



Full Length Article

# Under-coverage in high-statistics counting experiments with finite MC samples

C.A. Alexe <sup>a,c</sup>, J. Bendavid <sup>d</sup>, L. Bianchini <sup>b,c,\*</sup>, D. Bruschini <sup>a,c</sup><sup>a</sup> Scuola Normale Superiore, P.zza dei Cavalieri, 7, Pisa, 56126, Italy<sup>b</sup> Department of Physics E. Fermi, University of Pisa, Largo B. Pontecorvo, 3, Pisa, 56127, Italy<sup>c</sup> Sezione di Pisa, INFN, Largo B. Pontecorvo, 3, Pisa, 56127, Italy<sup>d</sup> European Organization for Nuclear Research (CERN), Geneva, Switzerland

## ARTICLE INFO

## Keywords:

HEP

Confidence interval

Coverage

Barlow–Beeston

## ABSTRACT

We consider the problem of setting confidence intervals on a parameter of interest from the maximum-likelihood fit of a physics model to a binned data set with a large number of bins, large event-counts per bin, and in the presence of systematic uncertainties modeled as nuisance parameters. We use the profile-likelihood ratio for statistical inference and focus on the case in which the model is determined from Monte Carlo simulated samples of finite size. We start by presenting a toy model in which the properties of widely used approximations of the profile-likelihood ratio in the asymptotic limit, which are commonly expected to hold in the high-statistics regime, are manifestly broken even if the numbers of events per bin in both the data and simulated samples are seemingly large enough to warrant their validity. We then move to the general setting to show how statistical uncertainties in the Monte Carlo predictions can affect the coverage of confidence intervals constructed in the asymptotic approximation always in the same direction, namely they lead to systematic under-coverage.

## 1. Introduction

Experiments in particle physics are often concerned with the problem of estimating a parameter of interest ( $\mu$ ) from the maximum-likelihood fit of a physics model to sampling distributions constructed from a large number of identically repeated events, such as decays or collisions of particles. When the number of events is large enough to populate the full space of experimental outcomes, sampling distributions can be more conveniently approximated by histograms with a finite number of bins, resulting in a so-called *binned analysis*. Best-fit values for the parameter of interest (POI) are generally of limited use unless presented together with a *confidence interval* (CI), that is the range of allowed values in which the *true* value  $\mu_t$  of the POI is likely to be found within some conventional level of confidence.

Let us then assume that the experiment under consideration aims at measuring the number of occurrences of some particle physics reaction in the  $n$  bins of a histogram, and that the expected number of events in each bin is related to  $\mu$  according to a given physics model. Under fairly general assumptions, the event counts  $\mathbf{y} = (y_1, \dots, y_n)$  can be modeled as a vector of  $n$  independent and Poisson-distributed random variables whose expectation value  $\langle \mathbf{y} \rangle = \mathbf{f}(\mu)$  is thus sufficient to fully define the joint probability density function of the data. The model  $\mathbf{f}(\mu)$

might be affected by other sources of systematic uncertainties, related for example to imperfect knowledge of the detector response, the presence of backgrounds, or additional unknown parameters. We will assume that these uncertainties can be accounted for by extending the parameter space to include additional variables  $\theta$ , commonly known as *nuisance parameters*.

Due to the complexity of modeling particle physics processes from generator up to detector-level, the function  $\mathbf{f}(\mu, \theta)$  can rarely be computed analytically, leaving Monte Carlo (MC) techniques as the only viable way for its determination. Given that MC samples necessarily comprise a finite number of events, a MC-simulated model prediction will never be *exact*, thus introducing an additional source of systematic uncertainty. A general rule is that the size of the samples should exceed that of the data by at least a few times. However, such a request can turn out computationally expensive or even unfeasible for analyses performed on high-statistics samples of events, especially when CPU-intensive MC simulations are needed to achieve the desired level of theoretical accuracy. In such cases, one has to live with samples of MC-simulated events of size comparable or smaller than the data. The situation can become even more problematic when the simulated events are weighted [1].

\* Corresponding author at: Department of Physics E. Fermi, University of Pisa, Largo B. Pontecorvo, 3, Pisa, 56127, Italy.  
E-mail address: [lorenzo.bianchini@unipi.it](mailto:lorenzo.bianchini@unipi.it) (L. Bianchini).

The correct frequentist approach to deal with the finite size of MC samples was first introduced by Barlow and Beeston in Ref. [2]. In their seminal paper, the authors proposed the idea of treating the uncertain MC predictions as additional nuisance parameters of which the available MC sample represents an ancillary measurement. Since then, several improvements in the treatment of MC uncertainties have been proposed in the literature, with particular emphasis on the case of low-statistic samples and the proper handling of event weights [1–4].

In this paper, we focus on the problem of setting confidence intervals on a parameter of interest in the presence of nuisance parameters and in the case of high-statistics counting experiments, for which asymptotic formulas from the theory of maximum likelihood estimators [5–7] are expected to hold. We notice that the high-statistics case has received less attention in the literature compared to the low-statistics limit, where asymptotic formulas are more likely invalidated by low event counts. Furthermore, the high-statistics case is presently of great interest for precision measurements performed on large-scale data samples, such as those collected at the LHC, see e.g. Refs. [8,9] for a recent review on this subject.

A relevant example of a precision measurement performed on a large-scale data sample, and with a large number of nuisance parameters, is represented by the measurement of the  $W$  boson mass by the CMS Collaboration [9]: the CMS measurement analyzed a sample of LHC collision data consisting of about  $10^8$  single-muon events, binned in the kinematics and charge of the muons, resulting in approximately  $3 \times 10^3$  bins, which are used to simultaneously measure the mass of the  $W$  boson together with a few thousands of nuisance parameters via a maximum-likelihood fit. The model is constructed by using samples of MC-simulated events whose overall statistical power exceeds that of the data by a factor of about 4. Alternate MC predictions corresponding to  $\pm 1\sigma$  variations of each nuisance parameter are constructed by reweighting the nominal MC simulation. The effect of statistical fluctuations in the alternate templates used to construct the systematic variations was observed to lead to a moderate under-coverage of Hessian uncertainties, which required an inflation of the statistical uncertainty of the MC templates to restore the correct coverage [9]. The main scope of this paper is to study the origin of such an effect from a more general perspective.

This paper is organized as follows. We start by presenting a hypothetical experiment which can be considered paradigmatic of our case of interest: the data set consists in a histogram with a large number of bins; the event counts in each bin are large; the physics model used to describe the data includes nuisance parameters and is determined from MC-simulated samples of size comparable to the data. We compute numerically the coverage of intervals constructed from CI-setting methods that rely on the asymptotic properties of maximum-likelihood estimators, such as those based on the Hessian matrix or the profile-likelihood ratio, and find that such intervals cover systematically less than their claimed confidence level. Then, we try alternative CI-setting methods that do not rely on asymptotic properties, such as those inspired by the unified approach by Feldman and Cousins [10] based on Neyman's construction [11] and again find similar issues. Inspired by our findings, we propose a heuristic definition of a confidence interval that better approaches the *correct* Feldman–Cousins interval. In the second part of the paper, we study the same problem from a more general perspective to ascertain in which circumstances statistical fluctuations of the MC model can lead to a systematic under-coverage of asymptotic confidence intervals. We conclude by suggesting numerical tests to quantify the impact of MC statistical fluctuations for a generic problem, remarking that the validity of the asymptotic approximation implied by Wilks' theorem should be carefully verified in the context of binned profile-likelihood analyses even when the number of measured events per bin is large and the statistical power of the simulated samples comparable to the data.

## 2. A paradigmatic toy model

Consider a hypothetical experiment measuring a set of event counts  $y_i$ , with  $i = 1, \dots, n$ , collected in the  $n$  bins of a histogram. The measured events can be produced by either a signal or a background process, labeled by the indexes  $j = 1$  and  $2$ , respectively. The two processes are assumed to generate independent Poisson-distributed event counts  $y_{ji}$  with expectation values in the  $i$ th bin given by

$$\langle y_{1i} \rangle \equiv v_{1i} = K, \quad \langle y_{2i} \rangle \equiv v_{2i} = \begin{cases} (1 + \epsilon)K & \text{for } i < \frac{n}{2} \\ K & \text{for } i \geq \frac{n}{2} \end{cases} \quad (1)$$

where  $|\epsilon| \ll 1$  and  $K \gg 1$  are constant. For simplicity, we take  $n$  as an even integer. In a real experiment, the constant  $K$  would be related to e.g. beam flux, detector efficiency, cross sections, branching ratios, etc. In our study, we find it more convenient to express  $K$  in terms of the total number of expected events ( $N$ ) in the data sample, which can be easily determined in terms of  $n$ ,  $K$ , and  $\epsilon$  from Eq. (1).

Let the purpose of this hypothetical experiment be a measurement of the unknown total number of signal events, henceforth denoted as *parameter of interest* (POI), with the unknown normalization of the background process playing the role of a *nuisance parameter* (NP). For this purpose, we will make use of a physics model which can predict the probability for a signal or background event to fall in a given bin. If this model could be derived directly from Eq. (1), the data histogram could be re-binned in just two bins with no loss of information on either the POI or NP. However, the crucial assumption made here is that the physics model is not known *analytically* but can only be predicted from finite-size MC samples consisting of  $t_{ji}$  simulated events in the  $i$ th bin and for the  $j$ th process. Since  $t_{ji}$  are also random variables, they will differ from their *true values* by statistical fluctuations introduced by the MC generation, which could, in fact, justify preserving the full granularity of the data.

Given a set of MC predictions  $t_{ji}$ , the signal normalization can be extracted by parameterizing the unknown expectation value  $\langle y_i \rangle$  in the  $i$ th bin as:

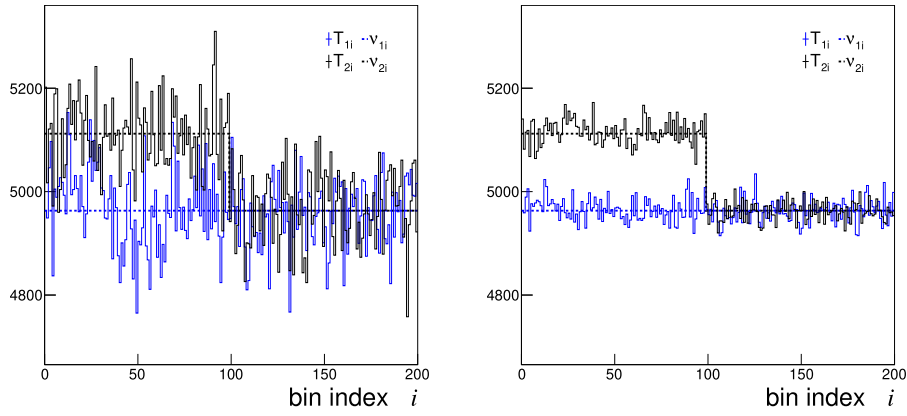
$$f_i(\mu_1, \mu_2; t_{1i}, t_{2i}) = \sum_{j=1}^2 \mu_j \frac{t_{ji}}{k_j} \quad (2)$$

where *strength modifiers*  $\mu_1$  and  $\mu_2$  have been conveniently introduced for the two processes in place of their respective total number of events, and  $k_j$  are constants that rescale the normalization of the MC samples to some reference value, which we are free to choose. In the following, we conveniently set  $k_j$  equal to the ratio between the expectation value of  $t_{ji}$  and its true value  $v_{ji}$ , so that by construction  $f_i$  is an unbiased estimator of  $\langle y_i \rangle$  for  $\mu_1 \equiv \mu_{1,i} = 1$  and  $\mu_2 \equiv \mu_{2,i} = 1$ . With this choice, the constants  $k_j$  represent the statistical power of the simulated samples relative to the data. For example,  $k_1 = k_2 = 1$  means that the total number of simulated events equals on average the number of data events, while values greater (smaller) than one indicate that more (less) events have been generated by the MC simulation than expected in data. Consistently with this assumption, numerical values for  $y_{ji}$  and  $t_{ji}$  in pseudo-experiments can be generated by drawing random integer numbers from the Poisson density with mean values  $v_{ji}$  and  $k_j v_{ji}$ , respectively, that is:

$$\text{Prob}[y_{ij}] = \frac{e^{-v_{ji}} (v_{ji})^{y_{ji}}}{y_{ji}!} \quad \text{and} \quad \text{Prob}[t_{ij}] = \frac{e^{-k_j v_{ji}} (k_j v_{ji})^{t_{ij}}}{t_{ij}!}. \quad (3)$$

We notice that the rescaled MC predictions  $T_{1i} = t_{1i}/k_1$  and  $T_{2i} = t_{2i}/k_2$ , commonly referred to as *templates*, also define the Jacobian matrix of the physics model  $\mathbf{f} = (f_1, \dots, f_n)$ , with respect to its parameters  $(\mu_1, \mu_2)$ , that is

$$\frac{\partial(f_1, \dots, f_n)}{\partial(\mu_1, \mu_2)} = \begin{pmatrix} T_{11} & T_{21} \\ \vdots & \vdots \\ T_{1n} & T_{2n} \end{pmatrix}. \quad (4)$$



**Fig. 1.** Examples of templates  $T_{ji}$  generated as described in the text for  $N = 2 \times 10^6$ ,  $\epsilon = 0.03$ ,  $n = 200$ , and for two values of  $k_1 = k_2 = k$ , namely  $k = 1$  (left) and  $k = 10$  (right). The horizontal dashed lines in each plot provide the expected number of events per bin for the signal and background processes.

The concept of Jacobian will be pivotal to Section 3. Examples of templates  $T_{ji}$  are shown in Fig. 1 for representative values of  $N$ ,  $\epsilon$ ,  $n$ , and  $k_j$ .

In the following, we will use the notation  $\mu = \mu_1$  and  $\theta = \mu_2$  to highlight the different role played by each parameter. In addition to a point estimator for the POI ( $\hat{\mu}$ ), we also want to calculate confidence intervals  $I$  at confidence level  $\alpha$  with the aim that  $\text{Prob}[\mu_t \in I] = \alpha$ , regardless of the true value  $\mu_t$ . In particular, we decide to report two-sided central confidence intervals at the conventional confidence level  $\alpha = 68.3\%$ . We take the half-width  $\hat{\sigma}_I$  of the intervals as an effective “ $1\sigma$ ” standard deviation, and determine the mean and median of its sampling distribution from a large number of identically repeated pseudo-experiments. The coverage of  $I$  is defined as the fraction of pseudo-experiments for which  $\mu_t \in I$ .

In order to verify that the coverage of  $I$  does not depend on  $\mu_t$ , two additional independent data sets  $y_{\pm}$  are also considered. The latter are generated similarly to the left of Eq. (3) with the replacement

$$v_{1i} \rightarrow v_{1i}^{\pm} = v_{1i} \cdot (1 \pm 5\bar{\sigma}_H), \quad (5)$$

where  $\bar{\sigma}_H$  is defined as the standard deviation of  $\hat{\mu}$  when the MC templates  $T_{ji}$  are replaced by their *true* values  $v_{ji}$ , or, equivalently, in the limit of infinite-size MC samples, that is  $k_j \rightarrow \infty$ . Clearly,  $\mu_t = 1 \pm 5\bar{\sigma}_H$  for these alternate data sets.

### 2.1. Likelihood functions for the toy model

The total event counts per bin  $y_i$  can be treated as  $n$  independent random variables drawn from the Poisson density  $\mathcal{P}$ , leading to the definition of the likelihood as:

$$L_P(\mu, \theta) = \prod_{i=1}^n \mathcal{P}(y_i | f_i(\mu, \theta)). \quad (6)$$

The maximum-likelihood estimator (MLE) for the likelihood parameters is the point  $(\hat{\mu}, \hat{\theta})$  that minimizes the negative log-likelihood function (NLL), which, apart from an irrelevant constant, can be written as

$$-\ln L_P(\mu, \theta) = \sum_{i=1}^n [f_i(\mu, \theta; t_{1i}, t_{2i}) - y_i \ln f_i(\mu, \theta; t_{1i}, t_{2i})]. \quad (7)$$

Eq. (7) neglects the fact that  $t_{ji}$  are also random variables. A straightforward way to account for their stochastic nature is the Barlow–Beeston approach [2], which consists in treating the vectors  $\mathbf{t}_j = (t_{j1}, \dots, t_{jn})$  as independent Poisson-distributed random variables with unknown mean values  $\theta_j$ , therefore extending the likelihood to include an extra contribution:

$$-\ln L_{P+BB}(\mu, \theta, \theta_1, \theta_2) = \sum_{i=1}^n [f_i(\mu, \theta; \theta_{1i}, \theta_{2i}) - y_i \ln f_i(\mu, \theta; \theta_{1i}, \theta_{2i})]$$

$$+ \sum_{i=1}^n \sum_{j=1}^2 (\theta_{ji} - t_{ji} \ln \theta_{ji}). \quad (8)$$

A major drawback of Eq. (8) is the growth in dimensionality: the number of likelihood parameters increases from  $p = 2$  to  $p = 2(n + 1)$ . However, as discussed in Ref. [2], numerical minimization of Eq. (8) can be still recast as a two-dimensional optimization problem, at the cost of solving a non-linear system of equations [2] at each step of the minimization, which can be challenging numerically [12]. For this reason, a simplification to Eq. (8) has been proposed by Conway [12], leading to the so-called Barlow–Beeston lite likelihood:

$$-\ln L_{P+BB\text{lite}}(\mu, \theta, \beta) = \sum_{i=1}^n [f_i(\mu, \theta; \beta_i t_{1i}, \beta_i t_{2i}) - y_i \ln f_i(\mu, \theta; \beta_i t_{1i}, \beta_i t_{2i})] + \sum_{i=1}^n m_i [\beta_i - \ln \beta_i], \quad (9)$$

where the last term accounts for the total statistical uncertainty of the nominal MC simulation in the  $i$ th bin as represented by the *effective* number of MC events  $m_i$ , defined as<sup>1</sup>

$$m_i = \left( \frac{t_{1i}}{k_1} + \frac{t_{2i}}{k_2} \right)^2 \left( \frac{t_{1i}}{k_1^2} + \frac{t_{2i}}{k_2^2} \right)^{-1}. \quad (11)$$

Eq. (9) still implies an increase in dimensionality, but this time the additional parameters  $\beta$  can be profiled analytically with a negligible computational overhead. This feature, together with its capability of catching the dominant effect of MC uncertainties, has made the Barlow–Beeston lite approach a very popular and successful solution in HEP analyses.

A summary of the bin-dependent variables introduced in the text is provided in Table 1.

### 2.2. Alternate likelihood parametrization

A data set distributed according to Eq. (1) is clearly not sensitive to the signal normalization when  $\epsilon = 0$ , which translates into a degeneracy

<sup>1</sup> The effective number of MC events can be defined such that its *relative* uncertainty equals the expectation for  $m_i$  unweighted events:

$$\frac{1}{\sqrt{m_i}} = \frac{\sqrt{\text{Var}[T_{1i} + T_{2i}]}}{T_{1i} + T_{2i}} = \frac{\sqrt{\frac{t_{1i}}{k_1^2} + \frac{t_{2i}}{k_2^2}}}{\frac{t_{1i}}{k_1} + \frac{t_{2i}}{k_2}} \quad (10)$$

For a MC sample consisting of  $M$  weighted events, each one having weight  $w_k$ , Eq. (11) can be generalized to  $m_i = \left( \sum_{k=1}^M w_k \right)^2 / \left( \sum_{k=1}^M w_k^2 \right)$ .

**Table 1**

Glossary of the variables defined for the  $i$ th bin, as discussed in the text. In the last line, the model prediction  $f_i$  should be considered as a fully parametric model (i.e. “exact”) for the Barlow–Beeston likelihood and “random” for the Barlow–Beeston lite.

| Variable         | Definition   | Role         |
|------------------|--|--------------|
| $v_{ji}^{(\pm)}$ | True num. of events from the $j$ th process              | exact        |
| $y_i$            | Num. of data events                                      | random       |
| $t_{ji}$         | Num. of MC events from the $j$ th process                | random       |
| $T_{ji}$         | Template value for the $j$ th process ( $= t_{ji}/k_j$ ) | random       |
| $m_i$            | Effective num. of MC events                              | random       |
| $\theta_{ji}$    | Barlow–Beeston NP for the $j$ th process                 | parameter    |
| $\beta_i$        | Barlow–Beeston lite NP                                   | parameter    |
| $f_i$            | Parametric model prediction for $\langle y_i \rangle$    | exact/random |

of the likelihood function with respect to its parameters. For  $\epsilon > 0$ , the problem is mathematically well defined, but the separation between the two processes becomes increasingly more difficult as  $\epsilon$  approaches zero due to the large anti-correlation between  $\mu_1$  and  $\mu_2$ . For example, for values of the  $\epsilon$  parameter of  $\mathcal{O}(1\%)$ , the linear correlation  $\rho$  between the two parameters is almost complete, e.g.  $\rho = -0.99989$  for  $\epsilon = 0.03$ . In fact, one might argue that the individual measurement of  $\mu_1$  and  $\mu_2$  is perhaps a poorly defined problem, given that the data is much more sensitive to their sum than to the difference.

The role played by the choice of parameters is investigated by considering an alternate parametrization of the likelihood defined by the linear transformation

$$\mu'_{1,2} = \frac{1}{2} (\mu_1 \mp \mu_2), \quad (12)$$

with  $\mu'_1$  ( $\mu'_2$ ) taken as POI (NP). Indeed, after applying the transformation in Eq. (12), the linear correlation coefficient<sup>2</sup> between the transformed parameters changes to 0.71 for the same value of  $\epsilon$ . However, we remark that the reduced correlation between parameters is not for free, given that the effective MC statistical power of the transformed templates

$$T'_{ji} = T_{ji} + (-1)^{j+1} T_{2i} \quad (13)$$

has also changed. The obvious solution would be to directly generate templates  $T'_{ji}$  for the transformed parameters, although this might not always be feasible for generic problems, for example because the MC generator is not flexible enough to allow for it.

When considering the alternate parametrization of Eq. (12), we thus set  $\mu = \mu'_1$  and  $\theta = \mu'_2$ , and change the definition of the alternate data sets accordingly, that is:

$$v_{1i}^{\pm} = v_{1i} \cdot (1 \pm 5\bar{\sigma}'_H) \quad \text{and} \quad v_{2i}^{\pm} = v_{2i} \cdot (1 \mp 5\bar{\sigma}'_H), \quad (14)$$

where  $\bar{\sigma}'_H$  is defined as the standard deviation of  $\hat{\mu}'_1$  for the nominal model in the limit of infinite-size MC samples. Given that  $\mu_1$  and  $\mu_2$  are highly correlated, and that their variances are also similar for  $\epsilon \ll 1$ , it is easy to see that  $\bar{\sigma}'_H \approx \bar{\sigma}_H$ , so for the sake of simplicity we will use  $\bar{\sigma}'_H = \bar{\sigma}_H$  in Eq. (14). For this other parametrization, we have  $\mu'_1 = 0$  for the nominal data and  $\mu'_1 = \pm 5\bar{\sigma}_H$  for the alternate.

The experiment can now be interpreted as measuring simultaneously the total number of events together with the asymmetry  $\mathcal{A}$  between the two halves of the histogram. Indeed, by denoting the number of expected events in the first (second) half of the bins as  $N_{<}$  ( $N_{>}$ ), it is easy to prove that for the true model

$$\mathcal{A} = \frac{N_{<} - N_{>}}{N_{<} + N_{>}} = \frac{\epsilon}{4} \left( 1 - \frac{\mu'_1}{\mu'_2} + \mathcal{O}(\epsilon) \right). \quad (15)$$

<sup>2</sup> We remark that the condition number of the covariance matrix is invariant under a linear transformation such as Eq. (12), which indeed offers a better metric to characterize the problem.

The statistical uncertainty on  $\mu'_2$  is of  $\mathcal{O}(N^{-1/2})$ , so it will be highly constrained when  $N$  is large. Hence, the uncertainty on  $\mathcal{A}$ , which is directly proportional to the uncertainty on the ratio  $\mu'_1/\mu'_2$  as for Eq. (15), is mostly driven by the limited precision on  $\mu'_1$ . Measurements of asymmetries between two portions of a distribution occur frequently in Particle Physics. The determination of the  $W$  boson mass at hadron colliders from the distribution of transverse momenta of the decaying leptons, already mentioned in Section 1, is a relevant example of this kind. Indeed, this measurement can be interpreted, at least to a first approximation, as the determination of the asymmetry of the transverse momentum distribution around the Jacobian peak [13].

### 2.3. Likelihood functions in the Gaussian approximation

In the limit  $y_i \gg 1$ , numerical approximations of the negative log-likelihood function can be made. Indeed, the latter consists in a sum of terms of the same form which can be expanded around  $f_i = y_i$  yielding:

$$f_i - y_i \ln f_i = \frac{(f_i - y_i)^2}{y_i} + \frac{2}{3} \frac{(f_i - y_i)^3}{y_i^2} + \frac{1}{2} \frac{(f_i - y_i)^4}{y_i^3} + \dots, \quad (16)$$

where the dots indicate terms that either do not depend on  $f_i$  or are of higher-order in the expansion. Neglecting terms beyond the quadratic order amounts to approximate the joint Poisson density as a Gaussian multivariate with covariance matrix  $\mathbf{V} = \text{diag}(y)$ . For the toy model, it has been verified that a joint Poisson density and its Gaussian approximation result in consistent point estimators and confidence intervals for all values of  $N$  and  $n$  considered in this study. For example, for  $N = 2 \times 10^6$  and  $n = 200$ , the number of events per bin is  $y_i \sim 10^4$ , and the second term in the expansion represents a correction of order  $(y_i - v_i)/y_i \sim \mathcal{O}(\sqrt{y_i})/y_i \sim 1\%$ . The complete list of likelihood functions in the Gaussian approximation used in the study is presented in Appendix A.

Another practical advantage of the Barlow–Beeston lite approach discussed in Section 2.1 becomes evident in the Gaussian approximation. To see this, we first notice that the right-hand side of Eq. (7), which would be the correct NLL function in the case of infinite-size MC samples, converges to a quadratic function of  $(\theta, \mu)$  in the limit  $y_i \gg 1$ , with error matrix given by  $\mathbf{V}$ . Profiling the  $\beta$  parameters in Eq. (9) introduces a nonlinearity in the parameters, which now appear also in the squared errors at the denominator of the profile likelihood function, see Eq. (A.6) in the appendix. Formally, this function would be the same quadratic function as in the infinite-size MC case, provided that the error matrix is replaced by  $\mathbf{V} + \mathbf{V}_{\text{MC}}$ , where  $\mathbf{V}_{\text{MC}}$  is a diagonal matrix with  $i$ th diagonal element equal to  $f_i^2/m_i$ . This additional contribution to the error matrix depends on  $(\mu, \theta)$ , thus formally breaking the quadratic dependence. However, we notice that, for values of  $(\mu, \theta)$  around the minimum,  $f_i$  is constrained to take values of order  $y_i$  by the penalty term  $(y_i - f_i)^2$  appearing at the numerator of the profile likelihood function, so that  $f_i^2/m_i \approx y_i/k$  to a very good approximation, where  $k \sim m_i/y_i$  represents the total statistical power of the MC simulation in the  $i$ th bin relative to the data.

This observation allows us to conclude that the profiled likelihood for the Barlow–Beeston lite case, can be approximated, at least numerically, as a quadratic function of  $\mu$  and  $\theta$  with error matrix given by  $\mathbf{V} + \mathbf{V}_{\text{MC}} \approx \mathbf{V} \cdot (1 + k^{-1})$ . In the following, we will denote the latter function as “Gaus. + MC stat”. As expected, numerical differences between minimizing the Barlow–Beeston lite likelihood or its Gaus. + MC stat. approximation are found to be negligible in the context of the toy study.

### 2.4. Confidence intervals in the asymptotic limit

Based on the properties of the maximum-likelihood estimators and of the profile-likelihood ratio in the asymptotic limit [6], we consider the following methods to set confidence intervals on  $\mu$ .

- **Hessian uncertainty**

The Hessian uncertainty  $\hat{\sigma}_H$  is defined as the square root of the last diagonal element of the inverse Hessian matrix of the NLL function at the minimum:

$$C = \left( -\frac{\partial^2 \ln L}{\partial^2(\theta, \mu)} \right)^{-1}, \quad (17)$$

where  $\theta$  is a shorthand for the complete set of nuisance parameters. In the asymptotic limit, the maximum-likelihood estimators are known to be Gaussian-distributed around their true values [5], with covariance matrix given by the expectation value of  $C$ , so that

$$[\hat{\mu} - \hat{\sigma}_H, \hat{\mu} + \hat{\sigma}_H], \quad (18)$$

is (asymptotically) a 68.3% CL interval for  $\mu$ .

- **Profile-Likelihood ratio**

The profile-likelihood ratio (PLR) is a function of  $\mu$  defined as [6]

$$t_\mu = -2 \ln \frac{L(\mu, \hat{\theta}_\mu)}{L(\hat{\mu}, \hat{\theta})}, \quad (19)$$

where  $\hat{\mu}$  and  $\hat{\theta}$  are the maximum-likelihood estimators for  $\mu$  and  $\theta$ , respectively, while  $\hat{\theta}_\mu$  is the NP point that minimizes the NLL at a fixed value of  $\mu$ . In the asymptotic limit, the distribution of  $t_\mu$  tends to a chi-square with one degree of freedom as for Wilks' theorem [14], with deviations expected at the  $\mathcal{O}(1/\sqrt{N})$  level, with  $N$  related to the size of the data. If this high-order correction can be neglected, the interval

$$\{\mu : t_\mu \leq 1\}. \quad (20)$$

is again a 68.3% CL interval for  $\mu$ .

Given that  $\hat{\mu}$  is also asymptotically Gaussian-distributed, it follows that the confidence intervals defined by Eqs. (17) and (19) coincide in the asymptotic limit.

## 2.5. Numerical results

The minimization of the negative log-likelihood function is performed numerically using the MIGRAD minimizer implemented in Minuit2 [15], with PLR scans performed by MINOS and Hessian-based uncertainties computed by HESSE. Linear algebra operations are performed using the Eigen library [16] in double precision. The nominal configuration parameters for the toy model are reported in Table 2. The effect of changing their values is studied in Section 2.7. For simplicity,  $k_1 = k_2 = k$  has been assumed.<sup>3</sup> Coverage, mean, and median of the  $1\sigma$  confidence interval for  $\mu = \mu_1$  are obtained from an ensemble of  $10^4$  identically repeated pseudo-experiments, and reported in Table 3, while the results for the other choice  $\mu = \mu_1'$  are shown in Table 4.

The expected Hessian uncertainty  $\bar{\sigma}_H$  for the nominal model and in the limit of infinite-size MC samples, that is when  $T_{ji} = v_{ji}$ , can be computed analytically. Using well-known results, which will be further elaborated in Section 3, it is easy to prove that

$$\bar{\sigma}_H = \frac{4}{|\epsilon| \sqrt{N}} (1 + \mathcal{O}(\epsilon)) \approx 0.094, \quad (21)$$

in very good agreement with the numerical result shown in the first row of Tables 3 and 4. Also, one can easily verify that the linear correlation between  $\mu_1$  and  $\mu_2$  differs from  $-1$  by terms of  $\mathcal{O}(\epsilon^2)$ . Differences between the Hessian uncertainty for the nominal and alternate models in Table 3 are mostly due to the different number of expected events when  $\mu_1$  is changed.

The outcome of the numerical study can be summarized as follows.

<sup>3</sup> For example, this would be appropriate when the signal and background can be defined as sub-processes of a same MC sample. The generalization to arbitrary  $k_j$  is also straightforward and is not a critical aspect for this study.

**Table 2**

Nominal (in bold) and alternate configuration parameters for the toy model.

| $N$ [ $\times 10^6$ ] | $\epsilon$ [ $\times 0.01$ ] | $n$                 | $k_1 = k_2 = k$ |
|-----------------------|------------------------------|---------------------|-----------------|
| 2, 10, 100            | 3, 6, 12                     | 20, 100, <b>200</b> | 1, 10, 40       |

1. The Poisson (2<sup>nd</sup> row) and Gaussian (3<sup>rd</sup> row) likelihood provide consistent results in terms of coverage and  $1\sigma$  uncertainty when using the definition of confidence intervals based on the Hessian matrix. In the Gaussian approximation, numerical minimization (3<sup>th</sup> row) or an analytical solution (4<sup>th</sup> row) provide identical results. As expected, by neglecting the stochastic nature of the MC templates, these likelihoods bring to severe under-coverage: the uncertainties are about twice smaller than  $\bar{\sigma}_H$  (1<sup>st</sup> row).
2. The Hessian uncertainty for the Gauss + MC stat. likelihood (5<sup>th</sup> row) gives the correct coverage for the nominal data set, but under-covers for the alternate data sets. The  $1\sigma$  uncertainty is still smaller than  $\bar{\sigma}_H$ , further suggesting that these confidence intervals will generally under-cover. Results obtained by using either the Gauss + MC stat. (5<sup>th</sup> row) or Gauss + BBlite likelihood (6<sup>th</sup> row) are quantitatively similar.
3. The Gauss + BB likelihood with Hessian uncertainty (7<sup>th</sup> row) provides uncertainties that are finally larger than in the asymptotic limit, but intervals still under-cover for both the nominal and alternate models. The level of under-coverage is similar for the three cases, suggesting that this method can at least provide confidence intervals with uniform coverage. It has been verified for several pseudo-data sets that a numerical minimization of the Poisson or Gaussian + BB likelihoods leads to identical numerical results. Motivated by the practical equivalence between the two, in the following we will concentrate on the latter since it can be minimized more efficiently.
4. Uncertainties based on either the Hessian matrix (7<sup>th</sup> row) or the profile-likelihood ratio (8<sup>th</sup> row) in the Gaussian approximation are similar, showing that for each pseudo-data the log-likelihood function is reasonably parabolic around the minimum. However, the sampling distribution of  $t_\mu$  is remarkably different from a chi-square with one degree of freedom, as illustrated by Fig. 2 for  $\mu \equiv \mu_1' = 0$  on pseudo-data generated at the true value  $\mu_1 = 0$ . The sampling distributions of the MLE estimator  $\hat{\mu}$  and of the estimated Hessian uncertainty  $\hat{\sigma}_H$  are shown in Fig. 3. They both deviate visibly from a Gaussian distribution, with the latter featuring a long tail towards the right. Similar results are obtained when considering the other parametrization.
5. Fitting  $(\mu_1, \mu_2)$  or  $(\mu_1', \mu_2')$  leads to qualitatively very similar results in terms of coverage. Numerical differences for  $\mu_1 = 1 \pm 5\bar{\sigma}_H$  are mostly due to the different definition of alternate data, see e.g. Eq. (5) and Eq. (14). Hence, we conclude that under-coverage occurs independently from the parametrization being used.

The main conclusion of this study is that CI-setting methods based on the asymptotic properties of the MLE or PLR are clearly invalidated even though the sample sizes are large in an absolute sense for both data and simulation. This motivates us to go beyond the asymptotic approximation.

## 2.6. Beyond the asymptotic approximation

We have considered the following CI-setting methods inspired by the classical Feldman and Cousins (FC) unified approach [10]. The main feature common to all methods is their use of the profile-likelihood ratio as an ordering principle with no *a priori* assumption on its statistical distribution. The latter is indeed determined *a posteriori* from a collection of randomly generated pseudo-experiments. At variance with the original Feldman–Cousins formulation, the main challenge here is represented by the presence of nuisance parameters.

**Table 3**

Coverage, mean, and median of the  $1\sigma$  confidence interval for  $\mu = \mu_1$  for the toy model with  $N = 2 \times 10^6$ ,  $n = 200$ ,  $k = 1$ , and  $\epsilon = 0.03$ , obtained from an ensemble of  $10^4$  identically repeated pseudo-experiments. Cases where the observed coverage agrees with the expectation of 68.3% within a relative  $\pm 5\%$  tolerance are highlighted with a bold font. When present, the number within parenthesis refers to the statistical error on the last digit.

| Likelihood           | Minimim. | CI method | $\mu_t - 1$        | Coverage         | $\hat{\sigma}$ (mean, median) |
|----------------------|----------|-----------|--------------------|------------------|-------------------------------|
| (1) Gauss (asympt.)  | Analytic | Hessian   | 0                  | 0.683            | 0.095, 0.095                  |
|                      |          |           | $+5\bar{\sigma}_H$ | <b>0.683</b>     | 0.106, 0.106                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.683            | 0.083, 0.083                  |
| (2) Poisson          | Numeric  | Hessian   | 0                  | 0.521(5)         | 0.058, 0.057                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.160(4)         | 0.064, 0.064                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.064(2)         | 0.050, 0.050                  |
| (3) Gauss            | Numeric  | Hessian   | 0                  | 0.520(5)         | 0.057, 0.057                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.161(4)         | 0.064, 0.064                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.063(2)         | 0.050, 0.050                  |
| (4) Gauss            | Analytic | Hessian   | 0                  | 0.520(5)         | 0.057, 0.057                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.161(4)         | 0.064, 0.064                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.063(2)         | 0.050, 0.050                  |
| (5) Gauss + MC stat. | Analytic | Hessian   | 0                  | <b>0.682</b> (5) | 0.081, 0.081                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.230(4)         | 0.086, 0.085                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.124(3)         | 0.076, 0.076                  |
| (6) Gauss + BB-lite  | Numeric  | Hessian   | 0                  | <b>0.679</b> (5) | 0.080, 0.080                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.259(4)         | 0.094, 0.094                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.097(3)         | 0.066, 0.066                  |
| (7) Gauss + BB       | Numeric  | Hessian   | 0                  | 0.439(5)         | 0.137, 0.132                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.449(5)         | 0.163, 0.157                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.454(5)         | 0.116, 0.111                  |
| (8) Gauss + BB       | Numeric  | PLR       | 0                  | 0.454(5)         | 0.141, 0.135                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.461(5)         | 0.170, 0.162                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.469(5)         | 0.116, 0.111                  |
| (9) Gauss + BB       | Numeric  | sPFC      | 0                  | 0.471(5)         | 0.160, 0.146                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.464(5)         | 0.184, 0.169                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.614(5)         | 0.166, 0.155                  |
| (10) Gauss + BB      | Numeric  | PFC       | 0                  | 0.484(5)         | 0.172, 0.157                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.546(5)         | 0.211, 0.188                  |
|                      |          |           | $-5\bar{\sigma}_H$ | <b>0.656</b> (5) | 0.176, 0.160                  |
| (11) Gauss + BB      | Numeric  | Barlett   | 0                  | 0.480(5)         | 0.168, 0.155                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.513(5)         | 0.205, 0.184                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.615(5)         | 0.166, 0.152                  |
| (12) Gauss + BB      | Numeric  | CH        | 0                  | <b>0.681</b> (5) | 0.236, 0.221                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.776(4)         | 0.296, 0.283                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.847(4)         | 0.281, 0.276                  |
| (13) Gauss + BB      | Numeric  | FC Cheat  | 0                  | <b>0.664</b> (5) | 0.229, 0.207                  |
|                      |          |           | $+5\bar{\sigma}_H$ | <b>0.673</b> (5) | 0.258, 0.236                  |
|                      |          |           | $-5\bar{\sigma}_H$ | <b>0.680</b> (5) | 0.190, 0.171                  |
| (14) Gauss + BB      | Numeric  | Heuristic | 0                  | <b>0.678</b> (5) | 0.228, 0.221                  |
|                      |          |           | $+5\bar{\sigma}_H$ | <b>0.660</b> (5) | 0.256, 0.248                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.720(4)         | 0.204, 0.197                  |

#### • Profiled Feldman–Cousins

If the asymptotic approximation is not valid, the profile-likelihood ratio  $t_\mu$  in Eq. (19) is not necessarily chi-square distributed and the coverage properties of Eq. (20) are not guaranteed. In the so-called profiled Feldman–Cousins approach [17], a best guess of the nuisance parameters  $\hat{\theta}_\mu$  is first obtained by minimizing the NLL function at a fixed value of  $\mu$ . Then, for each point  $\mu$  sampling the range of values that we wish to include/exclude from the confidence interval, an ensemble of so-called Feldman–Cousins pseudo-experiments [10] are generated by drawing event counts  $y_i$  around  $f_i(\mu, \hat{\theta}_\mu)$ . The sampling distribution  $P_{FC}$  of  $t_\mu$  is then used to determine a critical value  $c_\mu(\alpha)$  at significance level  $\alpha$ , such that

$$\int_0^{c_\mu(\alpha)} dt_\mu P_{FC}(t_\mu) = \alpha. \quad (22)$$

The confidence interval  $I_{FC}$  is defined as the set of points for which  $t_\mu$  is less than the critical value  $c_\mu(\alpha)$ . Here, we are interested in understanding whether an interval based on Eq. (22)

would contain  $\mu_t$  with the right probability for at least the nominal and alternate toy models. For this purpose, we determine the critical values  $c_\mu(\alpha)$  for  $\mu_k \in \{1, 1 \pm 5\bar{\sigma}_H\}$  and evaluate the coverage for pseudo-data distributed according to  $\mu_t = \mu_k$ . Since we mostly care about coverage, we do not perform the full Feldman–Cousins procedure to determine an exact confidence interval for each pseudo-experiment, which would be very time demanding, but follow instead a simplified procedure, namely, for the true and alternate models we define the confidence interval as

$$\{\mu : t_\mu \leq c_{\mu_k}\}, \quad \text{with } \mu_k \in \{1, 1 \pm 5\bar{\sigma}_H\}. \quad (23)$$

This construction leads to what we denote by *profiled FC* intervals. Notice that Eq. (23) is not equivalent to the full Feldman–Cousins construction because the critical values  $c_\mu$  are not computed for all possible values of  $\mu$ . The two definitions would eventually lead to the same intervals if the sampling distribution  $P_{FC}$  were independent from  $\mu$ .

#### • Simplified Feldman–Cousins

We considered an alternative construction of the FC interval where the best-fit estimators  $(\hat{\mu}, \hat{\theta})$  are first determined for all

**Table 4**

Coverage, mean, and median of the  $1\sigma$  confidence interval for  $\mu = \mu'_1$  for the toy model with  $N = 2 \times 10^6$ ,  $n = 200$ ,  $k = 1$ , and  $\epsilon = 0.03$ , obtained from an ensemble of  $10^4$  identically repeated pseudo-experiments. Cases where the observed coverage agrees with the expectation of 68.3% within a relative  $\pm 5\%$  tolerance are highlighted with a bold font. When present, the number within parenthesis refers to the statistical error on the last digit.

| Likelihood           | Minimim. | CI method | $\mu_t$            | Coverage         | $\hat{\sigma}$ (mean, median) |
|----------------------|----------|-----------|--------------------|------------------|-------------------------------|
| (1) Gauss (asympt.)  | Analytic | Hessian   | 0                  | 0.683            | 0.094, 0.094                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.683            | 0.094, 0.094                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.683            | 0.094, 0.094                  |
| (2) Poisson          | Numeric  | Hessian   | 0                  | 0.516(5)         | 0.057, 0.057                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.001(1)         | 0.057, 0.057                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.002(1)         | 0.057, 0.057                  |
| (3) Gauss            | Numeric  | Hessian   | 0                  | 0.516(5)         | 0.057, 0.057                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.001(1)         | 0.057, 0.057                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.002(1)         | 0.057, 0.057                  |
| (4) Gauss            | Analytic | Hessian   | 0                  | 0.516(5)         | 0.057, 0.057                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.001(1)         | 0.057, 0.057                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.002(1)         | 0.057, 0.057                  |
| (5) Gauss + MC stat. | Analytic | Hessian   | 0                  | <b>0.675</b> (5) | 0.080, 0.080                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.004(1)         | 0.080, 0.080                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.004(1)         | 0.080, 0.080                  |
| (6) Gauss + BB-lite  | Numeric  | Hessian   | 0                  | <b>0.674</b> (5) | 0.080, 0.080                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.004(1)         | 0.080, 0.080                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.004(1)         | 0.081, 0.080                  |
| (7) Gauss + BB       | Numeric  | Hessian   | 0                  | 0.443(5)         | 0.139, 0.133                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.468(5)         | 0.147, 0.140                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.465(5)         | 0.147, 0.141                  |
| (8) Gauss + BB       | Numeric  | PLR       | 0                  | 0.448(5)         | 0.140, 0.134                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.474(5)         | 0.152, 0.144                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.470(5)         | 0.143, 0.137                  |
| (9) Gauss + BB       | Numeric  | sPFC      | 0                  | 0.474(5)         | 0.158, 0.144                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.597(5)         | 0.211, 0.200                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.590(5)         | 0.191, 0.184                  |
| (10) Gauss + BB      | Numeric  | PFC       | 0                  | 0.471(5)         | 0.172, 0.156                  |
|                      |          |           | $+5\bar{\sigma}_H$ | <b>0.671</b> (5) | 0.222, 0.196                  |
|                      |          |           | $-5\bar{\sigma}_H$ | <b>0.675</b> (5) | 0.212, 0.192                  |
| (11) Gauss + BB      | Numeric  | Barlett   | 0                  | 0.466(5)         | 0.167, 0.154                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.633(5)         | 0.213, 0.188                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.635(5)         | 0.201, 0.181                  |
| (12) Gauss + BB      | Numeric  | CH        | 0                  | <b>0.675</b> (5) | 0.235, 0.220                  |
|                      |          |           | $+5\bar{\sigma}_H$ | 0.884(3)         | 0.345, 0.352                  |
|                      |          |           | $-5\bar{\sigma}_H$ | 0.884(3)         | 0.347, 0.354                  |
| (13) Gauss + BB      | Numeric  | FC Cheat  | 0                  | <b>0.677</b> (5) | 0.228, 0.206                  |
|                      |          |           | $+5\bar{\sigma}_H$ | <b>0.682</b> (5) | 0.232, 0.210                  |
|                      |          |           | $-5\bar{\sigma}_H$ | <b>0.672</b> (5) | 0.221, 0.199                  |
| (14) Gauss + BB      | Numeric  | Heuristic | 0                  | <b>0.675</b> (5) | 0.229, 0.222                  |
|                      |          |           | $+5\bar{\sigma}_H$ | <b>0.703</b> (5) | 0.241, 0.232                  |
|                      |          |           | $-5\bar{\sigma}_H$ | <b>0.700</b> (5) | 0.242, 0.233                  |

parameters, including the POI; then, an ensemble of FC pseudo-experiments are generated by drawing Poisson distributed event counts around  $f_i(\hat{\mu}, \hat{\theta})$ ; finally, the sampling distribution of  $t_\mu$  and critical value  $c_\mu$  are computed for  $\mu = \hat{\mu}$ . The resulting interval, which will be denoted as *simplified FC*, is then defined similarly to Eq. (23) with  $\mu_k = \hat{\mu}$ .

- **Profile-likelihood-ratio with Barlett's correction**

In case the asymptotic approximation is not valid, it has been suggested by Barlett [18,19] to modify the profile-likelihood ratio  $t_\mu$  such that its expectation value equals its asymptotic value of one (for one degree of freedom), that is  $t_\mu \rightarrow t'_\mu = t_\mu / \langle t_\mu \rangle$ . Thus, starting from the sampling distributions obtained from the FC pseudo-experiments,  $\langle t_\mu \rangle$  is determined for  $\mu \in \{1, 1 \pm 5\bar{\sigma}_H\}$ , and the corresponding confidence interval defined as

$$\{\mu : t'_\mu \leq 1\}. \quad (24)$$

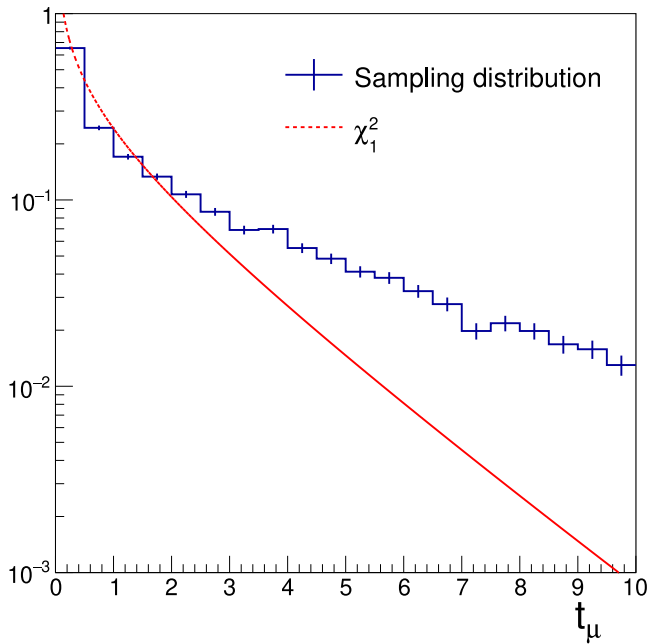
- **Cousins-Highlands**

In the profiled FC approach, all nuisance parameters, including those related to the MC statistical uncertainties, are fixed to their best-fit to the data, but their post-fit uncertainty is neglected

when the FC pseudo-experiments are generated. This can be partly accounted for by following an hybrid Bayesian-frequentist approach first proposed by Cousins and Highland [20]. In the Cousins-Highland (CH) approach, the bin contents  $y_i$  in each FC pseudo-experiment are drawn from different central values  $f_i(\mu, \hat{\theta}_\mu)$ , where  $\hat{\theta}_\mu$  is sampled from a multivariate normal distribution with mean  $\hat{\theta}_\mu$  and covariance matrix given by the *a posteriori* covariance matrix of  $\hat{\theta}_\mu$ . The resulting sampling distribution  $P_{\text{HC}}$  differs from  $P_{\text{FC}}$ , generally leading to more conservative intervals.

- **Feldman-Cousins "cheat"**

Finally, we considered an ideal, albeit unrealistic, construction of confidence intervals which is still based on the PLR, but with all parameters fixed to their *true* values for pseudo-experiments generation. In this case, the sampling distribution of  $t_\mu$  from the FC pseudo-experiments matches the sampling distribution of the pseudo-data so that the coverage should come out right by construction. We will use this definition as a reference for the true  $1\sigma$  uncertainty.



**Fig. 2.** Sampling distribution of  $t_\mu$ , the profile-likelihood function in the Gaussian approximation, evaluated at  $\mu \equiv \mu'_i = 0$  for data distributed according to the nominal model (i.e. with  $\mu'_i = 0$ ) with  $N = 2 \times 10^6$ ,  $\epsilon = 0.03$ ,  $n = 200$ , and  $k = 1$ , obtained from an ensemble of  $10^4$  identically repeated pseudo-experiments.

Using the numerical setup described in Section 2.4, we studied the performances of the newly introduced CI-setting methods. All results are reported in Tables 3 and 4. The main observations can be summarized as follows.

1. The profiled Feldman–Cousins approaches (9th and 10th rows) provide intervals that under-cover in general, although less than from using the PLR. However, the coverage is found to depend on the value of  $\mu_t$ . As expected, the profiled method (10th row) performs better than the simplified one (9th row), suggesting that the distribution  $P_{FC}(t_\mu)$  depends on  $\mu$ .
2. The Barlett correction to the PLR method (11th row) provides results similar to the profiled FC construction (10th row), showing that the sampling distribution  $P_{FC}$  from the FC pseudo-experiments is reasonably close to a chi-square with one degree of freedom. This is also in agreement with the previous observation that intervals based on either the Hessian matrix or the PLR function are almost identical.
3. The Cousins–Highlands method (12th row) leads to the largest uncertainty, but the intervals now over-cover, with a coverage level that depends on  $\mu_t$ .
4. The Feldman–Cousins “cheat” method provides the right coverage by construction: the right CI size for  $\mu$  would be about twice as large as the one obtained from the Hessian or PLR method on the full Barlow–Beeston likelihood, i.e. what one would have obtained based on the asymptotic properties of maximum-likelihood estimators.

The main conclusion of the study is that none of the realistic CI-setting methods considered here can provide the right coverage for the toy model: the asymptotic properties of the PLR are clearly invalidated by non-negligible corrections implied by Wilks’ theorem, which might have been unexpected given the large size of the data sample or the number of bins; on the other hand, methods that do not rely on asymptotic properties suffer from the problem of handling nuisance

parameters, most critically, those related to statistical fluctuations of the MC templates.

However, one could reasonably expect that both issues will be solved in the limit of negligible MC statistical uncertainties, and the right coverage will eventually be restored. This can be achieved in multiple ways, the most obvious being by increasing the size of the simulated samples relative to the data or, if possible, by using larger samples for both. Reducing the number of bins might also help. Alternatively, we might conclude that the problem is related to some intrinsic feature of the model. We test these hypotheses by studying alternate model configurations in the next subsection.

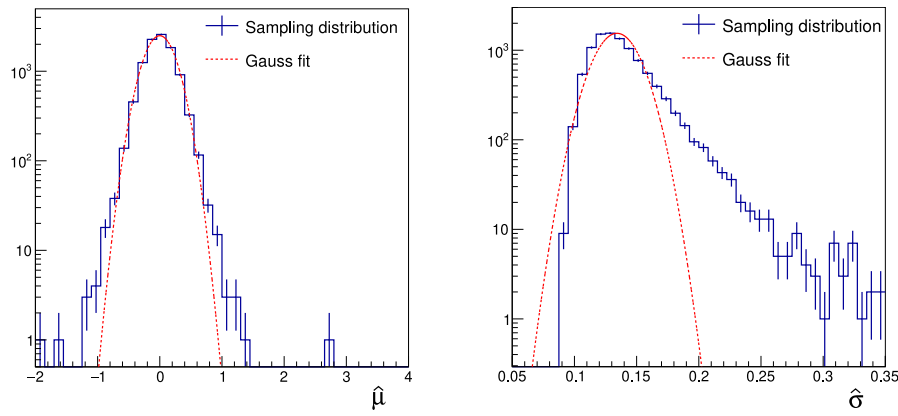
## 2.7. Recovering the correct coverage

The results obtained by using alternate values for  $N$ ,  $\epsilon$ ,  $n$ , and  $k$  are collected in Appendix B. The main observations from this study can be summarized as follows.

1. **Increasing the size of the MC samples.** Tables B.7 and B.8 collect the results obtained by increasing the statistical power of the MC simulation relative to the data ( $k$ ) by factors of 10 and 40, respectively. As one would have naively expected, the impact of the limited MC statistics becomes increasingly less relevant as the MC sample size grows compared to the data. For  $k = 10$ , the correct behavior is still not fully achieved, while for  $k = 40$  all likelihood functions and CI setting methods provide roughly the same intervals and the right coverage is recovered for all cases. This shows that the asymptotic properties of the MLE can be ultimately recovered once the MC template fluctuations become sufficiently small.
2. **Improving the signal-to-background separation.** The results reported in Tables B.9 and B.10 have been obtained by choosing  $\epsilon = 0.06$  and  $\epsilon = 0.12$ , respectively. As the value of  $\epsilon$  increases, the model becomes better behaved, as also reflected by a reduction in the condition number of the Hessian matrix. The coverage of the Barlow–Beeston likelihood with Hessian or PLR uncertainties improves, but the correct coverage cannot be recovered even for  $\epsilon \sim \mathcal{O}(10\%)$ . The level of under-coverage from the FC methods remains stable.
3. **Reducing the number of bins.** The results reported in Tables B.11 and B.12 have been obtained by rebinning the histograms by factors of 2 and 10, respectively. As  $n$  gets smaller, the coverage of the Hessian- or PLR-method improves steadily, but the intervals slightly under-cover even for  $n = 20$ . As expected, fluctuations of MC templates across bins tend to average out with a coarser binning. Again, the level of under-coverage from the FC method does not change significantly.
4. **Increasing the size of both data and simulation.** The results in Tables B.13 and B.14 have been obtained by increasing the number of events in data by factors of 10 and 100, respectively, while keeping the same proportion between data and simulation. For sufficiently large values of  $N$ , the asymptotic limit is ultimately reached when using the full Barlow–Beeston likelihood function (8), so that CI-setting methods based on the Hessian uncertainty or the PLR can finally provide the right coverage. The FC-methods still suffer from systematic under-coverage.

## 2.8. Discussion

The main results of the toy study are summarized in Table 5. The correct likelihood function for the problem under consideration is given by Eq. (8). Neglecting the stochastic nature of the MC templates  $t_{ji}$ , as done, for example, in Eq. (7), has a sizable impact in terms of both the coverage and the size of confidence intervals already with an equal amount of data and simulated events. The “Barlow–Beeston lite” or “MC stat” approaches treat the uncertain MC predictions as an



**Fig. 3.** Sampling distributions of the point-estimator  $\hat{\mu} \equiv \hat{\mu}'_1$  and of its Hessian uncertainty  $\hat{\sigma}_H$  obtained from the Gauss + Barlow–Beeston likelihood applied to data distributed according to the nominal model (i.e. with  $\mu_t = 0$ ) with  $N = 2 \times 10^6$ ,  $\epsilon = 0.03$ ,  $n = 200$ , and  $k = 1$ , obtained from an ensemble of  $10^4$  identically repeated pseudo-experiments. For both distributions, a Gaussian fit is overlaid for reference.

additional source of statistical noise to the data. This accounts in part for the random fluctuations of MC templates, but is not sufficient to restore the right coverage or interval size, as demonstrated by the difference between using the lite vs. full Barlow–Beeston methods. When using the correct likelihood function in the Gaussian approximation, the asymptotic properties of maximum-likelihood estimators and of the profile-likelihood ratio are manifestly broken even if the number of events per bin is large enough to justify, for example, the approximation of the NLL as a chi-square function. We found that the asymptotic properties can eventually be recovered by applying a sufficiently large scaling of the data and/or simulation sizes, but we could not find a first-principle indication of which scale can be considered sufficiently “large”.

Approaches based on the post-fit distribution of the profile-likelihood ratio in the Gaussian approximation suffer from a related problem: the post-fit nuisance parameters, in particular those related with the fluctuations of MC templates, are never exactly set to their true value. The footprint of the fluctuated MC templates survives in the post-fit distribution of  $t_\mu$ , leading to an artificially augmented sensitivity to  $\mu$ . We remark that the problem of under-coverage of confidence intervals in the presence of nuisance parameters is well-known already [17,21–23]. It can be ultimately ascribed to the fact that the standard Neyman’s construction [11] for confidence intervals on one parameter of interest is not uniquely defined when the likelihood depends on multiple nuisance parameters, since some prescription on how to project a larger multi-dimensional space onto a one-dimensional closed interval must be supplied [22,23]. Approaches based on the profile-likelihood function are just instances of the infinitely many ways to achieve such a dimensional reduction. While this matter has been mostly addressed in low-statistics counting experiments and for relatively simple problems, we have shown that the problem can also arise in high-statistics measurements.

### 2.8.1. An heuristic confidence interval

Motivated by these considerations, we propose an *heuristic* confidence interval which combines Hessian uncertainties in the limit of infinite-size MC samples with those obtained from two variants of the Barlow–Beeston method. More precisely, we define

$$I_{\text{heur}} = [\hat{\mu} - \hat{\sigma}_{\text{heur}}, \hat{\mu} + \hat{\sigma}_{\text{heur}}], \quad (25)$$

where  $\hat{\mu}$  is the maximum-likelihood estimator obtained by minimizing the full Barlow–Beeston likelihood from Eq. (8), while  $\hat{\sigma}_{\text{heur}}$  is defined as

$$\hat{\sigma}_{\text{heur}} = \bar{\sigma}_H \cdot \sqrt{1 + \frac{1}{k} \cdot \left( \frac{\hat{\sigma}_H^{\text{BBfull}}}{\hat{\sigma}_H^{\text{BBlite}}} \right)}, \quad (26)$$

where  $\bar{\sigma}_H$  is the Hessian uncertainty in the limit of infinite MC statistics,  $k$  is the statistical power of the MC simulation relative to the data, and  $\hat{\sigma}_H^{\text{BBfull}}$  ( $\hat{\sigma}_H^{\text{BBlite}}$ ) is the Hessian uncertainty derived from the full (lite) Barlow–Beeston likelihoods. We remark the importance of using the maximum-likelihood estimator  $\hat{\mu}$  based on the *full* Barlow–Beeston likelihood rather than its lite approximation, since the latter does not provide an unbiased estimator of  $\mu$ .

The rationale behind Eq. (26) is the following. The Hessian uncertainty derived from the full Barlow–Beeston likelihood underestimates the real statistical spread of  $\hat{\mu}$  due to the presence of spurious constraints which break the asymptotic approximation. Conversely, the Hessian uncertainty  $\bar{\sigma}_H$  would match the standard deviation of  $\hat{\mu}$ , if the MC templates were exact. We take it as our starting point. Our intuition is that statistical uncertainties on the MC templates should at least add in quadrature to the Poisson uncertainty of the data, which corresponds to the square root in Eq. (26). However, assigning an overall uncertainty to the total MC prediction per bin is not fully correct as it ignores the interplay between the  $\mu$  parameter and MC statistical fluctuations pertaining to different templates. In fact, the existence of such an interplay is made evident by the breaking of the Barlow–Beeston lite approximation. The ratio between Barlow–Beeston *full* and *lite* Hessian uncertainties, corresponding to the last term on the right-hand side of Eq. (26), provides an effective correction factor to account for the increased statistical uncertainty when the MC templates  $\theta_{ji}$  are allowed to float in the fit independently from each other.

This heuristic definition (14th row of Tables 3 and 4) provides good coverage properties for all values of  $\mu_t$ , as well as for different choices of model configurations. This suggests that  $\hat{\sigma}_{\text{heur}}$  could be at least used as a proxy for the real statistical spread of the  $\hat{\mu}$  estimator in the presence of fluctuations of both the data and the simulated samples, and so it provides a practical test to gauge the goodness of asymptotic confidence intervals. This observation is intriguing and perhaps worth being further investigated, although a rigorous proof of its validity for a generic problem is beyond the scope of this work. We should remark, however, that  $\hat{\sigma}_{\text{heur}}$  is defined in terms of  $\bar{\sigma}_H$ , which for a realistic problem might not be known *a priori*. In Section 3.6 we will address the problem of estimating  $\bar{\sigma}_H$  when only finite-size MC samples are available.

Two remarks are worth making here. First, we note that Eq. (26) might prove difficult to compute for realistic models that are more complex than our simple toy model, especially when the likelihood function has to handle correlations between different templates. In these cases, more sophisticated solutions should be found. Second, we have restricted our discussion to confidence intervals in one dimension, namely for a single parameter of interest, although in some cases one

**Table 5**

Summary of the coverage properties for different model configurations. A thick (cross) indicates that the right coverage, within a relative  $\pm 5\%$  tolerance, is (not) attained for both the nominal and alternate data sets. In this table, “BB” is a shorthand for Barlow–Beeston.

| MC stat. | CI          | Nominal | $k = 40$ | $\epsilon = 0.12$ | $n = 20$ | $N = 2 \times 10^8$ |
|----------|-------------|---------|----------|-------------------|----------|---------------------|
| none     | Hessian     | ×       | ✓        | ×                 | ×        | ×                   |
| MC stat. | Hessian     | ×       | ✓        | ×                 | ×        | ✓                   |
| BB-lite  | Hessian     | ×       | ✓        | ×                 | ×        | ✓                   |
| BB       | Hessian/PLR | ×       | ✓        | ✓                 | ×        | ✓                   |
| BB       | PFC/Barlett | ×       | ✓        | ×                 | ×        | ×                   |
| BB       | CH          | ×       | ✓        | ✓                 | ×        | ✓                   |
| BB       | FC Cheat    | ✓       | ✓        | ✓                 | ✓        | ✓                   |
| BB       | Heuristic   | ✓       | ✓        | ✓                 | ✓        | ✓                   |

might be interested in setting confidence intervals in more dimensions. How to assess (and possibly correct) under-coverage in more dimensions is beyond the scope of this work.

In the remainder of the paper, we will study the problem of under-coverage of asymptotic confidence intervals from a more general perspective with the goal of identifying the origin of the observed effect.

### 3. The general case

The main scope of this section is to characterize in more generality the role played by statistical fluctuations of MC templates in the determination of confidence intervals constructed in the asymptotic approximation. For this purpose, the following approach is followed.

We start by assuming that the MC templates are *exact*. This approximation allows us to derive a set of well-known analytic results for point estimators and confidence intervals based on the asymptotic properties of maximum-likelihood estimators [24]. The same formalism can also accommodate extensions of the likelihood function that account for finite-size effects in an approximate way, such as the Gauss + MC stat. or Barlow–Beeston *lite* method as discussed in Section 2, which are in practice equivalent to an overall rescaling of the statistical power of the data. Albeit approximate, these methods are still relevant in light of their diffusion in HEP analyses. We will then treat statistical fluctuations of MC templates as perturbations on top of the true unknown templates and collect their contributions to asymptotic formulas as an overall bias term, whose sign and magnitude will then be the main subject of study.

We remark that our treatment is not directly applicable to the *full* Barlow–Beeston method, because the introduction of additional nuisance parameters (associated with the individual MC predictions) breaks the quadratic dependence of the NLL on the parameters, which is a crucial assumption in our derivation. For example, in the context of the toy study discussed in Section 2, the implementation of the full Barlow–Beeston method has been found to affect the coverage of asymptotic confidence intervals in a significant way, although still insufficient to restore the right coverage. We will come back to the validity of the Barlow–Beeston *lite* approximation at the end of this section. We can argue, however, that the main results derived in this section are expected to be relevant even for the full Barlow–Beeston method. Indeed, the main source of under-coverage that will be identified by our study is related to the occurrence of *low-statistics* regimes along particular linear combinations of the data, represented in this context by a finite number of event counts, despite the latter being large in absolute: in such situations, asymptotic formulas for maximum-likelihood estimators are likely to be just not applicable even if the exact likelihood functions were considered.

#### 3.1. Methodology

Consider a binned data set represented by independent Poisson-distributed event counts  $\mathbf{y}$  with expectation value parametrized by the functional model  $\mathbf{f}(\mu, \theta)$ , where  $\mu$  is the parameter of interest and  $\theta = (\theta_1, \dots, \theta_p)$  is a vector of nuisance parameters, which are assumed to be

not constrained by auxiliary measurements. The generalization to the case where both constrained and unconstrained parameters are present is discussed in Appendix C, where an interesting connection between the present work and the main subject of Ref. [25] is pointed out.

As discussed previously, in the limit of large event counts, that is, for  $y_i \gg 1$  for  $i = 1, \dots, n$ , the negative log-likelihood function can be approximated as

$$-\ln L(\mu, \theta) = \frac{1}{2} (\mathbf{y} - \mathbf{f}(\mu, \theta))^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{f}(\mu, \theta)) + \dots \quad (27)$$

where  $\mathbf{V} = \text{diag}(\mathbf{y})$ , and the dots stand for an overall constant plus terms that are of higher-order in  $(y_i - f_i)$ . Aside from a factor of  $1/2$ , the right-hand side of Eq. (27) coincides the well-known Neyman’s  $\chi^2$  test-statistic [26]. We take  $\mathbf{f}$  to be a differentiable function of both  $\mu$  and  $\theta$  and expand it around a suitable initial point  $(\mu_0, \theta_0)$ :

$$\mathbf{f}(\mu, \theta) = \mathbf{f}_0 + \mathbf{j}(\mu - \mu_0) + \mathbf{J}(\theta - \theta_0) + \dots, \quad (28)$$

where  $\mathbf{j}$  and  $\mathbf{J}$  are the  $n \times 1$  and  $n \times p$  blocks of the Jacobian matrix of  $\mathbf{f}$  with respect to  $\mu$  and  $\theta$ , respectively, and the dots stand for higher-order terms in the Taylor expansion. While this approximation is exact for linear functions, such as the toy model of Section 2, for generic functions it should be iterated around the intermediate minima of Eq. (27), until convergence.<sup>4</sup> Hence, we will assume that  $\mathbf{j}$  and  $\mathbf{J}$  are evaluated at a point in the parameter space which is sufficiently close to the global minimum of the NLL function. Without loss of generality, we take  $\mathbf{J}$  as a full-rank matrix, otherwise the model should be redefined to remove the degeneracy.

##### 3.1.1. The profile-likelihood ratio

Thanks to Eq. (28), the function on the right-hand side of Eq. (27) can be minimized by solving a linear system of equations, leading to the well-known *linear least-square estimator* [5]. To this aim, we can first put the right-hand side of Eq. (27) in the more concise form:

$$-2 \ln L(\mu, \theta) = \|\mathbf{d} - \mathbf{b}(\mu - \mu_0) - \mathbf{A}(\theta - \theta_0)\|^2, \quad (29)$$

where  $\|\cdot\|$  indicates the Euclidean norm, while the newly introduced quantities are defined as

$$\mathbf{d} = \mathbf{V}^{-\frac{1}{2}} (\mathbf{y} - \mathbf{f}_0) \quad (30)$$

$$\mathbf{b} = \mathbf{V}^{-\frac{1}{2}} \mathbf{j} \quad (31)$$

$$\mathbf{A} = \mathbf{V}^{-\frac{1}{2}} \mathbf{J} \quad (32)$$

In the following, we will equivalently refer to  $\mathbf{b}$  and  $\mathbf{A}$  as parts of the *Jacobian matrix*, although they formally differ from  $\mathbf{j}$  and  $\mathbf{J}$  for a left multiplication by  $\mathbf{V}^{-\frac{1}{2}}$ . As we shall see, the effect of finite-size MC

<sup>4</sup> The validity of the linear approximation is model dependent and should be assessed case by case. In particular, functional models that change in a very asymmetric way around the true value of the parameters (e.g. due to non-negligible contributions from the second-order derivatives) should be handled with care.

samples on asymptotic confidence intervals can be fully characterized in terms of  $\mathbf{b}$  and  $\mathbf{A}$ .

The NP point that minimizes Eq. (29) at a fixed value of  $\mu$  is provided by the *profile likelihood estimator*  $\hat{\theta}_\mu$ . By inserting the expression for  $\hat{\theta}_\mu$  into Eq. (29), the log-likelihood function becomes a quadratic function of  $\mu$ , which can be again minimized yielding the maximum-likelihood estimator  $\hat{\mu}$ . Explicit formulas for  $\hat{\theta}_\mu$  and  $\hat{\mu}$  are provided in Appendix D.1. After some straightforward calculations, we can finally write the profile log-likelihood function<sup>5</sup> as

$$t_\mu = -2 \ln \frac{L(\mu, \hat{\theta}_\mu)}{L(\hat{\mu}, \hat{\theta})} = \mathbf{b}^T \mathbf{U} \mathbf{b} \cdot (\mu - \hat{\mu})^2, \quad (33)$$

where the matrix  $\mathbf{U}$  is defined as

$$\mathbf{U} = \mathbf{1} - \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T. \quad (34)$$

According to Wilks' theorem [14],  $t_\mu$  is asymptotically distributed as a chi-square with one degree-of-freedom, so that the quadratic form

$$S = \mathbf{b}^T \mathbf{U} \mathbf{b} = \mathbf{b}^T \mathbf{b} - \mathbf{b}^T \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b} = \sigma_{\text{H}}^{-2} \quad (35)$$

must be equal to the inverse variance of  $\hat{\mu}$ . We notice that  $\sigma_{\text{H}}$  can be equivalently obtained from the diagonal elements of the covariance matrix  $\mathbf{C}$ , defined as the inverse-Hessian matrix of the NLL function, see Eq. (17). A proof of this statement, together with an explicit expression of  $\mathbf{C}$  in terms of  $\mathbf{b}$  and  $\mathbf{A}$ , can be found in Appendix D.1.

The matrix  $\mathbf{U}$  introduced in Eq. (34) is a  $n \times n$  symmetric, positive semi-definite, and *idempotent* matrix.<sup>6</sup> The same properties pertain to the matrix  $\mathbf{1} - \mathbf{U}$ , where  $\mathbf{1}$  denotes the  $n \times n$  identity matrix. We exploit the fact that an idempotent matrix can be decomposed as the sum of the outer product of its normalized eigenvectors, and rewrite Eq. (34) in the equivalent form

$$\mathbf{U} = \mathbf{1} - \sum_{j=1}^p \mathbf{u}_j \mathbf{u}_j^T, \quad (36)$$

where  $\mathbf{u}_j$  are linearly independent vectors of unit norm associated with the  $p$  eigenvalues of the square matrix  $\mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$ , which are all equal to one. It can be easily proved that the  $\mathbf{u}_j$  eigenvectors provide an orthonormal basis to the linear space spanned by the  $p$  linearly-independent columns  $\mathbf{a}_1, \dots, \mathbf{a}_p$  of  $\mathbf{A}$ . Being associated with degenerate eigenvalues, the vectors  $\mathbf{u}_j$  obtained from the eigen-decomposition of  $\mathbf{1} - \mathbf{U}$  are not necessarily orthogonal, but can always be made such by using, for example, the Gram-Schmidt process, so they can be taken as mutually orthogonal with no loss of generality.

The formalism developed so far shows that the Hessian uncertainty  $\sigma_{\text{H}}$  is simply related to the norm of the projection of the vector  $\mathbf{b}$  onto the space *orthogonal* to the linear space spanned by the columns of  $\mathbf{A}$ . For brevity, we will denote the latter by  $\mathcal{V}_{\mathbf{A}}$ . Given that the  $\mathbf{u}_j$  vectors provide a complete basis of  $\mathcal{V}_{\mathbf{A}}$ , the  $\mathbf{b}$  vector can be always decomposed as

$$\mathbf{b} = \sum_{j=1}^p (\mathbf{b}^T \mathbf{u}_j) \mathbf{u}_j + \mathbf{b}_{\perp}, \quad (37)$$

where  $\mathbf{b}_{\perp}$  is vector of dimension  $n$  orthogonal to  $\mathcal{V}_{\mathbf{A}}$ , i.e.  $\mathbf{b}_{\perp}^T \mathbf{u}_j = 0$  for  $j = 1, \dots, p$ . It is convenient to consider two limiting cases, namely:

- the  $\mathbf{b}$  vector has a *large projection* onto  $\mathcal{V}_{\mathbf{A}}$ , that is  $\|\mathbf{b}_{\perp}\| \ll \|\mathbf{b}\|$ ;
- the  $\mathbf{b}$  vector is *almost orthogonal* to  $\mathcal{V}_{\mathbf{A}}$ , so that  $\|\mathbf{b}_{\perp}\| \approx \|\mathbf{b}\|$ .

The extent to which a given problem can be classified according to either of the two cases can be quantified by the positive-definite metric

<sup>5</sup> We remark that Eq. (33) would be the *exact* profile-likelihood ratio only in the Gaussian approximation.

<sup>6</sup> A symmetric square matrix  $\mathbf{U}$  is said to be idempotent if  $\mathbf{U}^T \mathbf{U} = \mathbf{U} = \mathbf{U}^T$ .

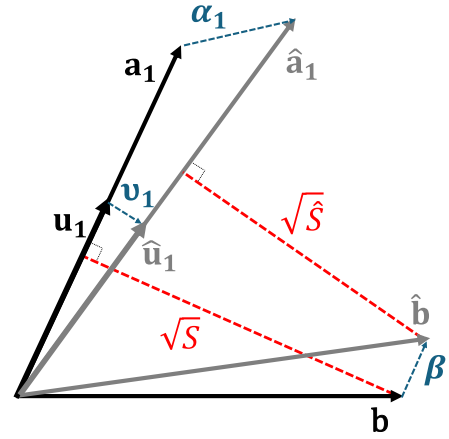


Fig. 4. A sketch showing the effect of statistical fluctuations on the relevant quantities introduced in the text. The simple case  $n = 2$  and  $p = 1$  has been considered.

$\rho_\mu$  defined as<sup>7</sup>

$$\rho_\mu = \sqrt{\frac{\mathbf{b}^T \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}}{\mathbf{b}^T \mathbf{b}}}, \quad (38)$$

which is always less or equal than one, so that the two cases would correspond to either  $\rho_\mu \rightarrow 1$  or 0, respectively.

### 3.2. Impact of MC fluctuations on $\sigma_{\text{H}}$

The results presented in Section 3.1.1 assume that the model  $\mathbf{f}$  is known exactly. Suppose instead that  $\mathbf{f}$  is determined from MC samples consisting of finite numbers of events, thus resulting in perturbed estimators  $\hat{\mathbf{b}} = \mathbf{b} + \beta$  and  $\hat{\mathbf{A}}$ , with the latter giving rise to a perturbed matrix  $\hat{\mathbf{U}}$ , defined as in Eq. (34) with the replacement  $\mathbf{A} \rightarrow \hat{\mathbf{A}}$ . Since  $\hat{\mathbf{U}}$  must also be idempotent, it admits an expansion in terms of a new basis of eigenvectors  $\hat{\mathbf{u}}_j = \mathbf{u}_j + \mathbf{v}_j$ . By expanding the *perturbed* value  $\hat{S}$  around its true value  $S$ , and taking the expectation value in the *space of MC predictions*, that is the mean value of  $\hat{S}$  if infinitely many identical replicas of the MC samples could be generated, we obtain:

$$\begin{aligned} \langle \hat{S} \rangle &= S + 2\langle \beta^T \rangle \mathbf{U} \mathbf{b} + \text{Tr} [\mathbf{U} \langle \beta \beta^T \rangle] \\ &\quad - 2 \sum_{j=1}^p (\mathbf{b}^T \mathbf{u}_j) (\mathbf{b}^T \langle \mathbf{v}_j \rangle) - \mathbf{b}^T \left( \sum_{j=1}^p \langle \mathbf{v}_j \mathbf{v}_j^T \rangle \right) \mathbf{b} \\ &\quad - 2 \sum_{j=1}^p (\text{Tr} [\langle \mathbf{v}_j \beta^T \rangle \mathbf{u}_j \mathbf{b}^T] + \text{Tr} [\langle \mathbf{v}_j \beta^T \rangle] \text{Tr} [\mathbf{u}_j \mathbf{b}^T]) + \dots \end{aligned} \quad (39)$$

The derivation of Eq. (39) from a perturbative expansion of  $\hat{S}$  is presented in Appendix D.2. Fig. 4 helps visualize the geometrical interpretation of  $\hat{S}$ , and the effect of statistical fluctuations on  $\mathbf{A}$  and  $\mathbf{b}$  for a representative problem with  $n = 2$ .

The correlation terms on the last line of Eq. (39) are model dependent, so additional assumptions are needed to proceed. We notice that in many cases the columns of the Jacobian of  $\mathbf{f}$  with respect to its  $p + 1$  parameters can be computed independently from each other. For example, this applies when  $\mu$  and  $\theta$  are strength modifiers relative to templates obtained from independently simulated processes, which is, for example, the case considered by Barlow and Beeston in their original work [2]. Conversely, there is at least another relevant

<sup>7</sup> It can be proved that  $\rho_\mu^2$  is equal to the so-called *global correlation coefficient*, which is used, for example, in Minuit [15] for fit diagnostics. See Appendix D.1 for further details.

case in which the assumption of uncorrelated templates might seem inappropriate, that is when nuisance parameters are associated with systematic uncertainties estimated from event-by-event re-weightings of the “nominal” MC sample. In this case, correlations between the columns of  $\mathbf{A}$  cannot be ignored *a priori* because of correlated Poisson fluctuations of the MC sample. However, if the dominant source of statistical uncertainty on the MC prediction originates from the variance of the event weights rather than from the purely Poisson uncertainty, different re-weightings of the same MC sample could be still treated as approximately independent. To proceed further, we make the assumption that the covariance matrix  $\langle \mathbf{v}_j \mathbf{v}_j^T \rangle$  is zero and consider the remaining terms in Eq. (39).

### 3.3. Contribution from the $\mathbf{A}$ matrix

Let us start by analyzing the bias terms in Eq. (39) related to fluctuations of the  $\mathbf{A}$  matrix. The expectation values  $\langle \mathbf{v}_j \rangle$  are generally different from zero. Indeed, from the normalization condition  $\|\hat{\mathbf{u}}_j\| = 1$  it follows that

$$2\mathbf{u}_j^T \langle \mathbf{v}_j \rangle = -\langle \mathbf{v}_j^2 \rangle \neq 0. \quad (40)$$

We notice that the effect of statistical fluctuations on the elements of  $\hat{\mathbf{A}}$  is to stretch and rotate its column vectors. However,  $S$  cannot depend on their normalization, since a global re-scaling of the columns can always be absorbed by a redefinition of their associated nuisance parameter.<sup>8</sup> Without loss of generality, we can then assume  $\|\mathbf{a}_j\| = 1$  and fix their orientation by requiring e.g.  $\mathbf{a}_j^T \mathbf{b} \geq 0$ . Random fluctuations on the normalized vectors  $\mathbf{a}_j$  generate perturbations  $\alpha_j$  corresponding to small rotations in a  $n$ -dimensional space, which must however average to zero when projected onto the sub-space orthogonal to  $\mathbf{a}_j$ , so that:

$$\langle \alpha_j \rangle = -\frac{\langle \alpha_j^2 \rangle}{2} \mathbf{a}_j, \quad (41)$$

where  $\langle \alpha_j^2 \rangle$  is the variance of the angle by which the vector  $\mathbf{a}_j$  is rotated. It is easy to see that  $\langle \alpha_j^2 \rangle$  must be of order  $n \cdot \Delta_j^2$ , with  $\Delta_j$  being the average relative MC uncertainty on the elements of  $\mathbf{a}_j$ . Under the same assumptions, one can easily prove that  $\langle \alpha_j \alpha_j^T \rangle$  must be a matrix that projects onto the subspace orthogonal to  $\mathbf{a}_j$  with equal probability along each of its  $n - 1$  directions, that is

$$\langle \alpha_j \alpha_j^T \rangle = \frac{\langle \alpha_j^2 \rangle}{n-1} (\mathbf{1} - \mathbf{a}_j \mathbf{a}_j^T). \quad (42)$$

By means of Eqs. (41) and (42), the right-hand side of Eq. (39) can be expanded further.

In the following, we will restrict to the case  $\mathbf{a}_j^T \mathbf{a}_k = \delta_{jk}$ , that is the column vectors of the matrix  $\mathbf{A}$  are orthogonal to each other. This hypothesis makes the analytic treatment simpler and its interpretation more intuitive, without precluding the general validity, as discussed in Appendix D.3. We also notice that this special case has some physical interest: for example, it is appropriate when nuisance parameters parametrize systematic effects that behave in an almost “orthogonal” way in the space of observables  $\mathbf{y}$ , for example when they are determined from a PCA analysis [27,28].

Under the assumption of orthonormal column vectors, we can exploit the arbitrariness in the definition of the  $\mathbf{u}_j$  vectors by choosing  $\mathbf{u}_j = \mathbf{a}_j$ . The second line on Eqs. (39) can be expanded by means of Eq. (41) and (42), obtaining after some straightforward calculation:

$$\frac{n}{n-1} \sum_{j=1}^p (\mathbf{b}^T \mathbf{a}_j)^2 \langle \alpha_j^2 \rangle - \frac{\mathbf{b}^T \mathbf{b}}{n-1} \left( \sum_{j=1}^p \langle \alpha_j^2 \rangle \right). \quad (43)$$

<sup>8</sup> This can be also proved by noticing that the matrix  $\mathbf{U}$  is invariant under a right multiplication of  $\mathbf{A}$  by an invertible matrix  $\mathbf{A}$ , that is  $\mathbf{A} \rightarrow \mathbf{A}\mathbf{A}$ , of which a column-wise re-scaling is a special case.

Let us first consider the case of a large value of  $\rho_\mu$ . Aside from numerical factors, the two pieces appearing in Eq. (43) are of the same order of magnitude, but of opposite sign. However, the former is scaled by a factor of  $n$ , thus giving a dominant *positive* contribution when the number of bins is large, which is our case of interest. This can be intuitively understood by noticing that when  $n \gg p$ , statistical fluctuations on  $\mathbf{A}$  have a larger probability of shuffling around the basis vectors by introducing new directions along which  $\mathbf{b}$  can have a finite projection, rather than rotate the vectors inside the same space, which would leave  $S$  unchanged. These extra contributions add quadratically to  $\|\mathbf{b}_\perp\|^2$  and can be numerically relevant if the latter is small. As discussed in Appendix D.4, the same conclusion can be proved for generic  $n$  under some more restrictive assumptions.

We now consider the opposite limiting case, that is  $\rho_\mu \rightarrow 0$ . The first term on the second line of Eq. (43) can be now neglected compared to the second term, which is instead always negative. Thus, we see that in this case the effect of random fluctuations of the MC templates is to reduce the value of  $\langle \hat{S} \rangle$  relative to  $S$  by a factor of

$$1 - \frac{1}{n-1} \left( \sum_{j=1}^p \langle \alpha_j^2 \rangle \right). \quad (44)$$

This result is also in line with our intuition: if  $\mathbf{b}$  is already orthogonal to the linear space spanned by the columns of  $\mathbf{A}$ , fluctuations on the latter can only generate spurious components of  $\mathbf{b}$  inside the linear space  $\mathcal{V}_\mathbf{A}$ , thus reducing the norm of  $\mathbf{b}_\perp$ . However, we notice that the relative decrease implied by Eq. (44) is only of order of the squared relative uncertainty on the elements of  $\hat{\mathbf{A}}$ , and is not scaled by  $n$ , so its overall impact might well be numerically less relevant than the effect discussed in the previous case.

For intermediate values of  $\rho_\mu$ , competition between positive and negative bias terms to  $\langle \hat{S} \rangle$  does not allow for a conclusive statement on the overall sign, and the bias term should be determined case by case. This can be well illustrated, for example, in a two-dimensional setting, as discussed in detail in Appendix D.5. In this simplified setup, the ratio  $\langle \hat{S} \rangle / S$  is found to depend on  $\rho_\mu$  as expected from Eqs. (43) and (44).

### 3.4. Contribution from the $\mathbf{b}$ vector

Finally, we consider the role of statistical fluctuations on the  $\mathbf{b}$  vector. Since the MC prediction for  $\hat{\mathbf{b}}$  is randomly distributed around the true value  $\mathbf{b}$ , the expectation value of  $\beta$  is zero, so that terms linear in  $\langle \beta \rangle$  do not contribute to the bias term in Eq. (39). The quadratic term on the first line of Eq. (43) is positive-definite.<sup>9</sup> If we make the reasonable assumption  $\langle \beta_i \beta_j \rangle = \langle \beta_i^2 \rangle \delta_{ij}$ , that is the MC predictions for different bins are statistically uncorrelated, the quadratic term can be further expanded as

$$\text{Tr} [\mathbf{U} \langle \beta \beta^T \rangle] = n \left( \sum_{i=1}^n \frac{1}{n} \langle \beta_i^2 \rangle \right) - \sum_{j=1}^p \left( \sum_{i=1}^n u_{j,i}^2 \langle \beta_i^2 \rangle \right), \quad (45)$$

where the two terms in parentheses correspond to different ways of averaging the bin-by-bin variances  $\langle \beta_i^2 \rangle$ . If the distribution of the latter is narrowly concentrated around an average value  $\overline{\langle \beta^2 \rangle}$ , the right-hand side of Eq. (45) is of order  $\sim (n-p) \overline{\langle \beta^2 \rangle}$ , so it grows linearly with the number of bins. Thus, we see that statistical fluctuations on  $\mathbf{b}$  can induce sizable positive biases to  $\langle \hat{S} \rangle$  for any value of  $\rho_\mu$ , provided that  $(n-p) \overline{\langle \beta^2 \rangle}$  is comparable to  $S$ .

<sup>9</sup> This can be also seen as consequence of Jensen’s inequality applied to  $\hat{S}$ , which is a convex function of  $\mathbf{b}$ .

### 3.5. Discussion and connection with the toy model

We now summarize the assumptions and the main implications of Eq. (39). We have assumed that the MC estimators for  $\mathbf{b}$  and  $\mathbf{A}$  are independent from each other and that statistical fluctuations of their elements can be treated as Gaussian-distributed noise around their true values. We have considered separately the impact of MC fluctuations on the blocks  $\mathbf{A}$  and  $\mathbf{b}$  of the Jacobian matrix in the form of a bias term to the quadratic form  $S$ , defined by Eq. (35). If  $\mathbf{b}$  has a large overlap with the linear space spanned by the columns of  $\mathbf{A}$ , or, equivalently, when  $\rho_\mu$  is close to one, then the bias terms associated with fluctuations of the  $\mathbf{A}$  matrix are positive-definite and increase with the number of bins  $n$ . Fluctuations of the  $\mathbf{b}$  vector also grow with  $n$  and are always positive-definite. In both cases,  $\hat{S}$  would differ from the true value  $S$  by a positive offset, implying that the Hessian uncertainty, estimated as  $\hat{\sigma}_H = \hat{S}^{-1/2}$ , would be underestimated.<sup>10</sup>

These findings can finally explain the observations made in the context of the toy model of Section 2: when the original parametrization is considered (i.e.  $\mu = \mu_1$ ), the linear correlation coefficient  $\rho = \rho_\mu$  is indeed close to unity, implying a large bias term from fluctuations of the matrix  $\mathbf{A} = (T_{21}/\sqrt{y_1}, \dots, T_{2n}/\sqrt{y_n})^T$ ; for  $\mu = \mu'_i$ , instead, the smaller correlation reduces the impact of statistical fluctuations on  $\mathbf{A}$ , and the bulk of finite-size MC effects are now propagated to  $\hat{\sigma}_H$  through a noisy vector  $\mathbf{b} = (T'_{11}/\sqrt{y_1}, \dots, T'_{1n}/\sqrt{y_n})^T$ .

Remarkably, none of the bias terms appearing in Eq. (43) is directly proportional to the true value  $S$  by a naive factor of  $\mathcal{O}(1) \cdot k^{-1}$ , with  $k$  related to the effective statistical power of the simulation relative to the data. The lack of proportionality also means that the bias and true value can stand in a ratio which is *a priori* arbitrary. This is one of the main observations of this work. Incidentally, this also explains how the Barlow–Beeston lite approach [12] can be insufficient to account for MC uncertainties. As discussed in Section 2.3, the lite version of the Barlow–Beeston likelihood is practically equivalent to a rescaling of the Poisson variance of the data by a factor of  $(1 + k^{-1})$ , implying an increase of the Hessian uncertainty  $\hat{\sigma}_H$  by the square root of the same factor. However, a bias in  $\hat{S}$ , which is both positive and comparable to  $S$ , might reduce  $\hat{\sigma}_H$  by a potentially larger factor, ultimately leading to the non-physical situation of confidence intervals that are smaller than they would be in the limit of infinite-size MC samples.

As a last step in our discussion, we address the question of providing quantitative estimates for the bias terms appearing in Eq. (39).

### 3.6. Asymptotic uncertainties from finite-size samples

The bias term generated by statistical fluctuations of the Jacobian matrix could be estimated in principle from Eq. (39) by computing the covariance matrices of  $\mathbf{b}$  and of the column vectors of  $\mathbf{A}$ , for example by bootstrap resampling, and using their nominal MC predictions  $\hat{\mathbf{b}}$  and  $\hat{\mathbf{A}}$  in place of their true, unknown values. The asymptotic Hessian uncertainty  $\hat{\sigma}_H = S^{-1/2}$ , i.e. the Hessian uncertainty in the limit of infinite MC statistics, could be determined from Eq. (39) after subtracting the bias contribution.

A possibly simpler way to estimate  $\hat{\sigma}_H$  consists in analyzing the scaling of the Hessian uncertainty  $\hat{\sigma}_H$  as a function of the MC-to-data ratio  $k$  at a fixed size  $N$  of the data. Indeed, the following scaling laws can be easily derived:

$$\mathbf{b} \sim \sqrt{N}, \quad \mathbf{U} \sim 1, \quad \langle \alpha_j^2 \rangle \sim \frac{1}{kN}, \quad \langle \beta\beta^T \rangle \sim \frac{1}{k}, \quad (46)$$

<sup>10</sup> More correctly, a positive bias on  $\hat{S}$  would also imply a negative bias on  $\hat{\sigma}_H$  if the variance of  $\hat{S}$  is small, so that  $\langle \hat{S}^{-1/2} \rangle \approx \langle \hat{S} \rangle^{-1/2}$ . In particular, by Taylor expansion, we find that  $\frac{3}{8} \frac{\langle \hat{S}^2 \rangle - \langle \hat{S} \rangle^2}{\langle \hat{S} \rangle^2} \ll 1$  should hold, a condition that is more likely to be satisfied when the number of bins is large, as for the law of large numbers.

from which one gets:

$$\bar{\sigma}_H = \hat{\sigma}_H \cdot \sqrt{1 + \frac{\Delta_N}{k}}, \quad (47)$$

where  $\Delta_N$  is a constant independent from  $k$ , and  $\hat{\sigma}_H$  is the Hessian uncertainty obtained from a MC sample with  $k$ -times the statistical power of the data. The right-hand side of Eq. (47) contains two unknowns:  $\bar{\sigma}_H$  and  $\Delta_N$ . If  $N$  is kept fixed, but a fraction  $1/\xi$ , with  $\xi > 1$  of the MC events is randomly selected and used to construct a new model prediction  $\hat{\mathbf{f}}$ , then the new Hessian uncertainty will be a fraction  $r_\xi < 1$  of the uncertainty from the full sample. Then, one can readily solve for  $\bar{\sigma}_H$ , obtaining:

$$\bar{\sigma}_H = \left( \frac{r_\xi \sqrt{\xi - 1}}{\sqrt{\xi r_\xi^2 - 1}} \right) \hat{\sigma}_H. \quad (48)$$

This result could be improved by taking the expectation values of  $\hat{\sigma}_H$  over several bootstrap re-samplings of the MC sample.

We can test the validity of Eq. (48) for the toy model of Section 2. Here, the mean Hessian uncertainty for  $N = 2 \times 10^6$ ,  $n = 200$ , and  $\alpha = 0.03$  is found to be  $\hat{\sigma}_H = 0.0572$  (0.0447) when  $k = 1$  ( $k = 0.5$ ), from which the asymptotic uncertainty can be predicted to be 0.0956 from Eq. (48), in good agreement with the true value of 0.0950. Extensions of Eq. (48) to the Barlow–Beeston lite likelihood are discussed in Appendix D.6.

Although  $\bar{\sigma}_H$  does not provide a reliable estimate of the actual statistical spread of  $\hat{\mu}$  induced by fluctuations of both the data and the MC samples, it can provide useful information nonetheless. For example, values of  $\bar{\sigma}_H$  significantly larger than  $\hat{\sigma}_H$  would signal the presence of spurious constraints on the POI which are likely to invalidate the asymptotic properties of maximum-likelihood estimators even in the presence of a proper account of MC statistical uncertainties, as done for example in the full Barlow–Beeston method. Moreover,  $\bar{\sigma}_H$  is one of the inputs needed to construct the heuristic confidence interval of Eq. (26).

### 3.7. Testing the Barlow–Beeston lite approximation

Suppose that  $\bar{\sigma}_H$  calculated from Eq. (47) is in good agreement with  $\hat{\sigma}_H$ , or, equivalently, that the Hessian uncertainty is stable when the size of the MC sample is reduced. One might still want to ensure that the lite version of the Barlow–Beeston method represents an adequate approximation for a given problem. For this purpose, we can use the formalism developed in the previous section. We start by noticing that the MLE estimators  $\hat{\theta}$  and  $\hat{\mu}$ , besides being functions of  $\mathbf{y}$ , also depend on  $\mathbf{j}$ ,  $\mathbf{J}$  and  $\mathbf{f}_0$  through the matrix  $\mathbf{B}$ , defined in Appendix D.1, and the vector  $\mathbf{d}$  defined by Eq. (30), which are all random variables of known variance. If the latter can be treated as independent and Gaussian-distributed variables with standard deviations  $\Delta B_{ik}$  and  $\Delta d_i$ , the contribution to the variance of the MLE estimator  $\hat{\mu}$  induced by fluctuations of the MC templates can be determined from standard error propagation as

$$\hat{\sigma}_{MC}^2 = \sum_{i=1}^n \left[ \sum_{k=1}^p \left( \frac{\partial \hat{\mu}}{\partial B_{ik}} \right)^2 (\Delta B_{ik})^2 + \left( \frac{\partial \hat{\mu}}{\partial d_i} \right)^2 (\Delta d_i)^2 \right]. \quad (49)$$

Analytical expressions for the two partial derivatives appearing in Eq. (49) have been derived in Appendix D.7. Eq. (49) can be easily extended to generic covariances for  $\hat{\mathbf{B}}$  and  $\hat{\mathbf{d}}$ . However, we remark that the right-hand side of Eq. (49) can only be evaluated using the MC prediction  $\hat{\mathbf{B}}$  and  $\hat{\mathbf{d}}$  derived from finite-size MC samples, so Eq. (49) might be a biased estimator of the real variance of  $\hat{\mu}$  under fluctuations of the MC templates. With this caveat in mind, we could still estimate the total variance of the maximum-likelihood estimator  $\hat{\mu}$  under fluctuations of both the data and the MC samples as

$$\hat{\sigma}_H^{\text{BBfull}} \approx \sqrt{\hat{\sigma}_H^2 + \hat{\sigma}_{MC}^2}. \quad (50)$$

**Table 6**

Ratio between  $\hat{\sigma}_{\text{MC}}$  and  $\hat{\sigma}_{\text{H}}/\sqrt{k}$  for the toy model of Section 2 by varying one configuration parameter at the time. The value for the nominal model is given by the number in bold.

| $k \times (1, 10, 40)$   | $n \times (\frac{1}{2}, \frac{1}{10}, \frac{1}{20})$ | $\epsilon \times (2, 4, 8)$ | $N \times (10, 10^2, 10^3)$ |
|--------------------------|--|-----------------------------|-----------------------------|
| <b>1.51</b> , 1.63, 1.65 | 1.40, 1.14, 1.07                                     | 1.27, 1.10, 1.03            | 1.14, 1.02, 1.00            |

Values of  $\hat{\sigma}_{\text{MC}}$  larger than the Barlow–Beeston *lite* expectation of  $\hat{\sigma}_{\text{H}}/\sqrt{k}$ , or, equivalently values of  $\hat{\sigma}_{\text{H}}^{\text{BBfull}}$  in excess of the naive expectation  $\hat{\sigma}_{\text{H}}\sqrt{1+k^{-1}}$ , should be taken as an indication that the *lite* approach is insufficient to provide a proper account of MC statistical uncertainties, suggesting that the full approach should be adopted instead. For example, for the toy study presented in Section 2, we find that  $\hat{\sigma}_{\text{MC}}$  estimated from Eq. (49) is a factor of  $\sim 1.5$  larger than  $\hat{\sigma}_{\text{H}}/\sqrt{k}$  for the nominal model configuration. Interestingly, this ratio decreases monotonically towards  $\sim 1$  when  $N$  or  $\epsilon$  are increased, or when  $n$  is reduced, corroborating the observation that the Barlow–Beeston *lite* approach becomes more efficient in these limits. Conversely, we find that the ratio approaches a plateau of  $\sim 1.6$  for  $k \gg 1$ , when  $n$ ,  $N$  and  $\epsilon$  are fixed to their default values, implying that the Barlow–Beeston *lite* version will never be equivalent to its full implementation for at least the nominal model configuration.<sup>11</sup> A summary of numerical results is reported in Table 6.

Finally, we remark that  $\hat{\sigma}_{\text{MC}}$  could also be estimated by considering the variance of  $\hat{\mu}$  over multiple bootstrap re-samplings of the MC samples. However, Eq. (49) offers the advantage that it is fully analytical and that, differently from resampling techniques, the estimators for  $\mathbf{B}$  and  $\mathbf{d}$  can be computed using the *full* statistical power of the simulated samples.

#### 4. Conclusions

In this work, we have studied the problem of setting confidence intervals on a parameter of interest measured from the binned analysis of high-statistics and high-granularity data sets, focusing on the case where the likelihood function depends on nuisance parameters and is constructed from finite-size MC samples. We have set ourselves in the limit of large event counts in all bins for both the data and the MC-simulated samples, such that asymptotic formulas from the theory of maximum-likelihood estimators would be naively expected to hold. In particular, by assuming Wilks’ theorem, central confidence intervals can be defined from either the profile-likelihood ratio or the inverse Hessian matrix of the likelihood function in the Gaussian approximation.

We have designed a simple toy model consisting of a signal and background process where the unknown signal (background) normalization is treated as the parameter of interest (nuisance parameter). We have further assumed that the likelihood function is constructed using Monte Carlo simulations of finite size. This toy model depends on configuration parameters that we are free to choose: the total number of expected events, the number of bins, a model parameter that increases the signal-to-background separation, and the statistical power of the MC samples relative to the data. By inspecting a large number of pseudo-experiments, we could establish numerically that asymptotic formulas are broken even when the statistical power of the simulated samples is a few times larger than that of the data. In particular, confidence intervals based on Wilks’ theorem suffer from systematic under-coverage. Consequently, we have tried alternative methods to set confidence intervals based on the post-fit distribution of the profile-likelihood ratio, such as the profiled Feldman–Cousins approach or

variants thereof [17,18,20], but none of them succeeded in recovering the correct coverage.

Methods based on the asymptotic properties of the maximum-likelihood estimator ultimately yield the right coverage, but only in the limit of a very large scaling of the data and/or simulated sample size, though we could not find a first-principle indication of how large such a scaling should be. Finally, by combining the Hessian uncertainties obtained from different likelihood functions, we have worked out a heuristic prescription for a confidence interval, defined in Eq. (25), which gets closer to the correct Feldman–Cousins interval for this problem, thus providing a practical test to quantify the level of under-coverage when other standard confidence intervals are used. This prescription is model-independent and could in principle be extended to a generic analysis, provided that the point-estimator and Hessian uncertainty from the full Barlow–Beeston method can be both computed.

Inspired by the toy model, we have studied the problem from a more general perspective, with the goal of understanding in which circumstances statistical fluctuations in the MC prediction can reduce the uncertainty predicted for the parameter of interest, ultimately leading to under-coverage of confidence intervals constructed in the asymptotic approximation. We found that this artificially augmented sensitivity can be expressed in terms of a bias term to the quadratic form  $S$  defined in Eq. (35) which, in the limit of infinite size MC samples, is equal to the inverse variance of the maximum-likelihood estimator of the POI. We have identified and characterized two main sources of bias which then leads to under-coverage: fluctuations on the Jacobian matrix of the model function with respect to the POI, and fluctuations on the Jacobian matrix with respect to the nuisance parameters. In the latter case, the effect can be greatly enhanced in the presence of a large global correlation of the parameter of interest, as defined by Eq. (38). Both effects can explain the results of the toy study. Interestingly, the bias term is not proportional to  $S$  by a naive factor of  $k^{-1}$ , suggesting that extensions of the likelihood functions that include the effect of finite MC samples in an effective way, such as the popular Barlow–Beeston *lite* method [12], can be insufficient to provide a full account of these effects.

Based on general scaling laws, we propose a practical test, described in Section 3.6, to estimate the standard deviation that the MLE for the parameter of interest would have in the limit of infinite-size MC samples. As the latter should always be smaller than the standard deviation for finite-size MC samples, it could be used as a reference to gauge how far an analysis is from the asymptotic regime. This test only requires the possibility of repeating the maximum-likelihood fit using a subset of the MC simulated samples. We recommend it to be performed in data analyses to ensure the validity of asymptotic formulas, especially in case the Barlow–Beeston *lite* approximation is used. Although the full Barlow–Beeston method should always be preferred, as it represents the correct likelihood function for finite-size MC samples, its implementation is often challenging in practice. In Section 3.7 we suggest a test, based on analytical formulas, to verify that the Barlow–Beeston *lite* approach provides an adequate approximation of the exact likelihood function.

While Wilks’ theorem for the profile-likelihood ratio guarantees that the desired asymptotic behavior will eventually be attained, the assumption of being in the asymptotic regime could be wrong even when the event counts are large in absolute terms for both data and simulation. This suggests that the MC statistical uncertainty that needs to be “small” to be safely in the asymptotic regime is not necessarily provided by total number of observed/simulated events bin, nor by the total number of bins. Care should be therefore taken when reporting the uncertainty on parameters of interest based on asymptotic formulas when the data model is derived from finite-size MC samples and additional nuisance parameters are profiled. This seems to be the case also for high-statistics experiments, and even when the statistical power of the simulated samples is comparable to that of the data.

<sup>11</sup> In practice this difference is irrelevant when  $k$  is large because the contribution from MC statistical uncertainties to the total variance will also be negligible in this limit.

## CRediT authorship contribution statement

**C.A. Alexe:** Writing – review & editing, Software, Methodology, Formal analysis, Conceptualization. **J. Bendavid:** Writing – review & editing, Software, Methodology, Investigation, Formal analysis, Conceptualization. **L. Bianchini:** Writing – original draft, Software, Methodology, Funding acquisition, Formal analysis. **D. Bruschini:** Writing – review & editing, Software, Methodology, Formal analysis, Conceptualization.

## Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Lorenzo Bianchini, Davide Bruschini, Cristina Alexe reports financial support was provided by European Research Council. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

C.A., L.B., and D.B. acknowledge financial support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant agreement N. 101001205).

## Appendix A. Approximate likelihood functions for the toy model

The likelihood functions described below encode different levels of approximating the exact log-likelihood function Eq. (8). For each definition, it should be understood that  $\mu_1$  and  $\mu_2$  are the appropriate functions of  $\mu$  and  $\theta$ , depending on whether the original or transformed parameters are being considered.

### • Poisson without MC stat.

$$-\ln L_P(\mu, \theta) = \sum_{i=1}^n \left[ \left( \mu_1 \frac{t_{1i}}{k_1} + \mu_2 \frac{t_{2i}}{k_2} \right) - y_i \ln \left( \mu_1 \frac{t_{1i}}{k_1} + \mu_2 \frac{t_{2i}}{k_2} \right) \right] \quad (\text{A.1})$$

This would be the correct likelihood function if the MC predictions  $t_{ji}$  were exact. For finite MC samples, it must be considered as an approximation. In the context of the toy study, it is mostly considered as a reference. It is expected to provide unbiased estimators for  $\mu$  and confidence intervals with the right coverage only in the limit  $k_j \rightarrow \infty$ .

### • Gauss without MC stat.

$$-\ln L_G(\mu, \theta) = \frac{1}{2} \sum_{i=1}^n \frac{\left( y_i - \left( \mu_1 \frac{t_{1i}}{k_1} + \mu_2 \frac{t_{2i}}{k_2} \right) \right)^2}{y_i}. \quad (\text{A.2})$$

This function provides an excellent approximation of the joint Poisson likelihood of Eq. (A.1) in the limit  $y_i \gg 1$ . An interesting feature of Eq. (A.2) is that its minimum can be found analytically by solving a linear system of equations [5]. Again, for finite MC samples, it represents an approximation of the correct likelihood. It is considered here mostly to justify the Gaussian approximation of the likelihood function.

### • Gauss + Barlow–Beeston

$$-\ln L_{G+BB}(\mu, \theta, \theta_1, \theta_2) = \frac{1}{2} \sum_{i=1}^n \left[ \frac{\left( y_i - \left( \mu_1 \frac{\theta_{1i}}{k_1} + \mu_2 \frac{\theta_{2i}}{k_2} \right) \right)^2}{y_i} + \sum_{j=1}^2 \frac{(\theta_{ji} - t_{ji})^2}{t_{ji}} \right]. \quad (\text{A.3})$$

This is the Barlow–Beeston likelihood of Eq. (8) in the Gaussian approximation. It is considered here to justify the Gaussian approximation of the likelihood function even in the presence of nuisance parameters for the uncertain MC predictions. Furthermore, it offers the advantage that the latter can be profiled analytically. Indeed, it is easy to prove that

$$(\hat{\mu}, \hat{\theta}) = \underset{\mu, \theta}{\operatorname{argmin}} \left[ \sum_{i=1}^n \frac{\left( y_i - \left( \mu_1 \frac{t_{1i}}{k_1} + \mu_2 \frac{t_{2i}}{k_2} \right) \right)^2}{y_i + \mu_1^2 \frac{t_{1i}}{k_1^2} + \mu_2^2 \frac{t_{2i}}{k_2^2}} \right] \quad (\text{A.4})$$

For the toy study, it has been verified that minimizing Eq. (A.3) over the full set of parameters or solving Eq. (A.4) leads to identical results within numerical accuracy.

### • Gauss + Barlow–Beeston lite

$$-\ln L_{G+BBlite}(\mu, \theta, \beta) = \frac{1}{2} \sum_{i=1}^n \left[ \frac{\left( y_i - \beta_i \cdot \left( \mu_1 \frac{t_{1i}}{k_1} + \mu_2 \frac{t_{2i}}{k_2} \right) \right)^2}{y_i} + (\beta_i - 1)^2 m_i \right]. \quad (\text{A.5})$$

This is a well-known and widely used approximation of the Barlow–Beeston likelihood [12]. It consists in assigning a unique parameter  $\beta_i$  for all MC processes in each bin constrained by the total effective relative uncertainty  $1/\sqrt{m_i}$ , as defined in Eq. (11). It is studied along with Eq. (A.3) as it represents a widely adopted approximation to the full problem. Again, the  $\beta$  parameters can be profiled analytically, leading to:

$$(\hat{\mu}, \hat{\theta}) = \underset{\mu, \theta}{\operatorname{argmin}} \left[ \sum_{i=1}^n \frac{\left( y_i - \left( \mu_1 \frac{t_{1i}}{k_1} + \mu_2 \frac{t_{2i}}{k_2} \right) \right)^2}{y_i + \frac{\left( \mu_1 \frac{t_{1i}}{k_1} + \mu_2 \frac{t_{2i}}{k_2} \right)^2}{m_i}} \right] \quad (\text{A.6})$$

### • Gauss + MC stat.

$$-\ln L_{G+MC}(\mu, \theta) = \frac{1}{2} \sum_{i=1}^n \frac{\left( y_i - \left( \mu_1 \frac{t_{1i}}{k_1} + \mu_2 \frac{t_{2i}}{k_2} \right) \right)^2}{y_i + \frac{t_{1i}}{k_1^2} + \frac{t_{2i}}{k_2^2}}. \quad (\text{A.7})$$

This is the simplest extension of Eq. (A.2) that accounts for statistical uncertainties on the MC templates: the total MC variance per bin is added to the Poisson variance of the data. The global minimum can still be found analytically. In the context of the toy study, it is considered to prove that the ‘‘Gauss + Barlow–Beeston lite’’ likelihood can be well approximated by a Neyman’s  $\chi^2$  test-statistic with augmented statistical uncertainties.

## Appendix B. Alternative model configurations

Numerical results for coverage, mean, and median of the  $1\sigma$  confidence interval for  $\mu = \mu'_1$ , obtained for alternative configurations of the toy model of Section 2, are reported in this Appendix.

The results reported in Table B.7 (B.8) have been obtained by increasing the data-to-simulation ratio  $k$  by a factor of 10 (40) compared to the nominal, while Table B.9 (B.10) corresponds to increasing the asymmetry factor  $\epsilon$  by a factor of 2 (4). The results reported in Table B.11 (B.12) has been obtained by reducing the number of bins  $n$  by a factor of 2 (10). Finally, the results reported in Table B.13 (B.14) have been obtained by increasing the total number of events  $N$  by a factor of 10 (100) compared to the nominal.

## Appendix C. Adding external constraints to the likelihood

In Section 3.1, all parameters of the likelihood were assumed to be unconstrained, though this is certainly not the most general case of interest. Indeed, it often happens that at least some of the nuisance parameters can be constrained by independent auxiliary measurements.

**Table B.7**

Coverage, mean, and median of the  $1\sigma$  confidence interval for  $\mu = \mu'$  for the toy model with  $N = 2 \times 10^6$ ,  $n = 200$ ,  $k = 10$ , and  $\epsilon = 0.03$ , obtained from an ensemble of  $10^4$  identically repeated pseudo-experiments. Cases where the observed coverage agrees with the expectation of 68.3% within a relative  $\pm 5\%$  tolerance are highlighted with a bold font. When present, the number within parenthesis refers to the statistical error on the last digit.

| Likelihood       | Minimim. | CI method | $\mu_t$                     | Coverage        | $\hat{\sigma}$ (mean, median) |
|------------------|----------|-----------|-----------------------------|-----------------|-------------------------------|
| Gauss (asympt.)  | Analytic | Hessian   | 0                           | 0.683           | 0.094, 0.094                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.683           | 0.094, 0.094                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.683           | 0.094, 0.094                  |
| Poisson          | Numeric  | Hessian   | 0                           | <b>0.661(5)</b> | 0.087, 0.087                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.529(5)        | 0.087, 0.087                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.532(5)        | 0.087, 0.087                  |
| Gauss            | Numeric  | Hessian   | 0                           | <b>0.663(5)</b> | 0.087, 0.087                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.528(5)        | 0.087, 0.087                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.532(5)        | 0.087, 0.087                  |
| Gauss            | Analytic | Hessian   | 0                           | <b>0.663(5)</b> | 0.087, 0.087                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.528(5)        | 0.087, 0.087                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.532(5)        | 0.087, 0.087                  |
| Gauss + MC stat. | Analytic | Hessian   | 0                           | <b>0.683(5)</b> | 0.091, 0.091                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.552(5)        | 0.091, 0.091                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.554(5)        | 0.091, 0.091                  |
| Gauss + BB-lite  | Numeric  | Hessian   | 0                           | <b>0.683(5)</b> | 0.091, 0.091                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.552(5)        | 0.091, 0.091                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.554(5)        | 0.091, 0.091                  |
| Gauss + BB       | Numeric  | Hessian   | 0                           | 0.647(5)        | 0.099, 0.099                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | <b>0.649(5)</b> | 0.100, 0.100                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.638(5)        | 0.100, 0.100                  |
| Gauss + BB       | Numeric  | PLR       | 0                           | 0.647(5)        | 0.099, 0.099                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | <b>0.650(5)</b> | 0.100, 0.100                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.639(5)        | 0.100, 0.100                  |
| Gauss + BB       | Numeric  | sPFC      | 0                           | <b>0.662(5)</b> | 0.103, 0.103                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.745(4)        | 0.124, 0.123                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.746(4)        | 0.123, 0.122                  |
| Gauss + BB       | Numeric  | PFC       | 0                           | <b>0.654(5)</b> | 0.102, 0.102                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.753(4)        | 0.124, 0.123                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.747(4)        | 0.123, 0.122                  |
| Gauss + BB       | Numeric  | Barlett   | 0                           | <b>0.653(5)</b> | 0.102, 0.102                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.740(4)        | 0.121, 0.121                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.738(4)        | 0.121, 0.120                  |
| Gauss + BB       | Numeric  | CH        | 0                           | <b>0.679(5)</b> | 0.107, 0.107                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.788(4)        | 0.135, 0.134                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.785(4)        | 0.134, 0.133                  |
| Gauss + BB       | Numeric  | FC Cheat  | 0                           | <b>0.680(5)</b> | 0.108, 0.106                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | <b>0.684(5)</b> | 0.109, 0.107                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | <b>0.690(5)</b> | 0.108, 0.106                  |
| Gauss + BB       | Numeric  | Heuristic | 0                           | <b>0.683(5)</b> | 0.107, 0.107                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | <b>0.686(5)</b> | 0.108, 0.108                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | <b>0.677(5)</b> | 0.109, 0.109                  |

Let us then assume that the likelihood function includes both constrained and unconstrained nuisance parameters. If the set of  $q$  constrained nuisance parameters is denoted by  $\tau$ , the right-hand side of Eq. (28) gets modified to

$$\mathbf{f}(\mu, \theta, \tau) = \mathbf{f}_0 + \mathbf{j}(\mu - \mu_0) + \mathbf{J}(\theta - \theta_0) + \mathbf{K}(\tau - \tau_0) + \dots, \quad (\text{C.1})$$

where  $\mathbf{K}$  is the  $n \times q$  Jacobian matrix associated with the NP subject to a constraint. In some cases, the effect of a constrained nuisance parameter relative to the nominal prediction is known exactly: for example, a luminosity uncertainty on the total MC normalization can be modeled as a multiplicative correction, i.e.  $\mathbf{K} \propto \mathbf{f}_0$ . In other cases, however,  $\mathbf{K}$  needs to be estimated numerically, giving rise to an additional source of systematic uncertainty, if MC samples of finite size are used to determine its elements. The negative log-likelihood function in Eq. (27) should be then extended accordingly to

$$- \ln L(\mu, \theta, \tau) \approx \frac{1}{2} (\mathbf{y} - \mathbf{f}(\mu, \theta, \tau))^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{f}(\mu, \theta, \tau))$$

$$+ \frac{1}{2} (\boldsymbol{\tau} - \boldsymbol{\tau}_0)^T (\boldsymbol{\tau} - \boldsymbol{\tau}_0) + \text{const.} \quad (\text{C.2})$$

where the estimators from the auxiliary measurements have been assumed as normally-distributed and independent from each other [29]. The additional parameters  $\tau$  can be profiled at a fixed value of  $(\mu, \theta)$  yielding the well-known result [24,29]:

$$- 2 \ln \frac{L(\mu, \theta, \hat{\tau}_{(\mu, \theta)})}{L(\hat{\mu}, \hat{\theta}, \hat{\tau})} = (\mathbf{y} - \mathbf{f}(\mu, \theta, \boldsymbol{\tau}_0))^T (\mathbf{V} + \mathbf{K}\mathbf{K}^T)^{-1} (\mathbf{y} - \mathbf{f}(\mu, \theta, \boldsymbol{\tau}_0)), \quad (\text{C.3})$$

that is, the Poisson covariance matrix of the data  $\mathbf{V}$  is augmented by the non-diagonal matrix  $\mathbf{K}\mathbf{K}^T$ , with their sum still symmetric and positive-definite. By means of Woodbury's matrix identity, we can expand

$$\begin{aligned} (\mathbf{V} + \mathbf{K}\mathbf{K}^T)^{-1} &= \mathbf{V}^{-\frac{1}{2}} \left( \mathbf{1} - \mathbf{H} (\mathbf{1} + \mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \right) \mathbf{V}^{-\frac{1}{2}} \\ &= \mathbf{V}^{-\frac{1}{2}} \mathbf{Z}^{-1} \mathbf{V}^{-\frac{1}{2}}, \end{aligned} \quad (\text{C.4})$$

**Table B.8**

Coverage, mean, and median of the  $1\sigma$  confidence interval for  $\mu = \mu'_i$  for the toy model with  $N = 2 \times 10^6$ ,  $n = 200$ ,  $k = 40$ , and  $\epsilon = 0.03$ , obtained from an ensemble of  $10^4$  identically repeated pseudo-experiments. Cases where the observed coverage agrees with the expectation of 68.3% within a relative  $\pm 5\%$  tolerance are highlighted with a bold font. When present, the number within parenthesis refers to the statistical error on the last digit.

| Likelihood       | Minimim. | CI method | $\mu_i$            | Coverage        | $\hat{\sigma}$ (mean, median) |
|------------------|----------|-----------|--------------------|-----------------|-------------------------------|
| Gauss (asympt.)  | Analytic | Hessian   | 0                  | 0.683           | 0.094, 0.094                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.683           | 0.094, 0.094                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.683           | 0.094, 0.094                  |
| Poisson          | Numeric  | Hessian   | 0                  | <b>0.669(5)</b> | 0.092, 0.092                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.672(5)</b> | 0.092, 0.092                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.665(5)</b> | 0.092, 0.092                  |
| Gauss            | Numeric  | Hessian   | 0                  | <b>0.669(5)</b> | 0.092, 0.092                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.672(5)</b> | 0.092, 0.092                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.667(5)</b> | 0.092, 0.092                  |
| Gauss            | Analytic | Hessian   | 0                  | <b>0.669(5)</b> | 0.092, 0.092                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.672(5)</b> | 0.092, 0.092                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.667(5)</b> | 0.092, 0.092                  |
| Gauss + MC stat. | Analytic | Hessian   | 0                  | <b>0.675(5)</b> | 0.093, 0.093                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.679(5)</b> | 0.093, 0.093                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.672(5)</b> | 0.094, 0.094                  |
| Gauss + BB-lite  | Numeric  | Hessian   | 0                  | <b>0.675(5)</b> | 0.093, 0.093                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.678(5)</b> | 0.093, 0.093                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.672(5)</b> | 0.094, 0.094                  |
| Gauss + BB       | Numeric  | Hessian   | 0                  | <b>0.665(5)</b> | 0.095, 0.095                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.674(5)</b> | 0.096, 0.096                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.668(5)</b> | 0.096, 0.096                  |
| Gauss + BB       | Numeric  | PLR       | 0                  | <b>0.665(5)</b> | 0.095, 0.095                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.675(5)</b> | 0.096, 0.096                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.669(5)</b> | 0.096, 0.096                  |
| Gauss + BB       | Numeric  | sPFC      | 0                  | <b>0.679(5)</b> | 0.096, 0.096                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.689(5)</b> | 0.098, 0.098                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.679(5)</b> | 0.098, 0.098                  |
| Gauss + BB       | Numeric  | PFC       | 0                  | <b>0.675(5)</b> | 0.096, 0.096                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.682(5)</b> | 0.098, 0.098                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.690(5)</b> | 0.098, 0.098                  |
| Gauss + BB       | Numeric  | Barlett   | 0                  | <b>0.675(5)</b> | 0.096, 0.096                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.681(5)</b> | 0.098, 0.098                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.690(5)</b> | 0.098, 0.098                  |
| Gauss + BB       | Numeric  | CH        | 0                  | <b>0.681(5)</b> | 0.098, 0.098                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.688(5)</b> | 0.100, 0.100                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.698(5)</b> | 0.100, 0.100                  |
| Gauss + BB       | Numeric  | FC Cheat  | 0                  | <b>0.683(5)</b> | 0.098, 0.097                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.689(5)</b> | 0.098, 0.097                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.684(5)</b> | 0.098, 0.098                  |
| Gauss + BB       | Numeric  | Heuristic | 0                  | <b>0.675(5)</b> | 0.098, 0.098                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.683(5)</b> | 0.098, 0.098                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.678(5)</b> | 0.098, 0.098                  |

with  $\mathbf{H} = \mathbf{V}^{-\frac{1}{2}} \mathbf{K}$ . The newly introduced matrix  $\mathbf{Z}$  is a symmetric and positive-definite matrix, so it is invertible and has a unique square-root matrix  $\mathbf{Z}^{\frac{1}{2}}$ . By means of Eq. (C.4), Eq. (C.3) can be ultimately recast in the same form of Eq. (29), namely

$$\|\mathbf{r}'_{\mu} - \mathbf{A}'(\theta - \theta_0)\|^2, \quad (\text{C.5})$$

where  $\mathbf{r}'_{\mu}$  and  $\mathbf{A}'$  are defined as in Eq. (30) with the replacement  $\mathbf{V}^{-1/2} \rightarrow (\mathbf{VZ})^{-1/2}$ . In this framework, statistical fluctuations affecting the entries of the Jacobian matrix  $\mathbf{K}$  are propagated into the NLL function by the matrix  $\mathbf{Z}$  which mixes them with fluctuations of  $\mathbf{J}$  in a possibly intricate way. However, due to the formal identity between Eqs. (29) and (C.5), the same analysis pursued in Section 3.2 still holds. Notice, however, that in the case of Eq. (C.4) some of the scaling laws of Eq. (46) are broken due to a non-trivial combination of statistical and systematic uncertainties. Also, the inclusion of external constraints in the likelihood function is likely to reduce  $\rho_{\mu}$  and hence improve the validity of the asymptotic approximation.

Constrained nuisance parameters naturally arise in problems where the nominal model prediction is affected by sources of systematic uncertainty for which some *a priori* knowledge is available from independent measurements or theoretical prejudice. In this respect,  $K_{ij}$  can be seen as the “error” on the expectation value in the  $i$ th bin as due to a  $\pm 1\sigma$  “error” of the  $j$ th source of systematic uncertainty. Thus, when  $\mathbf{K}$  is modeled via MC simulation, statistical fluctuations of the MC sample induce “errors on the errors”. This subject has a connection with Ref. [25], where breaking of asymptotic properties of the MLE has been also observed.

#### Appendix D. Supplementary material for Section 3

This appendix contains intermediate steps and supplementary material to support the main results of Section 3.

**Table B.9**

Coverage, mean, and median of the  $1\sigma$  confidence interval for  $\mu = \mu'_1$  for the toy model with  $N = 2 \times 10^6$ ,  $n = 200$ ,  $k = 1$ , and  $\epsilon = 0.06$ , obtained from an ensemble of  $10^4$  identically repeated pseudo-experiments. Cases where the observed coverage agrees with the expectation of 68.3% within a relative  $\pm 5\%$  tolerance are highlighted with a bold font. When present, the number within parenthesis refers to the statistical error on the last digit.

| Likelihood       | Minimim. | CI method | $\mu_t$            | Coverage        | $\hat{\sigma}$ (mean, median) |
|------------------|----------|-----------|--------------------|-----------------|-------------------------------|
| Gauss (asympt.)  | Analytic | Hessian   | 0                  | 0.683           | 0.047, 0.047                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.683           | 0.047, 0.047                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.683           | 0.047, 0.047                  |
| Poisson          | Numeric  | Hessian   | 0                  | 0.519(5)        | 0.039, 0.039                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.248(4)        | 0.039, 0.039                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.254(4)        | 0.039, 0.039                  |
| Gauss            | Numeric  | Hessian   | 0                  | 0.520(5)        | 0.039, 0.039                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.249(4)        | 0.039, 0.039                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.255(4)        | 0.039, 0.039                  |
| Gauss            | Analytic | Hessian   | 0                  | 0.520(5)        | 0.039, 0.039                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.249(4)        | 0.039, 0.039                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.255(4)        | 0.039, 0.039                  |
| Gauss + MC stat. | Analytic | Hessian   | 0                  | <b>0.680(5)</b> | 0.056, 0.056                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.362(5)        | 0.056, 0.056                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.372(5)        | 0.056, 0.056                  |
| Gauss + BB-lite  | Numeric  | Hessian   | 0                  | <b>0.680(5)</b> | 0.056, 0.056                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.363(5)        | 0.055, 0.055                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.372(5)        | 0.056, 0.056                  |
| Gauss + BB       | Numeric  | Hessian   | 0                  | 0.592(5)        | 0.067, 0.067                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.600(5)        | 0.068, 0.067                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.598(5)        | 0.068, 0.068                  |
| Gauss + BB       | Numeric  | PLR       | 0                  | 0.593(5)        | 0.067, 0.067                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.602(5)        | 0.068, 0.068                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.599(5)        | 0.068, 0.067                  |
| Gauss + BB       | Numeric  | sPFC      | 0                  | 0.500(5)        | 0.055, 0.054                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.532(5)        | 0.060, 0.059                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.528(5)        | 0.059, 0.059                  |
| Gauss + BB       | Numeric  | PFC       | 0                  | 0.560(5)        | 0.067, 0.061                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.628(5)        | 0.072, 0.064                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.628(5)        | 0.071, 0.063                  |
| Gauss + BB       | Numeric  | Barlett   | 0                  | 0.553(5)        | 0.065, 0.061                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.598(5)        | 0.070, 0.063                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.595(5)        | 0.069, 0.062                  |
| Gauss + BB       | Numeric  | CH        | 0                  | <b>0.687(5)</b> | 0.081, 0.077                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.794(4)        | 0.099, 0.090                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.789(4)        | 0.097, 0.088                  |
| Gauss + BB       | Numeric  | FC Cheat  | 0                  | <b>0.691(5)</b> | 0.081, 0.071                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.681(5)</b> | 0.083, 0.072                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.678(5)</b> | 0.081, 0.070                  |
| Gauss + BB       | Numeric  | Heuristic | 0                  | <b>0.682(5)</b> | 0.080, 0.080                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.690(5)</b> | 0.081, 0.081                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.684(5)</b> | 0.082, 0.081                  |

#### D.1. Additional results for the profile-likelihood ratio

The profile likelihood estimator for the NP is given by:

$$\hat{\theta}_\mu = \theta_0 + (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{r}_\mu. \quad (\text{D.1})$$

where  $\mathbf{r}_\mu = \mathbf{d} - \mathbf{b}(\mu - \mu_0)$ . Notice that the matrix  $\mathbf{A}^T \mathbf{A}$  is positive-definite, hence invertible, because  $\mathbf{A}$  was assumed to have full rank. By inserting the right-hand side of Eq. (D.1) into Eq. (29), the log-likelihood function becomes a quadratic function of  $\mu$ , which can be again minimized analytically yielding the estimator:

$$\hat{\mu} = \mu_0 + (\mathbf{b}^T \mathbf{U} \mathbf{b})^{-1} \mathbf{b}^T \mathbf{U} \mathbf{d}. \quad (\text{D.2})$$

An alternative way to prove the relation between the  $S$  variable defined in Eq. (35) and the variance of the MLE estimator  $\hat{\mu}$  is to derive directly the full set of MLE estimators, that is the point in the parameter

space that minimizes Eq. (27) globally, obtaining:

$$\begin{pmatrix} \hat{\theta} \\ \hat{\mu} \end{pmatrix} = \begin{pmatrix} \theta_0 \\ \mu_0 \end{pmatrix} + (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{d}, \quad (\text{D.3})$$

where  $\mathbf{B}$  is a  $n \times (p+1)$  matrix obtained by adding the column  $\mathbf{j}$  on the right of  $\mathbf{J}$ , and  $\mathbf{C} = (\mathbf{B}^T \mathbf{B})^{-1}$  is the covariance matrix of  $(\hat{\theta}, \hat{\mu})$ . By using the formula for the inverse block matrix, one can easily verify that

$$\mathbf{C} = \begin{pmatrix} \mathbf{W} & \mathbf{A}^T \mathbf{b} \\ \mathbf{b}^T \mathbf{A} & \mathbf{b}^T \mathbf{b} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{W}^{-1} + \mathbf{S}^{-1} \mathbf{W}^{-1} \mathbf{A}^T \mathbf{b} \mathbf{b}^T \mathbf{A} \mathbf{W}^{-1} & -\mathbf{S}^{-1} \mathbf{W}^{-1} \mathbf{A}^T \mathbf{b} \\ -\mathbf{S}^{-1} \mathbf{b}^T \mathbf{A} \mathbf{W}^{-1} & \mathbf{S}^{-1} \end{pmatrix}. \quad (\text{D.4})$$

where  $\mathbf{W} = \mathbf{A}^T \mathbf{A}$ . Eq. (D.4) implies that

$$\sigma_H^2 = [\mathbf{C}]_{p+1, p+1} = \mathbf{S}^{-1}, \quad (\text{D.5})$$

**Table B.10**

Coverage, mean, and median of the  $1\sigma$  confidence interval for  $\mu = \mu'_1$  for the toy model with  $N = 2 \times 10^6$ ,  $n = 200$ ,  $k = 1$ , and  $\epsilon = 0.12$ , obtained from an ensemble of  $10^4$  identically repeated pseudo-experiments. Cases where the observed coverage agrees with the expectation of 68.3% within a relative  $\pm 5\%$  tolerance are highlighted with a bold font. When present, the number within parenthesis refers to the statistical error on the last digit.

| Likelihood       | Minimim. | CI method | $\mu_t$            | Coverage        | $\hat{\sigma}$ (mean, median) |
|------------------|----------|-----------|--------------------|-----------------|-------------------------------|
| Gauss (asympt.)  | Analytic | Hessian   | 0                  | 0.683           | 0.024, 0.024                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.683           | 0.024, 0.024                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.683           | 0.024, 0.024                  |
| Poisson          | Numeric  | Hessian   | 0                  | 0.523(5)        | 0.022, 0.022                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.485(5)        | 0.022, 0.022                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.498(5)        | 0.022, 0.022                  |
| Gauss            | Numeric  | Hessian   | 0                  | 0.522(5)        | 0.022, 0.022                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.486(5)        | 0.022, 0.022                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.496(5)        | 0.022, 0.022                  |
| Gauss            | Analytic | Hessian   | 0                  | 0.522(5)        | 0.022, 0.022                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.486(5)        | 0.022, 0.022                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.496(5)        | 0.022, 0.022                  |
| Gauss + MC stat. | Analytic | Hessian   | 0                  | <b>0.689(5)</b> | 0.032, 0.032                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.648(5)        | 0.032, 0.032                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.655(5)</b> | 0.032, 0.032                  |
| Gauss + BB-lite  | Numeric  | Hessian   | 0                  | <b>0.688(5)</b> | 0.032, 0.032                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.648(5)        | 0.032, 0.032                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.656(5)</b> | 0.032, 0.032                  |
| Gauss + BB       | Numeric  | Hessian   | 0                  | <b>0.663(5)</b> | 0.033, 0.033                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.658(5)</b> | 0.033, 0.033                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.661(5)</b> | 0.034, 0.034                  |
| Gauss + BB       | Numeric  | PLR       | 0                  | <b>0.664(5)</b> | 0.033, 0.033                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.658(5)</b> | 0.034, 0.033                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.662(5)</b> | 0.034, 0.034                  |
| Gauss + BB       | Numeric  | sPFC      | 0                  | 0.521(5)        | 0.025, 0.025                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.512(5)        | 0.025, 0.025                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.519(5)        | 0.025, 0.025                  |
| Gauss + BB       | Numeric  | PFC       | 0                  | 0.596(5)        | 0.030, 0.027                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.593(5)        | 0.030, 0.028                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.592(5)        | 0.031, 0.028                  |
| Gauss + BB       | Numeric  | Barlett   | 0                  | 0.588(5)        | 0.029, 0.027                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.580(5)        | 0.030, 0.027                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.579(5)        | 0.030, 0.027                  |
| Gauss + BB       | Numeric  | CH        | 0                  | <b>0.700(5)</b> | 0.035, 0.033                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.711(5)</b> | 0.036, 0.033                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.706(5)</b> | 0.036, 0.033                  |
| Gauss + BB       | Numeric  | FC Cheat  | 0                  | <b>0.683(5)</b> | 0.035, 0.031                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.683(5)</b> | 0.036, 0.031                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.684(5)</b> | 0.036, 0.031                  |
| Gauss + BB       | Numeric  | Heuristic | 0                  | <b>0.689(5)</b> | 0.035, 0.035                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.685(5)</b> | 0.035, 0.035                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.689(5)</b> | 0.035, 0.035                  |

thus showing that, in the Gaussian approximation, the standard deviation of  $\hat{\mu}$  obtained from the Hessian matrix and from the concavity of the profile-likelihood ratio function coincide.<sup>12</sup>

An obvious consequence of Eq. (35) is that  $\mu$  cannot be measured (i.e.  $\sigma_H$  would be virtually infinite) if  $\mathbf{b}$  can be expressed as a linear combination of the columns of  $\mathbf{A}$ . In fact, in this case one could write  $\mathbf{b} = \mathbf{A}\mathbf{w}$  for some vector of weights  $\mathbf{w}$ , so that

$$\begin{aligned} \mathbf{b}^T \mathbf{U} \mathbf{b} &= \mathbf{b}^T (\mathbf{1} - \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T) \mathbf{A} \mathbf{w} \\ &= \mathbf{b}^T \mathbf{A} \mathbf{w} - \mathbf{b}^T \mathbf{A} (\mathbf{A}^T \mathbf{A})^{-1} (\mathbf{A}^T \mathbf{A}) \mathbf{w} \\ &= \mathbf{b}^T (\mathbf{A} \mathbf{w} - \mathbf{A} \mathbf{w}) = 0. \end{aligned} \quad (\text{D.6})$$

which would then imply  $\sigma_H \rightarrow \infty$ .

By means of Eq. (D.4), it can be easily proved that  $\rho_\mu^2$  is equivalent to the global correlation coefficient

$$1 - \left( [\mathbf{C}]_{p+1, p+1} \cdot [\mathbf{C}^{-1}]_{p+1, p+1} \right)^{-1}. \quad (\text{D.7})$$

Finally, we can use the results obtained so far to show that  $\rho_\mu^2 \sigma_H^2$  is also equal to the difference in quadrature between  $\sigma_H^2$  and the variance of  $\mu$  when  $\theta$  are fixed to their post-fit values. Hence,  $\rho_\mu$  can be in principle computed also from the *group impact* ( $I_\theta$ ) on  $\mu$  from all nuisance parameters [9,29], namely:

$$\rho_\mu = \frac{I_\theta}{\sigma_H}, \quad \text{with } I_\theta = \sqrt{[\mathbf{C}]_{\theta\theta}^T [\mathbf{C}]_{\theta\theta}^{-1} [\mathbf{C}]_{\theta\mu}}. \quad (\text{D.8})$$

<sup>12</sup> Formally, Eq. (D.4) implies that  $S$  is the Schur complement of the upper-left block  $\mathbf{W}$  of the Hessian matrix  $\mathbf{C}^{-1}$ .

**Table B.11**

Coverage, mean, and median of the  $1\sigma$  confidence interval for  $\mu = \mu'_1$  for the toy model with  $N = 2 \times 10^6$ ,  $n = 100$ ,  $k = 1$ , and  $\epsilon = 0.03$ , obtained from an ensemble of  $10^4$  identically repeated pseudo-experiments. Cases where the observed coverage agrees with the expectation of 68.3% within a relative  $\pm 5\%$  tolerance are highlighted with a bold font. When present, the number within parenthesis refers to the statistical error on the last digit.

| Likelihood       | Minimim. | CI method | $\mu_t$            | Coverage        | $\hat{\sigma}$ (mean, median) |
|------------------|----------|-----------|--------------------|-----------------|-------------------------------|
| Gauss (asympt.)  | Analytic | Hessian   | 0                  | 0.683           | 0.094, 0.094                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.683           | 0.094, 0.094                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.683           | 0.094, 0.094                  |
| Poisson          | Numeric  | Hessian   | 0                  | 0.530(5)        | 0.069, 0.069                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.060(2)        | 0.069, 0.069                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.067(2)        | 0.069, 0.069                  |
| Gauss            | Numeric  | Hessian   | 0                  | 0.531(5)        | 0.069, 0.069                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.060(2)        | 0.069, 0.069                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.067(3)        | 0.069, 0.069                  |
| Gauss            | Analytic | Hessian   | 0                  | 0.531(5)        | 0.069, 0.069                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.060(2)        | 0.069, 0.069                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.067(3)        | 0.069, 0.069                  |
| Gauss + MC stat. | Analytic | Hessian   | 0                  | <b>0.691(5)</b> | 0.098, 0.097                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.107(3)        | 0.098, 0.097                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.110(3)        | 0.098, 0.098                  |
| Gauss + BB-lite  | Numeric  | Hessian   | 0                  | <b>0.692(5)</b> | 0.098, 0.097                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.107(3)        | 0.097, 0.097                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.110(3)        | 0.098, 0.098                  |
| Gauss + BB       | Numeric  | Hessian   | 0                  | 0.545(5)        | 0.137, 0.134                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.549(5)        | 0.144, 0.140                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.548(5)        | 0.145, 0.141                  |
| Gauss + BB       | Numeric  | PLR       | 0                  | 0.552(5)        | 0.138, 0.134                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.557(5)        | 0.150, 0.145                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.556(5)        | 0.141, 0.137                  |
| Gauss + BB       | Numeric  | sPFC      | 0                  | 0.481(5)        | 0.126, 0.121                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.537(5)        | 0.149, 0.143                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.543(5)        | 0.138, 0.134                  |
| Gauss + BB       | Numeric  | PFC       | 0                  | 0.529(5)        | 0.151, 0.138                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.637(5)        | 0.181, 0.153                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.635(5)        | 0.169, 0.145                  |
| Gauss + BB       | Numeric  | Barlett   | 0                  | 0.523(5)        | 0.147, 0.136                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.599(5)        | 0.172, 0.149                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.594(5)        | 0.161, 0.141                  |
| Gauss + BB       | Numeric  | CH        | 0                  | <b>0.692(5)</b> | 0.193, 0.180                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.815(4)        | 0.266, 0.249                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.821(4)        | 0.253, 0.237                  |
| Gauss + BB       | Numeric  | FC Cheat  | 0                  | <b>0.688(5)</b> | 0.191, 0.170                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.683(5)</b> | 0.199, 0.174                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.683(5)</b> | 0.186, 0.164                  |
| Gauss + BB       | Numeric  | Heuristic | 0                  | <b>0.691(5)</b> | 0.186, 0.183                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.697(5)</b> | 0.196, 0.192                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.695(5)</b> | 0.197, 0.192                  |

where  $[C]_{\theta\theta}$  and  $[C]_{\theta\mu}$  stand for the upper-left  $p \times p$  and upper-right  $p \times 1$  blocks of the covariance matrix  $C$ .<sup>13</sup>

### D.2. Derivation of Eq. (39)

By expanding  $S$  around its true value, we obtain an *estimator*  $\hat{S}$  given by

$$\hat{S} = \hat{\mathbf{b}}^T \left( 1 - \sum_{j=1}^p \hat{\mathbf{u}}_j \hat{\mathbf{u}}_j^T \right) \hat{\mathbf{b}} = S + 2\beta^T \mathbf{U} \mathbf{b} + \beta^T \mathbf{U} \beta$$

<sup>13</sup> For a single parameter  $\theta_j$ , the impact can be more simply written as  $I_{\theta_j} = \frac{|C|_{\theta_j}}{\sqrt{|C|_{\theta_j \theta_j}}} = |\rho_{\mu \theta_j}| \cdot \sigma_H$ , with  $\rho_{\mu \theta_j}$  equal to the linear correlation between  $\mu$  and  $\theta_j$ .

$$\begin{aligned} & - 2 \sum_{j=1}^p (\mathbf{b}^T \mathbf{u}_j) (\mathbf{b}^T \mathbf{v}_j) - \mathbf{b}^T \left( \sum_{j=1}^p \mathbf{v}_j \mathbf{v}_j^T \right) \mathbf{b} \\ & + 2\beta^T \left( - \sum_{j=1}^p \mathbf{u}_j \mathbf{v}_j^T - \sum_{j=1}^p \mathbf{v}_j \mathbf{u}_j^T \right) \mathbf{b} + \dots, \end{aligned} \quad (\text{D.9})$$

where the dots stand for terms that are of higher-order in the perturbations. Eq. (39) follows by taking the expectation value of the right-hand side of Eq. (D.9).

### D.3. Contribution from the matrix $A$ for generic column vectors

We consider the more general case for the columns vectors  $\mathbf{a}_j$ . In this case, we can always find a set of  $p$  orthonormal vectors such that

$$\mathbf{u}_j = \sum_{k=1}^p w_{jk} \mathbf{a}_k \quad (\text{D.10})$$

**Table B.12**

Coverage, mean, and median of the  $1\sigma$  confidence interval for  $\mu = \mu'_1$  for the toy model with  $N = 2 \times 10^6$ ,  $n = 20$ ,  $k = 1$ , and  $\epsilon = 0.03$ , obtained from an ensemble of  $10^4$  identically repeated pseudo-experiments. Cases where the observed coverage agrees with the expectation of 68.3% within a relative  $\pm 5\%$  tolerance are highlighted with a bold font. When present, the number within parenthesis refers to the statistical error on the last digit.

| Likelihood       | Minimim. | CI method | $\mu_t$            | Coverage        | $\hat{\sigma}$ (mean, median) |
|------------------|----------|-----------|--------------------|-----------------|-------------------------------|
| Gauss (asympt.)  | Analytic | Hessian   | 0                  | 0.683           | 0.094, 0.094                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.683           | 0.094, 0.094                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.683           | 0.094, 0.094                  |
| Poisson          | Numeric  | Hessian   | 0                  | 0.514(5)        | 0.088, 0.088                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.450(5)        | 0.088, 0.087                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.442(5)        | 0.088, 0.088                  |
| Gauss            | Numeric  | Hessian   | 0                  | 0.514(5)        | 0.088, 0.088                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.450(5)        | 0.088, 0.087                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.441(5)        | 0.088, 0.088                  |
| Gauss            | Analytic | Hessian   | 0                  | 0.514(5)        | 0.088, 0.088                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.450(5)        | 0.088, 0.087                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.441(5)        | 0.088, 0.088                  |
| Gauss + MC stat. | Analytic | Hessian   | 0                  | <b>0.676(5)</b> | 0.125, 0.124                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.604(5)        | 0.125, 0.124                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.593(5)        | 0.125, 0.124                  |
| Gauss + BB-lite  | Numeric  | Hessian   | 0                  | <b>0.676(5)</b> | 0.125, 0.124                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.603(5)        | 0.124, 0.124                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.594(5)        | 0.125, 0.124                  |
| Gauss + BB       | Numeric  | Hessian   | 0                  | 0.641(5)        | 0.136, 0.134                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.653(5)</b> | 0.143, 0.141                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.653(5)</b> | 0.144, 0.141                  |
| Gauss + BB       | Numeric  | PLR       | 0                  | 0.649(5)        | 0.136, 0.135                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.660(5)</b> | 0.148, 0.145                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.660(5)</b> | 0.140, 0.138                  |
| Gauss + BB       | Numeric  | sPFC      | 0                  | 0.511(5)        | 0.101, 0.100                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.501(5)        | 0.104, 0.102                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.495(5)        | 0.100, 0.099                  |
| Gauss + BB       | Numeric  | PFC       | 0                  | 0.569(5)        | 0.127, 0.116                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.588(5)        | 0.138, 0.120                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.592(5)        | 0.130, 0.115                  |
| Gauss + BB       | Numeric  | Barlett   | 0                  | 0.559(5)        | 0.123, 0.115                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.568(5)        | 0.133, 0.118                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.572(5)        | 0.125, 0.114                  |
| Gauss + BB       | Numeric  | CH        | 0                  | <b>0.683(5)</b> | 0.149, 0.140                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.724(4)        | 0.166, 0.146                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.723(4)        | 0.155, 0.139                  |
| Gauss + BB       | Numeric  | FC Cheat  | 0                  | <b>0.688(5)</b> | 0.148, 0.130                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.681(5)</b> | 0.160, 0.137                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.685(5)</b> | 0.149, 0.129                  |
| Gauss + BB       | Numeric  | Heuristic | 0                  | <b>0.675(5)</b> | 0.145, 0.144                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.687(5)</b> | 0.152, 0.151                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.688(5)</b> | 0.153, 0.152                  |

where  $w_{jk}$  are the elements of a non-singular square matrix. After some straightforward algebra, the expectation value of the two last terms on the second line of Eq. (39) can be written as:

$$\sum_{j,k,l=1}^p w_{jk}w_{jl} (\mathbf{b}^T \mathbf{a}_k) (\mathbf{b}^T \mathbf{a}_l) \langle \alpha_l^2 \rangle \left( 1 + \frac{\delta_{kl}}{n-1} \right) - \frac{\mathbf{b}^T \mathbf{b}}{n-1} \sum_{j,k=1}^p w_{jk}^2 \langle \alpha_k^2 \rangle \quad (\text{D.11})$$

Again, we first focus on the case of large overlap between  $\mathbf{b}$  and the linear space  $\mathcal{V}_A$ , which allows us to consider only the first term on the right-hand side of Eq. (D.11) since the remaining terms are relatively suppressed by a factor of  $1/n$ . If we denote by  $\mathbf{v}$  a vector with  $k^{\text{th}}$ -element equal to  $\mathbf{b}^T \mathbf{a}_k$ , we see that the triple sum of Eq. (D.11) can be written as

$$\mathbf{v}^T \left( \sum_{j=1}^p \mathbf{w}_j \mathbf{w}_j^T \text{diag}(\langle \alpha_1^2 \rangle, \dots, \langle \alpha_p^2 \rangle) \right) \mathbf{v} \quad (\text{D.12})$$

The matrix defined between parentheses in Eq. (D.12) is positive (semi-)definite if  $\langle \alpha_k^2 \rangle > 0$  for all (at least one) index  $k$ , so that the quadratic form cannot be negative. In the limiting case of small  $\rho_\mu$ , the rightmost term in Eq. (D.11) dominates, and the same conclusions derived for the case of orthogonal columns holds.

#### D.4. The large correlation case for arbitrary $n$

By expanding  $\mathbf{b}$  as a linear combination of the vectors  $\mathbf{a}_j$ , and neglecting the  $\mathbf{b}_\perp$  term, we have

$$\frac{n}{n-1} \sum_{j=1}^p (\mathbf{b}^T \mathbf{a}_j)^2 \langle \alpha_j^2 \rangle - \frac{\mathbf{b}^T \mathbf{b}}{n-1} \left( \sum_{j=1}^p \langle \alpha_j^2 \rangle \right) > \frac{1}{n-1} \left[ p \sum_{j=1}^p (\mathbf{b}^T \mathbf{a}_j)^2 \langle \alpha_j^2 \rangle - \left( \sum_{j=1}^p (\mathbf{b}^T \mathbf{a}_j)^2 \right) \left( \sum_{j=1}^p \langle \alpha_j^2 \rangle \right) \right] \quad (\text{D.13})$$

where we have used the fact that  $n > p$ . The quantity within square brackets is positive-definite for all  $n$  as for Chebyshev's sum inequality,

**Table B.13**

Coverage, mean, and median of the  $1\sigma$  confidence interval for  $\mu = \mu'_1$  for the toy model with  $N = 2 \times 10^7$ ,  $n = 200$ ,  $k = 1$ , and  $\epsilon = 0.03$ , obtained from an ensemble of  $10^4$  identically repeated pseudo-experiments. Cases where the observed coverage agrees with the expectation of 68.3% within a relative  $\pm 5\%$  tolerance are highlighted with a bold font. When present, the number within parenthesis refers to the statistical error on the last digit.

| Likelihood       | Minimim. | CI method | $\mu_t$                     | Coverage        | $\hat{\sigma}$ (mean, median) |
|------------------|----------|-----------|-----------------------------|-----------------|-------------------------------|
| Gauss (asympt.)  | Analytic | Hessian   | 0                           | 0.683           | 0.030, 0.030                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.683           | 0.030, 0.030                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.683           | 0.030, 0.030                  |
| Poisson          | Numeric  | Hessian   | 0                           | 0.518(5)        | 0.028, 0.028                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.443(5)        | 0.028, 0.027                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.452(5)        | 0.028, 0.028                  |
| Gauss            | Numeric  | Hessian   | 0                           | 0.518(5)        | 0.028, 0.028                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.443(5)        | 0.028, 0.027                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.451(5)        | 0.028, 0.028                  |
| Gauss            | Analytic | Hessian   | 0                           | 0.518(5)        | 0.028, 0.028                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.443(5)        | 0.028, 0.027                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.451(5)        | 0.028, 0.028                  |
| Gauss + MC stat. | Analytic | Hessian   | 0                           | <b>0.690(5)</b> | 0.039, 0.039                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.604(5)        | 0.039, 0.039                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.608(5)        | 0.039, 0.039                  |
| Gauss + BB-lite  | Numeric  | Hessian   | 0                           | <b>0.690(5)</b> | 0.039, 0.039                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.604(5)        | 0.039, 0.039                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.608(5)        | 0.039, 0.039                  |
| Gauss + BB       | Numeric  | Hessian   | 0                           | 0.646(5)        | 0.042, 0.042                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.642(5)        | 0.042, 0.042                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.644(5)        | 0.042, 0.042                  |
| Gauss + BB       | Numeric  | PLR       | 0                           | 0.647(5)        | 0.042, 0.042                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.643(5)        | 0.043, 0.043                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.646(5)        | 0.042, 0.042                  |
| Gauss + BB       | Numeric  | sPFC      | 0                           | 0.518(5)        | 0.032, 0.032                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.521(5)        | 0.032, 0.032                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.520(5)        | 0.032, 0.032                  |
| Gauss + BB       | Numeric  | PFC       | 0                           | 0.579(5)        | 0.039, 0.036                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.599(5)        | 0.040, 0.036                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.594(5)        | 0.040, 0.036                  |
| Gauss + BB       | Numeric  | Barlett   | 0                           | 0.571(5)        | 0.038, 0.035                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.582(5)        | 0.039, 0.036                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.576(5)        | 0.039, 0.036                  |
| Gauss + BB       | Numeric  | CH        | 0                           | <b>0.693(5)</b> | 0.046, 0.043                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | 0.737(4)        | 0.048, 0.044                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | 0.729(4)        | 0.048, 0.044                  |
| Gauss + BB       | Numeric  | FC Cheat  | 0                           | <b>0.687(5)</b> | 0.046, 0.040                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | <b>0.677(5)</b> | 0.046, 0.040                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | <b>0.685(5)</b> | 0.046, 0.040                  |
| Gauss + BB       | Numeric  | Heuristic | 0                           | <b>0.690(5)</b> | 0.046, 0.046                  |
|                  |          |           | $+5\bar{\sigma}_{\text{H}}$ | <b>0.682(5)</b> | 0.046, 0.046                  |
|                  |          |           | $-5\bar{\sigma}_{\text{H}}$ | <b>0.681(5)</b> | 0.046, 0.046                  |

provided that  $\mathbf{b}^T \mathbf{a}_1 \geq \dots \geq \mathbf{b}^T \mathbf{a}_p$  and  $\langle \alpha_1^2 \rangle \geq \dots \geq \langle \alpha_p^2 \rangle$ , which includes both the case of uniform uncertainties on the columns, as well as the case of increasingly larger uncertainty for column vectors that have larger overlap with  $\mathbf{b}$ , which is also an intuitive result.

#### D.5. The two-dimensional problem

Eqs. (43) and (44) can be verified, for example, in a simple two-dimensional setup, namely for  $n = 2$  and  $p = 1$ . Modulo an overall normalization factor to the vector  $\mathbf{b}$  and to the single-column matrix  $\mathbf{A}$ , we can define

$$\mathbf{b} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{A} = \begin{pmatrix} s_\gamma \\ c_\gamma \end{pmatrix} \Rightarrow \mathbf{U} = \begin{pmatrix} c_\gamma^2 & 1 - s_\gamma c_\gamma \\ 1 - s_\gamma c_\gamma & s_\gamma^2 \end{pmatrix}, \quad (\text{D.14})$$

where  $s_\gamma \equiv \sin \gamma$ ,  $c_\gamma \equiv \cos \gamma$ , with  $\gamma \in [0, \pi]$  an angle that parametrizes the degree of correlation between  $\mu$  and the nuisance parameter  $\theta$ . From Eq. (D.14) we have that  $S = U_{11} = c_\gamma^2$  and  $\rho_\mu^2 = s_\gamma^2 = 1 - S$ . We

treat statistical fluctuations on  $\mathbf{A}$  as a pair of independent and Gaussian-distributed random variables  $(\alpha_1, \alpha_2)$ , which, for simplicity, are assumed to have the same variance  $\langle \alpha^2 \rangle$ . Under fluctuations, the matrix  $\mathbf{U}$  is then changed to

$$\mathbf{U} \rightarrow \hat{\mathbf{U}} = \begin{pmatrix} \frac{(c_\gamma + \alpha_2)^2}{(s_\gamma + \alpha_1)^2 + (c_\gamma + \alpha_2)^2} & \frac{(s_\gamma + \alpha_1 - c_\gamma - \alpha_2)^2 + (s_\gamma + \alpha_1)(c_\gamma + \alpha_2)}{(s_\gamma + \alpha_1)^2 + (c_\gamma + \alpha_2)^2} \\ - & \frac{(s_\gamma + \alpha_1)^2}{(s_\gamma + \alpha_1)^2 + (c_\gamma + \alpha_2)^2} \end{pmatrix} \quad (\text{D.15})$$

Using Eq. (D.15) evaluated at randomly generated points  $(\alpha_1, \alpha_2)$ , the estimator  $\hat{S} = \hat{U}_{11}$  can be calculated and its mean value  $\langle \hat{S} \rangle$  extracted from a large number of random samplings. The results are shown in Fig. D.5 as a function of  $\rho_\mu^2$  and for three representative values of  $\langle \alpha^2 \rangle$ . For small values of  $\rho_\mu^2$ , the ratio  $\langle \hat{S} \rangle / S$  is slightly less than one, i.e. the bias on  $\hat{S}$  is negative, in agreement with Eq. (44). Conversely, when  $\rho_\mu^2$  increases, the ratio becomes larger than one (hence, the bias gets positive), eventually growing as  $\langle \alpha^2 \rangle / (1 - \rho_\mu^2)$  for  $\rho_\mu^2 \rightarrow 1$ , in agreement with Eq. (43).

**Table B.14**

Coverage, mean, and median of the  $1\sigma$  confidence interval for  $\mu = \mu'_1$  for the toy model with  $N = 2 \times 10^8$ ,  $n = 200$ ,  $k = 1$ , and  $\epsilon = 0.03$ , obtained from an ensemble of  $10^4$  identically repeated pseudo-experiments. Cases where the observed coverage agrees with the expectation of 68.3% within a relative  $\pm 5\%$  tolerance are highlighted with a bold font. When present, the number within parenthesis refers to the statistical error on the last digit.

| Likelihood       | Minimim. | CI method | $\mu_t$            | Coverage        | $\hat{\sigma}$ (mean, median) |
|------------------|----------|-----------|--------------------|-----------------|-------------------------------|
| Gauss (asympt.)  | Analytic | Hessian   | 0                  | 0.683           | 0.009, 0.009                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.683           | 0.009, 0.009                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.683           | 0.009, 0.009                  |
| Poisson          | Numeric  | Hessian   | 0                  | 0.529(5)        | 0.009, 0.010                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.516(5)        | 0.009, 0.010                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.514(5)        | 0.009, 0.010                  |
| Gauss            | Numeric  | Hessian   | 0                  | 0.529(5)        | 0.009, 0.010                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.516(5)        | 0.009, 0.010                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.514(5)        | 0.009, 0.010                  |
| Gauss            | Analytic | Hessian   | 0                  | 0.529(5)        | 0.009, 0.010                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.516(5)        | 0.009, 0.010                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.514(5)        | 0.009, 0.010                  |
| Gauss + MC stat. | Analytic | Hessian   | 0                  | <b>0.693(5)</b> | 0.013, 0.013                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.678(5)</b> | 0.013, 0.013                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.679(5)</b> | 0.013, 0.013                  |
| Gauss + BB-lite  | Numeric  | Hessian   | 0                  | <b>0.693(5)</b> | 0.013, 0.013                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.678(5)</b> | 0.013, 0.013                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.678(5)</b> | 0.013, 0.013                  |
| Gauss + BB       | Numeric  | Hessian   | 0                  | <b>0.688(5)</b> | 0.013, 0.013                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.675(5)</b> | 0.013, 0.013                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.674(5)</b> | 0.013, 0.013                  |
| Gauss + BB       | Numeric  | PLR       | 0                  | <b>0.688(5)</b> | 0.013, 0.013                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.675(5)</b> | 0.013, 0.013                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.674(5)</b> | 0.013, 0.013                  |
| Gauss + BB       | Numeric  | sPFC      | 0                  | 0.518(5)        | 0.009, 0.009                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.522(5)        | 0.010, 0.010                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.518(5)        | 0.010, 0.010                  |
| Gauss + BB       | Numeric  | PFC       | 0                  | 0.590(5)        | 0.012, 0.011                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.591(5)        | 0.012, 0.011                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.602(5)        | 0.012, 0.011                  |
| Gauss + BB       | Numeric  | Barlett   | 0                  | 0.583(5)        | 0.011, 0.011                  |
|                  |          |           | $+5\bar{\sigma}_H$ | 0.582(5)        | 0.011, 0.011                  |
|                  |          |           | $-5\bar{\sigma}_H$ | 0.593(5)        | 0.011, 0.011                  |
| Gauss + BB       | Numeric  | CH        | 0                  | <b>0.686(5)</b> | 0.014, 0.013                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.694(5)</b> | 0.014, 0.013                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.701(5)</b> | 0.014, 0.013                  |
| Gauss + BB       | Numeric  | FC Cheat  | 0                  | <b>0.683(5)</b> | 0.013, 0.012                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.671(5)</b> | 0.014, 0.012                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.680(5)</b> | 0.013, 0.012                  |
| Gauss + BB       | Numeric  | Heuristic | 0                  | <b>0.692(5)</b> | 0.013, 0.014                  |
|                  |          |           | $+5\bar{\sigma}_H$ | <b>0.677(5)</b> | 0.013, 0.014                  |
|                  |          |           | $-5\bar{\sigma}_H$ | <b>0.679(5)</b> | 0.013, 0.014                  |

#### D.6. Scaling laws for the Barlow–Beeston lite likelihood

Eq. (48) assumes that the likelihood does not contain any explicit term to account for the statistical uncertainties on the MC templates. It can be easily extended to the case where the MC statistical uncertainties are treated *à la* Barlow–Beeston lite by exploiting the Gauss + MC stat. approximation (see Section 2.3), thus modifying the scaling of  $\mathbf{b}$  via an additional multiplicative factor  $\left(1 + \frac{1}{k}\right)^{-1/2}$ , yielding:

$$\bar{\sigma}_H = \left( \frac{r_\xi \sqrt{\xi - 1}}{\sqrt{\alpha \left(1 + \frac{1}{k}\right) r_\xi^2 - \left(1 + \frac{\xi}{k}\right)}} \right) \hat{\sigma}_H. \quad (\text{D.16})$$

#### D.7. Gradient of $\hat{\mu}$ with respect to $\mathbf{B}$ and $\mathbf{d}$

By applying the chain rule for differentiation, and using the known formula for the derivative of an inverse matrix, the Jacobian of  $(\hat{\theta}, \hat{\mu})$

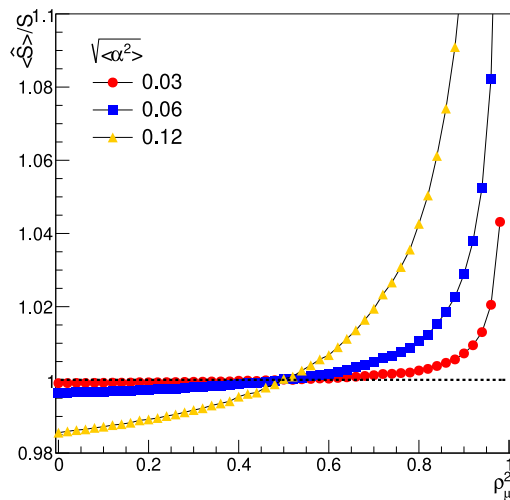
with respect to  $B_{ik}$  can be computed as

$$\begin{aligned} \frac{\partial}{\partial B_{ik}} \begin{pmatrix} \hat{\theta} \\ \hat{\mu} \end{pmatrix} &= \begin{pmatrix} \frac{\partial (\mathbf{B}^T \mathbf{B})^{-1}}{\partial B_{ik}} \end{pmatrix} \mathbf{B}^T \mathbf{d} + (\mathbf{B}^T \mathbf{B})^{-1} \begin{pmatrix} \frac{\partial \mathbf{B}^T}{\partial B_{ik}} \\ \frac{\partial \mathbf{B}}{\partial B_{ik}} \end{pmatrix} \mathbf{d} \\ &= -(\mathbf{B}^T \mathbf{B})^{-1} \frac{\partial (\mathbf{B}^T \mathbf{B})}{\partial B_{ik}} (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{d} + (\mathbf{B}^T \mathbf{B})^{-1} \begin{pmatrix} \frac{\partial \mathbf{B}^T}{\partial B_{ik}} \\ \frac{\partial \mathbf{B}}{\partial B_{ik}} \end{pmatrix} \mathbf{d} \\ &= -(\mathbf{B}^T \mathbf{B})^{-1} (\mathbf{\Pi}_{ik}^T \mathbf{B} + \mathbf{B}^T \mathbf{\Pi}_{ik}) \begin{pmatrix} \hat{\theta} \\ \hat{\mu} \end{pmatrix} + (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{\Pi}_{ik}^T \mathbf{d} \\ &= \mathbf{C} \begin{pmatrix} \mathbf{\Pi}_{ik}^T \left( \mathbf{d} - \mathbf{B} \begin{pmatrix} \hat{\theta} \\ \hat{\mu} \end{pmatrix} \right) \\ -\mathbf{B}^T \mathbf{\Pi}_{ik} \begin{pmatrix} \hat{\theta} \\ \hat{\mu} \end{pmatrix} \end{pmatrix}, \end{aligned} \quad (\text{D.17})$$

where  $\mathbf{\Pi}_{ik} = \mathbf{e}_i \mathbf{e}_k^T$ , with  $\mathbf{e}_l$  the  $l$ th vector of the canonical basis. The gradient  $\partial \hat{\mu} / \partial B_{ik}$  is provided by the last element of Eq. (D.17).

The gradient of  $\hat{\mu}$  with respect to  $\mathbf{d}$  can be readily obtained, giving:

$$\frac{\partial \hat{\mu}}{\partial d_i} = \sum_{k=1}^{p+1} C_{p+1k} B_{ik}. \quad (\text{D.18})$$



**Fig. D.5.** The ratio between the expectation value  $\langle \hat{S} \rangle$  and the true value  $S$  for the two-dimensional problem of Eq. (D.14) as a function of the squared correlation coefficient  $\rho_{\mu}^2$ , and for three representative values of  $\langle \alpha^2 \rangle$ .

## Data availability

Data will be made available on request.

## References

- [1] T. Glusenkamp, Probabilistic treatment of the uncertainty from the finite size of weighted Monte Carlo data, *Eur. Phys. J. Plus* 133 (218) (2018) <http://dx.doi.org/10.1140/epjp/i2018-12042-x>.
- [2] R. Barlow, C. Beeston, Fitting using finite Monte Carlo samples, *Comput. Phys. Comm.* 77 (2) (1993) 219–228, [http://dx.doi.org/10.1016/0010-4655\(93\)90005-W](http://dx.doi.org/10.1016/0010-4655(93)90005-W).
- [3] D. Chirkin, Likelihood description for comparing data with simulation of limited statistics, 2013, <http://dx.doi.org/10.48550/arXiv.1304.073>, [arXiv:1304.0735](https://arxiv.org/abs/1304.0735).
- [4] C. Argüelles, A. Schneider, T. Yuan, A binned likelihood for stochastic models, *J. High Energ. Phys.* 06 (030) (2019) [http://dx.doi.org/10.1007/JHEP06\(2019\)030](http://dx.doi.org/10.1007/JHEP06(2019)030).
- [5] G. Cowan (Ed.), *Statistical Data Analysis*, Oxford University Press, Oxford UK, 1998.
- [6] A.S. Stuart A. (Ed.), *Kendall's Advanced Theory of Statistics, Classical Inference and the Linear Model, vol. 2A*, Oxford University Press, London, 1999.
- [7] G. Cowan, et al., Asymptotic formulae for likelihood-based tests of new physics, *Eur. Phys. J. C* 71 (1554) (2011) <http://dx.doi.org/10.1140/epjc/s10052-011-1554-0>.
- [8] Electroweak, QCD and flavour physics studies with ATLAS data from run 2 of the LHC, *Phys. Rep.* 1116 (2025) 57–126, <http://dx.doi.org/10.1016/j.physrep.2024.12.003>.
- [9] Stairway to discovery: A report on the CMS programme of cross section measurements from millibarns to femtobarns, *Phys. Rep.* 1115 (2025) 3–115, <http://dx.doi.org/10.1016/j.physrep.2024.11.005>.
- [10] G.J. Feldman, R.D. Cousins, Unified approach to the classical statistical analysis of small signals, *Phys. Rev. D* 57 (1998) 3873–3889, <http://dx.doi.org/10.1103/PhysRevD.57.3873>.
- [11] J. Neyman, Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability, *Philos. Trans. R. Soc. Lond. Ser. A* 236 (767) (1937) 333–380, <http://dx.doi.org/10.1098/rsta.1937.0005>.
- [12] J. Conway, Incorporating nuisance parameters in likelihoods for multisource spectra, 2011, Proceedings of the PHYSTAT 2011, [physics.data-an. arXiv:1103.0354](https://arxiv.org/abs/1103.0354).
- [13] L. Rottoli, P. Torrielli, A. Vicini, Determination of the W-boson mass at hadron colliders, *Eur. Phys. J. C* 83 (10) (2023) 948, <http://dx.doi.org/10.1140/epjc/s10052-023-12128-z>, [arXiv:2301.04059](https://arxiv.org/abs/2301.04059).
- [14] S.S. Wilks, The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses, *Annals Math. Stat.* 9 (1) (1938) 60–62, <http://dx.doi.org/10.1214/aoms/1177732360>.
- [15] F. James, M. Roos, Minuit - a system for function minimization and analysis of the parameter errors and correlations, *Comput. Phys. Comm.* 10 (6) (1975) 343–367, [http://dx.doi.org/10.1016/0010-4655\(75\)90039-9](http://dx.doi.org/10.1016/0010-4655(75)90039-9).
- [16] G. Guennebaud, B. Jacob, et al., Eigen v3, 2010, <http://eigen.tuxfamily.org>.
- [17] NOvA Collaboration Collaboration, Monte Carlo method for constructing confidence intervals with unconstrained and constrained nuisance parameters in the nova experiment, *J. Instrum.* 20 (02) (2025) T02001, <http://dx.doi.org/10.1088/1748-0221/20/02/T02001>.
- [18] B.M. Stevenson, Properties of sufficiency and statistical tests, *Proc. R. Soc. Lond. A* 160 (1937) 268–282, <http://dx.doi.org/10.1098/rspa.1937.0109>.
- [19] G. Cowan, Statistical models with uncertain error parameter, *Eur. Phys. J. C* 79 (2019) 133, <http://dx.doi.org/10.1140/epjc/s10052-019-6644-4>.
- [20] R.D. Cousins, V.L. Highland, Incorporating systematic uncertainties into an upper limit, *Nucl. Instrum. Methods Phys. Res. A* 320 (1) (1992) 331–335, [http://dx.doi.org/10.1016/0168-9002\(92\)90794-5](http://dx.doi.org/10.1016/0168-9002(92)90794-5).
- [21] J. Conrad, O. Botner, A. Hallgren, C.P. de los Heros, Coverage of confidence intervals for Poisson statistics in presence of systematic uncertainties, 2002, [arXiv:hep-ex/0206034](https://arxiv.org/abs/hep-ex/0206034).
- [22] G. Punzi, Ordering algorithms and confidence intervals in the presence of nuisance parameters, in: *Statistical Problems in Particle Physics, Astrophysics and Cosmology*, 2005, pp. 88–92, [http://dx.doi.org/10.1142/9781860948985\\_0019](http://dx.doi.org/10.1142/9781860948985_0019).
- [23] K.S. Cranmer, Frequentist hypothesis testing with background uncertainty, 2003, URL <https://arxiv.org/abs/physics/0310108>. [arXiv:physics/0310108](https://arxiv.org/abs/physics/0310108).
- [24] O. Behnke, L. Moneta, Parameter estimation, in: *Data Analysis in High Energy Physics*, John Wiley and Sons, Ltd, 2013, pp. 27–73, <http://dx.doi.org/10.1002/9783527653416.ch2>.
- [25] G. Cowan, Statistical Models with Uncertain Error Parameters, *Eur. Phys. J. C* 79 (2) (2019) 133, <http://dx.doi.org/10.1140/epjc/s10052-019-6644-4>, [arXiv:1809.05778](https://arxiv.org/abs/1809.05778).
- [26] S. Baker, R.D. Cousins, Clarification of the Use of Chi Square and Likelihood Functions in Fits to Histograms, *Nucl. Instrum. Meth.* 221 (1984) 437–442, [http://dx.doi.org/10.1016/0167-5087\(84\)90016-4](http://dx.doi.org/10.1016/0167-5087(84)90016-4).
- [27] K. Pearson, On lines and planes of closest fit to systems of points in space, *Lond. Edinb. Dublin Philos. Mag. J. Sci.* 2 (11) (1901) 559–572, <http://dx.doi.org/10.1080/14786440109462720>.
- [28] H. Hotelling, Analysis of a complex of statistical variables into principal components, *J. Educ. Psychol.* 417 (1933) <http://dx.doi.org/10.1037/h0071325>.
- [29] A. Pinto, Z. Wu, F. Balli, N. Berger, M. Boonekamp, É. Chapon, T. Kawamoto, B. Malaescu, Uncertainty components in profile likelihood fits, *Eur. Phys. J. C* 84 (6) (2024) 593, <http://dx.doi.org/10.1140/epjc/s10052-024-12877-5>, [arXiv:2307.04007](https://arxiv.org/abs/2307.04007).