((i,j,k,l)\)-th cumulant\(^6\) of the first four derivatives of \( \ell(\alpha) \) evaluated at \( \alpha = 0 \), \( \{\ell_1(0), \ldots, \ell_4(0)\} \). The calculation of these cumulants is detailed in appendix A.

3.1 Approximations to \( p \)-values

For testing \( H_0 \) vs. \( H_1 \) we compute higher-order Edgeworth expansions for the tail probabilities of signed versions of the statistics in table 1. First, starting from the approximated \( R_{LR} \) cumulants given in ([10], section 5.4), straightforward calculations give the corresponding approximated cumulants

\(^6\)An abbreviated notation will be adhered to by omitting zeros, e.g., \( \nu_{1020} \equiv \nu_{102} \).
for model (1.2) listed in the first row of table 2. Substituting these expressions into (2.5) gives immediately

\[ P(R_{LR} \leq z) = \Phi(z) - \phi(z) \left[ \left( \frac{\gamma}{6} \right) n^{-1/2} + \left( \frac{3\rho - 2\gamma^2 z}{12} \right) n^{-1} + O(n^{-3/2}) \right]. \quad (3.3) \]

Next, we give analogous results for the expected and observed versions of Wald. This entails first approximating the appropriate cumulants, expressions for which are given in ([10], section 5.3) to third order accuracy for \( R_W \). The first four cumulants appear to be correctly stated except for the 2nd, the correct version of which should be:

\[ n[k_2(R_W) - 1] = (2v_21 + 3v_{101} + 3v_{02} + v_{0001}) v_2^{-2} + \left( v_{001}v_3 + \frac{7v_{100}^2}{2} + 5v_1^2 + 11v_{11}v_{001} \right) v_2^{-3}. \]

Straightforward calculations then give the values in the second row of table 2. Substitution of these into (2.5) gives eventually

\[ P(R_W \leq z) = \Phi(z) - \phi(z) \left[ \left( \frac{\gamma H_2(z)}{6} \right) n^{-1/2} + \left( \frac{\rho - \gamma^2 z}{2} + \frac{3\rho - 3H_2(z)}{24} + \frac{\gamma^2 H_2(z)}{72} \right) n^{-1} + O(n^{-3/2}) \right]. \quad (3.4) \]

The observed version of Wald is approximated in ([10], section 5.3) to second order accuracy only. Working from first principles (using Taylor series expansions of \( \ell(\cdot) \) with an appropriate number of terms) and following the general procedure outlined in ([10], chapter 5) for Edgeworth expansions, we obtain the following third-order accurate expressions for the first four cumulants of \( R_{W2} \):

\[ \hat{k}_1(R_{W2}) = \frac{v_{11}}{2v_2^{3/2} \sqrt{n}}, \quad \hat{k}_3(R_{W2}) = -\frac{v_{001}}{v_2^{3/2} \sqrt{n}}, \]

\[ n[k_2(R_{W2}) - 1] = \left( v_{21} + v_{02} - \frac{v_{0001}}{2} \right) v_2^{-2} + \left( \frac{7v_{11}^2}{4} - \frac{3v_{001}^2}{4} \right) v_2^{-3}, \]

and

\[ n\hat{k}_4(R_{W2}) = (v_4 + 6v_{21} + 3v_{02} - 2v_{0001}) v_2^{-2} + \left( 6v_{11}v_3 + 18v_1^2 - 3v_{001}^2 \right) v_2^{-3}. \]

The conversion of these expressions in terms of our log-likelihood function (1.2) is listed in table 2. Substitution of these values into (2.5) gives

\[ P(R_{W2} \leq z) = \Phi(z) - \phi(z) \left[ -\left( \frac{3\gamma + 2\gamma H_2(z)}{6} \right) n^{-1/2} + \left( \frac{3\rho - \gamma^2 z}{2} + \frac{5\rho + 2\gamma^2 H_3(z)}{12} + \frac{\gamma^2 H_2(z)}{18} \right) n^{-1} + O(n^{-3/2}) \right]. \quad (3.5) \]

Following the same procedure as for \( R_{W2} \), we obtain analogous third-order accurate expressions for the first four cumulants of \( R_S \):

\[ \hat{k}_1(R_S) = 0, \quad \hat{k}_2(R_S) = 1, \quad \hat{k}_3(R_S) = \frac{v_3}{v_2^{3/2} \sqrt{n}}, \quad \hat{k}_4(R_S) = \frac{v_4}{nv_2^2}. \]
Table 2. Approximations to the first four cumulants of the signed versions of the statistics in table 1, in the context of log-likelihood expression (1.2). The error in these approximations is $O(n^{-3/2})$.

| $R$  | $\hat{k}_1$         | $\hat{k}_2 - 1$ | $\hat{k}_3$ | $\hat{k}_4$ |
|------|--------------------|----------------|-------------|-------------|
| $R_{LR}$ | $-\frac{\gamma}{6\sqrt{n}}$ | $\frac{18\rho-135\gamma^2}{36n}$ | 0           | 0           |
| $R_W$   | 0                 | $\frac{\rho-\gamma^2-1}{n}$ | $\frac{\gamma}{\sqrt{n}}$ | $\frac{\rho-3}{n}$ |
| $R_{W2}$ | $-\frac{\gamma}{2\sqrt{n}}$ | $\frac{12\rho-5\gamma^2}{4n}$ | $-\frac{2\gamma}{\sqrt{n}}$ | $\frac{10\rho}{n}$ |
| $R_S$   | 0                 | 0              | $\frac{\gamma}{\sqrt{n}}$ | $\frac{\rho-3}{n}$ |
| $R_{W3}$ | $-\frac{\gamma}{2\sqrt{n}}$ | $\frac{126\rho-65\gamma^2}{36n}$ | $-\frac{2\gamma}{\sqrt{n}}$ | $\frac{10\rho}{n}$ |
| $R_{W4}$ | $-\frac{\gamma}{2\sqrt{n}} - \frac{3\rho-\gamma^2}{3n}$ | $\frac{2\gamma}{3\sqrt{n}} + \frac{126\rho-53\gamma^2}{36n}$ | $-\frac{2\gamma}{\sqrt{n}} - \frac{4\rho-\gamma^2}{n}$ | $\frac{10\rho}{n}$ |

The conversion of these expressions into our log-likelihood function appears in table 2. Substitution of these $R_S$ cumulants into (2.5) gives

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

Finally, table 2 also includes $O(n^{-3/2})$ cumulants for $R_{W3}$ and $R_{W4}$, obtained by similar calculations. The tail probabilities for these statistics are given in appendix B.

3.2 Normalizing transformations

A general transformation for “normalizing” the distribution of a statistic $R$ that is already approximately normal, is

$$\tilde{r} = \Phi^{-1}(P(R \leq r)),$$

where $r$ denotes the observed value of $R$ computed from the sample at hand and $\tilde{r}$ denotes the corresponding value of the normalized statistic $\tilde{R}$. The key idea is that by invoking an accurate (higher-order) approximation to the cumulative probabilities that comprise the argument of the standard normal quantile function $\Phi^{-1}(\cdot)$, the probability integral transform will then ensure better compliance with a Gaussian distribution. For typical expansions considered in this study,

$$P(R \leq r) = \Phi(r) - \phi(r) \left[ \frac{a(r)}{\sqrt{n}} + \frac{b(r)}{n} + O(n^{-3/2}) \right],$$

numerically stable evaluation of (3.7) to $O(n^{-3/2})$ can be performed at large $r$ by calculating

$$\tilde{r} = S_{\Phi}^{-1} \left[ S_\Phi(r) + \phi(r) \left[ \frac{a(r)}{\sqrt{n}} + \frac{b(r)}{n} \right] \right],$$

where $S_{\Phi}(z) \equiv 1 - \Phi(z)$ is the survival function of a $\mathcal{N}(0,1)$. This technique would therefore be immediately applicable to any of the statistics from table 2, and we therefore advocate usage of the normalized statistic (3.9). The resulting $\tilde{R}$ will follow a $\mathcal{N}(0,1)$ to an accuracy of $O(n^{-3/2})$ under $\mathcal{H}_0$. However, it must be kept in mind that this is an asymptotic statement regarding the order of the
error as \( n \) grows. For a given \( n \) the magnitude of the error will still be governed by an appropriate constant that could in some cases be quite large.

To quantify the deviations of \( R \) from normality as a function of \( r \), we investigate the \textit{normal approximation error} quantity
\[
\Delta R(r) = r - \bar{r}.
\] (3.10)

In combination with (3.8), the first order Taylor expansion of (3.7) yields a simple approximation
\[
\Delta R(r) \approx \frac{a(r)}{\sqrt{n}} + \frac{b(r)}{n}, \text{ valid to } O(n^{-1}).
\]

Without loss of generality, for the rest of the paper we will focus on the properties of the un-normalized \( \tilde{R} \), but in an actual application it would behoove the practitioner to use \( \tilde{R} \) since computation of required quantiles for hypothesis testing will be straightforward. Important results such as type I and type II error rates studied in section 6 for \( R \) will be identical for \( \tilde{R} \).

4 An illustrative example

In this section we choose a particular model in order to illustrate the consequences of applying the Edgeworth expansions for the statistics in table 2. As choices for the background and signal under model (1.1), we select a relatively simple configuration by letting \( b(x) \) follow a uniform distribution on \([0, 1]\), and \( s(x) \) a truncated Gaussian on \([0, 1]\):
\[
b(x) = \begin{cases} 1, & \text{if } x \in [0, 1] \\ 0, & \text{if } x \notin [0, 1] \end{cases},
\]
\[
s(x) = \begin{cases} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \int_0^1 e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy, & \text{if } x \in [0, 1] \\ 0, & \text{if } x \notin [0, 1] \end{cases}.
\]

This gives a flat background superimposed with a Gaussian signal. For this and the remaining sections, whenever specific settings of the signal are needed, we use
\[
\mu = 0.5, \quad \text{and} \quad \sigma = 0.1,
\] (4.3)

implying that the values of the dimensionless parameters defined in (3.1) and (3.2) are \( \gamma \approx 1.1 \) and \( \rho \approx 2.7 \).

Figure 1 gives an idea of the impact of the signal location and scale parameters \( \mu \) and \( \sigma \) on \( \gamma \) and \( \rho \). In each respective panel, the value of the parameter not being varied is as specified in (4.3). While the curves for \( \mu \) remain fairly flat, it is noteworthy that \( \sigma \) has a dramatic effect on both quantities as it approaches the origin. This effect is better understood when we note that for small \( \sigma \),
\[
\rho \approx \frac{\mathbb{E}_s \left[ \frac{\sigma^3}{\sigma^2} \right]}{(\mathbb{E}_s \left[ \frac{\sigma^2}{\sigma^2} \right])^3} = \frac{1}{\sqrt{2\pi}} \sigma^{-1}, \quad \text{and} \quad \gamma \approx \frac{\mathbb{E}_s \left[ \frac{\sigma^2}{\sigma^2} \right]}{(\mathbb{E}_s \left[ \frac{\sigma^2}{\sigma^2} \right])^{3/2}} = \sqrt{\frac{2}{3}} \pi^{-1/4} \sigma^{-1/2}.
\]

To give a sense of the effective normal approximation error for some of the statistics constructed from random samples of size \( n \) from model (4.1)–(4.3), the value \( \Delta R(r) \) defined in (3.10) is plotted in figure 2 as a function of the observed statistic value \( r \), for \( R_W \) and \( R_{LR} \). (A standard normal distribution would have a deviation of \( \Delta R(r) = 0 \).) The large deviations of the un-normalized
Figure 1. Impact of the signal location and scale parameters $\mu$ and $\sigma$ on the quantities $\gamma$ and $\rho$. In the left/right panel, the scale/location parameter is fixed at $\sigma = 0.1/\mu = 0.5$.

$R_W$ values from its third-order normal-corrected version are particularly striking, especially at the lower sample sizes. To understand what this means, note that the deviation of $R_W$ at $r = 5$ for $n = 200$ is approximately $\Delta R_W(5) = 0.25$, which, for the $Z \sim N(0, 1)$ reference distribution under $\mathcal{H}_0$, translates into the $p$-value being smaller by a factor of $P(Z > 5)/P(Z > 5.25) \approx 3.8$ (a higher signal significance would be claimed than what is supported by the data).

In stark contrast to this is the (by comparison) exceptionally low normal approximation error of $R_{LR}$, e.g., at $r = 4$ for $n = 200$, $\Delta R_{LR}(4) = -0.005$, which translates into the $p$-value being off by only a factor of $P(Z > 4)/P(Z > 3.995) \approx 0.98$. A glance at table 2 reveals a possible reason for the good performance of the signed LR statistic: the $O(n^{-3/2})$ values for the 3rd and 4th cumulants are zero. Since these cumulants are precisely the ones appearing with the high order Hermite polynomials in (2.5), they can contribute substantially to the magnitude of the Edgeworth approximation at large values of $r$. It was conjectured in [16] that this effect may be largely responsible for the near-optimality of $R_{LR}$.

The minimal sample size for which signal discovery in a mixture model remains possible can be estimated as follows. For our example, the Cramer-Rao uncertainty of $\alpha$ under $\mathcal{H}_0$ is $\sigma_{CR} = I(0)^{-1/2} \approx 0.741/\sqrt{n}$. Therefore, for $n \lesssim 14$, no methodology would allow one to distinguish a pure signal sample ($\alpha = 1$) from a pure background sample ($\alpha = 0$) at the level of $5\sigma$ which is a traditional threshold for signal discovery claims in high energy physics [17]. For $n = 14$, the approximated $R_{LR}$ cumulants in table 2 become $\hat{\kappa}_1 \approx -0.049$ and $\hat{\kappa}_2 \approx 1.064$, illustrating the magnitude of the expected $R_{LR}$ departure from normality in this extreme case. The smallest sample size illustrated in figure 2 is $n = 20$ and, indeed, for this $n$ significant departures from the $N(0, 1)$ behavior is apparent for all statistics, including $R_{LR}$.

5 Numerical simulations

This section undertakes an extensive investigation of the properties of the statistics in table 2. As the data generating process, we take the same flat background superimposed with Gaussian signal model as specified by (4.1), (4.2) and (4.3). For each of the sample sizes $n = \{20, 40, 200, 1000, 5000, 25000\}$, a total of $m = 10^9$ pseudo-experiments (replications) are car-
Figure 2. Effective normal approximation error $\Delta R(r)$ for the signed versions of the expected Wald ($R_W$), observed Wald ($R_{W2}$), Score ($R_S$), and LR ($R_{LR}$) statistics, constructed from random samples of size $n$ from model (1.1) under background and signal densities (4.1), (4.2) with $\mu = 0.5$ and $\sigma = 0.1$.

ried out. In each pseudo-experiment, a dataset of size $n$ is independently drawn from the model (with $\alpha = 0$) via Monte Carlo methods, in order to calculate the statistics in question.

Remark 1. There is an “absolute” constraint on $\alpha$, namely $\alpha \in (\alpha_{\text{min}}, \alpha_{\text{max}})$, stemming from the requirement that model (1.1) density must remain positive. For the example described by equations (4.1)–(4.3), $\alpha_{\text{min}} \approx -0.3345$ and $\alpha_{\text{max}} \approx 1.0$. The lower endpoint is about $6.4 \sigma_{C_R}$ away from 0 for $n = 200$, resulting in a negligible probability that $\hat{\alpha}$ would be affected by the constraint. For $n = 20$ (40), however, this endpoint is only $2.0$ (2.9) $\sigma_{C_R}$ away from 0, so a sizable $\hat{\alpha}$ pile-up is present at $\alpha_{\text{min}}$. This pile-up prohibits cumulant-based moment comparisons. In addition, when $\hat{\alpha}$ is at the boundary, statistics $R_{W3}$ and $R_{W4}$ can no longer be evaluated in accordance with their original definitions. For these reasons, $n = \{20, 40\}$ sample sizes are excluded from a number of tables presented in sections 5 and 6.

Tables 3 and 4 demonstrate how closely the mean, standard deviation, skewness, and kurtosis of the statistics in table 2 agree with the corresponding values predicted by $\mathcal{N}(0, 1)$ and $O(n^{-3/2})$
models. If \( r_1, \ldots, r_m \) denote the \( m \) simulated values of a particular statistic with empirical mean \( \bar{r} = \frac{1}{m} \sum r_i/m \), define the empirical \( k \)-th central moment as \( \mu_k = \frac{1}{m} \sum (r_i - \bar{r})^k/m \). The “simulated value” (SV) and “simulation uncertainty” (SU) quantities are, respectively, the appropriate empirical moment estimate, and the standard deviation of the estimate, determined as follows (see [18] for details).

- Table 3 for the mean: \( SV = \bar{r} \), and \( SU = \frac{1}{\sqrt{m}} \).
- Table 3 for the standard deviation minus 1. \( SU \) assumes the sample is drawn from a \( N(0, 1) \):

\[
SV = \sqrt{\mu_2} - 1, \quad SU = 1/\sqrt{2(m-1)}.
\]

- Table 4 for the skewness coefficient. \( SU \) assumes the sample is drawn from a \( N(0, 1) \):

\[
SV = \frac{\mu_3}{\mu_2^{3/2}}, \quad SU = \sqrt{\frac{6m(m-1)}{(m-2)(m+1)(m+3)}}.
\]

- Table 4 for the kurtosis coefficient. \( SU \) assumes the sample is drawn from a \( N(0, 1) \):

\[
SV = \frac{\mu_4}{\mu_2^2} - 3, \quad SU = \sqrt{\frac{24m(m-1)^2}{(m-3)(m-2)(3+m)(5+m)}}.
\]

If the statistics in question were exactly \( N(0, 1) \), the simulated values would all be zero (within simulation uncertainty). Sample sizes \( n = \{20, 40\} \) are omitted from these tables for the reason pointed out in Remark 1.

For the three classical statistics \( R_W, R_{LR}, \) and \( R_S \), simulation results are in excellent agreement with \( O(n^{-3/2}) \) predictions for \( n \geq 1000 \); whereas for the remainder, values of \( n \geq 25000 \) are generally needed. Agreement with predictions is markedly worse for \( R_{W4} \).

An alternative assessment via a chi-square goodness-of-fit test is carried out in table 5. Using a bin width of 0.1 and utilizing all bins with 25 or more predicted counts, the proportions of the respective \( R \) statistic values falling in each bin over the \( m = 10^9 \) replications are converted to a single \( \chi^2 \)-value comparing the observed proportions with predicted probabilities for a \( N(0, 1) \), and second and third order Edgeworth approximations. Each \( \chi^2 \)-value is then converted to a corresponding \( p \)-value for the test by calculating the survival probability at the value under the \( \chi^2 \) distribution with the indicated degrees of freedom. We see substantial disagreements with normality for all (unapproximated) statistics and all sample sizes. The agreement of simulations with Edgeworth predictions improves as one goes from \( O(n^{-1}) \) to \( O(n^{-3/2}) \), and from lower to higher \( n \), particularly for \( R_W, R_{LR}, \) and \( R_S \).

Table 6 proceeds in a similar fashion, but compares instead the Edgeworth-predicted tail probabilities for exceeding the value of \( r = 5 \) with the corresponding simulation-based exceedances. Taking the case of table entries corresponding to \( R_W \) for example, the \( O(n^{-3/2}) \) prediction is computed as \( P(R_W > 5) \) using \((3,4)\), whereas the simulated value is the empirical proportion (say \( \hat{p} \)) of the \( m = 10^9 \) \( R_W \) values exceeding 5. The simulation uncertainty is the standard error of the empirical proportion, \( \sqrt{\hat{p}(1-\hat{p})/m} \). Taking twice the simulation uncertainty as the metric
Table 3. Comparison of the $O(n^{-3/2})$ predictions to the first two cumulants from table 2 with simulation-based empirical estimates constructed from $10^9$ Monte Carlo replications. Predictions that differ from simulated values by more than twice the simulation uncertainty appear in bold face. The corresponding $N(0, 1)$ value is 0 for both quantities. The simulation uncertainty is $3.2 \times 10^{-5}$ for the mean and $2.2 \times 10^{-5}$ for the standard deviation minus 1.

|       | $R_W$               | $R_{W2}$              | $R_{W3}$              | $R_{W4}$              | $R_{LR}$             | $R_S$               |
|-------|---------------------|-----------------------|-----------------------|-----------------------|----------------------|---------------------|
| $n$   | Predicted           | Simulated             | Predicted             | Simulated             | Predicted            | Simulated           |
|       |                     |                       |                       |                       |                      |                     |
| 200   | 0                   | $-9.0 \times 10^{-5}$ | $11.46 \times 10^{-4}$| 1.97 \times 10^{-4}  |                      |
| 1000  | 0                   | $-0.9 \times 10^{-5}$ | $2.29 \times 10^{-4}$ | 2.50 \times 10^{-4}  |                      |
| 5000  | 0                   | $-1.6 \times 10^{-5}$ | $0.46 \times 10^{-4}$ | 0.76 \times 10^{-4}  |                      |
| 25000 | 0                   | $4.8 \times 10^{-5}$  | $0.09 \times 10^{-4}$ | $-0.30 \times 10^{-4}$|                      |
| 200   | $-39.222 \times 10^{-3}$ | $-40.724 \times 10^{-3}$ | $161.93 \times 10^{-4}$ | $174.00 \times 10^{-4}$ |                      |
| 1000  | $-17.541 \times 10^{-3}$ | $-17.670 \times 10^{-3}$ | $32.59 \times 10^{-4}$ | 33.31 \times 10^{-4}  |                      |
| 5000  | $-7.844 \times 10^{-3}$ | $-7.872 \times 10^{-3}$ | 6.53 \times 10^{-4}   | 6.86 \times 10^{-4}  |                      |
| 25000 | $-3.508 \times 10^{-3}$ | $-3.461 \times 10^{-3}$ | 1.31 \times 10^{-4}   | 0.92 \times 10^{-4}  |                      |
| 200   | $-39.222 \times 10^{-3}$ | $-40.817 \times 10^{-3}$ | 178.17 \times 10^{-4} | 191.06 \times 10^{-4} |                      |
| 1000  | $-17.541 \times 10^{-3}$ | $-17.678 \times 10^{-3}$ | 35.89 \times 10^{-4}  | 36.63 \times 10^{-4}  |                      |
| 5000  | $-7.844 \times 10^{-3}$ | $-7.872 \times 10^{-3}$ | 7.19 \times 10^{-4}   | 7.52 \times 10^{-4}  |                      |
| 25000 | $-3.508 \times 10^{-3}$ | $-3.461 \times 10^{-3}$ | 1.44 \times 10^{-4}   | 1.05 \times 10^{-4}  |                      |
| 200   | $-50.617 \times 10^{-3}$ | $-48.508 \times 10^{-3}$ | $441.74 \times 10^{-4}$ | $491.83 \times 10^{-4}$ |                      |
| 1000  | $-19.820 \times 10^{-3}$ | $-19.106 \times 10^{-3}$ | $153.76 \times 10^{-4}$ | $157.40 \times 10^{-4}$ |                      |
| 5000  | $-8.300 \times 10^{-3}$ | $-8.152 \times 10^{-3}$ | $59.72 \times 10^{-4}$ | $60.30 \times 10^{-4}$ |                      |
| 25000 | $-3.599 \times 10^{-3}$ | $-3.516 \times 10^{-3}$ | $24.88 \times 10^{-4}$ | $24.51 \times 10^{-4}$ |                      |
| 200   | $-13.074 \times 10^{-3}$ | $-13.199 \times 10^{-3}$ | 22.48 \times 10^{-4}  | 22.86 \times 10^{-4}  |                      |
| 1000  | $-5.847 \times 10^{-3}$ | $-5.860 \times 10^{-3}$ | 4.50 \times 10^{-4}   | 4.79 \times 10^{-4}  |                      |
| 5000  | $-2.615 \times 10^{-3}$ | $-2.631 \times 10^{-3}$ | 0.90 \times 10^{-4}   | 1.21 \times 10^{-4}  |                      |
| 25000 | $-1.169 \times 10^{-3}$ | $-1.121 \times 10^{-3}$ | 0.19 \times 10^{-4}   | $-0.21 \times 10^{-4}$|                      |
| 200   | 0                   | $2.7 \times 10^{-5}$  | 0                     | $-1.5 \times 10^{-5}$ |                      |
| 1000  | 0                   | $-1.6 \times 10^{-5}$ | 0                     | $0.1 \times 10^{-5}$  |                      |
| 5000  | 0                   | $1.3 \times 10^{-5}$  | 0                     | $3.8 \times 10^{-5}$  |                      |
| 25000 | 0                   | $0.1 \times 10^{-5}$  | 0                     | $4.7 \times 10^{-5}$  |                      |

(corresponding to 95% confidence), the last column indicates that predictions for $R_W$, $R_S$, and $R_{LR}$ are generally in agreement with simulations (especially $R_{LR}$), but the remaining statistics show significant differences at lower $n$. Consequently, it appears that one would need higher than third order based predictions in order to obtain nominally correct $p$ -values for these cases. Finally, note that predictions for $R_{LR}$ are also the closest to $N(0, 1)$ values, in agreement with the conclusions drawn from figure 2.
Table 4. Comparison of the $O(n^{-3/2})$ predictions to the skewness and excess kurtosis coefficients constructed from the cumulants in table 2, with simulation-based empirical estimates derived from $10^6$ Monte Carlo replications. Predictions that differ from simulated values by more than twice the simulation uncertainty appear in bold face. The corresponding $N(0,1)$ value is 0 for both coefficients. The simulation uncertainty is $8.0 \times 10^{-5}$ for skewness and $1.5 \times 10^{-4}$ for excess kurtosis.

|     | $n$  | Skewness      | Excess Kurtosis   |
|-----|------|---------------|-------------------|
|     |      | Predicted     | Simulated         | Predicted | Simulated |
| $R_W$ | 200  | 0.07817       | 0.07747           | $-15.5 \times 10^{-4}$ | $-10.6 \times 10^{-4}$ |
|     | 1000 | 0.03506       | 0.03513           | $-3.1 \times 10^{-4}$  | $-3.6 \times 10^{-4}$  |
|     | 5000 | 0.01569       | 0.01562           | $-0.6 \times 10^{-4}$  | $-2.7 \times 10^{-4}$  |
|     | 25000| 0.00702       | 0.00701           | $-0.1 \times 10^{-4}$  | $2.6 \times 10^{-4}$   |
| $R_{W2}$ | 200 | $-0.14951$ | $-0.17543$ | $126.10 \times 10^{-3}$ | $163.24 \times 10^{-3}$ |
|     | 1000 | $-0.06948$ | $-0.07150$ | $26.54 \times 10^{-3}$ | $27.76 \times 10^{-3}$ |
|     | 5000 | $-0.03132$ | $-0.03157$ | $5.36 \times 10^{-3}$  | $5.24 \times 10^{-3}$  |
|     | 25000| $-0.01402$ | $-0.01405$ | $1.08 \times 10^{-3}$  | $1.35 \times 10^{-3}$  |
| $R_{W3}$ | 200 | $-0.14879$ | $-0.17563$ | $125.29 \times 10^{-3}$ | $163.60 \times 10^{-3}$ |
|     | 1000 | $-0.06941$ | $-0.07152$ | $26.51 \times 10^{-3}$ | $27.77 \times 10^{-3}$ |
|     | 5000 | $-0.03131$ | $-0.03157$ | $5.36 \times 10^{-3}$  | $5.24 \times 10^{-3}$  |
|     | 25000| $-0.01402$ | $-0.01405$ | $1.08 \times 10^{-3}$  | $1.35 \times 10^{-3}$  |
| $R_{W4}$ | 200 | $-0.17965$ | $-0.21724$ | $113.11 \times 10^{-3}$ | $207.78 \times 10^{-3}$ |
|     | 1000 | $-0.07612$ | $-0.07883$ | $25.30 \times 10^{-3}$ | $30.56 \times 10^{-3}$ |
|     | 5000 | $-0.03269$ | $-0.03300$ | $5.25 \times 10^{-3}$  | $5.47 \times 10^{-3}$  |
|     | 25000| $-0.01431$ | $-0.01434$ | $1.07 \times 10^{-3}$  | $1.37 \times 10^{-3}$  |
| $R_{LR}$ | 200 | 0 | $-0.00084$ | 0 | $3.7 \times 10^{-4}$ |
|     | 1000 | 0 | 0.00006 | 0 | $-0.6 \times 10^{-4}$ |
|     | 5000 | 0 | $-0.00007$ | 0 | $-2.0 \times 10^{-4}$ |
|     | 25000| 0 | $-0.00001$ | 0 | $2.7 \times 10^{-4}$ |
| $R_S$   | 200 | 0.07844       | 0.07853           | $-15.5 \times 10^{-4}$ | $-14.6 \times 10^{-4}$ |
|     | 1000 | 0.03508       | 0.03501           | $-3.1 \times 10^{-4}$  | $-2.7 \times 10^{-4}$  |
|     | 5000 | 0.01569       | 0.01567           | $-0.6 \times 10^{-4}$  | $0.7 \times 10^{-4}$   |
|     | 25000| 0.00702       | 0.00693           | $-0.1 \times 10^{-4}$  | $2.3 \times 10^{-4}$   |

Figure 3 presents the logarithm (base 10) of the survival probabilities over the quantile range $4.5 \leq r \leq 5.5$ relevant for establishing credibility of signal discovery claims. The panels in each row correspond to different statistics: $R_W$, $R_{W2}$, $R_S$. The left and right panels show the effect of increasing the sample size from 200 to 1000, with $O(n^{-1})$ and $O(n^{-3/2})$ Edgeworth predictions identified by dashed and dotted lines, respectively, in each panel. Note that both predictions drift toward the 95% simulation uncertainty envelope (gray band) as $n$ increases; all of these converging to the solid $N(0,1)$ line as expected. An interesting aberration occurs in the middle.
Table 5. $\chi^2$ goodness-of-fit test for the statistics in table 2 with $\mathcal{N}(0, 1)$, and second and third order Edgeworth predictions. For each $n$, the test is constructed from $10^9$ replications using a bin width of 0.1 and all bins with 25 or more predicted counts. Entries are formatted as $\{a/b, c\}$, where $a$ is the $\chi^2$ statistic value, $b$ is the number of degrees of freedom (i.e., the number of bins used), and $c$ is the corresponding $p$-value.

|     | $n$     | $\mathcal{N}(0, 1)$ prediction | $O(n^{-1})$ prediction | $O(n^{-3/2})$ prediction |
|-----|---------|-------------------------------|------------------------|--------------------------|
| $R_W$ | 200     | 1012631/107, 0.00              | 12920/100, 0.00        | 358.7/107, 0.00          |
|      | 1000    | 205959/107, 0.00               | 534.4/106, 0.00        | 124.6/106, 0.10          |
|      | 5000    | 40702/107, 0.00                | 110.2/107, 0.40        | 90.29/107, 0.88          |
|      | 25000   | 8254/107, 0.00                 | 81.34/107, 0.97        | 82.37/107, 0.96          |
| $R_{W2}$ | 200     | 11789846/107, 0.00             | 3710820/92, 0.00       | 402208/116, 0.00         |
|      | 1000    | 1266791/107, 0.00              | 105137/101, 0.00       | 4104/110, 0.00           |
|      | 5000    | 231468/107, 0.00               | 2967/106, 0.00         | 169.5/107, 0.00          |
|      | 25000   | 44997/107, 0.00                | 196.9/107, 0.00        | 78.53/107, 0.98          |
| $R_{W3}$ | 200     | 12178306/107, 0.00             | 3961404/92, 0.00       | 426251/116, 0.00         |
|      | 1000    | 1275312/107, 0.00              | 110857/101, 0.00       | 4292/110, 0.00           |
|      | 5000    | 231797/107, 0.00               | 3169/106, 0.00         | 171.2/107, 0.00          |
|      | 25000   | 45006/107, 0.00                | 201.8/107, 0.00        | 77.90/107, 0.98          |
| $R_{W4}$ | 200     | 31650126/107, 0.00             | 5860538/95, 0.00       | 1352453/116, 0.00        |
|      | 1000    | 2166332/107, 0.00              | 116183/103, 0.00       | 8056/110, 0.00           |
|      | 5000    | 333907/107, 0.00               | 4365/106, 0.00         | 212.1/108, 0.00          |
|      | 25000   | 59517/107, 0.00                | 248.1/107, 0.00        | 87.67/107, 0.92          |
| $R_{LR}$ | 200     | 185597/107, 0.00               | 11493/107, 0.00        | 292.0/107, 0.00          |
|      | 1000    | 34917/107, 0.00                | 616.9/107, 0.00        | 126.5/107, 0.10          |
|      | 5000    | 7068/107, 0.00                 | 142.5/107, 0.01        | 113.8/107, 0.31          |
|      | 25000   | 1339/107, 0.00                 | 80.78/107, 0.97        | 82.95/107, 0.96          |
| $R_S$  | 200     | 1030516/107, 0.00              | 8697/100, 0.00         | 299.3/106, 0.00          |
|      | 1000    | 204171/107, 0.00               | 314.2/106, 0.00        | 116.5/106, 0.23          |
|      | 5000    | 40879/107, 0.00                | 95.88/107, 0.77        | 92.46/107, 0.84          |
|      | 25000   | 8100/107, 0.00                 | 107.2/107, 0.48        | 108.5/107, 0.44          |

panels corresponding to $R_{W2}$: not only is the $O(n^{-3/2})$ prediction in the “wrong” direction, but the $O(n^{-1})$ one became negative (and is thus absent from the plots). Such unexpected behavior far out in the tails, as well as the possibility for a lower order approximation to be more accurate than a higher order one, are documented in the literature ([10], section 5.3).

Figure 4 complements figure 3 by displaying the corresponding results only for $R_{LR}$, but focusing on sample sizes ranging from 20 to 1,000. Since both axes in both figures are on the same scale, comparisons are straightforward. As noted earlier, the results for $R_{LR}$ are remarkable in terms of: (i) close agreement with both the $\mathcal{N}(0, 1)$ line as well as the simulation envelope, and (ii) the relatively small magnitude of the Edgeworth corrections. The performance of $R_{W2}$ on the other hand appears to be markedly inferior to that of the rest.
Table 6. Comparison of Edgeworth-predicted survival probabilities at $r = 5$ with simulation-based empirical exceedance proportions computed over $10^9$ replications. The corresponding $\mathcal{N}(0, 1)$ value is $2.87 \times 10^{-7}$. Predictions that differ from simulated values by more than twice the simulation uncertainty appear in bold face.

|     | $O(n^{-3/2})$ Prediction | Simulated Value | Simulation Uncertainty |
|-----|--------------------------|-----------------|------------------------|
| $R_W$ |                          |                 |                        |
| 20  | $42.2 \times 10^{-7}$    | 49.4 $\times 10^{-7}$ | 0.70 $\times 10^{-7}$ |
| 40  | $25.6 \times 10^{-7}$    | 26.1 $\times 10^{-7}$ | 0.51 $\times 10^{-7}$ |
| 200 | $9.98 \times 10^{-7}$    | 10.55 $\times 10^{-7}$ | 0.32 $\times 10^{-7}$ |
| 1000| $5.44 \times 10^{-7}$    | 5.23 $\times 10^{-7}$ | 0.23 $\times 10^{-7}$ |
| 5000| $3.90 \times 10^{-7}$    | 4.04 $\times 10^{-7}$ | 0.20 $\times 10^{-7}$ |
| 25000| $3.30 \times 10^{-7}$   | 3.19 $\times 10^{-7}$ | 0.18 $\times 10^{-7}$ |
| $R_{W2}$ |                      |                 |                        |
| 20  | $192 \times 10^{-7}$     | 204 $\times 10^{-7}$ | 1.4 $\times 10^{-7}$  |
| 40  | $90.8 \times 10^{-7}$    | 12.0 $\times 10^{-7}$ | 0.35 $\times 10^{-7}$ |
| 200 | $15.0 \times 10^{-7}$    | 0.61 $\times 10^{-7}$ | 0.08 $\times 10^{-7}$ |
| 1000| $2.83 \times 10^{-7}$    | 0.90 $\times 10^{-7}$ | 0.09 $\times 10^{-7}$ |
| 5000| $1.76 \times 10^{-7}$    | 1.63 $\times 10^{-7}$ | 0.13 $\times 10^{-7}$ |
| 25000| $2.16 \times 10^{-7}$   | 2.10 $\times 10^{-7}$ | 0.14 $\times 10^{-7}$ |
| $R_{W3}$ |                      |                 |                        |
| 200 | $15.1 \times 10^{-7}$    | 0.68 $\times 10^{-7}$ | 0.08 $\times 10^{-7}$ |
| 1000| $2.86 \times 10^{-7}$    | 0.93 $\times 10^{-7}$ | 0.10 $\times 10^{-7}$ |
| 5000| $1.77 \times 10^{-7}$    | 1.63 $\times 10^{-7}$ | 0.13 $\times 10^{-7}$ |
| 25000| $2.16 \times 10^{-7}$   | 2.10 $\times 10^{-7}$ | 0.14 $\times 10^{-7}$ |
| $R_{W4}$ |                      |                 |                        |
| 200 | $9.45 \times 10^{-7}$    | 0.71 $\times 10^{-7}$ | 0.08 $\times 10^{-7}$ |
| 1000| $2.21 \times 10^{-7}$    | 1.07 $\times 10^{-7}$ | 0.10 $\times 10^{-7}$ |
| 5000| $1.85 \times 10^{-7}$    | 1.87 $\times 10^{-7}$ | 0.14 $\times 10^{-7}$ |
| 25000| $2.27 \times 10^{-7}$   | 2.26 $\times 10^{-7}$ | 0.15 $\times 10^{-7}$ |
| $R_{LR}$ |                       |                 |                        |
| 20  | $3.99 \times 10^{-7}$    | 4.25 $\times 10^{-7}$ | 0.21 $\times 10^{-7}$ |
| 40  | $3.30 \times 10^{-7}$    | 3.00 $\times 10^{-7}$ | 0.17 $\times 10^{-7}$ |
| 200 | $2.85 \times 10^{-7}$    | 2.82 $\times 10^{-7}$ | 0.17 $\times 10^{-7}$ |
| 1000| $2.81 \times 10^{-7}$    | 2.70 $\times 10^{-7}$ | 0.16 $\times 10^{-7}$ |
| 5000| $2.83 \times 10^{-7}$    | 2.94 $\times 10^{-7}$ | 0.17 $\times 10^{-7}$ |
| 25000| $2.85 \times 10^{-7}$   | 2.79 $\times 10^{-7}$ | 0.17 $\times 10^{-7}$ |
| $R_S$ |                          |                 |                        |
| 20  | $41.3 \times 10^{-7}$    | 35.7 $\times 10^{-7}$ | 0.60 $\times 10^{-7}$ |
| 40  | $25.2 \times 10^{-7}$    | 23.1 $\times 10^{-7}$ | 0.48 $\times 10^{-7}$ |
| 200 | $9.90 \times 10^{-7}$    | 10.26 $\times 10^{-7}$ | 0.32 $\times 10^{-7}$ |
| 1000| $5.43 \times 10^{-7}$    | 4.82 $\times 10^{-7}$ | 0.22 $\times 10^{-7}$ |
| 5000| $3.89 \times 10^{-7}$    | 3.59 $\times 10^{-7}$ | 0.19 $\times 10^{-7}$ |
| 25000| $3.30 \times 10^{-7}$   | 3.18 $\times 10^{-7}$ | 0.18 $\times 10^{-7}$ |
Figure 3. Second and third order approximated Edgeworth log (base 10) survival probabilities of \( R_W \), \( R_{W2} \), and \( R_{S} \), computed at quantiles \( r \) for \( n = 200 \) (left panels) and \( n = 1000 \) (right panels). The survival probability corresponding to \( N(0,1) \) is plotted with a solid line. The gray band is a 95% simulation uncertainty envelope, corresponding to the true survival probability.
Figure 4. Second and third order approximated Edgeworth log (base 10) survival probabilities of $R_{LR}$, computed at quantiles $r$ for four different sample sizes $n$, along with the $N(0,1)$ reference. The gray band is a 95% simulation uncertainty envelope, corresponding to the true survival probability.

6 Study of type I and type II errors

This section undertakes a study of the type I and type II error rates for testing the null $H_0$ vs. the alternative $H_1$. In the context of signal detection, these errors would correspond to false positive and false negative probabilities, respectively. Throughout the study we fix the nominal (or target) probability of type I error at $q_0 \equiv 1 - \Phi(5) \approx 2.87 \times 10^{-7}$. With the model settings once again specified by (4.1)–(4.3), we generate $m = 10^9$ replicates, each comprised of a random sample of size $n$ from the model. At low $n$ (20 and 40) the boundary effects described in Remark 1 start playing an important role. This leads to problems in the computation of $R_{W3}$ and $R_{W4}$, which are therefore absent from the relevant tables in this section.

To characterize the error rates, let $c_n$ be the quantile which corresponds to the predicted survival function value $q_0$. For any statistic $R$ in table 2, this quantile is the solution of the equation $q_0 = P_{\text{pre}}(R > c_n)$, where $P_{\text{pre}}(\cdot)$ is the probability predicted by the third-order Edgeworth expansions of section 3.1. Table 7 gives the values of the predicted $(1 - q_0)$-th quantiles $c_n$,.
as well as their simulated counterparts. The uncertainty in the former is approximated by the asymptotic standard error of the empirical \((1 - q_0)\)-th quantile, \(\sqrt{(1 - q_0)q_0/[m f(c_n)^2]}\), where \(f(z) = -dS_{\text{pre}}(z)/dz\). Since this density function is only available for the predicted quantiles, we attach the uncertainty to these values in the form of plus and minus the standard error.

Table 7. \(O(n^{-3/2})\) Edgeworth predictions for the \((1 - q_0)\)-th quantiles of the statistics in table 2, compared to their corresponding values computed based on \(10^5\) simulations. The uncertainty in the predicted values appears plus and minus the standard error for the empirical quantile. Predictions that deviate by more than twice the uncertainty from their simulated values are bolded.

| \(R\) | Method | Predicted/Simulated \((1 - q_0)\)-th Quantile |
|-------|--------|---------------------------------------------|
| \(R_{W}\) | Predicted | \(5.601 \pm 0.012\) |
|        | Simulated | \(5.486 \pm 0.012\) |
| \(R_{W2}\) | Predicted | \(5.917 \pm 0.012\) |
|        | Simulated | \(5.771 \pm 0.012\) |
| \(R_{W3}\) | Predicted | \(5.394 \pm 0.013\) |
|        | Simulated | \(5.000 \pm 0.013\) |
| \(R_{W4}\) | Predicted | \(5.292 \pm 0.014\) |
|        | Simulated | \(4.945 \pm 0.013\) |
| \(R_{LR}\) | Predicted | \(5.065 \pm 0.012\) |
|        | Simulated | \(5.027 \pm 0.012\) |
| \(R_{S}\) | Predicted | \(5.599 \pm 0.012\) |
|        | Simulated | \(5.483 \pm 0.012\) |

The \(c_n\) predictions from table 7 can now be used as the basis for a type I and II error assessment. We start with the former. Assuming that \(\mathcal{H}_0\) is rejected if \(R > c_n\),

\[
q_1(n) = P_{\text{sim}}(R > c_n | \alpha = 0) \tag{6.1}
\]

is the probability of a type I error (i.e., of falsely rejecting \(\mathcal{H}_0\)), estimated from simulations, as defined in (6.1). The results, computed empirically from the \(m\) Monte Carlo replicates (notation \(P_{\text{sim}}(\cdot)\)), are presented in table 8. If the \(O(n^{-3/2})\) predictions were exact, we would expect approximately 95% of all values to be within twice the simulation uncertainty of \(0.17 \times 10^{-7}\) from the nominal value of \(q_0\). We once again note that the three classical statistics, \(R_{W}, R_{LR},\) and \(R_{S}\), are the only ones to perform at nominal levels (to within Monte Carlo uncertainty) for most sample sizes.

For the statistical power study, we fix the probability of type I error at \(q_0\), and determine the probability of type II error, denoted by

\[
q_2(n, \alpha_1) = P_{\text{sim}}(R \leq c_n | \alpha = \alpha_1) \tag{6.2}
\]

Implicit in the notation is the fact that \(q_2(\cdot)\) will vary with both \(n\) and the actual signal fraction of \(\alpha_1\) under \(\mathcal{H}_1\). (Also note that \(q_1(n) = 1 - q_2(n, 0)\)). For each \(n\), we set the signal fraction at \(\alpha_1 = 5\sigma_{CR}\), i.e., five times the Cramer-Rao uncertainty. As all statistics have the same asymptotic distribution (to first order), this strategy ensures that \(q_2(n, \alpha_1) \to \Phi(0) = 0.5\) as \(n \to \infty\), whence the “difficulty” of signal discovery remains approximately constant as \(n\) varies.

The values of \(\sigma_{CR}\) for the \((s(x)\) and \(b(x)\) used in this study are given in table 9. Table 10 gives the resulting probability of type II error. In line with earlier results, it is no longer a surprise that \(R_{LR}\)
was on computing Edgeworth approximations to a mixture model with unknown signal fraction. The focus of the study presented in this manuscript is at most $2 \times 10^{-5}$. The smallest value for each $n$ appears in bold. The simulation uncertainty in these results is at most $2 \times 10^{-5}$.

Table 10. Type II error probabilities $q_2(n, \alpha_1)$ as defined in (6.2), determined empirically from $10^9$ simulations with $\alpha_1$ as given in table 9. The smallest value for each $n$ appears in bold. The simulation uncertainty in these results is at most $2 \times 10^{-5}$.

| $n$  | $R_W$  | $R_{W2}$ | $R_{W3}$ | $R_{W4}$ | $R_{LR}$ | $R_S$  |
|------|--------|----------|----------|----------|----------|--------|
| 20   | 0.85825 | 0.46558  | —        | —        | 0.68905  | 0.76135 |
| 40   | 0.67401 | 0.71384  | —        | —        | 0.63606  | 0.66509 |
| 200  | 0.59592 | 0.80891  | 0.80695  | 0.77782  | 0.58873  | 0.59510 |
| 1000 | 0.55021 | 0.63069  | 0.63082  | 0.59609  | 0.54827  | 0.55011 |
| 5000 | 0.52433 | 0.53250  | 0.53257  | 0.52830  | 0.52385  | 0.52432 |
| 25000| 0.51130 | 0.51192  | 0.51193  | 0.51156  | 0.51119  | 0.51129 |

yields the smallest type II errors at each setting of $n$ except $n = 20$. At $n = 20$, the boundary effects substantially distort the distribution of $R_{W2}$, so the $R_{W2}$ type II error rate becomes the smallest at the cost of the largest type I error rate.

7 Discussion

A wide gamut of near-optimal statistics can be used in testing for the presence of a signal under a mixture model with unknown signal fraction. The focus of the study presented in this manuscript was on computing Edgeworth approximations to $p$-values of the asymptotic distributions of such statistics under the null hypothesis of no signal, so as to provide more accurate inferences for finite samples. Comparisons were made between approximations of different orders, highlighting striking

Table 9. Values of the Cramér-Rao uncertainty for $\mathcal{H}_0$, $\sigma_{CR}$, and corresponding values of $\alpha = \alpha_1$ used as the actual model signal fraction under $\mathcal{H}_1$.

| $n$  | $\sigma_{CR}$ | $\alpha_1 = 5\sigma_{CR}$ |
|------|---------------|--------------------------|
| 20   | 1.6571 x 10^{-1} | 0.82853 |
| 40   | 1.1717 x 10^{-1} | 0.58586 |
| 200  | 5.2401 x 10^{-2} | 0.26200 |
| 1000 | 2.3434 x 10^{-2} | 0.11717 |
| 5000 | 1.0480 x 10^{-2} | 0.05240 |
| 25000| 4.6868 x 10^{-3} | 0.02343 |

Table 8. Type I error probabilities $q_1(n)$ as defined in (6.1). The survival function values are obtained based on $10^6$ simulated replicates for the predicted quantiles from table 7. Values that deviate by more than twice the simulation uncertainty of $1.7 \times 10^{-7}$ from the nominal value of $q_0 = 2.87 \times 10^{-7}$ are bolded.

| $n$  | $R_W$  | $R_{W2}$ | $R_{W3}$ | $R_{W4}$ | $R_{LR}$ | $R_S$  |
|------|--------|----------|----------|----------|----------|--------|
| 20   | $7.62 \times 10^{-7}$ | $46.5 \times 10^{-7}$ | —        | —        | $3.10 \times 10^{-7}$ | $2.74 \times 10^{-7}$ |
| 40   | $3.30 \times 10^{-7}$ | $1.11 \times 10^{-7}$ | —        | —        | $2.68 \times 10^{-7}$ | $3.12 \times 10^{-7}$ |
| 200  | $3.03 \times 10^{-7}$ | $0.07 \times 10^{-7}$ | $0.07 \times 10^{-7}$ | $0.13 \times 10^{-7}$ | $2.68 \times 10^{-7}$ | $2.73 \times 10^{-7}$ |
| 1000 | $2.71 \times 10^{-7}$ | $0.78 \times 10^{-7}$ | $0.78 \times 10^{-7}$ | $1.35 \times 10^{-7}$ | $2.77 \times 10^{-7}$ | $2.72 \times 10^{-7}$ |
| 5000 | $2.65 \times 10^{-7}$ | $2.44 \times 10^{-7}$ | $2.44 \times 10^{-7}$ | $2.56 \times 10^{-7}$ | $2.72 \times 10^{-7}$ | $2.77 \times 10^{-7}$ |
| 25000| $2.87 \times 10^{-7}$ | $2.82 \times 10^{-7}$ | $2.82 \times 10^{-7}$ | $2.83 \times 10^{-7}$ | $2.83 \times 10^{-7}$ | $2.86 \times 10^{-7}$ |
deviations from the target standard normal reference distribution in some cases, with consequent 
p-value inflation/deflation. Finally, the performance of the corrected statistics was examined in 
terms of false positive and false negative signal detection error rates.

Given the insights gained from this study, the following summary remarks can be offered. For 
small and moderate sample sizes, deviations from normality for some likelihood-based statistics used 
in signal searches, in particular the observed version of Wald and its variants, can be significant, thus 
altering the false discovery error rate. Moreover, for narrow signals these deviations are unbounded.

It is desirable to assess the magnitude of these deviations from normality. Their influence 
can be quantified by comparing the approximations developed in this study, to various orders of 
accuracy, with the asymptotic formulae, for all possible nuisance parameter values of the model. The 
third-order approximated versions of the test statistics manifest a substantially improved agreement 
with the behavior found in simulations.

In comparison with the Wald and Score tests, deviations from normality are substantially milder 
for the signed LR statistic. For the example model considered in this study, we find that the $O(n^{-3/2})$ 
approximation to the $R_{LR}$ distribution reliably predicts type I error rates down to $n = 20$, nearing 
the sample size limit at which signal/background separation at 5σ becomes impossible. On the 
other hand, at $n = 20$ and $r = 5$, the $N(0, 1)$ approximation underestimates the $p$-value by a factor 
of about 0.7. The $R_{LR}$ statistic also has better capabilities of detecting actual signals (higher power) 
while maintaining the false discovery rate at acceptably low levels. Tests based on the likelihood 
ratio should therefore be preferred.

As speculated by Mykland in [16], who proves that the $k$-th cumulant of $R_{LR}$ vanishes to 
$O(n^{-k/2})$ for all $k \geq 3$, this fact “...would seem to be the main asymptotic property governing 
the accuracy behavior...” of $R_{LR}$. To further elucidate this statement, we derive the fourth order 
Edgeworth-approximated tail probabilities for $R_{LR}$ in appendix C. As predicted, the $\kappa_3$ value is 
$O(n^{-3/2})$ while $\kappa_4 = \kappa_5 = 0$ to $O(n^{-2})$. The vanishing of the highest order cumulants results in 
the suppression of high order Hermite polynomials in the Edgeworth series, $H_k(z)$ with $k > 2$, and 
it is precisely these terms that affect the tail behavior the most. Other statistics do not enjoy this 
property, containing terms up to $H_8(z)$ in their $O(n^{-2})$ Edgeworth expansions.

Thus, concerning signal strength determination via a two-component mixture model (signal 
and background), the main message of the paper is two-fold:

- Always use the signed likelihood ratio (LR) test statistic, or truncated LR statistic as in [9].

- For low sample sizes (i.e., when the magnitudes of $\hat{k}_1$ or $\hat{k}_2 - 1$ for $R_{LR}$ defined in table 2 reach 
a few percent level), it may be important to either (a) use direct simulation, or (b) invoke an 
Edgeworth approximation, for the resulting local $p$-value computation.

If the inference goal consists in determination of the global $p$-value, we offer the following 
remarks. First, treat the nuisance parameters in $s(x)$ via GVM (Gross-Vitells Method). Then, 
in cases when there are no nuisance parameters in $b(x)$, or the practitioner is willing to treat 
all nuisance parameters via GVM, $O(n^{-3/2})$ normalized versions of the signed likelihood-based 
statistics via transformation (3.7) are suggested, facilitating standard inferences. Alternatively, the 
global significance of the signal test statistic could be adjusted conservatively, leading to a subsequent
(conservative) estimate of the global $p$-value, by proceeding according to the algorithm detailed in appendix D.

Acknowledgments

I. Volobouev thanks the Statistics Committee of the CMS Collaboration for comments and productive discussions. This work was supported in part by the United States Department of Energy grant DE-SC0015592.

A Calculation of cumulants of the log-likelihood function

Recalling that $\mathbb{E}[]$ and $\mathbb{V}[]$ denote, respectively, expectation and variance under $b(x)$, and that $n\nu_{ijkl}$ represents the $(i, j, k, l)$ joint cumulant of $(\ell_1(0), \ldots, \ell_4(0))$, we compute initially

$$n\nu_1 = \mathbb{E}[\ell_1(0)] = \mathbb{E} \left[ \sum_{i=1}^n \frac{s(x_i) - b(x_i)}{b(x_i)} \right] = \sum_{i=1}^n \int s(x) dx - \sum_{i=1}^n \int b(x) dx = n - n = 0,$$

whence $\nu_1 = 0$. Similar calculations now yield:

$$\nu_2 = \frac{1}{n} \mathbb{E}[\ell_1(0)^2] = \mathbb{E}_s[s/b] - 1.$$

$$\nu_3 = \frac{1}{n} \mathbb{E}[\ell_1(0)^3] = \mathbb{E}_s[s^2/b^2] - 3\mathbb{E}_s[s/b] + 2 = \gamma \nu_2^{3/2}.$$

$$\nu_{11} = \frac{1}{n} \mathbb{E}[\ell_1(0)\ell_2(0)] = -\gamma \nu_2^{3/2}.$$

$$\nu_{001} = \frac{1}{n} \mathbb{E}[\ell_3(0)] = 2\gamma \nu_2^{3/2}.$$

$$\nu_{101} = \frac{1}{n} \mathbb{E}[\ell_1(0)\ell_3(0)] = 2[\mathbb{E}_s(s^3/b^3) - 4\mathbb{E}_s(s^2/b^2) + 6\mathbb{E}_s(s/b) - 3] = 2\rho \nu_2^2.$$

$$\nu_4 = \frac{1}{n} \mathbb{E}[\ell_1(0)^4] - \frac{3}{n} [\mathbb{E}[\ell_1(0)^2]^2] = (\rho - 3)\nu_2^2.$$

$$\nu_{0001} = \frac{1}{n} \mathbb{E}[\ell_4(0)] = -6\rho \nu_2^2.$$

$$\nu_{02} = \frac{1}{n} \mathbb{V}[\ell_2(0)] = (\rho - 1)\nu_2^2.$$

$$\nu_{21} = \frac{1}{n} \mathbb{E}[\ell_1(0)^2\ell_2(0)] - \frac{1}{n} \mathbb{E}[\ell_1(0)^2] \mathbb{E}[\ell_2(0)] = (1 - \rho)\nu_2^2.$$

These results use the fact that the expressions for $\gamma$ and $\rho$ defined in (3.1) and (3.2) become $\gamma = \nu_3/\nu_2^{3/2}$ and $\rho - 3 = \nu_4/\nu_2^2$. Straightforward computations also yield the following expressions for the information numbers under model (1.1):

$$J(\alpha) = \sum_{i=1}^n \frac{(s(x_i) - b(x_i))^2}{p(x_i|\alpha)^2}, \quad \text{and} \quad I(\alpha) = n\mathbb{E}_s \left[ \frac{(s(x) - b(x))^2}{s(x)p(x|\alpha)} \right].$$
B Third-order edgeworth expansions for $R_{W3}$ and $R_{W4}$

For $R_{W3}$, substitution of the cumulants in Table 2 into (2.5) gives:

$$P(R_{W3} \leq z) = \Phi(z) - \phi(z) \left[ n^{-1/2} \gamma \left( -\frac{1}{2} - \frac{1}{3} H_2(z) \right) + n^{-1} \left( \frac{1}{36} (63 \rho - 28 \gamma^2) z + \frac{1}{12} (5 \rho + 2 \gamma^2) H_3(z) + \frac{\gamma^2}{18} H_5(z) \right) + O(n^{-3/2}) \right].$$

For $R_{W4}$, the second cumulant behavior differs from the other cases considered in that $\kappa_2 = 1 + O(n^{-1/2})$, so that the Edgeworth expansion in (2.5) must be rederived:

$$F(z) = \Phi(z) - \phi(z) \left[ \kappa_1 + \frac{1}{2} (\kappa_1^2 + \kappa_2 - 1) z + \left( \frac{1}{6} \kappa_3 + \frac{1}{2} \kappa_1 (\kappa_2 - 1) \right) H_2(z) \right. \left. + \left( \frac{1}{6} \kappa_1 \kappa_3 + \frac{1}{24} \kappa_4 + \frac{1}{8} (\kappa_2 - 1)^2 \right) H_3(z) + \frac{1}{12} (\kappa_2 - 1) \kappa_3 H_4(z) + \frac{1}{72} \kappa_2^2 H_5(z) + O(n^{-3/2}) \right].$$

Substitution of the cumulants in Table 2 into this expression then yields:

$$P(R_{W4} \leq z) = \Phi(z) - \phi(z) \left[ n^{-1/2} \gamma \left( \frac{z}{3} - \frac{1}{2} - \frac{1}{3} H_2(z) \right) + n^{-1} \left( \frac{\gamma^2}{3} - \rho + \frac{1}{36} (63 \rho - 22 \gamma^2) z - \frac{2}{3} \rho H_2(z) + \frac{1}{36} (15 \rho + 8 \gamma^2) H_3(z) \right. \right. \left. - \frac{\gamma^2}{9} H_4(z) + \frac{\gamma^2}{18} H_5(z) \right) + O(n^{-3/2}) \right].$$

C Fourth-order edgeworth expansion for $R_{LR}$

We extend (2.5) by computing the $O(n^{-2})$ Edgeworth expansion for the normal density:

$$F(z) = \Phi(z) - \phi(z) \left[ \kappa_1 + \frac{1}{2} (\kappa_1^2 + \kappa_2 - 1) z + \left( \frac{1}{6} \kappa_3 + \frac{1}{2} \kappa_1 (\kappa_2 - 1) \right) H_2(z) \right. \left. + \left( \frac{1}{6} \kappa_1 \kappa_3 + \frac{1}{24} \kappa_4 \right) H_3(z) + \left( \frac{1}{12} \kappa_3 (\kappa_2 + \kappa_2 - 1) + \frac{1}{24} \kappa_1 \kappa_4 + \frac{1}{120} \kappa_5 \right) H_4(z) \right. \left. + \frac{1}{72} \kappa_2^2 H_5(z) + \frac{1}{144} \kappa_3 (2 \kappa_1 \kappa_3 + \kappa_4) H_6(z) + \frac{1}{1296} \kappa_2^2 H_7(z) + O(n^{-2}) \right]. \quad (C.1)$$

This expansion assumes, as is the case for $R_{LR}$, that $\kappa_2 = 1 + O(n^{-1})$. (Expansions for statistics in which $\kappa_2 = 1 + O(n^{-1/2})$, e.g., $R_{W4}$, are more complicated.) To compute the cumulants of $R_{LR}$, introduce, similarly to $\rho$ and $\gamma$ defined as before, the (dimensionless) quantity

$$\xi = \frac{V_5}{24 (-V_2)^{5/2}} = \frac{\mathbb{E}_s [\frac{x^4}{b^4}] - 5 \mathbb{E}_s [\frac{x^2}{b^2}] - 10 \mathbb{E}_s [\frac{x^4}{b^4}] - 10 \mathbb{E}_s [\frac{x^2}{b^2}] - 4}{(\mathbb{E}_s [\frac{x^2}{b^2}] - 1)^{5/2}}.$$ 

In terms of these parameters, and to an accuracy of $O(n^{-2})$, we obtain the following expressions for the first five cumulants of $R_{LR}$:

$$\kappa_1 = \frac{\gamma}{6} n^{-1/2} + \left( \frac{\gamma}{16} - \frac{\gamma^3}{12} + \frac{5 \gamma \rho}{16} - \frac{11 \xi}{40} \right) n^{-3/2},$$

$$\kappa_2 = 1 + \frac{18 \rho - 13 \gamma^2}{36 n}, \quad \kappa_3 = \left( \frac{11 \gamma \rho}{4} - \frac{251 \gamma^3}{216} - \frac{9 \xi}{5} \right) n^{-3/2}.$$
with \( \kappa_4 = \kappa_5 = 0 \). Substitution of these cumulants into (C.1) gives:

\[
Pr(R_{LR} \leq z) = \Phi(z) - \phi(z) \left\{ n^{-1/2} \left( -\frac{2}{6} \right) + n^{-1} \left( \frac{1}{12} (3\rho - 2\gamma^2)z \right) + n^{-3/2} \left( \frac{\gamma}{16} - \frac{\gamma^3}{12} + \frac{5\gamma \rho}{16} - \frac{11\xi}{40} + \left[ \frac{5\gamma \rho}{12} - \frac{71\gamma^3}{432} - \frac{3\xi}{10} \right] H_2(z) \right) + O(n^{-2}) \right\}.
\]

(C.2)

For the model settings as in the illustrative example of section 4, \( \xi \approx 5.0 \). The \( O(n^{-2}) \) prediction agrees with the simulations better than the \( O(n^{-3/2}) \) prediction in terms of the mean, standard deviation, skewness, kurtosis, and \( \chi^2 \) test values listed in tables 3–5. However, the \( O(n^{-3/2}) \) approximation works better than \( O(n^{-2}) \) for high values of \( z \) and, in particular, for predicting the exceedances at \( z = 5 \) listed in table 6.

D Algorithm for global \( p \)-value adjustment

Letting \( \theta \in \Theta \) denote the vector of all nuisance parameters, obtain the (global) \( p \)-value for signal significance from GVM:

\[
p_{\text{global}} = Pr(R_{LR}(\hat{\theta}) > r(\hat{\theta})),
\]

where \( r(\theta) \) is the observed (local) value of the \( R_{LR}(\theta) \) statistic (with implicit dependence on \( \theta \) emphasized) computed from the sample on hand, and \( \hat{\theta} = \arg \max_{\theta \in \Theta} r(\theta) \). We now propose to adjust this \( p \)-value as follows.

(i) Recalling the normal approximation error for \( r(\theta) \) defined in (3.10), \( \Delta R(r(\theta)) = r(\theta) - \bar{r}(\theta) \), locate the value of \( \theta \) that yields the largest such error:

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

(Note that this search can use the same grid as the above search for \( p_{\text{global}} \).)

(ii) Express the global \( p \)-value in terms of the equivalent global \( r \), the \( N(0,1) \) quantile corresponding to a right-tail probability of \( 1 - p_{\text{global}} \):

\[
r_{\text{global}} = \Phi^{-1}(1 - p_{\text{global}}).
\]

(iii) Now adjust the global \( r \),

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

(iv) Finally the global adjusted \( p \)-value is then:

\[
p_{\text{adj}} = 1 - \Phi(r_{\text{adj}}).
\]

Note that the \( \Delta R \) adjustment calculated at \( \theta^* \) will normally exceed the adjustment not only at \( \hat{\theta} \) but also at the unknown “true” \( \theta \). It is in this sense that the procedure yields a conservative global \( p \)-value.

– 24 –
References

[1] S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, *Ann. Math. Stat.* 9 (1938) 60.

[2] R.B. Davies, Hypothesis testing when a nuisance parameter is present only under the alternative, *Biometrika* 74 (1987) 33.

[3] L. Lyons, Open Statistical Issues in Particle Physics, *Ann. Appl. Stat.* 2 (2008) 887.

[4] R.J. Adler, *The Geometry of Random Fields*, Wiley, New York U.S.A. (1981).

[5] R.J. Adler and J.E. Taylor, *Random Fields and Geometry*, Springer, New York U.S.A. (2007).

[6] E. Gross and O. Vitells, Trial factors for the look elsewhere effect in high energy physics, *Eur. Phys. J. C* 70 (2010) 525 [arXiv:1005.1891].

[7] O. Vitells and E. Gross, Estimating the significance of a signal in a multi-dimensional search, *Astropart. Phys.* 35 (2011) 230 [arXiv:1105.4355].

[8] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, Hoboken U.S.A. (2000).

[9] G. Cowan, K. Cranmer, E. Gross and O. Vitells, Asymptotic formulae for likelihood-based tests of new physics, *Eur. Phys. J. C* 71 (2011) 1554 [Erratum ibid. C 73 (2013) 2501] [arXiv:1007.1727].

[10] T. Severini, *Likelihood Methods in Statistics*, Oxford University Press, New York U.S.A. (2000).

[11] E. Lehmann and J. Romano, *Testing Statistical Hypotheses*, third edition, Springer, New York U.S.A. (2005).

[12] J. Shao, *Mathematical Statistics*, second edition, Springer, New York U.S.A. (2003).

[13] W.T. Eadie et al., *Statistical Methods in Experimental Physics*, North-Holland, Amsterdam The Netherlands (1971).

[14] F. James and M. Roos, Minuit: A System for Function Minimization and Analysis of the Parameter Errors and Correlations, *Comput. Phys. Commun.* 10 (1975) 343.

[15] D.R. Cox and O.E. Barndorff-Nielsen, *Inference and Asymptotics*, Chapman and Hall/CRC, London U.K. (1994).

[16] P.A. Mykland, Bartlett identities and large deviations in likelihood theory, *Ann. Stat.* 27 (1999) 1105.

[17] L. Lyons, Discovering the Significance of 5 sigma, arXiv:1310.1284.

[18] D.N. Joanes, C.A. Gill, Comparing measures of sample skewness and kurtosis, *J. Royal Stat. Soc. D* 47 (1998) 183.