HAL
open science

Modeling theoretical uncertainties in phenomenological analyses for particle physics

Jérôme Charles, Sebastien Descotes-Genon, Valentin Niess, Luiz Vale Silva

To cite this version:

Jérôme Charles, Sebastien Descotes-Genon, Valentin Niess, Luiz Vale Silva. Modeling theoretical uncertainties in phenomenological analyses for particle physics. European Physical Journal C: Particles and Fields, 2017, 77 (4), pp.214. 10.1140/epjc/s10052-017-4767-z . hal-01400814

HAL Id: hal-01400814

https://hal.science/hal-01400814v1

Submitted on 23 Apr 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LPT-Orsay-16-65

Modelling theoretical uncertainties in phenomenological analyses for particle physics

Jérôme Charles[a], Sébastien Descotes-Genon[b], Valentin Niess[c], Luiz Vale Silva[b,d,e]

[a] *CNRS, Aix-Marseille Univ, Université de Toulon, CPT UMR 7332, F-13288 Marseille Cedex 9, France*

[b] *Laboratoire de Physique Théorique (UMR 8627), CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91405 Orsay Cedex, France*

[c] *Laboratoire de Physique Corpusculaire, CNRS/IN2P3, UMR 6533, Campus des Cézeaux, 24 Av. des Landais, F-63177 Aubière Cedex, France*

[d] *Groupe de Physique Théorique, Institut de Physique Nucléaire, Univ. Paris-Sud, CNRS/IN2P3, Université Paris-Saclay, 91406 Orsay Cedex, France*

[e] *J. Stefan Institute, Jamova 39, P. O. Box 3000, 1001 Ljubljana, Slovenia*

arXiv:1611.04768v2 [hep-ph] 7 Apr 2017

Abstract

The determination of the fundamental parameters of the Standard Model (and its extensions) is often limited by the presence of statistical and theoretical uncertainties. We present several models for the latter uncertainties (random, nuisance, external) in the frequentist framework, and we derive the corresponding p-values. In the case of the nuisance approach where theoretical uncertainties are modeled as biases, we highlight the important, but arbitrary, issue of the range of variation chosen for the bias parameters. We introduce the concept of adaptive p-value, which is obtained by adjusting the range of variation for the bias according to the significance considered, and which allows us to tackle metrology and exclusion tests with a single and well-defined unified tool, which exhibits interesting frequentist properties. We discuss how the determination of fundamental parameters is impacted by the model chosen for theoretical uncertainties, illustrating several issues with examples from quark flavour physics.

Contents

1 Statistics concepts for particle physics **3**
1.1 p-values . . . 3
1.1.1 Data fitting and data reduction . . . 3
1.1.2 Model fitting . . . 4
1.2 Likelihood-ratio test statistic . . . 8

2 Comparing approaches to theoretical uncertainties **9**

3 Illustration of the approaches in the one-dimensional case **11**
3.1 Situation of the problem . . . 11
3.2 The random-δ approach . . . 12
3.3 The nuisance-δ approach . . . 12
3.4 The external-δ approach . . . 14

4 Comparison of the methods in the one-dimensional case **15**
4.1 p-values and confidence intervals . . . 15
4.2 Significance thresholds . . . 16
4.3 Coverage properties . . . 18
4.4 Conclusions of the uni-dimensional case . . . 22

5 Generalization to multi-dimensional cases **22**
5.1 General formulae . . . 22
5.2 Averaging measurements . . . 23
5.2.1 Averaging two measurements and the choice of a hypervolume . . . 23
5.2.2 Averaging n measurements with biases in a hyperball . . . 24
5.2.3 Averages with other approaches . . . 26
5.2.4 Other approaches in the literature . . . 26
5.3 Global fit . . . 27
5.3.1 Estimators and errors . . . 27
5.4 Goodness-of-fit . . . 29
5.5 Pull parameters . . . 29
5.6 Conclusions of the multi-dimensional case . . . 30

6 CKM-related examples **31**
6.1 Averaging theory-dominated measurements . . . 32
6.2 Averaging incompatible or barely compatible measurements . . . 36
6.3 Averaging quantities dominated by different types of uncertainties . . . 37
6.4 Global fits . . . 41

7 Conclusion **41**

A Singular covariance matrices **43**
A.1 Inversion of the covariance matrix . . . 43
A.2 Choice of a generalised inverse . . . 44
A.3 Examples . . . 45
A.3.1 Two measurements . . . 45
A.3.2 n fully correlated measurements . . . 47
A.3.3 Two fully correlated measurements with an uncorrelated measurement . . . 48
A.4 Choice of the inverse . . . 48

B Varying the biases in the presence of theoretical correlations **49**
B.1 Range of variations for the biases 49
B.2 Averaging measurements with theoretical correlations 50

C Definition of the test statistic in n dimensions **51**
C.1 Ambiguities in the definition of a 100% theoretical correlation 51
C.2 Reducing the problem to one bias parameter per observable 52

D Asymmetric uncertainties **53**

E 3-σ intervals for CKM-related examples **53**

In particle physics, an important part of the data analysis is devoted to the interpretation of the data with respect to the Standard Model (SM) or some of its extensions, with the aim of comparing different alternative models or determining the fundamental parameters of a given underlying theory [1–3]. In this activity, the role played by uncertainties is essential, since they constitute the limit for the accurate determination of these parameters, and they can prevent from reaching a definite conclusion when comparing several alternative models. In some cases, these uncertainties are from a statistical origin: they are related to the intrinsic variability of the phenomena observed, they decrease as the sample size increases and they can be modeled using random variables. A large part of the experimental uncertainties belong to this first category. However, another kind of uncertainties occurs when one wants to describe inherent limitations of the analysis process, for instance, uncertainties in the calibration or limits of the models used in the analysis. These uncertainties are very often encountered in theoretical computations, for instance when assessing the size of higher orders in perturbation theory or the validity of extrapolation formulae. Such uncertainties are often called "systematics", but they should be distinguished from less dangerous sources of systematic uncertainties, usually of experimental origin, that roughly scale with the size of the statistical sample and may be reasonably modeled by random variables [4]. In the following we will thus call them "theoretical" uncertainties: by construction, they lack both an unambiguous definition (leading to various recipes to determine these uncertainties) and a clear interpretation (beyond the fact that they are not from a statistical origin). It is thus a complicated issue to incorporate their effect properly, even in simple situations often encountered in particle physics [5–7] [1].

The relative importance of statistical and theoretical uncertainties might be different depending on the problem considered, and the progress made both by experimentalists and theorists. For instance, statistical uncertainties are the main issue in the analysis of electroweak precision observables [11, 12]. On the other hand, in the field of quark flavour physics, theoretical uncertainties play a very important role. Thanks to the B-factories and LHCb, many hadronic processes have been very accurately measured [13, 14], which can provide stringent constraints on the Cabibbo-Kobayashi-Maskawa matrix (in the Standard Model) [15–17], and on the scale and structure of New Physics (in SM extensions) [18–21]. However, the translation between hadronic processes and quark-level transitions requires information on hadronisation from strong interaction, encoded in decay constants, form factors, bag parameters... The latter are determined through lattice QCD simulations. The remarkable progress in computing power and in algorithms over the last 20 years has led to a decrease of statistical uncertainties and a dominance of purely theoretical uncertainties (chiral and heavy-quark extrapolations, scale chosen to set the lattice spacing, finite-volume effects, continuum limit...). As an illustration, the determination of the Wolfenstein parameters of the CKM matrix involves many constraints which are now limited by theoretical uncertainties (neutral-meson mixing, leptonic and

[1] The issue of theoretical uncertainties is naturally not the only question that arises in the context of statistical analyses. The statistical framework used to perform these analyses is also a matter of choice, with two main approaches, frequentist and Bayesian, adopted in different settings and for various problems in and beyond high-energy physics [1–3,8–10]. In this paper, we choose to focus on the frequentist approach to discuss how to model theoretical uncertainties.

semileptonic decays...) [22].

The purpose of this note is to discuss theoretical uncertainties in more detail in the context of particle physics phenomenology, comparing different models not only from a statistical point of view, but also in relation with the problems encountered in phenomenological analyses where they play a significant role. In Sec. 1, we summarise fundamental notions of statistics used in particle physics, in particular p-values and test statistics. In Sec. 2, we list properties that we seek in a good approach for theoretical uncertainties. In Sec. 3, we propose several approaches and in Sec. 4, we compare their properties in the most simple one-dimensional case. In Sec. 5, we consider multi-dimensional cases (propagation of theoretical uncertainties, average of several measurements, fits and pulls), which we illustrate using flavour physics examples related to the determination of the CKM matrix in Sec. 6, before concluding. An appendix is devoted to several issues connected with the treatment of correlations.

1 Statistics concepts for particle physics

We start by briefly recalling frequentist concepts used in particle physics, highlighting the role played by p-values in hypothesis testing and how they can be used to define confidence intervals.

1.1 p-values

1.1.1 Data fitting and data reduction

First, we would like to illustrate the concepts of data fitting and data reduction in particle physics, starting with a specific example, namely the observation of the time-dependent CP asymmetry in the decay channel $B^0(t) \to J/\psi K_S$ by the BaBar, Belle and LHCb experiments [23–25]. Each experiment collects a sample of observed decay times t_i corresponding to the B-meson events, where this sample is theoretically known to follow a PDF f. The PDF is parameterized in terms of a few physics parameters, among which we assume the ones of interest are the direct and mixing-induced C and S CP asymmetries. The functional form of this PDF is dictated on very general grounds by the CPT invariance and the formalism of two-state mixing (see, *e.g.*, [26]), and is independent of the particular underlying phenomenological model (*e.g.* the Standard Model of particle physics). In practice however, detector effects require to be modelled by additional parameters that modify the shape of the PDF. We denote by θ the set of parameters $\theta = (C, S, \ldots)$ that are needed to specify the PDF completely. The likelihood for the sample $\{t_i\}$ is defined by

$$\mathcal{L}_{\{t_i\}}(\theta) = \prod_{i=1}^{n} f(t_i; \theta) \tag{1}$$

and can be used as a test statistic to infer constraints on the parameters θ, and/or construct estimators for them, as will be discussed in more detail below. The combination of different samples/experiments can be done simply by multiplication of the corresponding likelihoods. On the other hand one can choose to work directly in the framework of a specific phenomenological model, by replacing in θ the quantities that are predicted by the model in terms of more fundamental parameters: for example in the Standard Model, and neglecting the "penguin" contributions, one has the famous relations $C = 0$, $S = \sin 2\beta$ where β is one of the angles of the Unitarity Triangle and can be further expressed in terms of the Cabibbo-Kobayashi-Maskawa couplings.

The latter choice of expressing the experimental likelihood in terms of model-dependent parameters such as β has however one technical drawback: the full statistical analysis has to be performed for each model one wants to investigate, e.g., the Standard Model, the Minimal Supersymmetric Standard Model, GUT models... In addition, building a statistical analysis directly on the initial likelihood requires one to deal with a very large parameter space, depending on the parameters

in θ that are needed to describe the detector response. One common solution to these technical difficulties is a two-step approach. In the first step, the data are *reduced* to a set of model- and detector-independent [2] random variables that contains the same information as the original likelihood (to a good approximation): in our example the likelihood-based estimators $\hat{C}$ and $\hat{S}$ of the parameters C and S can play the role of such variables (estimators are functions of the data and thus are random variables). In a second step, one can work in a particular model, e.g., in the Standard Model, to use $\hat{C}$ and $\hat{S}$ as inputs to a statistical analysis of the parameter β. This two-step procedure gives the same result as if the analysis were done in a single step through the expression of the original likelihood in terms of β. This technique is usually chosen if the PDF g of the estimators $\hat{C}$ and $\hat{S}$ can be parameterized in a simple way: for example, if the sample size is sufficiently large, then the PDF can often be modelled by a multivariate normal distribution, where the covariance matrix is approximately independent of the mean vector.

Let us now extend the above discussion to a more general case. A sample of random events is $\{E_i, i = 1 \dots n\}$, where each event corresponds to a set of directly measurable quantities (particle energies and momenta, interaction vertices, decay times...). The distribution of these events is described by a PDF, the functional form f of which is supposed to be known. In addition to the event value E, the PDF value depends on some fixed parameters θ, hence the notation $f(E;\theta)$. The likelihood for the sample $\{E_i\}$ is defined by $\mathcal{L}_{\{E_i\}}(\theta) = \prod_{i=1}^{n} f(E_i;\theta)$. We want to interpret the event observation in a given phenomenological scenario that predicts at least some of the parameters θ describing the PDF in terms of a set of more fundamental parameters χ.

To this aim we first reduce the event observation to a set of model- and detector-independent random variables X together with a PDF $g(X;\chi)$, in such a way that the information that one can get on χ from g is equivalent to the information one can get from f, once θ is expressed in terms of χ consistently with the phenomenological model of interest. Technically, it amounts to identifying a minimal set of variables x depending on θ that are independent of both the experimental context and the phenomenological model. One performs an analysis on the sample of events E_i to derive estimators $\hat{x}$ for x. The distribution of these estimators can be described in terms of a PDF that is written in the χ parametrization as $g(X;\chi)$, where we have replaced $\hat{x}$ by the notation X, to stress that in the following X will be considered as a new random variable, setting aside how it has been constructed from the original data $\{E_i\}$. Obviously, in our previous example for $B^0(t) \to J/\psi K_S$, $\{t_i\}$ correspond to $\{E_i\}$, C and S to x, and β to χ.

1.1.2 Model fitting

From now on we work with one or more observable(s) x, with associated random variable X, and an associated PDF $g(X;\chi)$ depending on purely theoretical parameters χ. With a slight abuse of notation we include in the symbol g not only the functional form, but also all the needed parameters that are kept fixed and independent of χ. In particular for a one-dimensional Gaussian PDF we have

$$g(X;\chi) \sim \exp\left[-\frac{1}{2}\left(\frac{X - x(\chi)}{\sigma}\right)^2\right] \tag{2}$$

where X is a potential value of the observable x and $x(\chi)$ corresponds to the theoretical prediction of x given χ. This PDF is obtained from the outcome of an experimental analysis yielding both a central value X_0 and an uncertainty σ, where σ is assumed to be independent of the realisation X_0 of the observable x and is thus included in the definition of g.

[2] It may happen that the detector and/or background effects have a sizeable impact on the fitted quantities $\hat{C}$ and $\hat{S}$; this can be viewed as uncertainties in the modelling of the event PDF f. These effects are reported as *systematic uncertainties* and in particle physics, it is customary to treat them on the same footing as the pure statistical uncertainties. Although we will not try to follow this avenue in the examples discussed here, it would be possible to consider these systematic uncertainties as theoretical uncertainties, to be modelled according to the methods that we describe in the following sections.

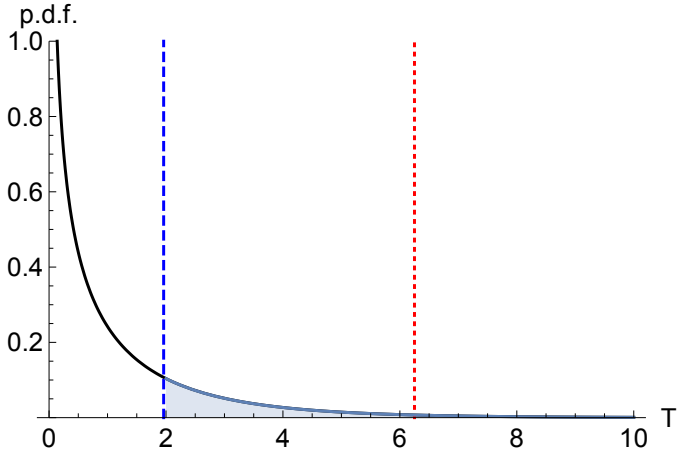


Figure 1: Illustration in the simple case where X is predicted as $x(\mu) = \mu$. Under the hypothesis $\mu_t = \mu$, and having measured $X = 0 \pm 1$, one can determine the associated p-value $p(0; \mu)$ by examining the distribution of the quadratic test statistic $T(X; \mu) = (X - \mu)^2$ assuming X is distributed as a Gaussian random variable with central value 0 and width 1. The blue dashed line corresponds to the value of T associated with the hypothesis $\mu = -1.4$, with a p-value obtained by considering the gray area. The red dotted line corresponds to the hypothesis $\mu = 2.5$.

Our aim is to derive constraints on the parameters χ, from the measurement $X_0 \pm \sigma$ of the observable x. One very general way to perform this task is *hypothesis testing*, where one wants to quantify how much the data are compatible with the null hypothesis that the true value of χ, χ_t, is equal to some fixed value χ:

$$\mathcal{H}_\chi : \chi_t = \chi \tag{3}$$

In order to interpret the observed data X_0 measured in a given experiment in light of the distribution of the observables X under the null hypothesis $\mathcal{H}_\chi$, one defines a *test statistic* $T(X; \chi)$, that is a scalar function of the data X that measures whether the data are in favour or not of the null hypothesis. We indicated the dependence of T on χ explicitly, i.e., the dependence on the null hypothesis $\mathcal{H}_\chi$. The test statistic is generally a definite positive function chosen in a way that large values indicate that the data present evidence against the null hypothesis. By comparing the actual data value $t = T(X_0; \chi)$ with the sampling distribution of $T = T(X; \chi)$ under the null hypothesis, one is able to quantify the degree of agreement of the data with the null hypothesis.

Mathematically it amounts to defining a *p-value*. One calculates the probability to obtain a value for the test statistic at least as large as the one that was actually observed, assuming that the null hypothesis is true. This tail probability is used to define the p-value of the test for this particular observation

$$1 - p(X_0; \chi) = \int_0^{T(X_0;\chi)} dT\, h(T|\mathcal{H}_\chi) = \mathcal{P}[T < T(X_0; \chi)] \tag{4}$$

where the PDF h of the test statistic is obtained from the PDF g of the data as

$$h(T|\mathcal{H}_\chi) = \int dX\, \delta\left[T - T(X; \chi)\right] g(X; \chi) \tag{5}$$

which can be obtained easily from comparing the convolution of $\frac{dT}{dX} h(T) = g(X)$ with a test function of T with the convolution of the r.h.s. of (5) with the same test function. A small value of the p-value means that $T(X_0; \chi)$ belongs to the "large" region, and thus provides evidence against the null hypothesis. This is illustrated for a simple example in Figs. 1 and 2.

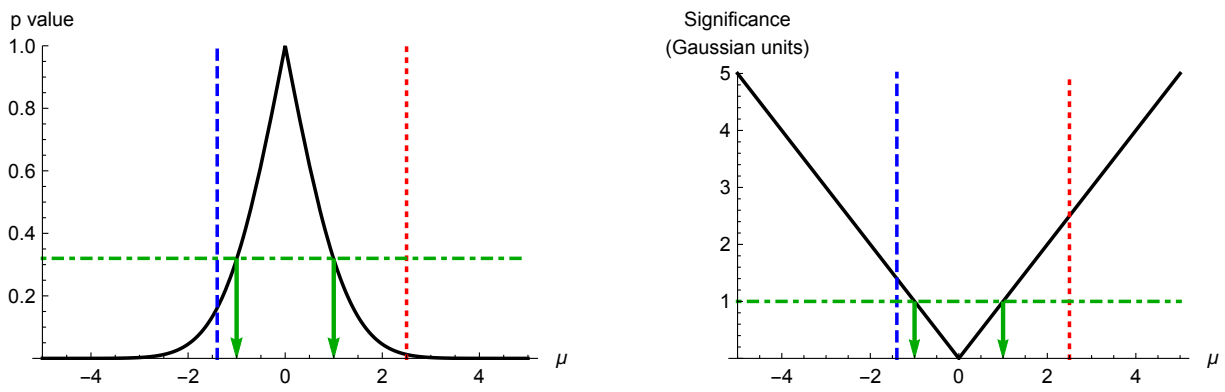


Figure 2: On the left: for a given observation $X = 0 \pm 1$, p-value $p(0;\mu)$ as a function of the value of μ being tested. Blue dashed and red dotted lines correspond to $\mu = -1.4$ and $\mu = 2.5$. A confidence interval for μ at 68% CL is obtained by considering the region of μ with a p-value larger than 0.32, as indicated by the green dotted dashed line and arrows. On the right: the same information is expressed in Gaussian units of σ, where the 68% CL interval corresponds to the region below the horizontal line of significance 1.

From its definition, one sees that $1 - p(X_0;\chi)$ is nothing else but the cumulative distribution function of the PDF h

$$\mathrm{CDF}[h](T(X_0;\chi)|\mathcal{H}_\chi) = \int dX\, \theta\left[T(X_0;\chi) - T(X;\chi)\right] g(X;\chi) \tag{6}$$

where θ is the Heaviside function. This expression corresponds to the probability for the test statistic to be smaller than a given value $T(X_0;\chi)$. The p-value in Eq. (4) is defined as a function of X_0 and as such, is a random variable.

Through the simple change of variable $\frac{dp}{dT}\frac{d\mathcal{P}}{dp} = \frac{d\mathcal{P}}{dT}$, one obtains that *the null distribution* (that is, the distribution when the null hypothesis is true) *of a p-value is uniform*, i.e., the distribution of values of the p-value is flat between 0 and 1. This uniformity is a fundamental property of p-values that is at the core of their various interpretations (hypothesis comparison, determination of confidence intervals...) [1,2].

In the frequentist approach, one wants to design a procedure to decide whether to accept or reject the null hypothesis $\mathcal{H}_\chi$, by avoiding as much as possible either incorrectly rejecting the null hypothesis (Type-I error) or incorrectly accepting it (Type-II error). The standard frequentist procedure consists in selecting a Type-I error α and determining a region of sample space that has the probability α of containing the data under the null hypothesis. If the data fall in this critical region, the hypothesis is rejected. This must be performed before data are known (in contrast to other interpretations, e.g, Fischer's approach of significance testing [1]). In the simplest case, the critical region is defined by a condition of the form $T \geq t_\alpha$, where t_α is a function of α only, which can be rephrased in terms of p-value as $p \leq \alpha$. The interest of the frequentist approach depends therefore on the ability to design p-values assessing the rate of Type-I error correctly (its understatement is clearly not desirable, but its overstatement yields often a reduction in the ability to determine the truth of an alternative hypothesis), as well as avoiding too large a Type-II error rate.

A major difficulty arises when the hypothesis to be tested is *composite*. In the case of numerical hypotheses like (3), one gets compositeness when one is only interested in a subset μ of the parameters χ. The remaining parameters are called *nuisance parameters* [3] and will be denoted by ν, thus

[3] "Nuisance" does not mean that these parameters are necessarily unphysical, "pollution" parameters. They can be fundamental constants of Nature, and interesting as such.

$\chi = (\mu, \nu)$. In this case the hypothesis $\mathcal{H}_\mu : \mu_t = \mu$ is composite, because determining the distribution of the observables requires the knowledge of the true value ν_t in addition to μ. In this situation, one has to devise a procedure to infer a "p-value" for $\mathcal{H}_\mu$ out of p-values built for the simple hypotheses where both μ and ν are fixed. Therefore, in contrast to a simple hypothesis, a composite hypothesis does not allow one to compute the distribution of the data [4].

At this stage, it is not necessarily guaranteed that the distribution of the p-value for $\mathcal{H}_\mu$ is uniform, and one may get different situations:

$$p \text{ exact} \quad : \quad P(p \leq \alpha | \mathcal{H}_\mu) = \alpha \tag{7}$$

$$p \text{ conservative} \quad : \quad P(p \leq \alpha | \mathcal{H}_\mu) < \alpha \tag{8}$$

$$p \text{ liberal} \quad : \quad P(p \leq \alpha | \mathcal{H}_\mu) > \alpha \tag{9}$$

which may depend on the value of α considered. Naturally, one would like to design as much as possible an exact p-value (exact coverage), or if this is not possible, a (reasonably) conservative one (overcoverage). Such p-values will be called "valid" p-values. In the case of composite hypotheses, the conservative or liberal nature of a p-value may depend not only on α, but also on the structure of the problem and of the procedure used to construct the p-value, and it has to be checked explicitly [1,2].

Once p-values are defined, one can build confidence intervals out of them by using the correspondence between acceptance regions of tests and confidence sets. Indeed, if we have an exact p-value, and the critical region $C_\alpha(X)$ is defined as the region where $p(X;\mu) < \alpha$, the complement of this region turns out to be a confidence set of level $1-\alpha$, i.e., $P[\mu \notin C_\alpha(X)] = 1-\alpha$. This justifies the general use of plotting the p-value as a function of μ, and reading the 68% or 95% CL intervals by looking at the ranges where the p-value curve is above 0.32 or 0.05. This is illustrated for a simple example in Figs. 2 and 3. Once again, this discussion is affected by issues of compositeness and nuisance parameters, as well as the requirement of checking the coverage of the p-value used to define these confidence intervals: an overcovering p-value will yield too large confidence intervals, which will prove indeed conservative.

A few words about the notation and the vocabulary are in order at this stage. A p-value necessarily refers to a null hypothesis, and when the null hypothesis is purely numerical such as (3) we can consider the p-value as a mathematical function of the fundamental parameter μ. This of course does not imply that μ is a random variable (in frequentist statistics, it is always a fixed, but unknown, number). When the p-value as a function of μ can be described in a simple way by a few parameters, we will often use the notation $\mu = \mu_0 \pm \sigma_\mu$. In this case, one can easily build the p-value and derive any desired confidence interval. Even though this notation is similar to the measurement of an observable, we stress that this does not mean that the fundamental parameter μ is a random variable, and it should not be seen as the definition of a PDF. In line with this discussion, we will call *uncertainties* the parameters like σ that can be given a frequentist meaning, e.g., they can be used to define the PDF of a random variable. On the other hand, we will call *errors* the intermediate quantities such as σ_μ that can be used to describe the p-value of a fundamental parameter, but cannot be given a statistical meaning for this parameter.

[4] We have defined compositeness for numerical hypotheses, since this is our case of interest in the following. More generally, compositeness also occurs in the case of non-numerical hypotheses such as "The Standard Model is true", for which it is not possible to compute the distribution of data either. Indeed assuming that the Standard Model is true does not imply anything on the value of its fundamental parameters, and thus one cannot compute the distribution of a given observable under this hypothesis.

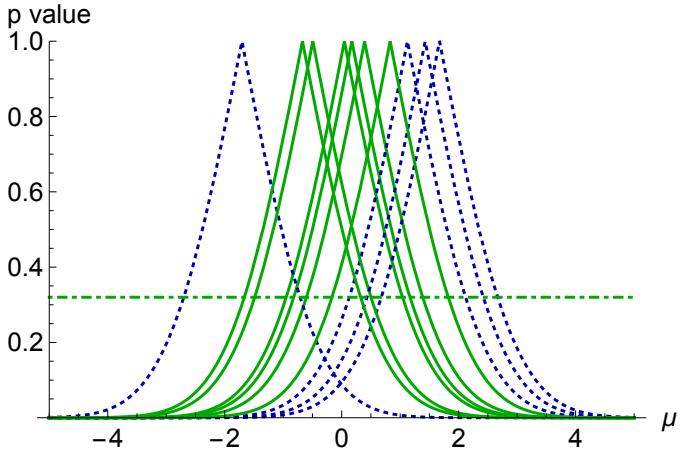


Figure 3: A α-CL interval built from a p-value with exact coverage has a probability of α of containing the true value. This is illustrated in the simple case of a quantity X which has a true value $\mu_t = 0$ but is measured with an uncertainty $\sigma = 1$. Each time a measurement is performed, it will yield a different value for X_0 and thus a different p-value curve as a function of the hypothesis tested $\mu_t = \mu$. From each measurement, a 68% CL interval can be determined by considering the part of the curve above the line $p = 0.32$, but this interval may or may not contain the true value $\mu_t = 0$. The curves corresponding to the first case (second case) are indicated with 6 green solid lines (4 blue dotted lines). Asymptotically, if the p-value has exact coverage, 68% of these confidence intervals will contain the true value.

1.2 Likelihood-ratio test statistic

Here we consider test statistics that are constructed from the logarithm of the likelihood [5]

$$T \sim -2 \ln \mathcal{L}_X(\chi) \qquad \mathcal{L}_X(\chi) \equiv g(X;\chi) \tag{10}$$

More precisely, one uses tests based on the likelihood ratio in many instances. Its use is justified by the Neyman-Pearson lemma [1, 2, 27] showing that this test has appealing features in a binary model with only two alternatives for χ_t, corresponding to the two simple hypotheses $\mathcal{H}_{\chi_1}$ and $\mathcal{H}_{\chi_2}$. Indeed one can introduce the likelihood ratio $\mathcal{L}_X(\chi_1)/\mathcal{L}_X(\chi_2)$, define the critical region where this likelihood ratio is smaller than a given α, and decide that one rejects $\mathcal{H}_{\chi_1}$ whenever the observation falls in this critical region. This test is the most powerful test that can be built [1, 2], in the sense that among all the tests with a given Type-I error α (probability of rejecting $\mathcal{H}_{\chi_1}$ when $\mathcal{H}_{\chi_1}$ is true), the likelihood ratio test has the smallest Type-II error (probability of accepting $\mathcal{H}_{\chi_1}$ when $\mathcal{H}_{\chi_2}$ is true). These two conditions are the two main criteria to determine the performance of a test.

In the case of a composite hypothesis, there is no such clear-cut approach to choose the most powerful test. The Maximum Likelihood Ratio (MLR) is inspired by the Neynman-Pearson lemma, comparing the most plausible configuration under $\mathcal{H}_\mu$ with the most plausible one in general:

$$\begin{aligned} T(X;\mu) &= -2 \ln \frac{\mathrm{Max}_{\nu_t} \mathcal{L}_X(\mu, \nu_t)}{\mathrm{Max}_{\mu_t,\nu_t} \mathcal{L}_X(\mu_t, \nu_t)} \\ &= \mathrm{Min}_{\nu_t}[-2 \ln \mathcal{L}_X(\mu, \nu_t)] - \mathrm{Min}_{\mu_t,\nu_t}[-2 \ln \mathcal{L}_X(\mu_t, \nu_t)] \end{aligned} \tag{11}$$

Let us emphasise that even though T is constructed not to depend on the nuisance parameters ν explicitly, its distribution Eq. (5) *a priori* depends on them (through the PDF g). Even though the

[5] Strictly speaking, the likelihood is only defined for the actually measured data X_0: $\mathcal{L}_0(\chi) \equiv g(X_0;\chi)$ and thus is only a function of the parameters χ. Nevertheless it is common practice to use the word "likelihood" for the object $g(X;\chi)$, considered as a function of both the observables X and the parameters χ.

Neyman-Pearson lemma does not apply here, there is empirical evidence that this test is powerful, and in some cases it exhibits good asymptotic properties (easy computation and distribution independent of nuisance parameters) [1, 2].

For the problems considered here, the MLR choice features alluring properties, and in the following we will use test statistics that are derived from this choice. First, if $g(X;\chi_t)$ is a multidimensional Gaussian function, then the quantity $-2\ln\mathcal{L}_X(\chi_t)$ is the sum of the squares of standard normal random variables, *i.e.,* is distributed as a χ^2 with a number of degrees of freedom (N_{dof}) that is given by $\dim(X)$. Secondly, for linear models, in which the observables X depend linearly on the parameters χ_t, the MLR Eq. (11) is again a sum of standard normal random variables, and is distributed as a χ^2 with $N_{dof} = \text{dimension}(\mu)$. Wilks' theorem [28] states that this property can be extended to non-Gaussian cases in the asymptotic limit: under regularity conditions and when the sample size tends to infinity, the distribution of Eq. (11) will converge to the same χ^2 distribution depending only on the number of parameters tested.

The great virtue of the χ^2-distribution is that it only depends on the number of degrees of freedom, which means in particular that the null-distribution of Eq. (11) is independent of the nuisance parameters ν, whenever the conditions of the Wilks' theorem apply. Furthermore the integral (4) can be computed straightforwardly in terms of complete and incomplete Γ functions:

$$\begin{aligned} p(X_0;\mu) &= \text{Prob}\left(T(X_0;\mu), N_{dof}\right) \\ &\equiv \frac{\Gamma(N_{dof}/2, T(X_0;\mu)/2)}{\Gamma(N_{dof}/2)} \end{aligned} \tag{12}$$

In practice the models we want to analyse, such as the Standard Model, predict non linear relations between the observables and the parameters. In this case one has to check whether Wilks' theorem applies, by considering whether the theoretical equations can be approximately linearized [6].

2 Comparing approaches to theoretical uncertainties

We have argued before that an appealing test statistic is provided by the likelihood ratio Eq. (11) due to its properties in limit cases (linearised theory, asymptotic limit). These properties rely on the fact that the likelihood ratio can be built as a function of random variables described by measurements involving only statistical uncertainties. However, in flavour physics (as in many other fields in particle physics), there are not only statistical but also theoretical uncertainties. Indeed, as already indicated in the introduction, these phenomenological analyses combine experimental information and theoretical estimates. In the case of flavour physics, the latter come mainly from QCD-based calculations, which are dominated by theoretical uncertainties.

Unfortunately, the very notion of theoretical uncertainty is ill-defined as "anything that is not due to the intrinsic variability of data". Theoretical uncertainties (model uncertainty) are thus of a different nature with respect to statistical uncertainties (stochastic uncertainty, i.e. variability in the data), but they can only be *modelled* (except in the somewhat academic case where a bound on the difference between the exact value and the approximately computed one can be proven). The choice of a model for theoretical uncertainties involves not only the study of its mathematical properties and its physical implications in specific cases, but also some personal taste. One can indeed imagine several ways of modelling/treating theoretical uncertainties:

- one can (contrarily to what has just been said) treat the theoretical uncertainty on the same footing as a statistical uncertainty; in this case, in order to follow a meaningful frequentist procedure, one has to assume that one lives in a world where the repeated calculation of a

[6] More precisely, the asymptotic limit is reached when the model can be linearized for all values of the data that contribute significantly to the integral (4). It corresponds to the situation where the errors on the parameters derived from computing p-values are small with respect to the typical parameter scales of the problem.

given quantity leads to a distribution of values around the exact one, with some variability that can be modelled as a PDF ("random-δ approach"),

- one can consider that theoretical uncertainties can be modelled as external parameters, and perform a purely statistical analysis for each point in the theoretical uncertainty parameter space; this leads to an infinite collection of p-values that will have to be combined in some arbitrary way, following a model averaging procedure ("external-δ approach"),
- one can take the theoretical uncertainties as fixed asymptotic biases [7], treating them as nuisance parameters that have to be varied in a reasonable region ("nuisance-δ approach").

There are some desirable properties for a convincing treatment of theoretical uncertainties:

- as general as possible, i.e., apply to as many "kinds" of theoretical uncertainties as possible (lattice uncertainties, scale uncertainties) and as many types of physical models as possible,
- leading to meaningful confidence intervals, in reasonable limit cases: obviously, in the absence of theoretical uncertainties, one must recover the standard result; one may also consider the type of constraint obtained in the absence of statistical uncertainties,
- exhibiting good coverage properties, as it benchmarks the quality of the statistical approach: the comparison of different models provides interesting information but does not shed light on their respective coverage,
- associated with a statistically meaningful goodness-of-fit,
- featuring reasonable asymptotic properties (large samples),
- yielding the errors as a function of the estimates easily (error propagation), in particular by disentangling the impact of theoretical and statistical contributions,
- leading to a reasonable procedure to average independent estimates – if possible, it should be equivalent for any analysis to include the independent estimates separately or the average alone (associativity). In addition, one may wonder whether the averaging procedure should be conservative or aggressive (i.e., the average of similar theoretical uncertainties should have a smaller uncertainty or not), and if the procedure should be stationary (the uncertainty of an average should be independent of the central values or not),
- leading to reasonable results in the case of averages of inconsistent measurements.

Finally a technical requirement is the computing power needed to calculate the best fit point and confidence intervals for a large parameter space with a large number of constraints. Even though it should not be the sole argument in favour of a model, it should be kept in mind (a very complicated model for theoretical uncertainties would not be particularly interesting if it yields very close results to a much simpler one).

We summarize some of the points mentioned above in Tab. 1. As it will be seen, it will however prove challenging to fulfill all these criteria at the same time, and we will have to make compromises along the way.

[7] A bias is defined as the difference between the average of the estimator among a large number of experiments with finite sample size and the true value. An estimator is said to be consistent if it converges to the true value when the size of the sample tends to infinity (e.g., maximum likelihood estimators). Consistency implies that the bias vanishes asymptotically, while inconsistency may stem from theoretical uncertainties.

3 Illustration of the approaches in the one-dimensional case

3.1 Situation of the problem

We will now discuss the three different approaches and some of their properties in the simplest case, i.e. with a single measurement (for an experimental quantity) or a single theoretical determination (for a theoretical quantity). Following a fairly conventional abuse of language, we will always refer to this piece of information as a "measurement" even though some modelling may be involved in its extraction through data reduction, as discussed in Sec. 1. The main, yet not alone, aim is to model/interpret/exploit a measurement like [8]

$$X = X_0 \pm \sigma \text{ (exp)} \pm \Delta(\text{th}) \tag{13}$$

to extract information on the value of the associated fundamental parameter μ. Without theoretical uncertainty ($\Delta = 0$), one would use this measurement to build a PDF

$$\text{PDF}_{\text{no th}}(X;\mu) = \mathcal{N}_{(\mu,\sigma)}(X) \tag{14}$$

yielding the MLR test statistic

$$T_{\text{no th}} = \frac{(X-\mu)^2}{\sigma^2} \tag{15}$$

and one can build a p-value easily from Eq. (4)

$$p_{\text{no th}}(\mu) = 1 - \text{Erf}\left[\frac{|\mu - X_0|}{\sqrt{2}\sigma}\right] \tag{16}$$

In the presence of a theoretical uncertainty Δ, the situation is more complicated, as there is no clear definition of what Δ corresponds to. A possible first step is to introduce a theoretical uncertainty parameter δ that describes the shift of the approximate theoretical computation from the exact value, and that is taken to vary in a region that is defined by the value of Δ. This leads to the PDF

$$\text{PDF}(X;\mu) = \mathcal{N}_{(\mu+\delta,\sigma)}(X) \tag{17}$$

in such a way that in the limit of an infinite sample size ($\sigma \to 0$), the measured value of X reduces to $\mu + \delta$. The challenge is to extract some information on μ, given the fact that the value of δ remains unknown.

The steps (to be spelt out below) to achieve this goal are:

- Take a model corresponding to the interpretation of δ: *random variable, external parameter, fixed bias as a nuisance parameter...*
- Choose a test statistic $T(X;\mu)$ that is consistent with the model and that discriminates the null hypothesis: *Rfit, quadratic, other ...*
- Compute, consistently with the model, the p-value that is in general a function of μ and δ
- Eliminate the dependence with respect to δ by some well-defined procedure
- Exploit the resulting p-value (coverage, confidence intervals, goodness-of-fit)

Since we focus on Gaussian experimental uncertainties (the generalization to other shapes is formally straightforward but may be technically more complicated), for all approaches that we discuss in this note we take the following PDF

$$\text{PDF}(X;\mu) = \mathcal{N}_{(\mu+\delta,\sigma)}(X) \tag{18}$$

where, in the limit of an infinite sample size ($\sigma \to 0$), μ can be interpreted as the exact value of the parameter of interest, and $\mu + \delta$ the approximately theoretically computed one. The interpretation of δ will differ depending on the approach considered, which we will discuss now.

[8] We discuss how the method can be adapted for asymmetric uncertainties in App. D.

Approach	Random-δ	Nuisance-δ	External-δ
Hypothesis	Random var.	Composite hyp.	Family of simple hyp.
	$\mathrm{PDF}_\Delta(\delta)$	$\mathcal{H}_\mu : \mu_t = \mu$	$\mathcal{H}_\mu^{(\delta)} : \mu_t = \mu + \delta$
Test	Likelihood ratio	Quadratic	Quadratic
Constraint on δ	–	Ω	Ω
Associativity	Yes if normal PDF	Yes if Ω hyperball	Yes if Ω hyperball
Splitting of errors	Yes if normal PDF	Yes for all Ω	Yes for all Ω
Stationarity	Yes	Yes	Yes
Simple asympt. lim.	Yes if normal PDF	Yes	Yes
Simple $\sigma \to 0$ limit	Depends on PDF	Ω	Ω
Particular cases	naive Gaussian	Fixed/adaptive nuis.	Scan
if we take	normal PDF	Fixed/adaptive Ω	Sup over fixed Ω

Table 1: Summary table of various approaches to theoretical uncertainties considered in the text.

3.2 The random-δ approach

In the random-δ approach, δ would be related to the variability of theoretical computations, that one can model with some PDF for δ, such as $\mathcal{N}_{(0,\Delta)}$ (normal) or $\mathcal{U}_{(-\Delta,+\Delta)}$ (uniform). The natural candidate for the test statistic $T(X;\mu)$ is the MLR built from the PDF. One considers a model where $X = s + \delta$ is the sum of two random variables, s being distributed as a Gaussian of mean μ and width σ, and δ as an additional random variable with a distribution depending on Δ.

One may often consider for δ a variable normally distributed with a mean zero and a width Δ (denoted naive Gaussian or "nG" in the following, corresponding to the most common procedure in the literature of particle physics phenomenology). The resulting PDF for X is then the convolution of two Gaussian PDFs, leading to

$$\mathrm{PDF}_{\mathrm{nG}}(X;\mu) = \mathcal{N}_{(\mu,\sqrt{\sigma^2+\Delta^2})}(X) \tag{19}$$

to which corresponds the usual quadratic test statistic (obtained from MLR)

$$T_{\mathrm{nG}} = \frac{(X-\mu)^2}{\sigma^2+\Delta^2} \tag{20}$$

recovering the p value that would be obtained when the two uncertainties are added in quadrature

$$p_{\mathrm{nG}}(\mu) = 1 - \mathrm{Erf}\left[\frac{|\mu - X|}{\sqrt{2}\sqrt{\sigma^2+\Delta^2}}\right] \tag{21}$$

We should stress that considering δ as a random variable corresponds to a rather strange frequentist world [9], and there is no strong argument that would help to choose the associated PDF (for instance, δ could be a variable uniformly distributed over $[-\Delta, \Delta]$). However for a general PDF, the p-value has no simple analytic formula and it must be computed numerically from Eq. (4). In the following, we will only consider the case of a Gaussian PDF when we discuss the random-δ approach.

3.3 The nuisance-δ approach

In the nuisance approach, δ is not interpreted as a random variable but as a fixed parameter so that in the limit of an infinite sample size, the estimator does not converge to the true value μ_t, but to $\mu_t + \delta$. The distinction between statistical and theoretical uncertainties is thus related to their effect as the sample size increases, statistical uncertainties decreasing while theoretical uncertainties

[9] On the other hand, this is natural in the Bayesian approach, where incomplete information is modelled as a PDF for the unknown parameters associated with a theoretical uncertainty [8–10].

remaining of the same size (see Refs. [29–31] for other illustrations in the context of particle physics). One works with the null hypothesis $\mathcal{H}_\mu : \mu_t = \mu$, and one has then to determine which test statistic is to be built.

In the frequentist approach, the choice of the test statistic is arbitrary as long as it models the null hypothesis correctly, i.e., the smaller the value of the test statistic, the better the agreement of the data with the hypothesis. A particularly simple possibility consists in the quadratic statistic already introduced earlier:

$$T_{\text{nuisance}} = \text{Min}_\delta \left[\left(\frac{X - \mu - \delta}{\sigma} \right)^2 + \left(\frac{\delta}{\Delta} \right)^2 \right] = \frac{(X-\mu)^2}{\sigma^2 + \Delta^2} \tag{22}$$

where the minimum is not taken over a fixed range, but on the whole space. The great virtue of the quadratic shape is that in linear models it remains quadratic after minimization over any subset of parameters, in contrast with alternative, non-quadratic, test statistics.

The PDF for X is normal, with mean $\mu + \delta$ and variance σ^2

$$\text{PDF}_{\text{nuisance}}(X; \mu) = \mathcal{N}_{(\mu+\delta,\sigma)}(X) \tag{23}$$

Although we choose test statistics for the random-δ and nuisance-δ of the same form, Eqs. (20) and (22), the different PDFs Eqs. (19) and (23) imply very different constructions for the p-values and the resulting statistical outcomes. Indeed, with this PDF for the nuisance-δ approach, T is distributed as a rescaled, non-central χ^2 distribution with a non-centrality parameter $(\delta/\sigma)^2$ (this non-centrality parameter illustrates that the test statistic is centered around μ whereas the distribution of X is centered around $\mu + \delta$). δ is then a genuine asymptotic bias, implying inconsistency: in the limit of an infinite sample size, the estimator constructed from T is μ, whereas the true value is $\mu + \delta$. Using the previous expressions, one can easily compute the cumulative distribution function of this test statistic

$$1 - \text{CDF}_\delta(\mu) = 1 + \frac{1}{2}\text{Erf}\left(\frac{\delta - |\mu - X|}{\sqrt{2}\sigma} \right) - \frac{1}{2}\text{Erf}\left(\frac{\delta + |\mu - X|}{\sqrt{2}\sigma} \right) \tag{24}$$

which depends explicitly on δ but not on Δ (as indicated before, even if T is built to be independent of nuisance parameters, its PDF depends on them *a priori*).

To infer the p-value one can take the supremum value for δ over some interval Ω

$$p_\Omega = \text{Max}_{\delta \in \Omega}[1 - \text{CDF}_\delta(\mu)] \tag{25}$$

The interpretation is the following: if the (unknown) true value of δ belongs to Ω, then p_Ω is a valid p-value for μ, from which one can infer confidence intervals for μ. This space cannot be the whole space (as one would get $p = 1$ trivially for all values of μ), but there is no natural candidate (i.e., coming from the derivation of the test statistic). More specifically, should the interval Ω be kept fixed or should it be rescaled when investigating confidence intervals at different levels (*e.g.* 68% vs 95%)?

- If one wants to keep it fixed, $\Omega_r = r[-\Delta, \Delta]$:

$$p_{\text{fixed } \Omega_r} = \text{Max}_{\delta \in \Omega_r}[1 - \text{CDF}_\delta(\mu)] \tag{26}$$

  One may wonder what is the best choice for r, as the p-value gets very large if one works with the reasonable $r = 3$, while the choice $r = 1$ may appear as non-conservative. We will call this treatment the *fixed* r-nuisance approach.

- One can then wonder whether one would like to let Ω depend on the value considered for p. In other words, if we are looking at a $k\,\sigma$ range, we could consider the equivalent range for δ. This would correspond to

$$p_{\text{adapt } \Omega} = \text{Max}_{\delta \in \Omega_{k\sigma(p)}}[1 - \text{CDF}_\delta(\mu)] \tag{27}$$

where $k_\sigma(p)$ is the "number of sigma" corresponding to p

$$k_\sigma(p)^2 = \mathrm{Prob}^{-1}(p, N_{\mathrm{dof}} = 1) \tag{28}$$

where the function Prob has been defined in Eq. (12). We will call this treatment the *adaptive* nuisance approach. The correct interpretation of this p-value is: p is a valid p-value if the true (unknown) value of δ/Δ belongs to the "would be" $1-p$ confidence interval around 0. This is not a standard coverage criterion: one can use *adaptive coverage*, and *adaptively valid p-value*, to name this new concept. Note that Eqs. (27)-(28) constitute a non-algebraic implicit equation, that has to be solved by numerical means.

Let us emphasise that the fixed interval is very close to the original 'Rfit' method of the CKMfitter group [15, 16] in spirit, but not numerically, as will be shown below by an explicit comparison. In contrast the adaptive choice is more aggressive in the region of δ close to zero, but allows this parameter to take large values, provided one is interested in computing small p-values accordingly. In this sense, the adaptive approach provides a unified approach to deal with two different issues of importance, namely the metrology of parameters (at 1 or 2 σ) and exclusion tests (at 3 or 5 σ).

3.4 The external-δ approach

In this approach, the parameter δ is also considered as a fixed parameter. The idea behind this approach is very simple, and it is close to what experimentalists often do to estimate systematic effects: in a first step one considers that δ is a fixed constant, and one performs a standard, purely statistical analysis that leads to a p-value that explicitly depends on δ. If one takes $X \sim \mathcal{N}_{(\mu+\delta,\sigma)}$ and T quadratic [either $(X-\mu-\delta)^2/\sigma^2$ or $(X-\mu-\delta)^2/(\sigma^2+\Delta^2)$] [10]:

$$p_\delta(\mu) = 1 - \mathrm{Erf}\left[\frac{|X-\mu-\delta|}{\sqrt{2}\sigma}\right] . \tag{29}$$

Note that this procedure actually corresponds to the simple null hypothesis $\mathcal{H}_\mu^{(\delta)} : \mu_t = \mu + \delta$ instead of $\mathcal{H}_\mu$: $\mu_t = \mu$, hence one gets an infinite collection of p-values instead of a single one related to the aimed constraint on μ.

Since δ is unknown one has to define a procedure to average all the $p_\delta(\mu)$ obtained. The simplest possibility is to take the envelope (i.e., the maximum) of $p_\delta(\mu)$ for δ in a definite interval (*e.g.* $[-\Delta, +\Delta]$), leading to:

$$p_{\mathrm{nRfit}} = 1 \qquad \text{if } |X-\mu| \leq \Delta \tag{30}$$

$$= 1 - \mathrm{Erf}\left[\frac{|X-\mu\pm\Delta|}{\sqrt{2}\sigma}\right] \qquad \text{otherwise} \tag{31}$$

By analogy with the previous case, we will call this treatment the *fixed* r-external approach for $\delta \in \Omega_r$. This is equivalent to the Rfit ansatz used by CKMfitter [15, 16] in the one-dimensional case (but not in higher dimensions), proposed to treat theoretical uncertainties in a different way from statistical uncertainties, treating all values within $[-\Delta, \Delta]$ on an equal footing. We recall that the Rfit ansatz was obtained starting from a well test statistic, with a flat bottom with a width given by the theoretical error and parabolic walls given by statististical uncertainty.

A related method, called the Scan method, has been developed in the context of flavour physics [32, 33]. It is however slightly different from the case discussed here. First, the test statistic chosen is not the same, since the Scan method uses the likelihood rather than the likelihood ratio, i.e. it relies on

[10] The choice of the weight in the denominator of the test statistic will be discussed in the multidimensional case in Sec. 5.2.3, but it does not impact the result for the p-value in one dimension where it plays only the role of an overall normalisation that cancels when computing the p-value.

the test statistic $T = -2 \log \mathcal{L}(\mu, \nu)$ which is interpreted assuming that T follows a χ^2-law with the corresponding number of degrees of freedom N, including both parameters of interest and nuisance parameters [11]. Then the $1 - \alpha$ confidence region is then determined by varying nuisance parameters in given intervals (typically Ω_1), but accepting only points where $T \leq T_c$, where T_c is a critical value so that $P(T \geq T_c; N|H_0) \geq \alpha$ (generally taken as $\alpha = 0.05$). This latter condition acts as a test of compatibility between a given choice of nuisance parameters and the data.

4 Comparison of the methods in the one-dimensional case

In the following, we will discuss properties of the different approaches in the case of one dimension. More specifically, we will consider:

- the random-δ approach with a Gaussian random variable, or naive Gaussian (nG), see Sec. 3.2,
- the nuisance-δ approach with quadratic statistic and fixed range, or fixed nuisance, see Sec. 3.3,
- the nuisance-δ approach with quadratic statistic and adaptive range, or adaptive nuisance, see Sec. 3.3,
- the external-δ approach with quadratic statistic and fixed range, equivalent to the Rfit approach in one dimension, see Sec. 3.4.

Note that we will not consider other (non quadratic) statistics. Finally, we consider

$$X = 0 \pm \sigma \pm \Delta \qquad \sigma^2 + \Delta^2 = 1 \tag{32}$$

varying Δ/σ as an indication of the relative size of the experimental and theoretical uncertainties.

4.1 p-values and confidence intervals

We can follow the discussion of the previous section and plot the results for the p-values obtained from the various methods discussed above in Fig. 4, where we compare nG, Rfit, fixed nuisance and adaptive nuisance approaches. From these p-values, we can infer confidence intervals at a given significance level and a given value of Δ/σ, and determine the length of the (symmetric) confidence interval (see Tab. 2). We notice the following points:

- by construction, nG always provides the same errors whatever the relative proportion of theoretical and statistical uncertainties, and all the approaches provide the same answer in the limit of no theoretical uncertainty $\Delta = 0$.
- by construction, for a given $n\sigma$ confidence level, the interval provided by the adaptive nuisance approach is identical to the one obtained using the fixed nuisance approach with a $[-n, n]$ interval. This explains why the adaptive nuisance approach yields identical results to the fixed 1-nuisance approach at 1 σ (and similarly for the fixed 3-nuisance approach at 3 σ). The corresponding curves cannot be distinguished on the upper and central panels of Fig. 5.
- the adaptive nuisance approach is numerically quite close to the nG method; the maximum difference occurs for $\Delta/\sigma = 1$ (up to 40% larger error size for 5 σ intervals).
- the p-value from the fixed-nuisance approach has a very wide plateau if one works with the 'reasonable' range $[-3\Delta, +3\Delta]$, while the choice of $[-\Delta, +\Delta]$ might be considered as non conservative.

[11] Such a test statistic tends typically to be less sensitive to discrepancies in a global fit than the likelihood ratio. In presence of quantities having no or little dependence on the scanned parameters, the impact of discrepancies is diluted in the case of the likelihood statistic.

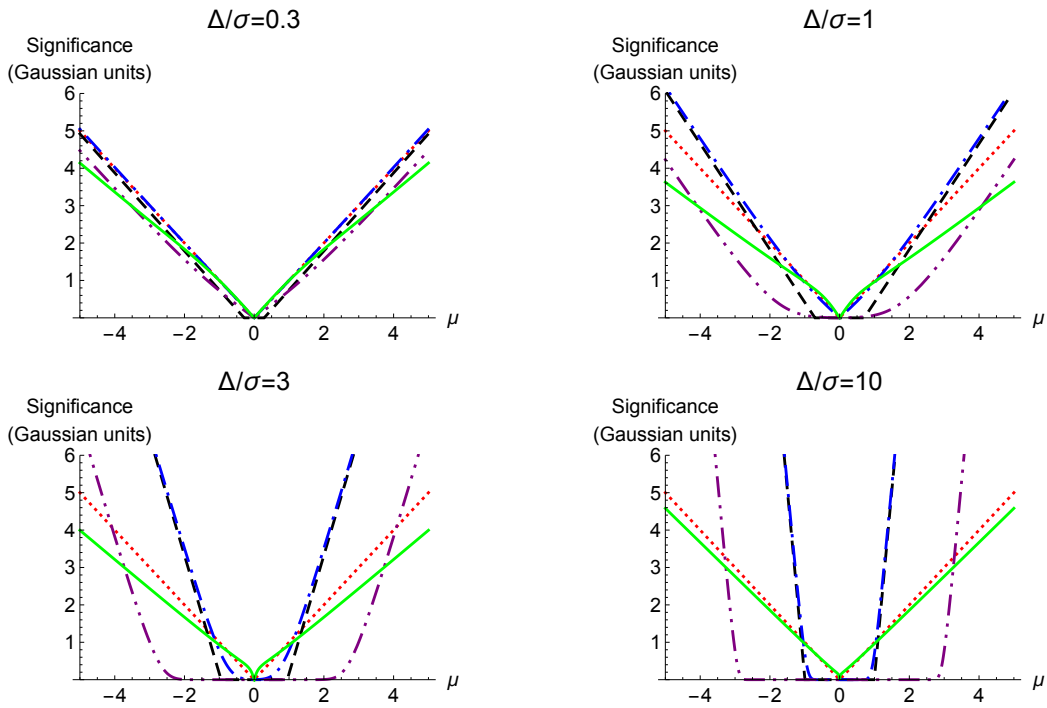


Figure 4: Comparison of different treatments of theoretical uncertainties of the measurement $X = 0 \pm \sigma$ (exp) $\pm\, \Delta$(th), with different values of Δ/σ (with the normalisation $\sqrt{\Delta^2 + \sigma^2} = 1$). The p-values have been converted into a significance in Gaussian units of σ following the particle physics conventions. The various approaches are: nG (dotted, red), Rfit or 1-external (dashed, black), fixed 1-nuisance (dotted-dashed, blue), fixed 3-nuisance (dotted-dotted-dashed, purple), adaptive nuisance (solid, green).

- the 1-external and fixed 1-nuisance approaches are close to each other and less conservative than the adaptive approach, which is expected, but also than nG, for confidence intervals at 3 or 5 σ when theory uncertainties dominate.
- when dominated by theoretical uncertainties (Δ/σ large), all approaches provide 3 and 5 σ errors smaller than the nG approach, apart from the adaptive nuisance approach.

4.2 Significance thresholds

Another way of comparing methods consists in taking the value of μ for which the p-value corresponds to $1, 3, 5$ σ (in significance scale) in a given method, and compute the corresponding p-values for the other methods. The results are gathered in Tabs. 3 and 4. Qualitatively, the comparison of significances can be seen from Fig. 4: if the size of the error is fixed, the different approaches quote different significances for this same error.

In agreement with the previous discussion, we see that fixed 1-nuisance and 1-external yield similar results for 3 and 5 σ, independently of the relative size of statistical and theoretical effects. Moreover, they are prompter to claim a tension than nG, the most conservative method in this respect being the adaptive nuisance approach.

As a physical illustration of this problem, we can consider the current situation for the anomalous magnetic moment of the muon, namely the difference between the experimental measurement and the theoretical computation in the Standard Model [34]:

$$(a_\mu^{exp} - a_\mu^{SM}) \times 10^{11} = 288 \pm 63_{stat} \pm 49_{th} \tag{33}$$

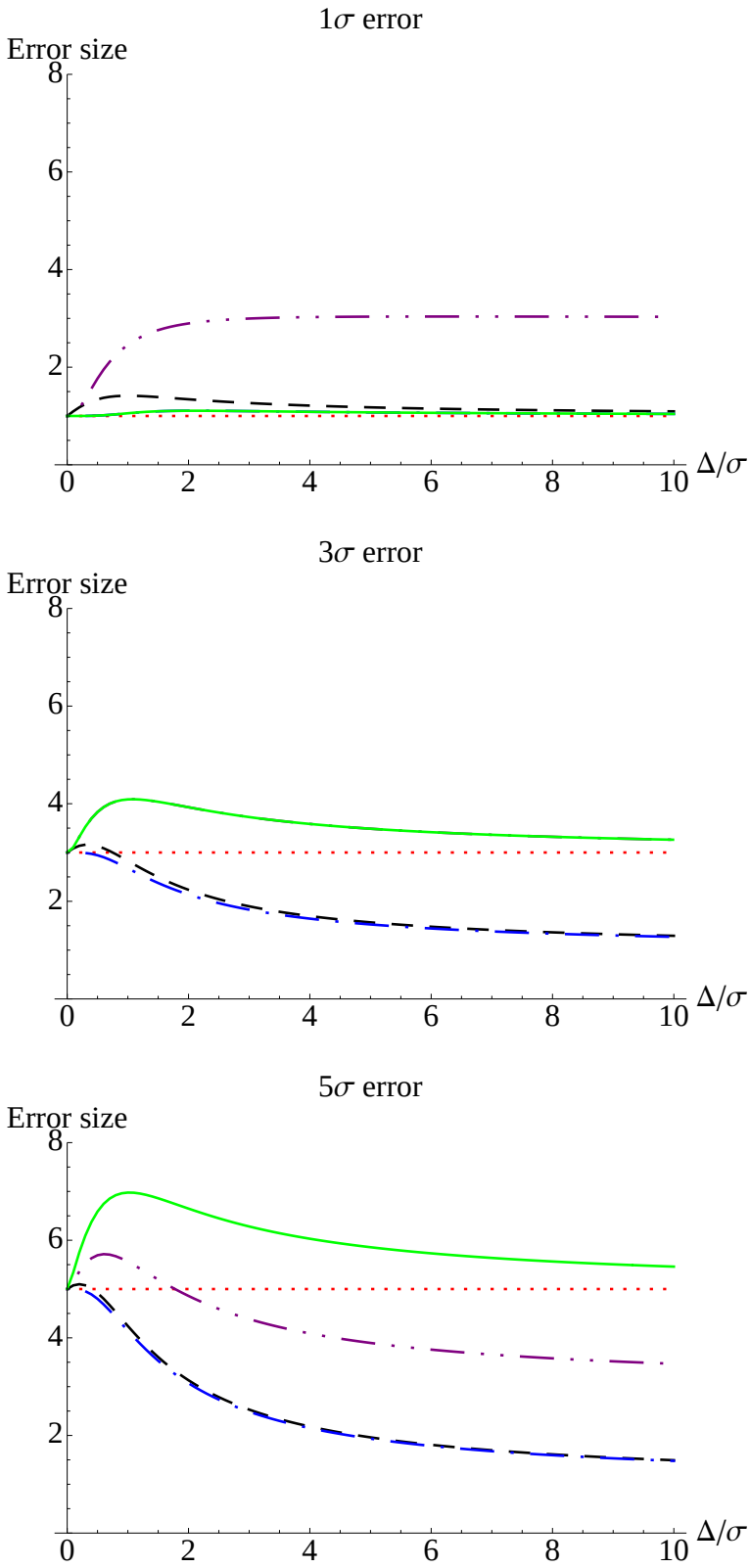


Figure 5: Comparison of the size of the $(1, 3, 5)\sigma$ errors (upper, central and lower panels respectively) as a function of Δ/σ. Different approaches are shown: nG (dotted, red), Rfit or 1-external (dashed, black), fixed 1-nuisance (dotted-dashed, blue), fixed 3-nuisance (dotted-dotted-dashed, purple), adaptive nuisance (solid, green). In the upper panel (1σ confidence level), the adaptive and fixed 1-nuisance approaches yield the same result by construction, and the two curves cannot be distinguished (only the adaptive one is shown). The same situation occurs in the central panel corresponding to 3 σ with the adaptive and fixed 3-nuisance approaches.

	nG	1-nuisance	adaptive nuisance	1-external
$\Delta/\sigma = 0.3$				
1σ	1.0	1.0	1.0	1.2
3σ	3.0	3.0	3.5	3.2
5σ	5.0	5.0	6.1	5.1
$\Delta/\sigma = 1$				
1σ	1.0	1.1	1.1	1.4
3σ	3.0	2.7	4.1	2.8
5σ	5.0	4.1	7.0	4.2
$\Delta/\sigma = 3$				
1σ	1.0	1.1	1.1	1.3
3σ	3.0	1.8	3.7	1.9
5σ	5.0	2.5	6.3	2.5
$\Delta/\sigma = 10$				
1σ	1.0	1.0	1.0	1.1
3σ	3.0	1.3	3.3	1.3
5σ	5.0	1.5	5.5	1.5

Table 2: Comparison of the size of one-dimensional confidence intervals at $1, 3, 5\sigma$ for various methods and various values of Δ/σ.

This discrepancy has a different significance depending on the model chosen for theoretical uncertainties, which can be computed from the associated p-value (under the hypothesis that the true value of $a_\mu^{SM} - a_\mu^{exp}$ is $\mu = 0$). [12] The nG method yields 3.6 σ, the 1-external approach 3.8 σ, the 1-nuisance approach 4.0 σ, and the adaptive nuisance approach 2.7 σ. The overall pattern is similar to what can be seen from the above tables, with a significance of the discrepancy which depends on the model used for theoretical uncertainties.

4.3 Coverage properties

As indicated in Sec. 1.1.2, p-values are interesting objects if they cover exactly or slightly overcover in the domain where they should be used corresponding to a given significance, see Eqs. (7)-(9). If coverage can be ensured for a simple hypothesis [1, 2], this property is far from trivial and should be checked explicitly in the case of composite hypotheses, where compositeness comes from nuisance parameters that can be related to theoretical uncertainties, or other parameters of the problem.

For all methods we study coverage properties in the standard way: one first fixes the true values of the parameters μ and δ (which are not assumed to be random variables), from which one generates a large sample of toy experiments X_i. Then for each toy experiment one computes the p-value at the true value of μ. The shape of the distribution of p-values indicates over, exact or under coverage. More specifically, one can determine $P(p \geq 1 - \alpha)$ for a CL of α: if it is larger (smaller) than α, the method overcovers (undercovers) for this particular CL, i.e. it is conservative (liberal). We emphasise that this property is *a priori* dependent on the chosen CL.

In order to compare the different situations, we take $\sigma^2 + \Delta^2 = 1$ for all methods, and compute for each method the coverage fraction (the number of times the confidence level interval includes the true value of the parameter being extracted) for various confidence levels and for various values of Δ/σ. Note that the coverage depends also on the true value of δ/Δ (the normalized bias). The

[12] In full generality, one should have kept the different sources of theoretical uncertainties separated, as their combination in a single theoretical uncertainty depends on the precise model used for theoretical uncertainties. We consider here the result of ref. [34] where all theoretical uncertainties are already combined.

1σ signif. threshold	nG	1-nuisance	adaptive nuisance	1-external
nG	1	0.9	1.0	0.4
1-nuisance	1.1	1	1.0	0.5
adaptive nuisance	1.1	1.0	1	0.5
1-external	1.4	1.4	1.2	1
3σ signif. threshold	nG	1-nuisance	adaptive nuisance	1-external
nG	3	3.4	2.3	3.2
1-nuisance	2.7	3	2.0	2.8
adaptive nuisance	4.1	4.9	3	4.8
1-external	2.8	3.2	2.1	3
5σ signif. threshold	nG	1-nuisance	adaptive nuisance	1-external
nG	5	6.2	3.6	6.1
1-nuisance	4.1	5	3.0	4.9
adaptive nuisance	7.0	∞	5	∞
1-external	4.2	5.1	3.1	5

Table 3: Comparison of 1D $1, 3, 5\sigma$ significance thresholds for $\Delta/\sigma = 1$. For instance, the first line should read: if with nG a p-value=1 σ is found, then the corresponding values for the three other methods are 0.9/1.0/0.4 σ. ∞ means that the corresponding p-value was numerically zero (corresponding to more than 8 σ).

1σ signif. threshold	nG	1-nuisance	adaptive nuisance	1-external
nG	1	0.8	0.9	0.2
1-nuisance	1.1	1	1.0	0.5
adaptive nuisance	1.1	1.0	1	0.5
1-external	1.3	1.4	1.1	1
3σ signif. threshold	nG	1-nuisance	adaptive nuisance	1-external
nG	3	6.6	2.4	6.5
1-nuisance	1.8	3	1.5	2.8
adaptive nuisance	3.7	∞	3	∞
1-external	1.9	3.2	1.6	3
5σ signif. threshold	nG	1-nuisance	adaptive nuisance	1-external
nG	5	∞	4.0	∞
1-nuisance	2.5	5	2.0	4.9
adaptive nuisance	6.3	∞	5	∞
1-external	2.5	5.1	2.1	5

Table 4: Comparison of 1D $1, 3, 5\sigma$ significance thresholds for $\Delta/\sigma = 3$. Same comments as in the previous table.

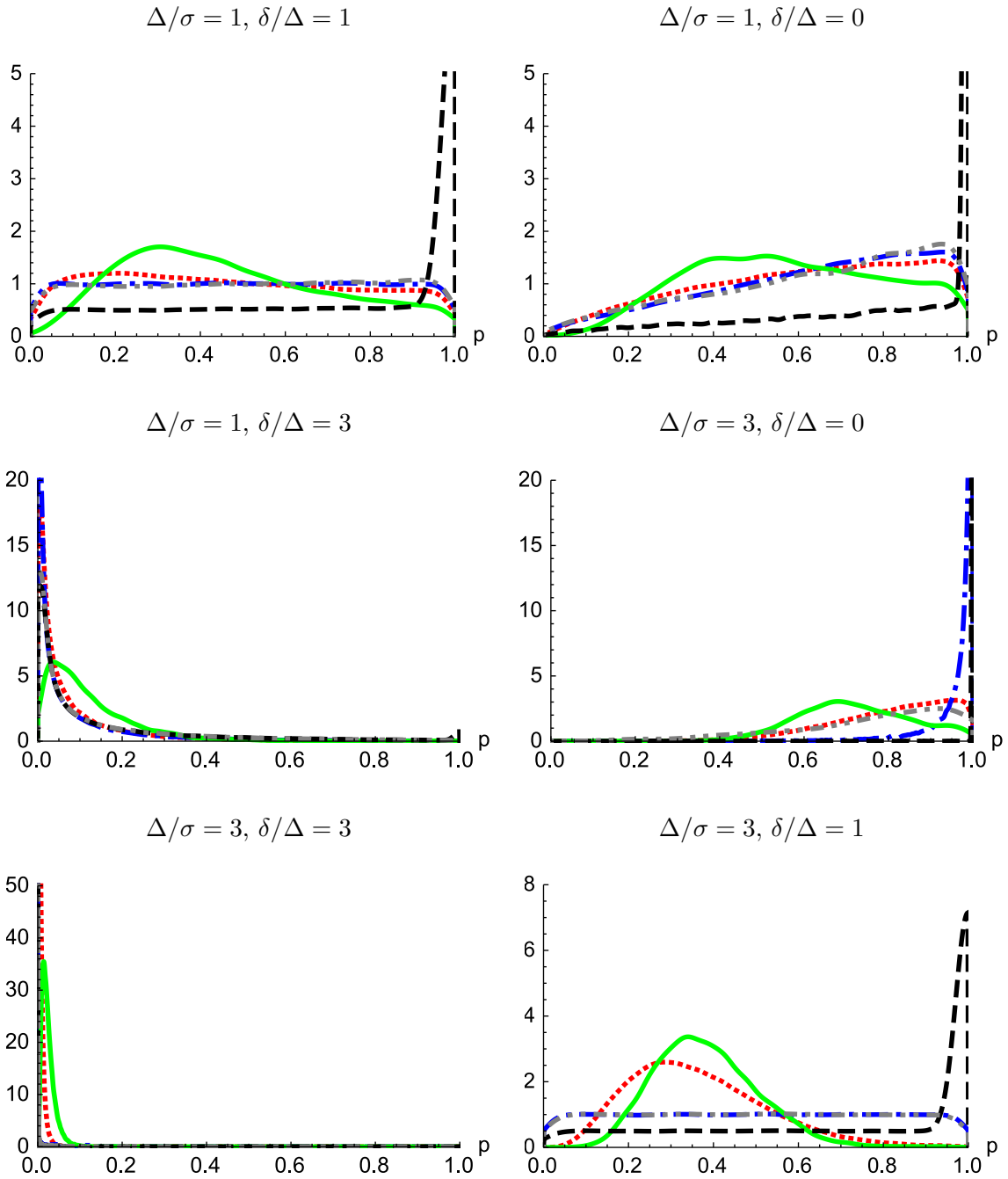


Figure 6: Distribution of p-value (for a fixed total number of events) for different true values δ/Δ and various relative sizes of statistical and theoretical uncertainties Δ/σ. The following approaches are shown: nG (dotted, red), Rfit or 1-external (dashed, black), fixed 1-nuisance (dotted-dashed, blue), adaptive nuisance (solid, green). Since the 1-external approach produces clusters of $p = 1$ p-values, the coverage values excluding these clusters are also shown, as well as the distribution of p-values (dotted-dotted-dashed, grey). Note that the behaviour of the 1-external p-value around $p = 1$ is smoothened by the graphical representation.

	68.27% CL	95.45% CL	99.73% CL	68.27% CL	95.45% CL	99.73% CL
	$\Delta/\sigma = 1$, $\delta/\Delta = 1$			$\Delta/\sigma = 1$, $\delta/\Delta = 0$		
nG	65.2%	96.6%	99.9%	84.1%	99.5%	100.0%
1-nuisance	68.2%	95.4%	99.7%	86.5%	99.3%	100.0%
adaptive nuisance	68.3%	99.6%	100.0%	86.4%	100.0%	100.0%
1-external	83.9%	97.8%	99.9%	95.4%	99.7%	100.0%
1-ext. (excl. $p \equiv 1$)	69.2%	95.7%	99.8 %	85.5%	99.1%	100.0 %
	$\Delta/\sigma = 1$, $\delta/\Delta = 3$			$\Delta/\sigma = 3$, $\delta/\Delta = 0$		
nG	5.76%	43.2%	89.1%	99.8%	100.0%	100.0%
1-nuisance	6.60%	38.0%	78.4%	100.0%	100.0%	100.0%
adaptive nuisance	6.53%	75.4%	99.8%	99.9%	100.0%	100.0%
1-external	16.0%	50.3%	84.2%	100.0%	100.0%	100.0%
1-ext. (excl. $p \equiv 1$)	14.0%	49.1%	83.8 %	98.5%	100.0%	100.0 %
	$\Delta/\sigma = 3$, $\delta/\Delta = 3$			$\Delta/\sigma = 3$, $\delta/\Delta = 1$		
nG	0.00%	0.35%	68.7%	56.3%	100.0%	100.0%
1-nuisance	0.00%	0.00%	0.07%	68.1%	95.5%	99.7%
adaptive nuisance	0.00%	9.60%	99.8%	68.2%	100.0%	100.0%
1-external	0.00%	0.00%	0.13%	84.1%	97.7%	99.9 %
1-ext. (excl. $p \equiv 1$)	0.00%	0.00%	0.13%	68.2%	95.4%	99.7%

Table 5: Coverage properties of the various methods at 68.27, 95.45 and 99.73% CL, for different true values of δ/Δ contained in, at the border of, or outside the fixed volume Ω, and for various relative sizes of statistical and theoretical uncertainties Δ/σ.

results are gathered in Tab. 5 and Fig. 6. We also indicate the distribution of p values obtained for the different methods.

One notices in particular that the 1-external approach has a cluster of values for $p = 1$, which is expected due to the presence of a plateau in the p-value. This behaviour makes the interpretation of the coverage more difficult, and as a comparison, we also include the results when we consider the same distribution with the $p = 1$ values removed. Indeed one could imagine a situation where reasonable coverage values could only be due to the $p = 1$ clustering, while other values of p would systematically undercover: such a behaviour would either yield no constraints or too liberal constraints on the parameters depending on the data.

The results are the following:

- if Ω is fixed and does not contain the true value of δ/Δ ("unfortunate" case), both external-δ and nuisance-δ approaches lead to undercoverage; the size of the effect depends on the distance of δ/Δ with respect to Ω. This is also the case for nG.

- if Ω is fixed and contains the true value of δ/Δ ("fortunate" case), both the external-δ and nuisance-δ approaches overcover. This is also the case for nG.

- if Ω is adaptive, for a fixed true value of δ, a p-value becomes valid if it is sufficiently small so that the corresponding interval contains δ. Therefore, for the adaptive nuisance-δ approach, there is always a maximum value of CL above which all p-values are conservative; this maximum value is given by $1 - \mathrm{Erf}[\delta/(\sqrt{2}\Delta)]$.

To interpret the pattern of coverage seen above in the external and nuisance approaches, note that one starts with a p-value that has exact coverage under the individual simple hypotheses when δ is fixed. Therefore, as long as the true value δ lies within the range over which one takes the supre-

mum, this procedure yields a conservative envelope. This explains the overcoverage/undercoverage properties for the external-δ and nuisance-δ approaches given above.

4.4 Conclusions of the uni-dimensional case

It should be stressed that, by construction, all methods are conservative if the true value of the δ parameter satisfy the assumption that has been made for the computation of the p-value. Therefore coverage properties are not the only criterion to investigate in this situation in order to assess the methods: in particular one has to study the robustness of the p-value when the assumption set on the true value of δ is not true. The adaptive approach provides a means to deal with *a priori* unexpected true values of δ, provided one is interested in a small enough p-value, that is, a large enough significance effect. Other considerations (size of confidence intervals, significance thresholds) suggest that the adaptive approach provides an interesting and fairly conservative framework to deal with theoretical uncertainties. We are going to consider the different approaches in the more general multi-dimensional case, putting emphasis on the adaptive nuisance-δ approach and the quadratic test statistic.

5 Generalization to multi-dimensional cases

Up to here we only have discussed the simplest example of a single measurement X linearly related to a single model parameter μ. Obviously the general case is multi-dimensional, where we deal with several observables, depending on several underlying parameters, possibly in a non-linear way, with several measurements involving different sources of theoretical uncertainty. Typical situations correspond to averaging different measurements of the same quantity, and performing fits to extract confidence regions for fundamental parameters from the measurement of observables. In this section we will discuss the case of an arbitrary number of observables in a linear model with an arbitrary number of parameters, where we are particularly interested in a one-dimensional or two-dimensional subset of these parameters.

5.1 General formulae

We start by defining the following quadratic test statistic

$$T(X;\chi,\delta) = (X - x(\chi) - \Delta\tilde{\delta})^T \cdot W_s \cdot (X - x(\chi) - \Delta\tilde{\delta}) + \tilde{\delta}^T \cdot \widetilde{W}_t \cdot \tilde{\delta} \tag{34}$$

where $X = (X_i,\ i = 1,\ldots,n)$ is the n-vector of measurements, $x = (x_i,\ i = 1,\ldots,n)$ is the n-vector of model predictions for the X_i that depends on $\chi = (\chi_j,\ j = 1,\ldots,n_\chi)$, the n_χ-vector of model parameters, $\tilde{\delta}$ is the m-vector of (dimensionless) theoretical biases, W_s is the (possibly non diagonal) $n \times n$ inverse of the statistical covariance matrix C_s, $\widetilde{W}_t$ is the inverse of the (possibly non diagonal) $m \times m$ theoretical correlation matrix $\widetilde{C}_t$, Δ is the $n \times m$-matrix of theoretical uncertainties $\Delta_{i\alpha}$, so that the reduced biases $\tilde{\delta}_\alpha$ have a range of variation within $[-1, 1]$ (this explains the notation with tildes for the reduced quantities rescaled to be dimensionless).

After minimization over the $\tilde{\delta}_\alpha$, T can be recast into the canonical form

$$T(X;\chi) = (X - x(\chi))^T \cdot \bar{W} \cdot (X - x(\chi)) \tag{35}$$

where

$$\bar{W} = W_s - B^T \cdot A^{-1} \cdot B \tag{36}$$

with

$$B = (W_s\Delta)^T \qquad\qquad A = \widetilde{W}_t + B\Delta \tag{37}$$

The definition of $\bar{W}$ involves the inverse of matrices that can be singular. This may occur in particular in cases where the statistical uncertainties are negligible and some of the theoretical uncertainties are assumed to be 100% correlated. This requires us to define a generalised inverse, including singular cases, which is described in detail in App. A and corresponds to a variation of the approach presented in Ref. [5]. Ambiguities and simplifications that can occur in the definition of T are further discussed in App. C. In particular, one can reduce the test statistic to the case $m = n$ with a diagonal Δ matrix without losing information. In the case where both correlation/covariance matrices are regular, Eq. (36) boils down to $\bar{W} = [C_s + C_t]^{-1}$ with $C_t = \Delta \widetilde{C}_t \Delta^T$. This structure is reminiscent of the discussion of theoretical uncertainties as biases and the corresponding weights given in Ref. [29], but it extends it to the case where correlations yield singular matrices.

We will focus here on the case where the model is *linear*, i.e., the predictions x_i depend linearly on the parameters χ_j:

$$x_i(\chi) = \sum_{k=1}^{n_\chi} a_{ik}\chi_k + b_i \tag{38}$$

where a_{ik} and b_i are constants. We leave the phenomenologically important non-linear case and its approximate linearisation for a dedicated discussion in a separate paper [35].

Following the one-dimensional examples in the previous sections, we always assume that the measurements X_i have Gaussian distributions for the statistical part. We will consider two main cases of interest in our field: averaging measurements and determining confidence intervals for several parameters.

5.2 Averaging measurements

We start by considering the averages of several measurements of a single quantity, each with both statistical and theoretical uncertainties, with possible correlations. We will focus mainly on the nuisance-δ approach, starting with two measurements before moving to other possibilities.

5.2.1 Averaging two measurements and the choice of a hypervolume

A first usual issue consists in the case of two uncorrelated measurements $X_1 \pm \sigma_1 \pm \Delta_1$ and $X_2 \pm \sigma_2 \pm \Delta_2$ that we want to combine. The procedure is well defined in the case of purely statistical uncertainties, but it depends obviously on the way theoretical uncertainties are treated. As discussed in Sec. 2, associativity is a particulary appealing property for such a problem as it allows one to replace a series of measurements by its average without loss of information.

Averaging two measurements amounts to combining them in the test statistic. The nuisance-δ approach, together with the quadratic statistic Eq. (34), in the absence of correlations yields:

$$T = \frac{(X_1 - \mu)^2}{\sigma_1^2 + \Delta_1^2} + \frac{(X_2 - \mu)^2}{\sigma_2^2 + \Delta_2^2} = (\mu - \hat{\mu})^2 (w_1 + w_2) + T_{\min} \tag{39}$$

with

$$\hat{\mu} = \frac{w_1 X_1 + w_2 X_2}{w_1 + w_2} \qquad w_i = \frac{1}{\sigma_i^2 + \Delta_i^2} \qquad T_{\min} = \frac{(X_1 - X_2)^2}{\sigma_1^2 + \Delta_1^2 + \sigma_2^2 + \Delta_2^2} \tag{40}$$

$\hat{\mu}$ is a linear combination of Gaussian random variables, and is thus distributed according to a Gaussian p.d.f, with mean $\mu + \delta_\mu$ and variance σ_μ^2

$$\delta_\mu = \frac{w_1 \delta_1 + w_2 \delta_2}{w_1 + w_2} \qquad \sigma_\mu^2 = \frac{w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2}{(w_1 + w_2)^2} \tag{41}$$

Therefore, $T - T_{\min}$ is distributed as a rescaled uni-dimensional non-central χ^2 distribution with non-centrality parameter $(\delta_\mu / \sigma_\mu)^2$.

σ_μ corresponds to the statistical part of the error on μ. δ_1 and δ_2 remain unknown by construction, and the combined theory error can only be obtained once a region of variation is chosen for the δ's (as a generalisation of the [-1,1] interval in the one-dimension case). If one maximises the p-value over a rectangle $\mathcal{C}$ (called "hypercube case" in the following, in reference to its multidimensional generalisation), δ_μ varies in Δ_μ, with

$$\Delta_\mu = \frac{w_1\Delta_1 + w_2\Delta_2}{w_1 + w_2} \tag{42}$$

recovering the proposal in Ref. [29] for the treatment of systematic uncertainties. In this case, δ_1 and δ_2 are allowed to be varied separately, without introducing any relation in their values, and can assume both extremal values. On the other hand, if one performs the maximisation over a disk (referred to as the "hyperball case" for the same reasons as above) one has the range

$$\Delta_\mu = \frac{\sqrt{w_1^2\Delta_1^2 + w_2^2\Delta_2^2}}{w_1 + w_2} \tag{43}$$

In this case, the values of δ_1 and δ_2 are somehow related, since they cannot both reach extremal values simultaneously.

Each choice of volume provides an average with different properties. As discussed earlier, associativity is a very desirable property: one can average different observations of the same quantity prior to the full fit, since it gives the same result as keeping all individual inputs. The hyperball choice indeed fulfills associativity. On the other hand, the hypercube case does not: the combination of the inputs 1 and 2 yields the following test statistic: $(w_1 + w_2)(\mu - \hat{\mu})^2$, whereas the resulting combination $\hat{\mu} \pm \sigma_\mu \pm \Delta_\mu$ has the statistic $(\mu - \hat{\mu})^2/(\sigma_\mu^2 + \Delta_\mu^2)$. The two statistics are proportional and hence lead to the same p-value, but they are not equivalent when added to other terms in a larger combination.

A comment is also in order concerning the size of the uncertainties for the average. In the case of the hypercube, the resulting linear addition scheme is the only one where the average of different determinations of the same quantity cannot lead to a weighted theoretical uncertainty that is smaller than the smallest uncertainty among all determinations [13]. In the case of the hyperball, it may occur that the average of different determinations of the same quantity yields a weighted theoretical uncertainty smaller than the smallest uncertainty among all determinations.

Whatever the choice of the volume, a very important and alluring property of our approach is the clean separation between the statistical and theoretical contribution to the uncertainty on the parameter of interest. This is actually a general property that directly follows from the choice of a quadratic statistic, and in the linear case it allows one to perform global fits while keeping a clear distinction between various sources of uncertainty.

5.2.2 Averaging n measurements with biases in a hyperball

We will now consider here the problem of averaging n, possibly correlated, determinations of the same quantity, each individual determination coming with both a Gaussian statistical uncertainty, and a number of different sources of theoretical uncertainty. We focus first on the nuisance$-\delta$ approach, as it is possible to provide closed analytic expressions in this case. We will first discuss the variation of the biases over a hyperball, before discussing other approaches, which will be illustrated and compared with examples from flavour physics in Sec. 6.

[13] This is true at least for approaches where theoretical errors are modelled by fixed bias parameters: the combined error on the quantity of interest is a weighted sum as in Eq. (42), and the maximal value of this quantity can only be made always larger than each individual contribution if the corners of the hypercube are included in the maximisation region.

We use the test statistic Eq. (34) for μ, with $x(\chi)$ simply replaced by μU, where U is the n-vector $(1,\ldots,1)$. After minimization over the $\tilde{\delta}_\alpha$, T can be recast into the canonical form

$$T(\mu) = (X - \mu U)^T \cdot \bar{W} \cdot (X - \mu U) \tag{44}$$

The minimization of Eq. (44) over μ leads to an estimator $\hat{\mu}$ of the average in terms of the measurements X_i

$$\hat{\mu} = \sum_i w_i X_i \qquad w_i = \sum_j \bar{W}_{ij} \times \left[\sum_{i,j} \bar{W}_{ij}\right]^{-1} \tag{45}$$

that allows one to compute the statistical uncertainty σ_μ in the following way

$$\sigma_\mu^2 = \sum_{i,j} (C_s)_{ij} w_i w_j \tag{46}$$

The theoretical bias is given by $\delta_\mu = \sum_{i,\alpha} w_i \Delta_{i\alpha} \tilde{\delta}_\alpha$. We would like to vary $\tilde{\delta}_\alpha$ in ranges required to infer the theoretical uncerainty, identifying the combination of biases that is uncorrelated. This is a well known problem of statistics, and it can be easily achieved in a linear manner by noticing that the relevant combination is $\Delta^T \widetilde{C}_t \Delta$, cf. Eq. (36), and by introducing the Cholesky decomposition for the theoretical correlation matrix $\widetilde{C}_t = P \cdot P^T$, with P a lower triangular matrix with diagonal positive entries. This yields the expression for the bias

$$\delta_\mu = \sum_{i,\alpha} w_i \Delta_{i\alpha} \tilde{\delta}_\alpha = \sum_{i,\alpha,\beta} w_i \Delta_{i\alpha} P_{\alpha\beta} (P^{-1}\tilde{\delta})_\beta \tag{47}$$

where $(P^{-1}\tilde{\delta})_\beta$ are uncorrelated biases. If the latter biases are varied over a hyperball, the biases $\tilde{\delta}$ are varied over a hyperellipsoid elongated along the directions corresponding to strong correlations (see App. B for illustrations) and one gets

$$\Delta_\mu = \sqrt{\sum_\beta \left(\sum_{i,\alpha} w_i \Delta_{i\alpha} P_{\alpha\beta}\right)^2} = \sqrt{w^T C_t w} \qquad \text{(hyperball)} \tag{48}$$

Known (linear) statistical correlations between two measurements are straightforward to implement, by using the full covariance matrix in the test statistic Eq. (46). On the other hand, in the physical problems considered here (involving hadronic inputs from lattice QCD simulations), it often happens that two *a priori* independent calculations of the same quantity are statistically correlated, because they use the same (completely or partially) ensemble of gauge configurations. The correlation is not perfect of course, since usually different non linear actions are used to perform the computation. However the accurate calculation of the full covariance matrix is difficult, and in many cases it is not available in the literature. For definiteness, we will assume that if two lattice calculations are statistically correlated, then the (linear) correlation coefficient is one. In such a case the covariance matrix is singular, and its inverse W_s is ill-defined, as well as all quantities that are defined above in terms of W_s. A similar question arises for fully correlated theoretical uncertainties (coming from the same method), leading to ambiguities in the definition of $\widetilde{W}_t$. Details on these issues are given in Apps. A and B.

Statistical uncertainties are assumed here to be strictly Gaussian and hence symmetric (see App. D for more detail on the asymmetric case). In contrast, in the nuisance approach, a theoretical uncertainty that is modelled by a bias parameter δ may be asymmetric: that is, the region in which δ is varied may depend on the sign of δ, *e.g.*, $\delta \in [-\Delta_-, +\Delta_+]$ in one dimension with the fixed hypercube approach ($\Delta_\pm \geq 0$). In order to keep the stationarity property that follows from the quadratic statistic, we take the conservative choice $\Delta = \mathrm{Max}(\Delta_+, \Delta_-)$ in the definition Eq. (34). Let us emphasise that this symmetrisation of the test statistic is independent of the range in which δ is varied: if theoretical uncertainties are asymmetric, one computes Eqs. (46)-(48) to express the asymmetric combined uncertainties $\Delta_{\mu,\pm}$ in terms of the $\Delta_{i\alpha,\pm}$.

5.2.3 Averages with other approaches

In Sec. 5.2.1, we indicated that other domains can be chosen in principle in order to perform the averages of measurements, for instance a hypercube rather than a hyperball. If we do not try to take into account theoretical correlations in the range of variation, it is quite easy to determine the result for Δ

$$\Delta_\mu = \sum_\alpha \left| \sum_i w_i \Delta_{i\alpha} \right| \qquad \text{(hypercube, no theoretical correlation)} \tag{49}$$

reminiscent of the formulae derived in Ref. [29]. However, we encountered severe difficulties when trying to include theoretical correlations in the discussion. Similarly to the hyperball case, it would be interesting to consider a linear transformation P of the biases (for instance, the Cholesky decomposition of C_t, but the discussion is more general), so that $(P^{-1}\tilde{\delta})_\beta$ are uncorrelated biases varied within a hypercube. This would lead to $\tilde{\delta}$ varied within a deformed hypercube, which corresponds to cutting the hypercube by a set of $(\tilde{\delta}_i, \tilde{\delta}_j)$ hyperplanes. It can take a rather complicated convex polygonal shape that is not symmetric along the diagonal in the $(\tilde{\delta}_i, \tilde{\delta}_j)$ plane, leading to the unpleasant feature that the order in which the measurements are considered in the average matters to define the range of variation of the biases (an illustration is given in App. B) [14].

As indicated before, this discussion occurs for any linear transformation P and is not limited to the Cholesky decomposition. We have not been able to find other procedures that would avoid these difficulties while paralleling the hypercube case. In the following, we will thus use Eq. (49) even in the presence of theoretical correlations: therefore, the latter will be taken into account in the definition of T through $\bar{W}$, but not in the definition of the range of variations to compute the error Δ. We also notice that the problems that we encounter are somehow due to contradicting expectations concerning the hypercube approach. In Sec. 5.2.1, the hypercube corresponds to values of δ_1 and δ_2 left free to vary without relation among them (contrary to the hyperball case). It seems therefore difficult to introduce correlations in this case which was designed to avoid them initially. Our failure to introduce correlations in this case might be related to the fact that the hypercube is somehow designed to avoid such correlations from the start and cannot accomodate them easily.

In the case of the external-δ approach, the scan method leads to the same discussion as for the nuisance case, provided that one uses the following statistic: $T = (X-\mu-\delta)^2/(\sigma^2+\Delta^2)$. This choice is different from Ref. [32] by the normalisation ($\sigma^2+\Delta^2$ rather than σ^2) in order to take into account of the importance of both uncertainties when combining measurements (damping measurements which are unprecise in one way or the other). As indicated in Sec. 3.4, the difference of normalisation of the test statistic does not affect the determination of the p-value in the uni-dimensional case, but it has an impact once several determinations are combined. The choice above corresponds to the usual one when Δ is of statistical nature. It gives a reasonable balance when two or more inputs are combined, that all come with both statistical and theoretical uncertainties.

A similar discussion holds for the random-δ approach. However, if the combined errors σ_μ and Δ_μ are the same between the nuisance-δ (with hyperball), the random-δ and the external-δ (with hyperball) approaches, we emphasise that the p-value for μ built from these errors is different and yields different uncertainties for a given confidence level for each approach, as discussed in Sec. 3.

5.2.4 Other approaches in the literature

There are other approaches available in the literature, often starting from the random-δ approach (i.e., modeling all uncertainties as random variables).

The Heavy Flavour Averaging Group [36] choose to perform the average including correlations. In the absence of knowledge on the correlation coefficient between uncertainties of two measurements

[14] This problem does not occur in the hyperball case, where the section of the hyperellipsoid by a hyperplane always yields an ellipse symmetric along the diagonal, with an elongation according to the theoretical correlation between the biases.

(typically coming from the same method), they tune the correlation coefficient so that the resulting uncertainty is maximal (which is not $\rho = 1$ in the case where the correlated uncertainties have a different size and are combined assuming a statistical origin, see App. A.2). This choice is certainly the most conservative one when there is no knowledge concerning correlations.

The Flavour Lattice Averaging Group [37] follows the proposal in Ref. [38]: they build a covariance matrix where correlated sources of uncertainties are included with 100% correlation, and they perform the average by choosing weights w_i that are not optimal but are well defined even in the presence of $\rho = \pm 1$ correlation coefficients. As discussed in App. A.2, our approach to singular covariance matrices is similar but more general and guarantees that we recover the weights advocated in Ref. [38] for averages of fully correlated measurements.

Finally, the PDG approach [34] combines all uncertainties in a single covariance matrix. In the case of inconsistent measurements, one may then obtain an average with an uncertainty that may be interpreted as 'too small' (notice however that the weighted uncertainty does not increase with the incompatibility of the measurements). This problem occurs quite often in particle physics and cannot be solved by purely statistical considerations (even in the absence of theoretical uncertainties). If the model is assumed to be correct, one may invoke an underestimation of the uncertainties. A (commonly used) recipe in the pure statistical case has been adopted by the Particle Data Group, which consists in computing a factor $S = \sqrt{\chi^2/(N_{dof} - 1)}$ and rescaling all uncertainties by this factor. A drawback of this approach is the lack of associativity: the inconsistency is either removed or kept as it is, depending on whether the average is performed before any further analysis, or inside a global fit. Furthermore since the ultimate goal of statistical analyses is indeed to exclude the null hypothesis (*e.g.* the Standard Model), it looks counter-intuitive to first wash out possible discrepancies by an *ad hoc* procedure. Therefore we refrain to define a S factor in presence of theoretical uncertainties, and leave the discussion of discrepancies between independent determinations of the same quantity to a case-by-case basis, based on physical (and not statistical) grounds.

In the case of the Rfit approach adopted by the CKMfitter group [15, 16], a specific recipe was chosen to avoid underestimating combined uncertainties in the case of marginally compatible values. The idea is first combine the statistical uncertainties by combining the likelihoods restricted to their statistical part, then assign to this combination the smallest of the individual theoretical uncertainties. This is justified by the following two points: the present state of the art is assumed not to allow one to reach a better theoretical accuracy than the best of all estimates, and this best estimate should not be penalized by less precise methods. In contrast with the plain (or naive) Rfit approach for averages (consisting in just combining Rfit likelihoods without further treatment), this method of combining uncertainties was called educated Rfit and is used by the CKMfitter group for averages [17, 19, 22]. Let us note finally that the calculation of pull values, discussed in Sec. 5.3, is a crucial step for assessing the size of discrepancies.

5.3 Global fit

5.3.1 Estimators and errors

Another prominent example of multi-dimensional problem is the extraction of a constraint on a particular parameter of the model from the measured observables. If the model is linear, Eq. (38), the discussion follows closely that of Sec. 5.2.2. In the case where there is a single parameter of interest μ, we do not write explicitly the calculations and refer to Sec. 6 for numerical examples.

We start from the test statistic Eq. (34) in the linear case defined in Eq. (38), reducing the number of theoretical biases to the case $m = n$ as indicated in App. C. Following the same discussion as in Sec. 5.2.2, we can minimise with respect to $\tilde{\delta}_\alpha$, leading to the canonical form

$$T(X;\chi) = (X - x(\chi))^T \cdot \bar{W} \cdot (X - x(\chi)) \tag{50}$$

The minimum of this function is found at the point $\hat{\chi}_k$ where

$$\left.\frac{\partial T}{\partial \chi_q}\right|_{\chi=\hat{\chi}} = 0, \qquad \hat{\chi} = (a^T \bar{W} a)^{-1} \cdot (a^T \bar{W}(X-b)) \tag{51}$$

so that we have

$$\hat{\chi}_q = \sum w_i^{(q)}(X_i - b_i), \; w_i^{(q)} = [(a^T \bar{W} a)^{-1} a^T \bar{W}]_{qi} \tag{52}$$

The minimum $\hat{\chi}_q$ is thus linearly related to the measured observables X_i and their statistical properties are closely related. The test statistic for a particular parameter $\mu = \chi_q$ will lead to $T(X;\mu) = (\mu - \hat{\chi}_q)^2 \times (a^T \bar{W} a)_{qq}$, so that the discussion of the p-value for μ follows exactly the discussion for uni-dimensional measurements [15].

For instance, if the observables X_i have central values X_{i0} and variances $\sigma^2_{X_i}$, the central value and the variance for $\hat{\chi}_q$ (corresponding also to the central value and statistical uncertainty for the p-value for $\mu = \chi_q$), can readily be obtained from

$$\mu_0 = \hat{\chi}_{q0} \;=\; \sum_{i,j,l=1}^{n} (a^T \bar{W} a)^{-1}_{qj} \times \left[a_{ij} \bar{W}_{il}\right] \times (X_{l0} - b_l) = w^{(q)T}(X_0 - b) \tag{53}$$

$$\sigma^2_\mu = \hat{\sigma}^2_{\chi_q 0} \;=\; \sum_{i,j,l=1}^{n} \left[(a^T \bar{W} a)^{-1}_{qj}\right]^2 \times \left[a_{ij} \bar{W}_{il}\right]^2 \times (\sigma_{X_l})^2 = w^{(q)T} C_s w^{(q)} \tag{54}$$

Similarly to what was presented in the previous section, the theoretical uncertainty on $\mu = \chi_q$ is obtained in the hyperball case as

$$\Delta_\mu = \sqrt{\sum_{i,j,l=1}^{n} \left[(a^T \bar{W} a)^{-1}_{qj}\right]^2 \times \left[a_{ij} \bar{W}_{il}\right]^2 \times (\Delta_{X_i})^2} = \sqrt{w^{(q)T} C_t w^{(q)}} \qquad \text{(hyperball)} \tag{55}$$

It remains to determine how to define the theoretical correlation in this framework, denoted κ_{qr} corresponding to the actual parameters of interest. This can be seen as trying to infer a scalar product on the vectors $[w^{(q)} \Delta P]_i$ from the knowledge of a norm, here L^2. We will thus define the theoretical correlation in the following way

$$\kappa_{qr} \;=\; \frac{w^{(q)T} C_t w^{(r)}}{\sqrt{w^{(q)T} C_t w^{(q)}} \sqrt{w^{(r)T} C_t w^{(r)}}} \qquad \text{(hyperball)} \tag{56}$$

In Sec. 5.2.2 we encountered difficulties in extending the discussion to the hypercube case. We can define errors varying the biases without correlations in the definition of the hypercube

$$\Delta_\mu = \sum_{i=1}^{n} \left| \sum_{j,l=1}^{n} (a^T \bar{W} a)^{-1}_{qj} \times \left[a_{ij} \bar{W}_{il}\right] \times \Delta_{X_i} \right| \qquad \text{(hypercube no correlation)} \tag{57}$$

but we could not determine a way of defining this hypercube taking into account theoretical correlations. Moreover, there is no obvious way to extend the definition of theoretical correlation for the hypercube in a similar way to Eq. (56), as there is no scalar product associated to the L^1-norm. We will thus not quote theoretical correlations for the hypercube case.

[15] As discussed in Sec. 3.4, the overall normalisation of $T(X;\mu)$ is irrelevant to derive uni-dimensional p-values.

5.4 Goodness-of-fit

We would like also to compute the distribution of $T_{\min}$ in presence of biases and extract a goodness-of-fit value. Coming back to the initial problem, we see that $T_{\min}$ can be written as

$$T_{\min} = (X-b)^T(\bar{W} - \bar{W}a(a^T\bar{W}a)^{-1}a^T\bar{W})(X-b) = (X-b)^T M(X-b) \tag{58}$$

where X are distributed following a multivariate normal distribution, with central value $a\chi + b + \Delta\tilde{\delta}$ and correlation matrix C_s. The CDF $H_{\tilde{\delta}}(t)$ for $T_{\min}$ at fixed $\tilde{\delta}$ can thus be rephrased in the following way: considering a vector Y distributed according to a multivariate normal distribution of covariance C_s centred around 0, $H_{\tilde{\delta}}(t)$ is the probability $P[(Y - a\chi - \Delta\tilde{\delta})^T M(Y - a\chi - \Delta\tilde{\delta}) \leq t]$.

We are able to reexpress this problem as a linear combination of non-central χ^2 distributions. Indeed, we can define

$$C_s = LL^T, \qquad L^T ML = K\alpha K^T, \qquad \beta = K^T L^{-1}(a\chi - \Delta\tilde{\delta}) \tag{59}$$

with L lower triangular (using the Cholesky decomposition), α is diagonal and K orthogonal (so that α are the (positive) eigenvalues of $L^T ML$ and thus of MC_s). Let us note that α does depend only on C_s and C_t, whereas the dependence on the true value of χ and $\tilde{\delta}$ is only present in β. The problem is then equivalent to considering a vector Z distributed according to a multivariate normal distribution of covariance identity centred around 0, and computing $P[(Z-\beta)^T\alpha(Z-\beta) \leq t]$. This is the CDF of a linear combination of the form $\sum_i \alpha_i X_i^2$ corresponding to non-central χ^2 distributions.

In the case where α is proportional to identity, the CDF can be expressed in terms of the generalised Marcum Q-function

$$H_{\tilde{\delta}}(t) = 1 - Q_{n/2}\left(\sqrt{\lambda}, \sqrt{t/\alpha}\right) \tag{60}$$

with the non-centrality parameter $\lambda = \sum_i \beta_i^2$. In the general case, the answer can be found in various articles, for instance in Ref. [39], as a linear combination of infinitely many (central or non-central) χ^2 distribution functions, and in Ref. [40], where an expansion in terms of Laguerre polynomials is provided for a fast numerical evaluation. We can thus infer the corresponding p value as

$$p_\Omega = \max_{\tilde{\delta}\in\Omega}[1 - H_{\tilde{\delta}}(t)] \tag{61}$$

where $\tilde{\delta}$ has to be varied in a hyperball or a hypercube depending on the volume chosen, and χ_q are replaced by their estimated values μ_{χ_q}.

5.5 Pull parameters

In addition to the general indication given by goodness-of-fit indicators, it is useful to determine the agreement between individual measurements and the model. One way of quantifying this agreement consists in determining the pull of each quantity. Indeed, the agreement between the indirect fit prediction and the direct determination of some observable X is measured by its pull, which can be determined by considering the difference of minimum values of the test statistic including or not the observables [22]. In the absence of non-Gaussian effects or correlations, the pulls are random variables of vanishing mean and unit variance.

The pull of an observable X_m can be conveniently computed by introducing an additional *pull parameter* p_{X_m} in the test statistic $T(X_0; \chi, p_{X_m})$

$$T = (X_0 - x(\chi) - P_m)^T \bar{W}(X_0 - x(\chi) - P_m) \qquad (P_m)_i = \delta_{mi}\, p_{X_m}/\sqrt{\bar{W}_{mm}} \tag{62}$$

The pull parameter p_{X_m} is a dimensionless fit parameter for which one can compute confidence intervals, or errors and uncertainties. Its best-fit value is a random variable that measures the distance of the indirect prediction (determined by the global fit) from the direct measurement, in units of σ.

The p-value for the null hypothesis $p_{X_m} = 0$ is by definition the pull for X_i. It can be understood as a comparison of the best-fit value of the test statistic reached letting p_{X_a} free (corresponding to a global fit without the measurement X_m) with the case setting $p_{X_m} = 0$ (corresponding to a global fit including the measurement X_m).

As far as the test statistic is concerned, the pull parameter can be treated on the same footing as the parameters χ, and it can be determined in the same way as in the previous section, first solving the minimisation condition $\partial T/\partial p_{X_m} = 0$, and plugging the result for p_{X_m} into T, leading to the same expression for T as in Eq. (50), but with $\bar{W}$ replaced by the matrix

$$\bar{W}_{ij}^{(m)} = \bar{W}_{ij} - \frac{\bar{W}_{im}\bar{W}_{jm}}{\bar{W}_{mm}} \tag{63}$$

which can be solved as before for $\hat{\chi}$, leading to the expression for $\hat{p}_{X_m}$

$$\hat{p}_{X_m} = \sum_i y_i^{(m)} X_i \qquad y_i^{(m)} = \frac{1}{\sqrt{\bar{W}_{mm}}}[\bar{W} - \bar{W}a(a^T\bar{W}^{(k)}a)^{-1}a^T\bar{W}^{(k)}]_{mi} \tag{64}$$

If the statistical method allows one to separate the statistical and theoretical contributions to the error on p_{X_i}, one can report the values for the errors $\Delta_{p_{X_i}}$ and $\sigma_{p_{X_i}}$ in addition to the pull itself: this gives an indication of how independent from theoretical uncertainties the underlying tested hypothesis is. One can also extend this notion for N parameters, introducing N distinct pull parameters and determining the p-value for the null hypothesis where all pull parameters vanish simultaneously.

As an illustration in a simple case, one can compute the pulls associated with the average of n measurements, introducing the modified test statistic compared to Eq. (44):

$$T(\mu, p_{X_m}) = (X - \mu U - P_m)^T \bar{W}(X - \mu U - P_m) \tag{65}$$

corresponding to the case with only one parameter $\chi = \mu$, $a = U$, $b = 0$. The minimisation with respect to both parameters yields an estimator of of the pull parameter in this particular case

$$\hat{p}_{X_m} = \sum_i y_i^{(m)} X_i \qquad y_i^{(m)} = \sqrt{\bar{W}_{mm}} \frac{\left(\sum_j \bar{W}_{mj}\right)\left(\sum_j \bar{W}_{ji}\right) - \bar{W}_{mi}\sum_{jl}\bar{W}_{jl}}{\left(\sum_j \bar{W}_{mj}\right)^2 - \bar{W}_{mm}\sum_{jl}\bar{W}_{jl}} \tag{66}$$

allowing a propagation of errors in a similar way to the average of several measurements discussed in secs. 5.2.1 and 5.2.2. Numerical examples are presented in Sec. 6.

5.6 Conclusions of the multi-dimensional case

We have discussed several situations where a multi-dimensional approach is needed in phenomenology analysis. In addition to the issues already encountered in one dimension, a further arbitrary choice must be performed in the multi-dimensional case for nuisance and external approaches concerning the shape of the volume in which the biases are varied: two simple cases are given by the hypercube and the hyperball, corresponding respectively to the well-known linear and quadratic combination of uncertainties. We have then discussed how to average two (or several) measurements, emphasising the case of the nuisance approach. We have finally illustrated how a fit could be performed in order to determine confidence regions. Beyond the metrology of the model, we can also determine the agreement between model and experiments thanks to the pull parameters associated with each observable.

The uni-dimensional case (stationarity of the quadratic test statistic under minimisation, coverage properties) has led us to prefer the adaptive nuisance approach, even though the fixed nuisance approach could also be considered. In the multidimensional case, the hyperball in conjunction with the quadratic test statistic allows us to keep associativity when performing averages, so that it is

Reference		N_f	Mean	Stat	Theo
ETMC10	[41]	2	0.532	$\pm$ 0.019	$\pm 0.003 \pm 0.007 \pm 0.003 \pm 0.008 \pm 0.005$
LVdW11	[42]	2+1	0.5572	$\pm$ 0.0028	$\pm 0.0045 \pm 0.0033 \pm 0.0039 \pm 0.0006 \pm 0.0134$
BMW11	[43]	2+1	0.5644	$\pm$ 0.0059	$\pm 0.0022 \pm 0.0008 \pm 0.0006 \pm 0.0006 \pm 0.0002 \pm 0.0056$
RBC-UKQCD12	[44]	2+1	0.554	$\pm$ 0.008	$\pm 0.007 \pm 0.003 \pm 0.012$
SWME14	[45]	2+1	0.5388	$\pm$ 0.0034	$\pm 0.0237 \pm 0.0048 \pm 0.0005 \pm 0.0108 \pm 0.0022 \pm 0.0016 \pm 0.0005$

Method	Average	1 σ CI	2 σ CI	3 σ CI	5 σ CI
nG	$0.5577 \pm 0.0063 \pm 0$	0.5577 ± 0.0063	0.5577 ± 0.0126	0.5577 ± 0.0189	0.5577 ± 0.0315
naive Rfit	$0.5562 \pm 0.0120 \pm 0.0018$	0.5562 ± 0.0138	0.5562 ± 0.0258	0.5562 ± 0.0379	0.5562 ± 0.0619
educ Rfit	$0.5562 \pm 0.0020 \pm 0.0100$	0.5562 ± 0.0120	0.5562 ± 0.0139	0.5562 ± 0.0159	0.5562 ± 0.0198
1-hypercube	$0.5577 \pm 0.0038 \pm 0.0176$	0.5577 ± 0.0193	0.5577 ± 0.0240	0.5577 ± 0.0281	0.5577 ± 0.0360
adapt hyperball	$0.5577 \pm 0.0038 \pm 0.0050$	0.5577 ± 0.0068	0.5577 ± 0.0165	0.5577 ± 0.0257	0.5577 ± 0.0436

Pull	nG	(e)Rfit	1-hypercube	adaptive hyperball
ETMC10	$-1.22 \pm 1.04 \pm 0$ (1.2σ)	(0.0σ)	$-1.22 \pm 0.85 \pm 1.88$ (0.3σ)	$-1.22 \pm 0.85 \pm 0.60$ (1.1σ)
LVdW11	$-0.04 \pm 1.10 \pm 0$ (0.0σ)	(0.0σ)	$-0.04 \pm 0.35 \pm 2.71$ (0.0σ)	$-0.04 \pm 0.35 \pm 1.04$ (0.1σ)
BMW11	$1.74 \pm 1.49 \pm 0$ (1.2σ)	(0.0σ)	$1.74 \pm 0.86 \pm 4.32$ (0.0σ)	$1.74 \pm 0.86 \pm 1.21$ (1.0σ)
RBC-UKQCD12	$-0.27 \pm 1.08 \pm 0$ (0.2σ)	(0.0σ)	$-0.27 \pm 0.55 \pm 2.38$ (0.0σ)	$-0.27 \pm 0.56 \pm 0.93$ (0.4σ)
SWME14	$-0.75 \pm 1.03 \pm 0$ (0.7σ)	(0.0σ)	$-0.75 \pm 0.19 \pm 2.24$ (0.0σ)	$-0.75 \pm 0.19 \pm 1.01$ (0.7σ)

Table 6: Top: Lattice determinations of the kaon bag parameter $B_K^{\overline{\rm MS}}(2\text{GeV})$. Middle: Averages according to the various methods, and corresponding confidence intervals for various significances. Bottom: Pulls associated to each measurement for each method. For Rfit methods, we quote only the significance of the pull, whereas other methods yield the pull parameter as well as the pull itself under the form $p \pm \sigma \pm \Delta$ (significance of the pull).

rigourously equivalent from the statistical point of view to keep several measurements of a given observable or to average them in a single value. We have also been able to discuss theoretical correlations using the hyperball case at two different stages: including the correlations among observables in the domain of variations of the biases when computing the errors Δ, and providing a meaningful definition for the theoretical correlation among parameters of the fit. We have not found a way to keep these properties in the case of the hypercube. Moreover, choosing the hypercube may favour best-fit configurations where all the biases are at the border of their allowed regions, whereas the hyperball prevents such 'fine-tuned' solutions from occurring.

For comparison, in the following we will focus on two nuisance approaches: fixed 1-hypercube and adaptive hyperball with a preference for the latter. The other combinations would yield far too conservative (adaptive hypercube) or too liberal (fixed 1-hyperball) ranges of variations for the biases.

6 CKM-related examples

We will now consider the differences between the various approaches considered using several examples from quark flavour physics. These examples will be only for illustrative purposes, and we refer the reader to other works [15, 16, 22, 35] for a more thorough discussion of the physics and the inputs involved. From the previous discussion, we could consider a large set of approaches for theoretical uncertainties.

We will restrict to a few cases compared to the previous sections. First, we will consider educated Rfit (Rfit with specific treatment of uncertainties for averages), as used by the CKMfitter analyses and described in Sec. 5.2.4, while the naive Rfit approach will only be shown for the sake of comparison and is not understood as an appropriate model. We will also consider two nuisance approaches,

Reference		N_f	Mean	Stat	Theo
ETMC09	[46]	2	244	$\pm$ 3	$\pm 2 \pm 7$
HPQCD10	[47]	2+1	248.0	$\pm$ 1.4	$\pm 0.4 \pm 1.4 \pm 1.0 \pm 0.8 \pm 0.3 \pm 0.3 \pm 0.3$
FNAL-MILC11	[48]	2+1	260.1	$\pm$ 8.9	$\pm 2.2 \pm 1.6 \pm 1.0 \pm 1.4 \pm 2.8 \pm 2.0 \pm 3.4 \pm 1.8$
FNAL-MILC14	[49]	2+1+1	248.8	$\pm$ 0.3	$\pm 1.2 \pm 0.2 \pm 0.1 \pm 0.4$
ETMC14	[50]	2+1+1	247.2	$\pm$ 3.9	$\pm 0.7 \pm 1.2 \pm 0.3$

Method	Average	1 σ CI	2 σ CI	3 σ CI	5 σ CI
nG	$248.5 \pm 1.1 \pm 0$	248.5 ± 1.1	248.5 ± 2.2	248.5 ± 3.3	248.5 ± 5.5
naive Rfit	$248.1 \pm 0.9 \pm 1.3$	248.1 ± 2.2	248.1 ± 3.1	248.1 ± 4.1	248.1 ± 5.9
educ Rfit	$248.1 \pm 0.3 \pm 1.9$	248.1 ± 2.2	248.1 ± 2.5	248.1 ± 2.8	248.1 ± 3.4
1-hypercube	$248.5 \pm 0.5 \pm 2.7$	248.5 ± 3.0	248.5 ± 3.5	248.5 ± 4.0	248.5 ± 5.0
adapt hyperball	$248.5 \pm 0.5 \pm 1.0$	248.5 ± 1.2	248.5 ± 2.8	248.5 ± 4.3	248.5 ± 7.2

Pull	nG	(e)Rfit	1-hypercube	adaptive hyperball
ETMC09	$-0.59 \pm 1.01 \pm 0$ (0.6σ)	(0.0σ)	$-0.59 \pm 0.39 \pm 1.47$ (0.0σ)	$-0.59 \pm 0.39 \pm 0.93$ (0.6σ)
HPQCD10	$-0.28 \pm 1.12 \pm 0$ (0.3σ)	(0.0σ)	$-0.28 \pm 0.60 \pm 2.77$ (0.0σ)	$-0.28 \pm 0.60 \pm 0.95$ (0.4σ)
FNAL-MILC11	$1.08 \pm 1.00 \pm 0$ (1.1σ)	(0.0σ)	$1.08 \pm 0.82 \pm 1.74$ (0.3σ)	$1.08 \pm 0.83 \pm 0.57$ (1.0σ)
FNAL-MILC14	$0.63 \pm 1.82 \pm 0$ (0.3σ)	(0.0σ)	$0.63 \pm 1.05 \pm 4.97$ (0.0σ)	$0.63 \pm 1.05 \pm 1.48$ (0.5σ)
ETMC14	$-0.35 \pm 1.04 \pm 0$ (0.3σ)	(0.0σ)	$-0.35 \pm 0.94 \pm 1.20$ (0.2σ)	$-0.35 \pm 0.94 \pm 0.43$ (0.4σ)

Table 7: Top: Lattice determinations of the D_s-meson decay constant f_{D_s} (in MeV). Middle: Averages according to the various methods, and corresponding confidence intervals for various significances. Bottom: Pull associated to each measurement for each method. For Rfit methods, we quote only the significance of the pull, whereas other methods yield the pull parameter as well as the pull itself under the form $p \pm \sigma \pm \Delta$ (significance of the pull).

namely the adaptive hyperball and the 1-hypercube cases. Our examples will be chosen in the context of CKM fits, and correspond approximately to the situation for Summer 2014 conferences. However, for pedagogical purposes, we have simplified intentionally some of the inputs compared to actual phenomenological analyses performed in flavour physics [35].

6.1 Averaging theory-dominated measurements

We start by illustrating the case of measurements dominated by theoretical uncertainties, which is the case for the lattice determinations. We take the case of B_K, which is needed to discuss $K\bar{K}$ mixing, and has been the subject of important debates concerning its agreement (or not) with the rest of the global fit. We have selected a particular list of lattice determinations given in Tab. 6 (top). For each measurement, we have kept the various theoretical uncertainties separate, since their combination (linear or quadratic) depends on the method used. For purposes of illustration, we perform an average over measurements performed with different lattice gauge actions, we symmetrise the results having asymmetric uncertainties [16] and we neglect all correlations. We stress that this is done only for purposes of illustration, and that an extended list of lattice QCD results with asymmetric uncertainties and correlations will be taken into account in forthcoming phenomenological applications [35].

The results for each method are given in Tab. 6 (middle). The first column corresponds to the outcome of the averaging procedure. In all the approaches considered, we can split statistical and theoretical uncertainties. In the case of naive Rfit, one combines the measurements by adding the well statistic corresponding to each measurement: the resulting test statistic T is a well with a bottom, the width of which can be interpreted as a theoretical uncertainty, whereas the width at $T_{\min} + 1$

[16] In this section, we will not deal with asymmetric uncertainties, and for illustrative purpose, we symmetrise all uncertainties, statistical and theoretical, following Eq. (105).

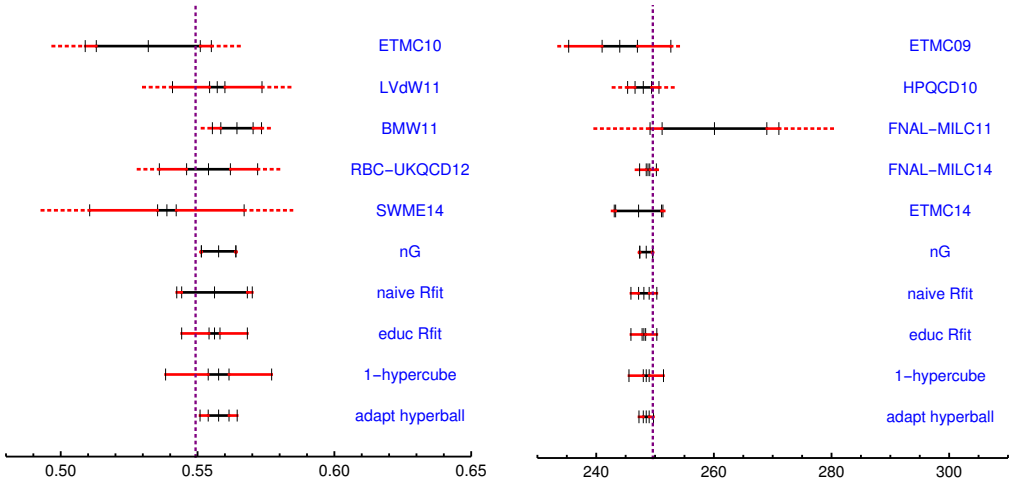


Figure 7: Left: Inputs for $B_K^{\overline{\rm MS}}(2\text{GeV})$ and the averages resulting from the different models considered here. Right: Same for the lattice determinations of the D_s-meson decay constant (in MeV). The black range gives the statistical error. For each individual input, the solid red range indicates the 1σ interval according to the adaptive hyperball approach (combining theoretical errors in quadratically) and the dashed red range according to the 1-fixed hypercube approach (combining theoretical errors linearly). For average according to the different approaches, the black range corresponds again to the statistical error, whereas the red range corresponds to the 1σ interval following the corresponding approach. The comparison between black and red ranges illustrates the relative importance of statistical and theoretical errors. Finally, for illustrative purposes, the vertical purple line gives the arithmetic average of the inputs (same weight for all central values).

determines the statistical uncertainty [17]. The case of educated Rfit was described in Sec. 5.2.4. The confidence intervals are obtained from the p-value determined from the "average" column.

We compute the pulls in the same way in both cases, interpreting the difference of T_{min} with and without the observables as a random variable distributed according to a χ^2 law with $N_{dof} = 1$. The propagation of uncertainties for the quadratic statistic was detailed in Secs. 5.2.1 and 5.2.2 where the separate extraction of statistical and theoretical uncertainties was described. The tables are obtained by plugging the average into the 1-dimensional p-value associated with the method, and reading from the p-value the corresponding confidence interval at the chosen significance. The associated pulls are given in Tab. 6 (bottom).

We present the same analysis in the case of the D_s-meson decay constant f_{D_s} in Tab. 7 (with the same caveat concerning the selected inputs, asymmetries and correlations), while graphical comparisons of the different averages in both cases can be seen at 1σ in Fig. 7 (a similar plot at 3σ is given in Fig. 12 in App. E).

For both quantities B_K and f_{D_s} at large confidence level (3 σ and above), the most conservative method is the adaptive hyperball nuisance approach, whereas the one leading to the smallest uncertainties is the educated Rfit approach. Below 3 σ, the 1-hypercube approach is more conservative than the adaptive hyperball nuisance approach, and it becomes less conservative above that threshold. The most important differences are observed at large CL/significance. The statistical uncertainty obtained in the nG approach is by construction identical to the combination in quadrature of the statistical and theoretical uncertainties obtained in the adaptive hyperball approach. However, one can notice that the confidence intervals for high significances in the two approaches are different, with nG being less conservative. The overall very good agreement of lattice determinations means

[17] In general, for naive Rfit, the tails of the resulting test statistic T are neither Gaussian nor symmetric. However, our approximation is valid to a good accuracy for our illustrative purposes and the examples discussed in this section.

Reference		Mean	Stat ($\times 10^{-3}$)	Theo ($\times 10^{-3}$)
ALEPH-j & s	[52]	0.1224	$\pm$ 0.9 $\pm$ 0.9 $\pm$ 1.2	$\pm$ 3.5
OPAL-j & s	[53]	0.1189	$\pm$ 0.8 $\pm$ 1.6 $\pm$ 1.0	$\pm$ 3.6
JADE-j & s	[54]	0.1172	$\pm$ 0.6 $\pm$ 2.0 $\pm$ 3.5	$\pm$ 3.0
Dissertori-3j	[55]	0.1175	$\pm$ 2.0	$\pm$ 1.5
JADE-3j	[56]	0.1199	$\pm$ 1.0 $\pm$ 2.1 $\pm$ 5.4	$\pm$ 0.7
BS-T	[57]	0.1172	$\pm$ 1.0 $\pm$ 0.8	$\pm$ 1.2 $\pm$ 1.2
DW-T	[58]	0.1165	$\pm$ 2.2	$\pm$ 1.7
AFHMS-T	[59]	0.1135	$\pm$ 0.2	$\pm$ 0.5 $\pm$ 0.9
GLM-T	[60]	0.1134	$\pm$ 2.5	$\pm$ 0.6
HKMS-C	[61]	0.1123	$\pm$ 0.2	$\pm$ 0.7 $\pm$ 1.4

Method	Average	1 σ CI	2 σ CI	3 σ CI	5 σ CI
nG	$0.1143 \pm 0.0010 \pm 0$	0.1143 ± 0.0010	0.1143 ± 0.0020	0.1143 ± 0.0030	0.1143 ± 0.0050
naive Rfit	$0.1145 \pm 0.0002 \pm 0$	0.1145 ± 0.0002	0.1145 ± 0.0004	0.1145 ± 0.0006	0.1145 ± 0.0011
educ Rfit	$0.1145 \pm 0.0001 \pm 0.0006$	0.1145 ± 0.0007	0.1145 ± 0.0009	0.1145 ± 0.0010	0.1145 ± 0.0013
1-hypercube	$0.1143 \pm 0.0005 \pm 0.0018$	0.1143 ± 0.0020	0.1143 ± 0.0026	0.1143 ± 0.0031	0.1143 ± 0.0041
adapt hyperball	$0.1143 \pm 0.0005 \pm 0.0009$	0.1143 ± 0.0011	0.1143 ± 0.0026	0.1143 ± 0.0039	0.1143 ± 0.0067

Pull	nG	(e)Rfit	1-hypercube	adaptive hyperball
ALEPH-j & s	$1.30 \pm 0.69 \pm 0$ (1.9σ)	(2.5σ)	$1.30 \pm 0.26 \pm 0.91$ (1.8σ)	$1.30 \pm 0.26 \pm 0.63$ (1.6σ)
OPAL-j & s	$0.93 \pm 0.69 \pm 0$ (1.3σ)	(0.4σ)	$0.93 \pm 0.29 \pm 0.89$ (0.7σ)	$0.93 \pm 0.29 \pm 0.63$ (1.2σ)
JADE-j & s	$0.76 \pm 0.79 \pm 0$ (0.9σ)	(0.0σ)	$0.76 \pm 0.55 \pm 0.84$ (0.6σ)	$0.76 \pm 0.55 \pm 0.57$ (0.9σ)
Dissertori-3j	$1.13 \pm 0.77 \pm 0$ (1.4σ)	(0.9σ)	$1.13 \pm 0.58 \pm 0.95$ (0.9σ)	$1.13 \pm 0.58 \pm 0.51$ (1.3σ)
JADE-3j	$1.10 \pm 1.00 \pm 0$ (1.1σ)	(0.8σ)	$1.10 \pm 0.98 \pm 0.46$ (1.0σ)	$1.10 \pm 0.98 \pm 0.22$ (1.1σ)
BS-T	$0.36 \pm 0.92 \pm 0$ (0.4σ)	(0.2σ)	$0.36 \pm 0.88 \pm 0.41$ (0.4σ)	$0.36 \pm 0.88 \pm 0.26$ (0.4σ)
DW-T	$0.15 \pm 0.97 \pm 0$ (0.2σ)	(0.1σ)	$0.15 \pm 0.96 \pm 0.18$ (0.2σ)	$0.15 \pm 0.96 \pm 0.10$ (0.2σ)
AFHMS-T	$-0.24 \pm 0.78 \pm 0$ (0.3σ)	(0.0σ)	$-0.24 \pm 0.57 \pm 1.00$ (0.1σ)	$-0.24 \pm 0.57 \pm 0.53$ (0.4σ)
GLM-T	$-0.29 \pm 0.95 \pm 0$ (0.3σ)	(0.2σ)	$-0.28 \pm 0.88 \pm 0.73$ (0.2σ)	$-0.28 \pm 0.88 \pm 0.36$ (0.3σ)
HKMS-C	$-2.27 \pm 1.35 \pm 0$ (1.7σ)	(1.4σ)	$-2.27 \pm 0.72 \pm 2.27$ (0.7σ)	$-2.27 \pm 0.72 \pm 1.14$ (1.4σ)

Table 8: Top: Determinations of $\alpha_S(M_Z)$ using e^+e^- annihilation, taken from Ref. [34]. Middle: Averages for $\alpha_S(M_Z)$ from e^+e^- annihilation according to the various methods, and corresponding confidence intervals for various significances. Bottom: Pull associated to each measurement for each method. For Rfit methods, we quote only the significance of the pull, whereas other methods yield the pull parameter as well as the pull itself under the form $p \pm \sigma \pm \Delta$.

vanishing pulls for Rfit methods (since all the wells have a common bottom with a vanishing $T_{\min}$). For the other methods, the pull parameter has statistical and theoretical errors of similar size in the adaptive hyperball case, whereas theoretical errors tend to dominate in the 1-hypercube method. This yields smaller pulls in the latter approach.

A last illustration, which does not come solely from lattice simulations, is provided by the determination the strong coupling constant $\alpha_S(M_Z)$. The subject is covered extensively by recent reviews [34, 51], and we stress that we do not claim to provide an accurate alternative average to these reviews which requires a careful assessment of the various determinations and their correlations. As a purely illustrative example, we will focus on the average of determinations from e^+e^- annihilation under a set of simplistic hypotheses for the separation between statistical and theoretical uncertainties. In order to allow for a closer comparison with Refs. [34,62], we try to assess correlations this time. We assume that theoretical uncertainties for the same set of observables ($j\&s$, $3j$, T), but from different experiments, are 100% correlated, and the statistical uncertainties for determinations from similar experimental data are 100% correlated (BS-T, DW-T, AFHMS-T) [18]

[18] In addition, we have made further choices concerning the separation of statistical and theoretical uncertainties based

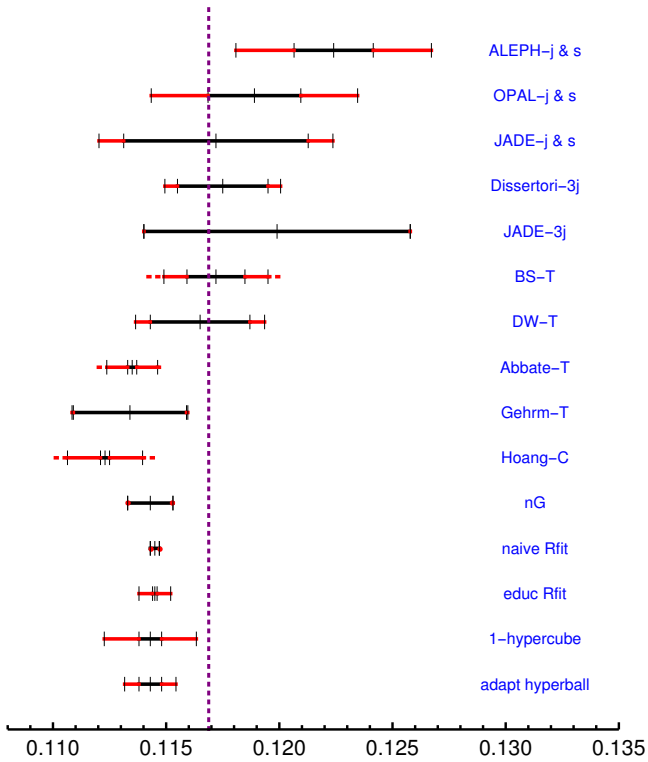


Figure 8: Determinations of the strong coupling constant at M_Z through e^+e^- annihilation, and the averages resulting from the different models considered here. The intervals are given at 1 σ. See Fig. 7 for the legend.

We perform the average in the different cases considered, see Tab. 8 (middle), which are represented graphically in Fig. 8 (a similar plot at 3σ is given in Fig. 13 in App. E). We notice that the various approaches yield results with similar central values to the nG case. The pulls for individual quantities are mostly around 1 σ, and they are smaller in the adaptive hyperball approach compared to the nG one, showing better consistency. Refs. [34,62] take a different approach, "range averaging", which amounts to considering the spread of the central values for the various determinations, leading to $\alpha_S(M_Z) = 0.1174 \pm 0.0051$ for the determination from e^+e^- annihilation data considered here [62]. This approach is motivated in Ref. [34] by the complicated pattern of correlations and the limited compatibility between some of the inputs and, more importantly, it does not take into account that the different determinations have different accuracies according to the uncertainties quoted. The approach in Refs. [34, 62] conservatively accounts for the possibility that some uncertainties are underestimated. On the contrary, our averages given in Tab. 8 and Fig. 8 assume that all the inputs should be taken into account and averaged according to the uncertainties given in the original articles. The difference in the underlying hypotheses for the averages explain the large difference observed between our results and the ones in Refs. [34, 62]. Note however that our numerics directly follow from the use of the different averaging methods, and lack the necessary critical assessment of the individual determinations of $\alpha_S(m_Z)$ performed in Refs. [34, 62].

on the following considerations. Ref. [60] discusses the sources of uncertainties (scales, function parameters, b-quark mass) within a fit leading to uncertainties assumed to be of statistical nature, with a further systematic uncertainty coming from the difference between the two different schemes. The systematic uncertainties in Ref. [57] are assumed to be of statistical nature in the absence of any opposite statement. For the first two classes (j & s and 3j) hadronisation is taken into account by Monte Carlo methods, while for the last two classes (T and C) analytic analyses are made: in the former (latter) case, the hadronic uncertainties are treated as statistical (theoretical).

Reference		Mean	Stat	Theo
Exclusive	CKMfitter Summer 14	3.28	$\pm$ 0.15	$\pm$ 0.26
Inclusive	CKMfitter Summer 14	4.359	$\pm$ 0.180	$\pm 0.013 \pm 0.027 \pm 0.037 \pm 0.161 \pm 0.200$

Method	Average	1 σ CI	2 σ CI	3 σ CI	5 σ CI
nG	$3.79 \pm 0.22 \pm 0$	3.79 ± 0.22	3.79 ± 0.44	3.79 ± 0.65	3.79 ± 1.1
naive Rfit	$3.70 \pm 0.12 \pm 0$	3.70 ± 0.12	3.70 ± 0.23	3.70 ± 0.35	3.70 ± 0.58
educ Rfit	$3.70 \pm 0.11 \pm 0.26$	3.70 ± 0.38	3.70 ± 0.49	3.70 ± 0.61	3.70 ± 0.84
1-hypercube	$3.79 \pm 0.12 \pm 0.34$	3.79 ± 0.40	3.79 ± 0.54	3.79 ± 0.67	3.79 ± 0.91
adapt hyperball	$3.79 \pm 0.12 \pm 0.18$	3.79 ± 0.24	3.79 ± 0.57	3.79 ± 0.88	3.79 ± 1.49

Pull	nG	(e)Rfit	1-hypercube	adaptive hyperball
Exclusive	$-3.60 \pm 1.46 \pm 0$ (2.5σ)	(1.6σ)	$-3.60 \pm 0.78 \pm 2.31$ (1.9σ)	$-3.60 \pm 0.78 \pm 1.23$ (1.9σ)
Inclusive	$3.40 \pm 1.38 \pm 0$ (2.5σ)	(1.6σ)	$3.40 \pm 0.74 \pm 2.20$ (1.9σ)	$3.40 \pm 0.74 \pm 1.16$ (1.9σ)

Table 9: Top: Determinations of $|V_{ub}| \cdot 10^3$ from semileptonic decays. Middle: Averages according to the various methods, and corresponding confidence intervals for various significances. Bottom: Pulls associated to each determination for each method. For Rfit methods, we quote only the significance of the pull, whereas other methods yield the pull parameter as well as the pull itself under the form $p \pm \sigma \pm \Delta$ (significance of the pull).

6.2 Averaging incompatible or barely compatible measurements

Another important issue occurs when one wants to combine barely compatible measurements. This is for instance the case for $|V_{ub}|$ and $|V_{cb}|$ from semileptonic decays, where inclusive and exclusive determinations are not in very good agreement. The list of determinations used for illustrative purposes and the results for each method are given in Tabs. 9 and 10, together with the corresponding graphical comparisons in Fig. 9 (a similar plot at 3σ is given in Fig. 14 in App. E). Our inputs are slightly different from Ref. [36] for several reasons. The inclusive determination of $|V_{ub}|$ corresponds to the BLNP approach [64], and we consider the theoretical uncertainties from shape functions (leading and subleading), weak annihilation, and heavy-quark expansion uncertainties on matching and m_b. We use only branching fractions measured for $B \to \pi\ell\nu$ and average the unquenched lattice calculations quoted in Ref. [36]. For $|V_{cb}|$ exclusive we also split the various sources of theoretical uncertainties coming from the determination of the form factors. We assume that there are no correlations among all these uncertainties.

The lack of compatibility between the two types of determination means in particular that the naive Rfit combined likelihood has not flat bottom, and thus no theoretical uncertainty. This behaviour was one of the reasons to propose the educated Rfit approach, where the theoretical uncertainty of the combination cannot be smaller than any of the individual measurements.

The same pattern of conservative and aggressive approaches can be observed, with a fairly good agreement at 3 σ level (apart from the naive Rfit approach, already discussed). At 5 σ, the adaptive hyperball proves again rather conservative, even though the theoretical error of the averages are smaller than the 1-hypercube nuisance and the educated Rfit approaches. The analysis of the pulls yields similar conclusions, with discrepancies at the 2 σ for $|V_{ub}|$ and between 2 and 3 σ for $|V_{cb}|$. Once again, theoretical errors for the pull parameters are larger in the 1-hypercube approach than in the adaptive hyperball case. Let us also notice that in both cases, there are only two quantities to combine, so that the two pull parameters are by construction opposite to each other up to an irrelevant scaling factor, leading to the same pull for both quantities.

Reference		Mean	Stat	Theo
Exclusive	CKMfitter Summer 14	38.99	± 0.49	$\pm 0.04 \pm 0.21 \pm 0.13 \pm 0.39 \pm 0.17 \pm 0.04 \pm 0.19$
Inclusive	CKMfitter Summer 14	42.42	± 0.44	± 0.74

Method	Average	1 σ CI	2 σ CI	3 σ CI	5 σ CI
nG	$40.41 \pm 0.55 \pm 0$	40.41 ± 0.55	40.41 ± 1.11	40.41 ± 1.66	40.41 ± 2.77
naive Rfit	$41.00 \pm 0.33 \pm 0$	41.00 ± 0.32	41.00 ± 0.65	41.00 ± 0.98	41.00 ± 1.64
educ Rfit	$41.00 \pm 0.33 \pm 0.74$	41.00 ± 1.07	41.00 ± 1.39	41.00 ± 1.72	41.00 ± 2.38
1-hypercube	$40.41 \pm 0.34 \pm 0.99$	40.41 ± 1.15	40.41 ± 1.57	40.41 ± 1.94	40.41 ± 2.65
adapt hyperball	$40.41 \pm 0.34 \pm 0.44$	40.41 ± 0.60	40.41 ± 1.45	40.41 ± 2.26	40.41 ± 3.84

Pull	nG	(e)Rfit	1-hypercube	adaptive hyperball
Exclusive	$-4.75 \pm 1.56 \pm 0$ (3.1σ)	(2.3σ)	$-4.75 \pm 0.91 \pm 2.65$ (2.6σ)	$-4.75 \pm 0.91 \pm 1.26$ (2.3σ)
Inclusive	$3.98 \pm 1.30 \pm 0$ (3.1σ)	(2.3σ)	$3.98 \pm 0.77 \pm 2.22$ (2.6σ)	$3.98 \pm 0.77 \pm 0.74$ (2.3σ)

Table 10: Top: Determinations of $|V_{cb}| \cdot 10^3$ from semileptonic decays. Middle: Averages according to the various methods, and corresponding confidence intervals for various significances. Bottom: Pulls associated to each determination for each method. For Rfit methods, we quote only the significance of the pull, whereas other methods yield the pull parameter as well as the pull itself under the form $p \pm \sigma \pm \Delta$ (significance of the pull).

6.3 Averaging quantities dominated by different types of uncertainties

In order to illustrate the role played by statistical and theoretical uncertainties, we consider the question of averaging quantities dominated by one or the other. This happens for instance when one wants to compare a theoretically clean determination with other determination potentially affected by large theoretical uncertainties. This situation occurs in flavour physics for instance when one compares the extraction of $\sin(2\beta)$ from time-dependent asymmetries in $b \to c\bar{c}s$ and $b \to q\bar{q}s$ decays (let us recall that for the CKM global fit, only charmonium input is used for $\sin(2\beta)$). The first have a very small penguin pollution, which we will neglect, whereas the latter is significantly affected by such a pollution. The corresponding estimates of $\sin(2\beta)$ have large theoretical uncertainties, and for illustration we use the computation done in Ref. [63].

The results are collected in Tab. 11, which were computed neglecting all possible correlations between the different extractions. One can see that the resulting theoretical uncertainty from the combination of the various inputs remains small, so that most of the approaches yield a very similar result for the confidence intervals. The corresponding pulls show a global consistency concerning the observables that deviate by 1σ.

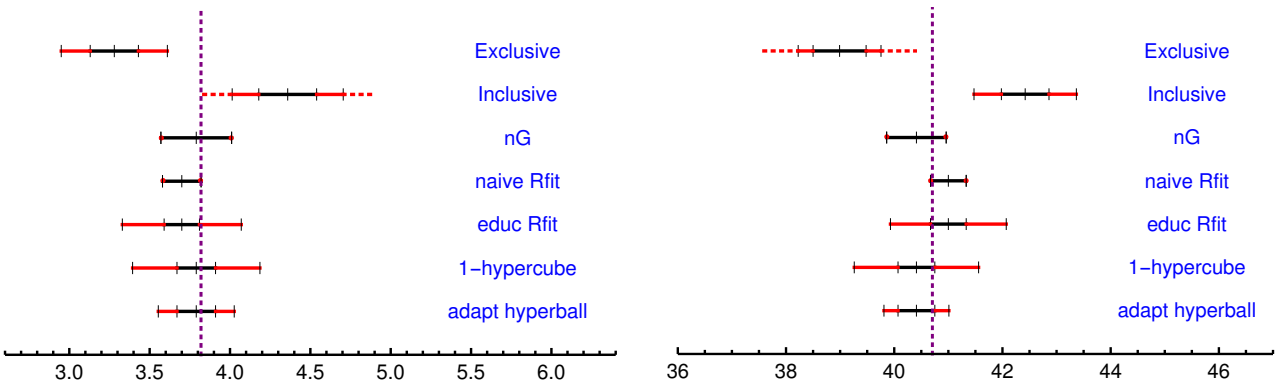


Figure 9: Left: Inclusive and exclusive inputs for the CKM matrix element $|V_{ub}|$ (times 10^3) and the averages resulting from the different models considered here. Right: Same for the determinations of $|V_{cb}|$ (times 10^3) CKM matrix element. The intervals are given at 1 σ. See Fig. 7 for the legend.

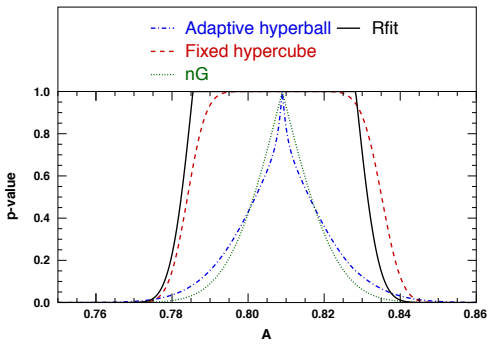
	$\sin(2\beta_{\rm eff})$	$\Delta S = \sin(2\beta_{\rm eff}) - \sin(2\beta)$	$\sin(2\beta)$
$\pi^0 K_S$	$0.57 \pm 0.17 \pm 0$	$0.085 \pm 0 \pm 0.065$	$0.485 \pm 0.17 \pm 0.065$
$\rho^0 K_S$	$0.525 \pm 0.195 \pm 0$	$-0.135 \pm 0 \pm 0.155$	$0.66 \pm 0.195 \pm 0.155$
$\eta' K_S$	$0.63 \pm 0.06 \pm 0$	$0.015 \pm 0 \pm 0.015$	$0.615 \pm 0.06 \pm 0.015$
ϕK_S	$0.73 \pm 0.12 \pm 0$	$0.03 \pm 0 \pm 0.02$	$0.7 \pm 0.12 \pm 0.02$
ωK_S	$0.71 \pm 0.21 \pm 0$	$0.11 \pm 0 \pm 0.10$	$0.6 \pm 0.21 \pm 0.10$
$(c\bar{c})K_S$	0.689 ± 0.018	0	$0.689 \pm 0.018 \pm 0$

Method	Average	1 σ CI	2 σ CI	3 σ CI	5 σ CI
nG	$0.681 \pm 0.017 \pm 0$	0.681 ± 0.017	0.681 ± 0.034	0.681 ± 0.051	0.681 ± 0.085
naive Rfit	$0.683 \pm 0.017 \pm 0$	0.683 ± 0.017	0.683 ± 0.034	0.683 ± 0.051	0.683 ± 0.085
educ Rfit	$0.683 \pm 0.017 \pm 0.$	0.683 ± 0.017	0.683 ± 0.034	0.683 ± 0.051	0.683 ± 0.084
1-hypercube	$0.681 \pm 0.017 \pm 0.003$	0.681 ± 0.017	0.681 ± 0.034	0.681 ± 0.052	0.681 ± 0.086
adapt hyperball	$0.681 \pm 0.017 \pm 0.002$	0.681 ± 0.017	0.681 ± 0.034	0.681 ± 0.052	0.681 ± 0.090

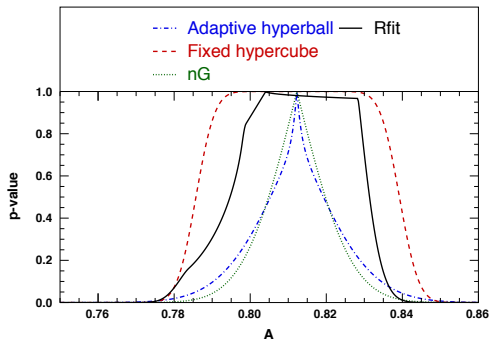
Pull	nG	(e)Rfit	1-hypercube	adaptive hyperball
$\pi^0 K_S$	$-1.09 \pm 1.00 \pm 0$ (1.1σ)	(0.8σ)	$-1.09 \pm 0.94 \pm 0.37$ (1.1σ)	$-1.09 \pm 0.94 \pm 0.36$ (1.1σ)
$\rho^0 K_S$	$-0.09 \pm 1.00 \pm 0$ (0.1σ)	(0.0σ)	$-0.09 \pm 0.79 \pm 0.63$ (0.1σ)	$-0.09 \pm 0.79 \pm 0.62$ (0.1σ)
$\eta' K_S$	$-1.16 \pm 1.04 \pm 0$ (1.1σ)	(0.9σ)	$-1.16 \pm 1.01 \pm 0.28$ (1.1σ)	$-1.16 \pm 1.01 \pm 0.24$ (1.1σ)
ϕK_S	$0.16 \pm 1.01 \pm 0$ (0.1σ)	(0.0σ)	$0.16 \pm 1.00 \pm 0.19$ (0.2σ)	$0.16 \pm 1.00 \pm 0.17$ (0.2σ)
ωK_S	$-0.35 \pm 1.00 \pm 0$ (0.3σ)	(0.0σ)	$-0.35 \pm 0.91 \pm 0.44$ (0.3σ)	$-0.35 \pm 0.91 \pm 0.43$ (0.4σ)
$(c\bar{c})K_S$	$3.79 \pm 2.97 \pm 0$ (1.3σ)	(1.1σ)	$3.79 \pm 2.87 \pm 1.63$ (1.1σ)	$3.79 \pm 2.87 \pm 0.78$ (1.2σ)

Table 11: Top: Symmetrised determinations of $\sin(2\beta_{\rm eff})$ from various penguin $b \to q\bar{q}s$ modes and from charmonia modes [36], and estimate within QCD factorisation of the correction from penguin pollution in the Standard Model (symmetrised range quoted in Tab. 1 in Ref. [63]). We neglect any penguin pollution in the case of the charmonium extraction of $\sin(2\beta)$. Middle: Averages according to the various methods, and corresponding confidence intervals for various significances. Bottom: Pulls associated to each determination for each method. For Rfit methods, we quote only the significance of the pull, whereas other methods yield the pull parameter as well as the pull itself under the form $p \pm \sigma \pm \Delta$ (significance of the pull).

Method	Fit result	1 σ	2 σ	3 σ
	A	Scenario A		
nG	0.809 ± 0.011	0.809 ± 0.011	0.809 ± 0.023	0.809 ± 0.034
Rfit	0.807 ± 0.026	0.807 ± 0.026	0.807 ± 0.031	0.807 ± 0.035
1-hypercube	$0.809 \pm 0.004 \pm 0.025$	0.809 ± 0.028	0.809 ± 0.033	0.809 ± 0.037
adaptive hyperball	$0.809 \pm 0.004 \pm 0.010$	0.809 ± 0.012	0.809 ± 0.029	0.809 ± 0.043
	A	Scenario B		
nG	0.812 ± 0.011	0.812 ± 0.011	0.812 ± 0.022	0.812 ± 0.033
Rfit	$0.804^{+0.029}_{-0.014}$	$0.804^{+0.029}_{-0.014}$	$0.804^{+0.033}_{-0.025}$	$0.804^{+0.038}_{-0.030}$
1-hypercube	$0.812 \pm 0.004 \pm 0.027$	0.812 ± 0.029	0.812 ± 0.034	0.812 ± 0.038
adaptive hyperball	$0.812 \pm 0.004 \pm 0.010$	0.812 ± 0.012	0.812 ± 0.027	0.812 ± 0.042

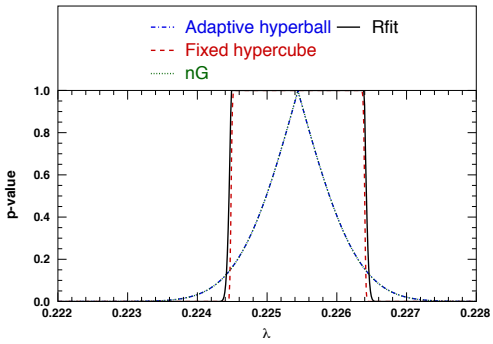


Scenario A

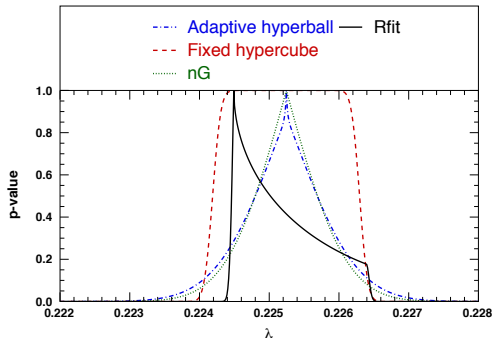


Scenario B

Method	Fit result	1 σ	2 σ	3 σ
	λ	Scenario A		
nG	0.2254 ± 0.0007	0.2254 ± 0.0007	0.225 ± 0.0013	0.2254 ± 0.0020
Rfit	0.2254 ± 0.0010	0.2254 ± 0.0010	0.2254 ± 0.0010	0.2254 ± 0.0010
1-hypercube	$0.2254 \pm 0.0000 \pm 0.0010$	0.2254 ± 0.0010	0.2254 ± 0.0010	0.2254 ± 0.0010
adaptive hyperball	$0.2254 \pm 0.0000 \pm 0.0007$	0.2254 ± 0.0007	0.2254 ± 0.0014	0.2254 ± 0.0020
	λ	Scenario B		
nG	0.2252 ± 0.0007	0.2252 ± 0.0007	0.2252 ± 0.0013	0.2252 ± 0.0020
Rfit	$0.2245^{+0.0011}_{-0.0001}$	$0.2245^{+0.0011}_{-0.0001}$	$0.2245^{+0.0020}_{-0.0001}$	$0.2245^{+0.0020}_{-0.0001}$
1-hypercube	$0.2252 \pm 0.0001 \pm 0.0011$	0.2252 ± 0.0011	0.2252 ± 0.0012	0.2252 ± 0.0013
adaptive hyperball	$0.2252 \pm 0.0001 \pm 0.0007$	0.22525 ± 0.00070	0.2252 ± 0.0015	0.2252 ± 0.0022



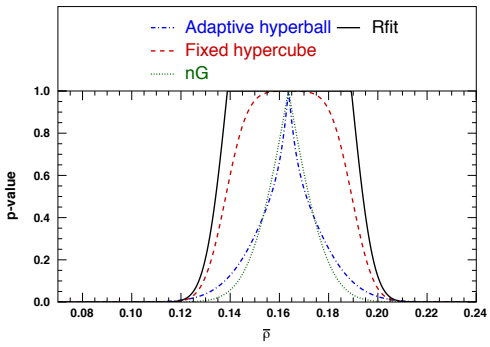
Scenario A



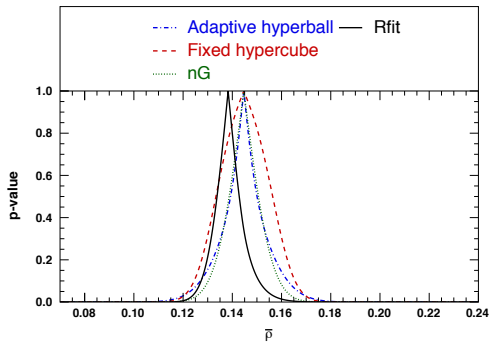
Scenario B

Table 12: Numerical results and p-values for the CKM parameters in A and λ for Scenarios A and B, depending on the method chosen. For each quantity, we provide the error budget, whenever possible, and the plots of the p-values for Scenarios A (left) and B (right).

Method	Fit result	1 σ	2 σ	3 σ
	$\bar{\rho}$	Scenario A		
nG	0.164 ± 0.012	0.164 ± 0.012	0.164 ± 0.025	0.164 ± 0.037
Rfit	0.164 ± 0.032	0.164 ± 0.032	0.164 ± 0.039	0.164 ± 0.046
1-hypercube	$0.164 \pm 0.007 \pm 0.026$	0.164 ± 0.029	0.164 ± 0.038	0.164 ± 0.045
adaptive hyperball	$0.164 \pm 0.007 \pm 0.010$	0.164 ± 0.014	0.164 ± 0.032	0.164 ± 0.051
	$\bar{\rho}$	Scenario B		
nG	0.145 ± 0.009	0.145 ± 0.009	0.145 ± 0.018	0.145 ± 0.027
Rfit	0.138 ± 0.007	0.138 ± 0.007	$0.138^{+0.016}_{-0.013}$	$0.138^{+0.028}_{-0.020}$
1-hypercube	$0.145 \pm 0.007 \pm 0.011$	0.145 ± 0.015	0.145 ± 0.024	0.145 ± 0.031
adaptive hyperball	$0.145 \pm 0.007 \pm 0.005$	0.145 ± 0.009	0.145 ± 0.023	0.145 ± 0.036

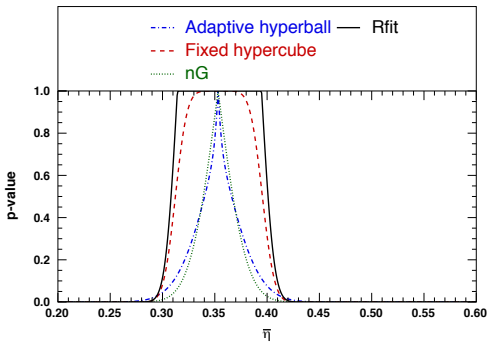


Scenario A

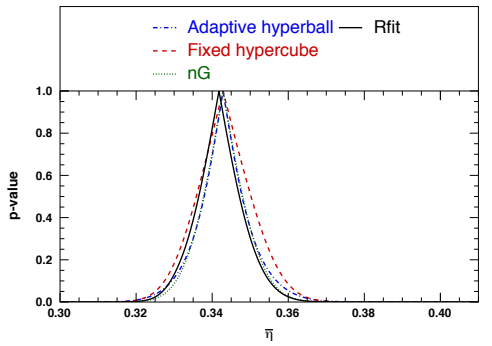


Scenario B

Method	Fit result	1 σ	2 σ	3 σ
	$\bar{\eta}$	Scenario A		
nG	0.353 ± 0.021	0.353 ± 0.021	0.353 ± 0.042	0.353 ± 0.063
Rfit	$0.354^{+0.050}_{-0.049}$	$0.354^{+0.050}_{-0.049}$	$0.354^{+0.059}_{-0.058}$	$0.354^{+0.068}_{-0.067}$
1-hypercube	$0.353 \pm 0.009 \pm 0.041$	0.353 ± 0.046	0.353 ± 0.057	0.353 ± 0.067
adaptive hyperball	$0.353 \pm 0.009 \pm 0.019$	0.353 ± 0.023	0.353 ± 0.054	0.353 ± 0.083
	$\bar{\eta}$	Scenario B		
nG	0.343 ± 0.008	0.343 ± 0.008	0.343 ± 0.016	0.343 ± 0.023
Rfit	0.342 ± 0.008	0.342 ± 0.008	$0.342^{+0.016}_{-0.015}$	$0.342^{+0.024}_{-0.022}$
1-hypercube	$0.343 \pm 0.007 \pm 0.007$	0.343 ± 0.011	0.343 ± 0.019	0.343 ± 0.027
adaptive hyperball	$0.343 \pm 0.007 \pm 0.003$	0.343 ± 0.008	0.343 ± 0.018	0.343 ± 0.028



Scenario A



Scenario B

Table 13: Numerical results and p-values for the CKM parameters in $\bar{\rho}$ and $\bar{\eta}$ for Scenarios A and B, depending on the method chosen. For each quantity, we provide the error budget, whenever possible, and the plots of the p-values for Scenarios A (left) and B (right).

Observable	Input
$\|V_{ud}\|$	$0.97425 \pm 0 \pm 0.00022$
$\|V_{ub}\|$	$(3.70 \pm 0.12 \pm 0.26) \times 10^{-3}$
$\|V_{cb}\|$	$(41.00 \pm 0.33 \pm 0.74) \times 10^{-3}$
Δm_d	(0.510 ± 0.003) ps^{-1}
Δm_s	(17.757 ± 0.021) ps^{-1}
B_s/B_d	$1.023 \pm 0.013 \pm 0.014$
B_s	$1.320 \pm 0.017 \pm 0.030$
f_{B_s}/f_{B_d}	$1.205 \pm 0.004 \pm 0.007$
f_{B_s}	$225.6 \pm 1.1 \pm 5.4$ MeV
η_B	$0.5510 \pm 0 \pm 0.0022$
$\bar{m}_t$	$165.95 \pm 0.35 \pm 0.64$ GeV
α	$(87.8 \pm 3.4)^\circ$
$\sin(2\beta)$	0.682 ± 0.019
γ	$(72.8 \pm 6.7)^\circ$

Table 14: Inputs for the theory-dominated CKM fits, inspired by the data available in Summer 2014. Scenario A is restricted to the upper part of the table, whereas Scenario B includes all inputs

6.4 Global fits

In order to illustrate the impact of the treatment of theoretical uncertainties, we consider a global fit including mainly observables that come with a theoretical uncertainty. The list of observables is given in Tab. 14. Their values are motivated by the CKMfitter inputs used in Summer 2014, but they are used only for purposes of illustration [19]. We consider two fits: Scenario A involves only constraints dominated by theoretical uncertainties whereas Scenario B includes also constraints from the angles (statistically dominated).

As far as the CKM matrix elements are concerned the Standard Model is linear but it is not linear in all the other fundamental parameters of the Standard Model. For the illustrative purposes of this note, the first step thus consists in determining the minimum of the full (non-linear) χ^2, and to linearise the Standard Model formulae for the various observables around this minimum (we choose the inputs of scenario B to determine this point): this define an exactly linear model, which at this stage should not be used for realistic phenomenology but is useful for the comparison of the methods presented here. One can use the results presented in the previous section in order to determine the p-value as a function of each of the parameters of interest. In the case of the nuisance-δ approach, we can describe this p-value using the same parameters as before, namely a central value, a statistical error and a theoretical error.

We provide the results for the 4 CKM parameters in both scenarios in Tabs. 12 and 13 (using the same linearised theory described above). We also indicate the profiles of the p-values. As before, we observe that the methods give similar results at the 2-3 σ level, although the adaptive hyperball method tends to be more conservative than the others.

7 Conclusion

A problem often encountered in particle physics consists in analysing data within the Standard Model (or some of its extensions) in order to extract information on the fundamental parameters

[19] In particular, most of the inputs have several sources of theoretical uncertainties, which should be combined together linearly or in quadrature according to the model of theoretical uncertainties chosen. Since we just want to illustrate the difference between the various approaches at the level of the fit, we take as inputs the values obtained in a given framework (Rfit) without recomputing the averages and uncertainties for each approach.

of the model. An essential role is played here by uncertainties, which can be classified in two categories, statistical and theoretical. If the former can be treated in a rigorous manner within a given statistical framework, the latter must be described through models. The problem is particularly acute in flavour physics, as theoretical uncertainties often play a central role in the determination of underlying parameters, such as the four parameters describing the CKM matrix in the Standard Model.

This article aims at describing and comparing several approaches that can be implemented in a frequentist framework. After recalling some elements of frequentist analysis, we have discussed three different approaches for theoretical uncertainties: the random-δ approach treats theoretical uncertainties as random variables, the external-δ approach considers them as external parameters leading to an infinity of p-values to be combined through model averaging, the nuisance-δ describes them through fixed biases which have to be varied over a reasonable region. These approaches have to be combined with particular choices for the test statistic used to compute the p-value. We have illustrated these approaches in the one-dimensional case, recovering the Rfit model used by CKMfitter as a particular case of the external-δ approach, and discussing the interesting alternative of a quadratic test statistic.

In the case of the nuisance-δ approach, one has to decide over which range the bias parameter should be varied. It is possible to compute the p-value by taking the supremum of the bias over a fixed range fixed by the size of the theoretical uncertainty to be modeled (fixed nuisance approach). An alluring alternative consists in adjusting the size of the range to the confidence level chosen: the range for a low confidence level can be obtained by varying the bias parameter in a small range, whereas a range for a high confidence level could require a more conservative (and thus larger) range for the bias parameter. We have designed such a scheme, called adaptive nuisance approach. It provides a unified statistical approach to deal with the metrology of the parameters (for low CL ranges) and the exclusion of models (for high CL ranges).

We have determined the p-values associated with each approach for a measurement involving both statistical and theoretical uncertainties. We have also studied the size of error bars, the significance of deviations and the coverage properties. In general, the most conservative approaches correspond to a naive Gaussian treatment (belonging to the random-δ approach) and the adaptive nuisance approach. The latter is better defined and more conservative than the former in the case where statistical and theoretical approaches are of similar size. Other approaches (fixed nuisance, external) turn out less conservative at large confidence level.

We have then considered extensions to multi-dimensional cases, focusing on the linear case where the quantity of interest is a linear combination of observables. Due to the presence of several bias parameters, one has to make another choice concerning the shape of the space over which the bias parameters are varied. Two simple examples are the hypercube and the hyperball, leading to a linear or quadratic combination of theoretical uncertainties respectively. The hypercube is more conservative, as it allows for sets of values of the bias parameters that cannot be reached within the hyperball. On the other hand, the hyperball has the great virtue of associativity, so that one can average different measurements of the same quantity or put all of them in a global fit, without changing its outcome. It also allows us to include theoretical correlations easily, both in the range of variation of biases to determine errors and in the definition of theoretical correlations for the outcome of a fit. We have discussed the average of several measurements using the various approaches, including correlations. We considered in detail the case of 100% correlations leading to a non-invertible covariance matrix. We also discussed global fits and pulls in a linearised context. We have then provided several comparisons between the different approaches using examples from flavour physics: averaging theory-dominated measurements, averaging incompatible measurements linear fits to a subset of flavour inputs.

It is now time to determine which choice seems preferable in our case. Random-δ has no strong statistical basis: its only advantage consists in its simplicity. External-δ is closer in spirit to the

determination of systematics as performed by experimentalists, but it starts with an inappropriate null hypothesis and tries to combine an infinite set of p-values in a single p-value. On the contrary, the nuisance-δ approach starts from the beginning with the correct null hypothesis and deals with a single p-value.

This choice is independent from another choice, i.e., the range of variation for the parameter δ. Indeed, when several bias parameters are involved, one may imagine different multidimensional spaces for their variations, in particular the hyperball and the hypercube. As said earlier, the hyperball has the interesting property of associativity when performing averages and avoids fine-tuned solutions where all parameters are pushed in a corner of phase space. The hypercube is closer in spirit to the Rfit model (even though the latter is not a bias model), but it cannot avoid fine-tuned situations and it does not seem well suited to deal with theoretical correlations, since it is designed from the start to avoid such correlations.

A third choice consists in determining whether one wants to keep the volume of variation fixed (fixed approach), or to modify it depending on the desired confidence level (adaptive approach). Adaptive hypercube is in principle the most conservative choice but in practice, it gives too large errors, whereas fixed hyperball would give very small errors. Fixed hypercube is more conservative at low confidence levels (large p-values), whereas adaptive hyperball is more conservative at large confidence levels (small p-values).

This overall discussion leads us to consider the nuisance approach with adaptive hyperball as a promising approach to deal with flavour physics problems, which we will investigate in more phenomenological analyses in forthcoming publications [35].

Acknowledgments

We would like to thank S. T'Jampens for collaboration at an early stage of this work, as well as all our collaborators from the CKMfitter group for many useful discussions on the statistical issues covered in this article. We would also like to express a special thanks to the Mainz Institute for Theoretical Physics (MITP) for its hospitality and support during the workshop "Fundamental parameters from lattice QCD" where part of this work was presented and discussed. LVS acknowledges financial support from the Labex P2IO (Physique des 2 Infinis et Origines). SDG acknowledges partial support from Contract FPA2014-61478-EXP. This project has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreements No 690575, No 674896 and No. 692194.

A Singular covariance matrices

A.1 Inversion of the covariance matrix

In Sec. 5.2.2, we perform the average of N measurements relying on a test statistic involving the inverse of the statistical covariance and the theoretical correlation matrices. In the case where at least two observables are fully correlated, these matrices are singular and they cannot be inverted naively. One must thus determine a generalised inverse for these matrices. For definiteness, we consider the case where only statistical uncertainties are involved. The statistical test reads

$$T = (X - \mu U)^T . \bar{W} . (X - \mu U) \tag{67}$$

where U is a vector containing N times the unit value, $\bar{W} = C_s^+$ is a generalised inverse of the covariance matrix C_s (identical to C_s^{-1} if the matrix C_s is not singular).

Minimising T yields

$$\begin{aligned}\hat{\mu} &= \frac{U^T.\bar{W}.X}{U^T.\bar{W}.U} = \sum_i w_i X_i \qquad w_i = \frac{(\bar{W}.U)_i}{U^T.\bar{W}.U} \\ \sigma_\mu^2 &= w^T.C_s.w = \frac{U^T.\bar{W}.C_s.\bar{W}.U}{(U^T.\bar{W}.U)^2}\end{aligned} \tag{68}$$

We have to choose a generalised inverse C_s^+. We cannot rely on arguments based on the case where C_s is invertible (for instance taking a correlation $0 < \rho < 1$, followed by the limit $\rho \to 1$) since this limit is singular. We can start by constraining the structure of C_s^+ due to the particular structure of C_s. We have

$$C_s = \Sigma.\Gamma.\Sigma = \Sigma.R.D.R^T.\Sigma \tag{69}$$

where Σ is a diagonal matrix with uncertainties as entries $\{\sigma_1, \ldots \sigma_n\}$, Γ the correlation matrix with entries between -1 and 1 (and diagonal entries equal to 1), R is an orthogonal matrix, and D is a diagonal matrix with entries in decreasing order

$$d_1 \geq d_2 \geq \ldots \geq d_m > 0 = d_{m+1} = \ldots = d_n \tag{70}$$

The entries of D are positive since C_s is assumed to be positive, with

$$\sum_{i=1}^{N} d_i = \mathrm{Tr}(D) = \mathrm{Tr}(\Gamma) = n \Longrightarrow d_1 \leq n \tag{71}$$

A generalised inverse for C_s can be expressed in terms of a generalised inverse for D, if we define

$$C_s^+ = \Sigma^{-1}.R.D^+.R^T.\Sigma^{-1} \tag{72}$$

Indeed a generalised inverse for C_s obeys $C_s C_s^+ C_s = C_s$, which is equivalent to the condition

$$D.D^+.D = D \quad \Longrightarrow \quad D^+ = \left[\begin{array}{c|c} 1/d & A \\ \hline A^T & B \end{array}\right] \tag{73}$$

where d is the $m \times m$ diagonal matrix with entries d_i, A is an $m \times (n-m)$ arbitrary matrix and B is an $(n-m) \times (n-m)$ arbitrary matrix. A and B can only depend on $d_1, \ldots d_m$, and each choice of A and B correspond to an admissible generalised inverse.

Under these conditions, we find for the weights and the variance

$$\begin{aligned} w_i &= \frac{(\Sigma^{-1}.R.D^+.R^T.\Sigma^{-1}.U)_i}{U^T.C_s^+.U} \\ \sigma_\mu^2 &= w^T.C_s.w = \frac{U^T.C_s^+.C_s.C_s^+.U}{(U^T.C_s^+.U)^2} \end{aligned} \tag{74}$$

A.2 Choice of a generalised inverse

The most common generalised inverse is the Moore-Penrose pseudoinverse, obtained by adding three other conditions on C_s^+ on top of the definition of a generalised inverse. The condition $C_s^+ C_s C_s^+ = C_s^+$ (reflexive generalised inverse) would translate as $D^+.D.D^+ = D^+$ leading to the condition $B = A^T.d.A$ in Eq. (73), whereas the two other conditions for the Moore-Penrose inverse of C_s^+ do not translate easily on D^+. Unfortunately, we will see in explicit examples that this pseudoinverse gives more weight to measurements with a poor accuracy, and is thus not appropriate in our case.

An alluring alternative to obey Eq. (73) consists in considering $A = 0$ and $B = \lambda \times 1_{(n-m)\times(n-m)}$ proportional to the identity, with λ a real number to be fixed. In this case, the weights read

$$w_i = \frac{1}{U^T C_s^+ U} \sum_{j=1}^{n} \frac{1}{\sigma_i \sigma_j} (RD^+R^T)_{ij} \tag{75}$$

Let us assume that σ_a becomes much smaller than the other σ_i, the weights are dominated by

$$w_i \sim \frac{1}{U^T C_s^+ U} \frac{1}{\sigma_i \sigma_a} (RD^+R^T)_{ia} \tag{76}$$

Since the first (normalisation) factor is the same for all the inputs, the dominant weight will be w_a, under the condition that

$$0 \neq (RD^+R^T)_{aa} = \sum_{j=1}^{n} (R_{aj})^2 \frac{1}{d_j} + \lambda \sum_{j=n+1}^{N} (R_{aj})^2 = \lambda + \sum_{j=1}^{n} (R_{aj})^2 \left(\frac{1}{d_j} - \lambda \right) \tag{77}$$

which is a condition fulfilled for $0 < \lambda \leq 1/d_1$. We see that the family of generalised inverses thus defined [20] has the following properties

- they can be computed in a very simple way
- for $0 < \lambda \leq 1/d_1$, if a determination is much more precise than the others, it will dominate the average

For $\lambda = 1/d_1$, we call C_s^+ the λ-inverse of C_s. For $\lambda = 0$, we recover the Moore-Penrose pseudoinverse for D, and call this generalised inverse the 0-inverse of C_s. As said earlier, one could also consider the possibility of taking the Moore-Penrose pseudoinverse of C_s directly. We will illustrate these three possibilities with a few simple examples.

A.3 Examples

A.3.1 Two measurements

In the case of two uncorrelated measurements, there is no problem with inversion, and we get for all methods

$$\begin{aligned} C_s^{-1} &= \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix} \qquad w = \frac{1}{\sigma_1^2 + \sigma_2^2} \begin{pmatrix} \sigma_2^2 \\ \sigma_1^2 \end{pmatrix} \\ \sigma_\mu^2 &= \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2} \end{aligned} \tag{78}$$

For partially correlated measurements ($|\rho| < 1$), the same inversion can be performed, leading to

$$\begin{aligned} C_s^{-1} &= \frac{1}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_1^2} & -\frac{\rho}{\sigma_1\sigma_2} \\ -\frac{\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{pmatrix} \\ w &= \frac{1}{\sigma_1^2 - 2\sigma_1\sigma_2\rho + \sigma_2^2} \begin{pmatrix} \sigma_2(\sigma_2 - \rho\sigma_1) \\ \sigma_1(\sigma_1 - \rho\sigma_2) \end{pmatrix} \sim \begin{pmatrix} 1 \\ -\rho\sigma_1/\sigma_2 \end{pmatrix} \end{aligned} \tag{79}$$

[20] The definition of C_s^+ can be extended for an arbitrary matrix C in the following way. Σ is defined as the diagonal matrix with entries $\{\sqrt{|C_{11}|}, \ldots \sqrt{|C_{NN}|}\}$ (if a diagonal entry is 0, one defines Σ with 1 in the corresponding entry). The matrix $\Gamma = \Sigma^{-1}.C.\Sigma^{-1}$ can be written according to a singular value decomposition $\Gamma = R.D.S$ with two rotation matrices R and S. Once the generalised inverse D^+ is defined, the corresponding generalised inverse of C is defined as $C^+ = \Sigma^{-1}.S^T.D^+.R^T.\Sigma^{-1}$.

and the expression for the uncertainty

$$\sigma_\mu^2 = \frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2} \sim \sigma_1^2(1-\rho^2) \tag{80}$$

In each case, we indicate the limit where σ_1 becomes much smaller than σ_2 with the $\sim$ symbol, i.e., one measurement is much more accurate than the other. A comment is in order with respect to the HFAG approach at this stage. As noticed in Ref. [36], the maximal uncertainty is $\min(\sigma_1^2, \sigma_2^2)$ and corresponds to the correlation coefficient $\rho = \min(\sigma_1/\sigma_2, \sigma_2/\sigma_1)$ (it is not $\rho = 1$).

In the case of two fully correlated measurements, we have

$$C_s = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2 \\ \sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{81}$$

with $d_1 = 2$, $d_2 = 0$. The λ-inverse for C_s yields

$$\begin{aligned} C_s^+ &= \begin{pmatrix} \frac{1}{2\sigma_1^2} & 0 \\ 0 & \frac{1}{2\sigma_2^2} \end{pmatrix} \\ w &= \frac{1}{\sigma_1^2 + \sigma_2^2} \begin{pmatrix} \sigma_2^2 \\ \sigma_1^2 \end{pmatrix} \sim \begin{pmatrix} 1 \\ \sigma_1^2/\sigma_2^2 \end{pmatrix} \\ \sigma_\mu^2 &= \frac{\sigma_1^2\sigma_2^2[\sigma_1+\sigma_2]^2}{[\sigma_1^2+\sigma_2^2]^2} \sim \sigma_1^2 \end{aligned} \tag{82}$$

where we indicated the limit when $\sigma_1 \to 0$. The 0-inverse yields

$$\begin{aligned} C_s^+ &= \begin{pmatrix} \frac{1}{4\sigma_1^2} & \frac{1}{4\sigma_1\sigma_2} \\ \frac{1}{4\sigma_1\sigma_2} & \frac{1}{4\sigma_2^2} \end{pmatrix} \\ w &= \frac{1}{\sigma_1 + \sigma_2} \begin{pmatrix} \sigma_2 \\ \sigma_1 \end{pmatrix} \sim \begin{pmatrix} 1 \\ \sigma_1/\sigma_2 \end{pmatrix} \\ \sigma_\mu^2 &= \frac{4\sigma_1^2\sigma_2^2}{[\sigma_1+\sigma_2]^2} \sim 4\sigma_1^2 \end{aligned} \tag{83}$$

and the Moore-Penrose pseudoinverse yields

$$\begin{aligned} C_s^+ &= \frac{1}{(\sigma_1^2+\sigma_2^2)^2} \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2 \\ \sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \\ w &= \frac{1}{(\sigma_1+\sigma_2)} \begin{pmatrix} \sigma_1 \\ \sigma_2 \end{pmatrix} \sim \begin{pmatrix} \sigma_1/\sigma_2 \\ 1 \end{pmatrix} \\ \sigma_\mu^2 &= \frac{(\sigma_1^2+\sigma_2^2)^2}{(\sigma_1+\sigma_2)^2} \sim \sigma_2^2 \end{aligned} \tag{84}$$

A.3.2 n fully correlated measurements

We have a correlation matrix $\tilde{C}_s$ with unit entries everywhere. This yields $d_1 = n$, $d_{i>1} = 0$. The λ-inverse yields

$$
\begin{aligned}
C_s^+ &= \begin{pmatrix} \frac{1}{n\sigma_1^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{n\sigma_n^2} \end{pmatrix} \\
w &= \frac{1}{\sum_i 1/\sigma_i^2} \begin{pmatrix} 1/\sigma_1^2 \\ \vdots \\ 1/\sigma_N^2 \end{pmatrix} \sim \begin{pmatrix} 1 \\ \sigma_1^2/\sigma_2^2 \\ \vdots \\ \sigma_1^2/\sigma_n^2 \end{pmatrix} \\
\sigma_\mu^2 &= \frac{(\sum 1/\sigma_i)^2}{(\sum 1/\sigma_i^2)^2} \sim \sigma_1^2
\end{aligned} \tag{85}
$$

where we indicated the limit when $\sigma_1 \to 0$. The 0-inverse yields

$$
\begin{aligned}
C_s^+ &= \begin{pmatrix} \frac{1}{n^2\sigma_1^2} & \frac{1}{n^2\sigma_1\sigma_2} & \cdots & \frac{1}{n^2\sigma_1\sigma_n} \\ \vdots & & & \vdots \\ \frac{1}{n^2\sigma_1\sigma_n} & \frac{1}{n^2\sigma_2\sigma_n} & \cdots & \frac{1}{n^2\sigma_n^2} \end{pmatrix} \\
w &= \frac{1}{\sum_i 1/\sigma_i} \begin{pmatrix} 1/\sigma_1 \\ \vdots \\ 1/\sigma_n \end{pmatrix} \sim \begin{pmatrix} 1 \\ \sigma_1/\sigma_2 \\ \vdots \\ \sigma_1/\sigma_n \end{pmatrix} \\
\sigma_\mu^2 &= \frac{n^2}{(\sum_i 1/\sigma_i)^2} \sim n^2\sigma_1^2
\end{aligned} \tag{86}
$$

The Moore-Penrose pseudoinverse yields

$$
\begin{aligned}
C_s^+ &= \frac{1}{(\sum_i \sigma_i^2)^2} \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_1\sigma_n \\ \vdots & \ddots & \vdots \\ \sigma_1\sigma_n & \cdots & \sigma_n^2 \end{pmatrix} \\
w &= \frac{1}{\sum \sigma_i} \begin{pmatrix} \sigma_1 \\ \vdots \\ \sigma_n \end{pmatrix} \sim \frac{1}{\sum_{i>1} \sigma_i} \begin{pmatrix} \sigma_1 \\ \vdots \\ \sigma_n \end{pmatrix} \\
\sigma_\mu^2 &= \frac{(\sum_i \sigma_i^2)^2}{(\sum \sigma_i)^2} \sim \frac{(\sum_{i>1} \sigma_i^2)^2}{(\sum_{i>1} \sigma_i)^2}
\end{aligned} \tag{87}
$$

We can actually show that in this situation, the choice of the λ-inverse is optimal in the family of generalised inverses defined in App. A.2. Indeed, there is only one non-vanishing eigenvalue $d_1 = n$, leading to

$$
\sigma_\mu^2 = \frac{(\sum 1/\sigma)^2/n^2}{[(\sum 1/\sigma)^2/n^2 + \lambda\,[\sum 1/\sigma^2 - (\sum 1/\sigma)^2/n]]^2} \tag{88}
$$

which is minimal for the maximal value $\lambda = 1/d_1$, corresponding to the λ-inverse.

A.3.3 Two fully correlated measurements with an uncorrelated measurement

Let us consider

$$C_s = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2 & 0 \\ \sigma_1\sigma_2 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix} \tag{89}$$

with $d_1 = 2$, $d_2 = 1$, $d_3 = 0$.

The λ-inverse for C_s yields

$$\begin{aligned} C^+ &= \begin{pmatrix} \frac{1}{2\sigma_1^2} & 0 & 0 \\ 0 & \frac{1}{2\sigma_2^2} & 0 \\ 0 & 0 & \frac{1}{\sigma_3^2} \end{pmatrix} \\ w &= \frac{1}{2\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \sigma_2^2\sigma_3^2} \begin{pmatrix} \sigma_2^2\sigma_3^2 \\ \sigma_1^2\sigma_3^2 \\ 2\sigma_1^2\sigma_2^2 \end{pmatrix} \sim \begin{pmatrix} 1 \\ \sigma_1^2/\sigma_2^2 \\ 2\sigma_1^2/\sigma_3^2 \end{pmatrix} \\ \sigma_\mu^2 &= \frac{\sigma_1^2\sigma_2^2\sigma_3^2[2\sigma_1\sigma_2\sigma_3^2 + 4\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \sigma_2^2\sigma_3^2]}{[2\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \sigma_2^2\sigma_3^2]^2} \sim \sigma_1^2 \end{aligned} \tag{90}$$

The 0-inverse for C_s yields

$$\begin{aligned} C_s^+ &= \begin{pmatrix} \frac{1}{4\sigma_1^2} & \frac{1}{4\sigma_1\sigma_2} & 0 \\ \frac{1}{4\sigma_1\sigma_2} & \frac{1}{4\sigma_2^2} & 0 \\ 0 & 0 & \frac{1}{\sigma_3^2} \end{pmatrix} \\ w &= \frac{1}{4\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \sigma_2^2\sigma_3^2 + 2\sigma_1\sigma_2\sigma_3^2} \begin{pmatrix} \sigma_2\sigma_3^2(\sigma_1+\sigma_2) \\ \sigma_1\sigma_3^2(\sigma_1+\sigma_2) \\ 4\sigma_1^2\sigma_2^2 \end{pmatrix} \sim \begin{pmatrix} 1 \\ \sigma_1/\sigma_2 \\ 4\sigma_1^2/\sigma_3^2 \end{pmatrix} \\ \sigma_\mu^2 &= \frac{4\sigma_1^2\sigma_2^2\sigma_3^2}{4\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \sigma_2^2\sigma_3^2 + 2\sigma_1\sigma_2\sigma_3^2} \sim 4\sigma_1^2 \end{aligned} \tag{91}$$

The Moore-Penrose pseudoinverse yields

$$\begin{aligned} C_s^+ &= \begin{pmatrix} \frac{\sigma_1^2}{(\sigma_1^2+\sigma_2^2)^2} & \frac{\sigma_1\sigma_2}{(\sigma_1^2+\sigma_2^2)^2} & 0 \\ \frac{\sigma_1\sigma_2}{(\sigma_1^2+\sigma_2^2)^2} & \frac{\sigma_2^2}{(\sigma_1^2+\sigma_2^2)^2} & 0 \\ 0 & 0 & \frac{1}{\sigma_3^2} \end{pmatrix} \\ w &= \frac{1}{(\sigma_1^2+\sigma_2^2)^2 + (\sigma_1+\sigma_2)^2\sigma_3^2} \begin{pmatrix} \sigma_1\sigma_3^2(\sigma_1+\sigma_2) \\ \sigma_2\sigma_3^2(\sigma_1+\sigma_2) \\ (\sigma_1^2+\sigma_2^2)^2 \end{pmatrix} \sim \frac{1}{(\sigma_2^2+\sigma_3^3)} \begin{pmatrix} \sigma_1\sigma_3^2/\sigma_2 \\ \sigma_3^2 \\ \sigma_2^2 \end{pmatrix} \\ \sigma_\mu^2 &= \frac{(\sigma_1^2+\sigma_2^2)^2\sigma_3^2}{(\sigma_1^2+\sigma_2^2)^2 + (\sigma_1+\sigma_2)^2\sigma_3^2} \sim \frac{\sigma_2^2\sigma_3^2}{\sigma_2^2+\sigma_3^2} \end{aligned} \tag{92}$$

A.4 Choice of the inverse

In the above examples, the λ-inverse yields interesting results for the generalised inverse in cases that are likely to be useful. In the limit where one measurement becomes very accurate, it dominates the average. In this situation, other generalised inverses of the same family, like the 0-inverse, yield results of the same order, but larger, for the combined uncertainty, whereas the Moore-Penrose pseudoinverse yields a combined uncertainty dominated by the least precise measurements. For 100% correlated uncertainties, the λ-inverse recovers Schmelling's proposal [38] used by the Flavour

Lattice Averaging Group [37], and it does not run into the danger of underestimating the resulting uncertainty as discussed by the Heavy Flavour Averaging Group [36].

For these reasons, we choose the λ-inverse to compute both the inverse statistical covariance matrix and the inverse theoretical correlation matrix when these matrices are singular (the regular case being trivial).

B Varying the biases in the presence of theoretical correlations

B.1 Range of variations for the biases

Another issue consists in implementing correlations for the biases describing theoretical uncertainties. Some differences occur compared to statistical uncertainties, since different models are used in both cases (random variables versus biases). As described in Sec. 5.2.2, once the weights w_i are determined, the theoretical uncertainty is given by $\delta_\mu = \sum_i w_i \Delta_{i\alpha} \tilde{\delta}_\alpha$, which requires one to determine the range of variation for the normalised biases $\tilde{\delta}_\alpha$. We want to describe their variation starting from variations of uncorrelated variables. This can be achieved through a linear transformation by introducing the Cholesky decomposition for the theoretical correlation matrix $C_t = P \cdot P^T$ with P a lower triangular matrix with diagonal positive entries. We obtain the expression for the theoretical uncertainty

$$\delta_\mu = \sum_i w_i \Delta_i \tilde{\delta}_i = \sum_{i,j} w_i \Delta_{i\alpha} P_{\alpha\beta} (P^{-1}\tilde{\delta})_\beta \tag{93}$$

where $(P^{-1}\tilde{\delta})_j$ are uncorrelated biases varied in a hyperball, leading to

$$\Delta_\mu = \sqrt{\sum_\beta \left(\sum_{i,\alpha} w_i \Delta_{i\alpha} P_{\alpha\beta} \right)^2} \quad \text{(hyperball)} \tag{94}$$

There is an ambiguity in the definition of P when C_t is only semi-definite positive (which occurs when C_t is singular due to 100% correlations, and exhibits not only positive but also vanishing eigenvalues). We define then P by computing $P(\epsilon)$ for the shifted matrix $C_t + \epsilon \times 1_{m\times m}$ and defining $P = \lim_{\epsilon \to 0^+} P(\epsilon)$. This limit is not singular, and it allows one to define the limit of two measurements fully correlated theoretically as a smooth limit of the general case with a partial correlation.

One should emphasise that in the case of a singular correlation matrix C_t for theoretical uncertainties, we may have to treat this singularity at two different stages: first when we build the test statistic involving $\bar{W}$ (depending on the structure of the statistical and theoretical correlation matrices), second when we consider the domain of variation for the parameters $\tilde{\delta}$. We stress that we used different procedures in both cases (λ-inverse for $\bar{W}$, Cholesky decomposition for $\tilde{\delta}$), which involves some arbitrariness, but reproduces desirable properties for the combined uncertainties and domains of variation of the biases in this singular limit.

In the case of a hypercube, we may want to follow the same procedure and define

$$\Delta_\mu = \sum_\beta \left| \sum_{i,\alpha} w_i \Delta_{i\alpha} P_{\alpha\beta} \right| \qquad \text{(hypercube with correlations ?)} \tag{95}$$

The question mark indicates that this definition is only tentative, and will not actually be used. Indeed as discussed in Sec. 5.2.3 and illustrated in the following sections, this definition has the rather unpleasant feature that the ranges of variations depend on the order of the inputs used, and we have not been able to identify an alternative choice for the range of variations that would avoid this problem, which does not occur in the hyperball case. These difficulties could be somehow expected from the properties of the hypercube case discussed in Sec. 5.2.1. Indeed, in the case of two

measurements, the hypercube corresponds to values of δ_1 and δ_2 left free to vary without relation among them (contrary to the hyperball case). It seems therefore difficult to introduce correlations in this case which was designed to avoid them initially. Our failure to introduce correlations in this case might be related to the fact that the hypercube is somehow designed to avoid such correlations from the start and cannot accommodate them easily.

We thus propose the alternative definition, ignoring theoretical correlations to determine the range of variations for the biases

$$\Delta_\mu = \sum_\alpha \left| \sum_i w_i \Delta_{i\alpha} \right| \quad \text{(hypercube no correlation)} \tag{96}$$

B.2 Averaging measurements with theoretical correlations

If we take two measurements $X_1 \pm \sigma_1 \pm \Delta_1$ and $X_2 \pm \sigma_2 \pm \Delta_2$ with σ_1 and σ_2 uncorrelated, but Δ_1 and Δ_2 correlated with a correlation ρ, one gets the Cholesky decomposition

$$C_t = P.P^T \qquad P = \begin{pmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix} \tag{97}$$

so that the variations for the two (normalised) biases $\tilde{\delta}_1$ and $\tilde{\delta}_2$ are given by

$$\tilde{\delta}_1 = d_1 \qquad \tilde{\delta}_2 = \rho d_1 + \sqrt{1-\rho^2} d_2 \tag{98}$$

where d_1 and d_2 are varied in a hyperball or a hypercube following Eqs. (94) and (95) respectively. Eq. (96) would correspond to neglecting correlations and setting $\rho = 0$ in Eq. (98).

In the case of a hypercube with correlations, $\tilde{\delta}_1, \tilde{\delta}_2$ are varied in a parallelogram with two sides parallel to the $\tilde{\delta}_2$ axis, whereas they are varied in a tilted ellipse in the hyperball case, as can be seen in Fig. 10. In both cases, the limiting case where $\rho \to \pm 1$ corresponds to $\tilde{\delta}_1$ and $\tilde{\delta}_2$ varied along a diagonal line, meeting our expectations for fully correlated theoretical uncertainties. We see that this treatment yields a symmetric domain for $\tilde{\delta}_1$ and $\tilde{\delta}_2$ in the hyperball case, but not in the hypercube case, which means that the two uncertainties are not treated in a symmetric way [21]. As indicated before, Eq. (96) corresponds to the hypercube with $\rho = 0$, i.e., a square domain for $\tilde{\delta}_1$ and $\tilde{\delta}_2$.

One can easily extend the same procedure to a larger number of correlated theoretical uncertainties. As indicated above, the hyperball with correlations yields domains of variations which are symmetric for any pair $(\tilde{\delta}_k, \tilde{\delta}_l)$ whereas the hypercube with correlations does not. This means that the range of variation chosen for the biases will depend on the order of the inputs: a mere reshuffling of the inputs will yield different ranges of variations for the biases and (in general) different outcomes for averages and fits. In addition, we should emphasise that a total correlation $(C_t)_{k,l} = 0$ between two biases does not have the same impact for the domain of variation in the $(\tilde{\delta}_k, \tilde{\delta}_l)$ plane in both approaches: in the hyperball case, one obtains an undeformed disk, whereas the hypercube case yields a complicated convex polytope depending on the other elements of the correlation matrix (see Fig. 11 in the case of three biases) (a symmetrisation of the Cholesky decomposition in the form $P + P^T$ or a different choice of linear transformation would yield similar results).

These features lead us to neglect correlations in the hypercube range of variations, whereas we keep them when considering the hyperball case. We thus discard Eq. (95) and consider only Eqs. (94) and (96) in our analyses.

[21] One could try to symmetrise the problem, but one would lose the connection with the Cholesky decomposition, with the unpleasant feature that all domains of variation would be identical and thus do not take into account correlations.

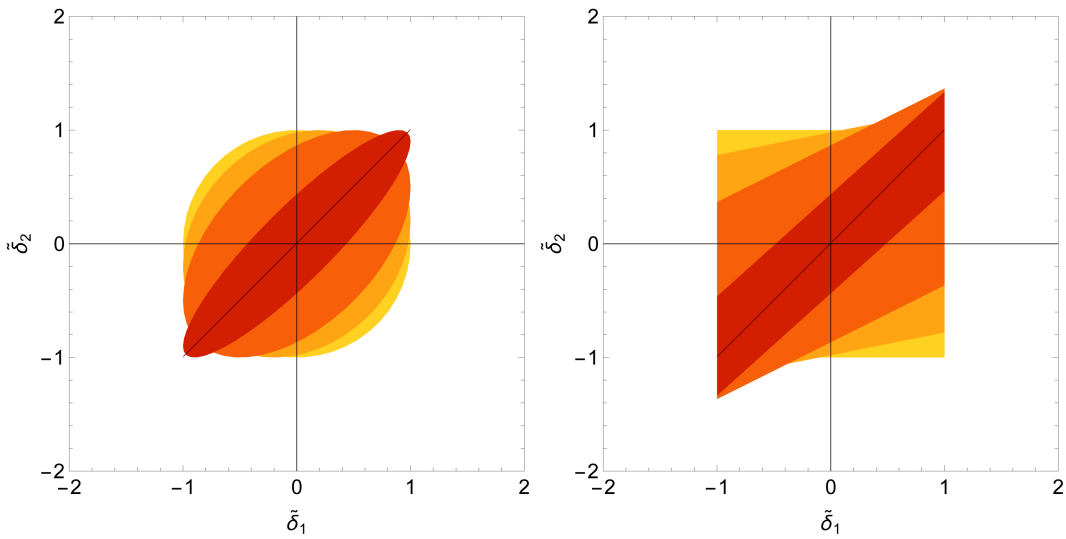


Figure 10: Ranges of variation for $\tilde{\delta}_1$ and $\tilde{\delta}_2$ for $\rho = 0, 0.2, 0.5, 0.9, 1$, going from light (yellow) to dark (red). The variation over a hyperball (left) or a hypercube (right) is considered.

C Definition of the test statistic in n dimensions

C.1 Ambiguities in the definition of a 100% theoretical correlation

In Eq. (34), one may be uncertain about the case where a theoretical uncertainty is fully correlated between two observables. Let us imagine that we have two quantities $X_1 = X_{10} \pm \sigma_1 \pm \Delta_1$ and $X_2 = X_{20} \pm \sigma_2 \pm \Delta_2$ with the two theoretical uncertainties being fully correlated. We can imagine describing the theoretical uncertainties either via $m = 2$ parameters fully correlated through $\widetilde{C}_t$:

$$I: \qquad \Delta = \begin{bmatrix} \Delta_1 & 0 \\ 0 & \Delta_2 \end{bmatrix} \qquad \tilde{C}_t = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{99}$$

or as $m = 1$ parameter intervening in the two quantities via Δ

$$II: \qquad \Delta = \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} \qquad \tilde{C}_t = [1] \tag{100}$$

We can see in the above discussion that the only relevant combination of Δ and $\widetilde{C}_t$ is actually ΔP, whether in the definition of $\bar{W}$ that involves $\Delta\widetilde{C}_t\Delta^T = (\Delta P)(\Delta P)^T$, or in the discussion of the theoretical uncertainty Δ_μ. We have

$$\begin{aligned} I \quad &: \quad P = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \quad \Delta P = \begin{bmatrix} \Delta_1 & 0 \\ \Delta_2 & 0 \end{bmatrix} \\ II \quad &: \quad P = [1] \quad \Delta P = \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} \end{aligned} \tag{101}$$

leading to the same $\Delta\widetilde{C}_t\Delta^T$ and showing that only one uncorrelated bias parameter is needed in both cases, even though we started from a different number of bias parameters. The discussion can be extended to an arbitrary number of fully correlated theoretical uncertainties. Obviously, for partial correlations, only $\widetilde{C}_t$ can be used with an unchanged number of bias parameters.

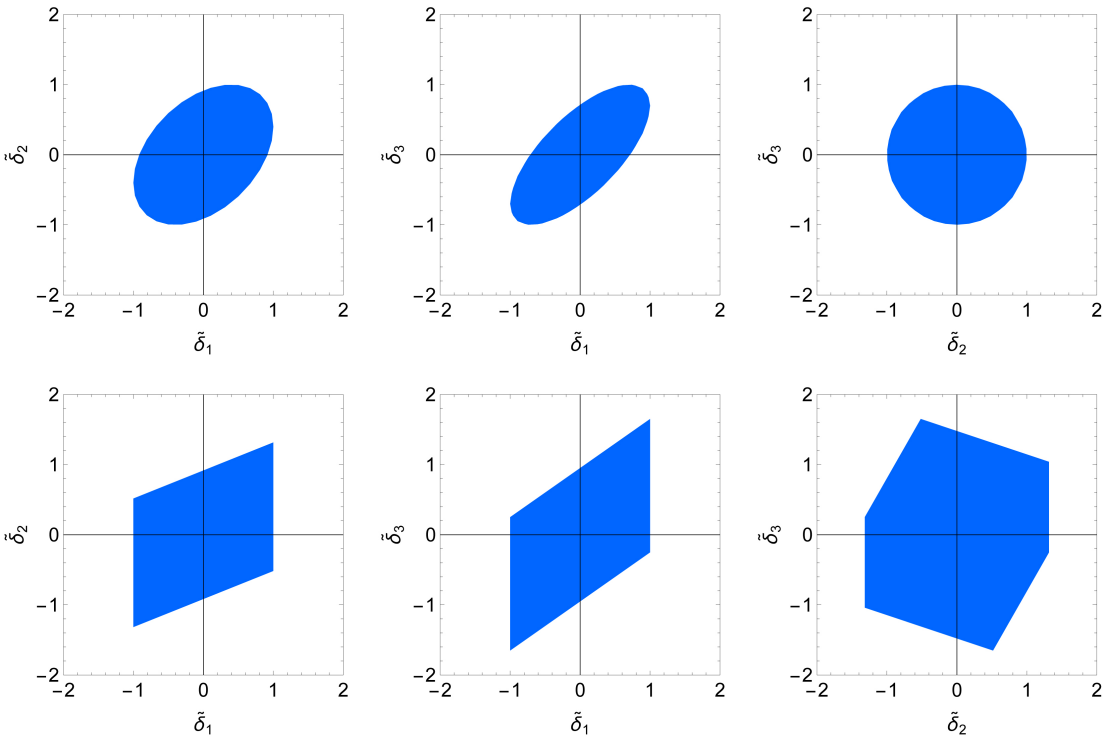


Figure 11: Ranges of variation for $\tilde{\delta}_1, \tilde{\delta}_2, \tilde{\delta}_3$ with $\rho_{12} = 0.4, \rho_{13} = 0.7, \rho_{23} = 0$. The variation over a hyperball (top) or a hypercube (bottom) with correlations is considered. Neglecting correlations would yield discs (top) and squares (bottom).

C.2 Reducing the problem to one bias parameter per observable

We can define a reduced version of the problem Eq. (34), with only n bias parameters rather than m. We have to determine an equivalent problem

$$T'(x, \tilde{\delta}') = [X - x - \Delta'\tilde{\delta}']^T W_s [X - x - \Delta'\tilde{\delta}'] + \tilde{\delta}^T \bar{W}'_t \tilde{\delta}' \tag{102}$$

where $\bar{W}'_t$ and Δ' are $n \times n$ matrices, and Δ' is diagonal. From what was discussed before, we see that we will obtain the same result for the weights $w^{(q)}$, the variances and the correlations, if we ensure that $\Delta P = \Delta' P'$.

This can be achieved by defining Δ' and the correlation matrix $\widetilde{C}'_t$ using

$$\Delta \widetilde{C}_t \Delta^T = \Delta' \widetilde{C}'_t \Delta' \tag{103}$$

$\widetilde{C}_t$ is positive semi-definite, which means that $\Delta \widetilde{C}_t \Delta^T$ will also be. The diagonal elements of a positive semi-definite matrix are positive, and therefore, one can define Δ' so that $\widetilde{C}'_t$ has 1 as a diagonal.

It could occur that $\Delta \widetilde{C}_t \Delta^T$ has 0 on the diagonal for some k^{th} entry. But since $\Delta \widetilde{C}_t \Delta^T$ is positive semi-definite, one can prove that the corresponding row and column then vanish, meaning that the corresponding bias parameter does not actually occur in the reduced problem. In such a case, one can define $\Delta'_k = 0$ and C'_t vanishing on the kth row and column, and $C'_{t,kk} = 1$ (this is the case for instance if there is no theoretical uncertainty for some of the observables).

Moreover, one can check that $\widetilde{C}'$ is indeed a correlation matrix by defining the scalar product $(x, y) = x^T \Delta \widetilde{C}_t \Delta^T y$. We can apply the Cauchy-Schwartz inequality to the basis vectors $u^{(i)}$ defined

so that $u_j^{(i)} = \delta_{ij}$ (i.e., only one non-vanishing component):

$$\begin{aligned} (u^{(i)}, u^{(j)})^2 &\leq (u^{(i)}, u^{(i)})(u^{(j)}, u^{(j)}) \\ (\Delta'_i)^2(\Delta'_j)^2(\widetilde{C}'_{t,ij})^2 &\leq (\Delta'_i)^2(\Delta'_j)^2\widetilde{C}'_{t,ii}\widetilde{C}'_{t,jj} \end{aligned} \tag{104}$$

so that $|\widetilde{C}'_{t,ij}| \leq 1$ and $\widetilde{C}'_{t,ii} = 1$, with the appropriate structure of a correlation matrix.

Finally, the Cholesky decomposition of $\widetilde{C}'_t$ corresponds to $P' = (\Delta')^{-1}\Delta P$. Therefore, the determination of the theoretical uncertainties for Δ_μ remains indeed the same with the new set of biases.

We have thus reduced the problem of n measurements and m theoretical biases to the case with n measurements, each of them having with a single bias parameter, with correlations among the biases. Without loss of generality we can consider that Δ is diagonal and $m = n$.

D Asymmetric uncertainties

In this article, statistical uncertainties are assumed to be strictly Gaussian and hence symmetric. In practice, if asymmetric uncertainties are quoted, we symmetrise in the following manner

$$X = \mu^{+\sigma_+}_{-\sigma_-} \to X = \left(\mu + \frac{\sigma_+ - \sigma_-}{2}\right) \pm \left(\frac{\sigma_+ + \sigma_-}{2}\right) \tag{105}$$

This is also the case for the theoretical uncertainties in the random-δ approach.

In contrast, it is perfectly possible to have asymmetric theoretical uncertainties in the nuisance-δ or external-δ approaches described above. A theoretical uncertainty that is modeled by a bias parameter δ may be asymmetric: that is, the region in which δ is varied may depends on the sign of δ, *e.g.* $\delta \in [-\Delta_-, +\Delta_+]$ in one dimension ($\Delta_\pm \geq 0$).

In the case of a quadratic test statistic, we want to keep the stationarity property stemming from the symmetric quadratic shape, by using a test statistic Eq. (22) with $(\Delta_+ + \Delta_-)/2$ or $\mathrm{Max}(\Delta_+, \Delta_-)$ in the definition, the second possibility being more conservative and our preferred choice in the following. As indicated in Sec. 5.2.2, this is independent of the range of variation Ω chosen, which will be kept asymmetric, e.g., $[-\Delta_-, \Delta_+]$ in the fixed nuisance approach.

In the case of the Rfit approach [15, 16], we can use the fact that the well test statistic has a shape that is independent of the central value chosen, as long as the position of the flat bottom remains unchanged. One can thus shift the central value by an arbitrary quantity if one remains at the bottom of the well. It is thus completely equivalent to take asymmetric theoretical ranges or to take symmetric theoretical ranges following Eq. (105) where $\sigma_\pm$ is replaced by $\Delta_\pm$.

E 3-σ intervals for CKM-related examples

We collect here the intervals at 3 σ for the various approaches applied to the CKM examples discussed in Sec. 6. Figs. 12, 13 and 14 are the 3-σ equivalents of Figs. 7, 8 and 9 showing 1 σ intervals. The comparison between the two series of plot shows how the intervals evolve with the confidence level. In particular, the adaptive hyperball approach appears more (less) conservative than the 1-hypercube approach at high (low) significance. This change of hierarchy explains why we choose a different convention to plot the 1 σ (dashed horizontal line) and 3 σ (vertical lines in the middle of the solid intervals) intervals for the 1-hypercube approach in Figs. 7, 8, 9 on one hand and Figs. 12, 13, 14 on the other hand.

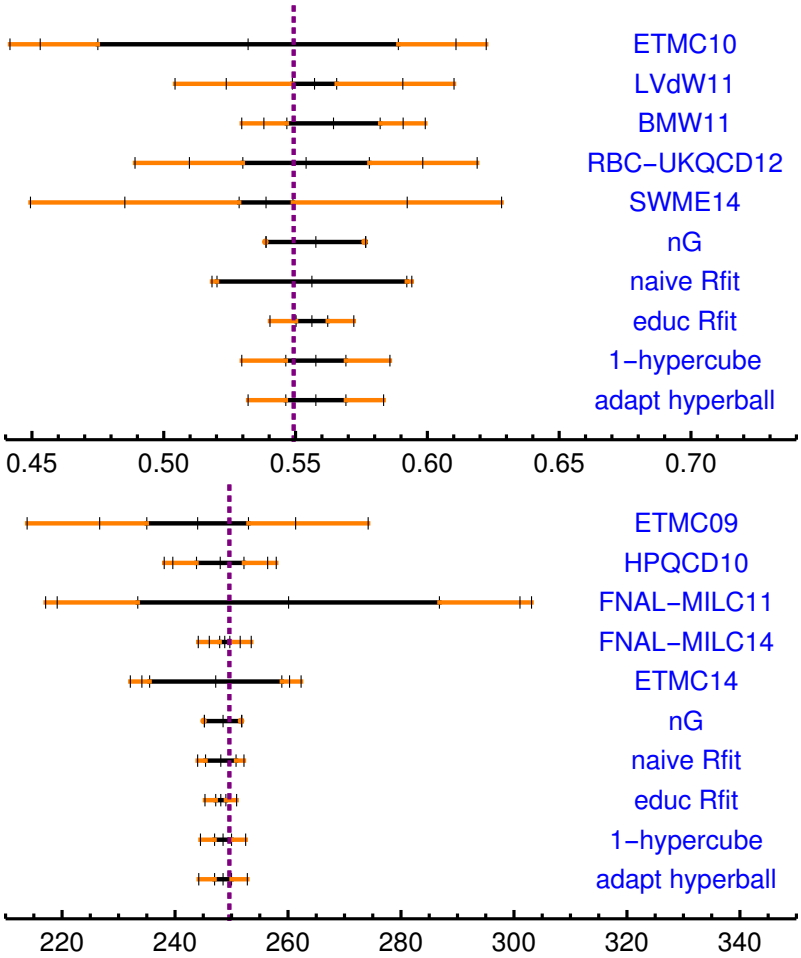


Figure 12: (Top) Inputs for $B_K^{\overline{\rm MS}}(2\text{GeV})$ and the averages resulting from the different models considered here. (Bottom) Same for the lattice determinations of the D_s-meson decay constant (in MeV). The black range gives the statistical error. For each individual input, the solid yellow range indicates the 3 σ interval according to the adaptive hyperball approach, whereas the interval corresponding to the 1-fixed hypercube approach is given by the vertical lines in the middle of the solid yellow intervals. For average according to the different approaches, the black range corresponds again to the 3 σ statistical error, whereas the yellow range corresponds to the 3 σ interval following the corresponding approach. The comparison between black and yellow ranges illustrates the relative importance of statistical and theoretical errors. Finally, for illustrative purposes, the vertical purple line gives the arithmetic average of the inputs (same weight for all central values).

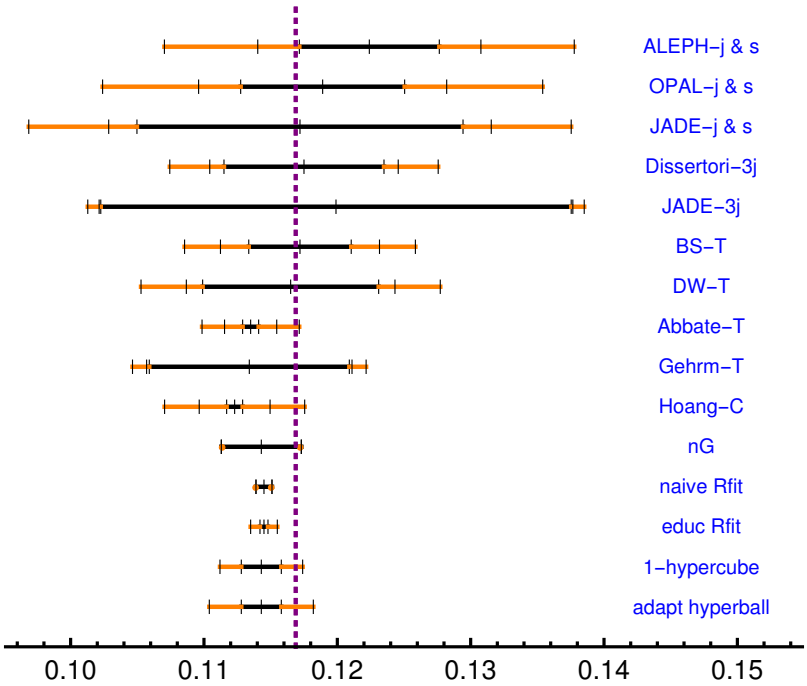


Figure 13: Determinations of the strong coupling constant at M_Z through e^+e^- annihilation, and the averages resulting from the different models considered. The intervals are given at 3 σ. See Fig. 12 for the legend.

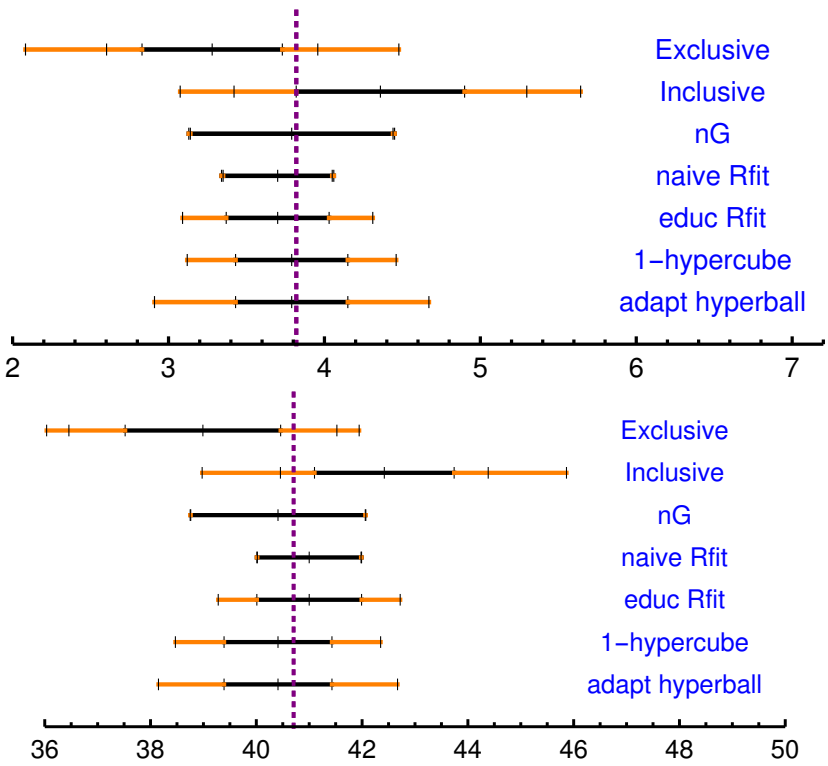


Figure 14: (Top) Inclusive and exclusive inputs for the CKM matrix element $|V_{ub}|$ (times 10^3) and the averages resulting from the different models considered here. (Bottom) Same for the determinations of $|V_{cb}|$ (times 10^3) CKM matrix element. The intervals are given at 3 σ. See Fig. 12 for the legend.

References

[1] F. James, "Statistical methods in experimental physics," Hackensack, USA: World Scientific (2006).

[2] G. Cowan, "Statistics for Searches at the LHC," doi:10.1007/978-3-319-05362-2_9 arXiv:1307.2487 [hep-ex].

[3] M. G. Kendall and A. Stuart, "The Advanced Theory of Statistics," Griffin (1969).

[4] P. Sinervo, eConf C **030908** (2003) TUAT004.

[5] M. Schmelling, hep-ex/0006004.

[6] W. A. Rolke, A. M. Lopez and J. Conrad, Nucl. Instrum. Meth. A **551** (2005) 493 [physics/0403059].

[7] R. D. Cousins, J. T. Linnemann and J. Tucker, Nucl. Instrum. Meth. A **595** (2008) 480.

[8] W. M. Bolstad and J. M. Curran, "Introduction to Bayesian Statistics," John Wiley & Sons (2016).

[9] G. D'Agostini, Rept. Prog. Phys. **66** (2003) 1383 doi:10.1088/0034-4885/66/9/201 [physics/0304102].

[10] A. J. Bevan, "Statistical Data Analysis for the Physical Sciences," Cambridge Press (2013).

[11] S. Schael *et al.* [ALEPH and DELPHI and L3 and OPAL and LEP Electroweak Collaborations], Phys. Rept. **532** (2013) 119 [arXiv:1302.3415 [hep-ex]].

[12] M. Baak *et al.* [Gfitter Group Collaboration], Eur. Phys. J. C **74** (2014) 3046 [arXiv:1407.3792 [hep-ph]].

[13] R. Aaij *et al.* [LHCb Collaboration], Eur. Phys. J. C **73** (2013) 4, 2373 [arXiv:1208.3355 [hep-ex]].

[14] A. J. Bevan *et al.* [BaBar and Belle Collaborations], Eur. Phys. J. C **74** (2014) 3026 [arXiv:1406.6311 [hep-ex]].

[15] A. Hocker, H. Lacker, S. Laplace and F. Le Diberder, Eur. Phys. J. C **21** (2001) 225 [hep-ph/0104062].

[16] J. Charles *et al.* [CKMfitter Group Collaboration], Eur. Phys. J. C **41** (2005) 1 [hep-ph/0406184].

[17] J. Charles *et al.*, Phys. Rev. D **84** (2011) 033005 [arXiv:1106.4041 [hep-ph]].

[18] O. Deschamps, S. Descotes-Genon, S. Monteil, V. Niess, S. T'Jampens and V. Tisserand, Phys. Rev. D **82** (2010) 073012 [arXiv:0907.5135 [hep-ph]].

[19] A. Lenz *et al.*, Phys. Rev. D **83** (2011) 036004 [arXiv:1008.1593 [hep-ph]].

[20] A. Lenz, U. Nierste, J. Charles, S. Descotes-Genon, H. Lacker, S. Monteil, V. Niess and S. T'Jampens, Phys. Rev. D **86** (2012) 033008 [arXiv:1203.0238 [hep-ph]].

[21] J. Charles, S. Descotes-Genon, Z. Ligeti, S. Monteil, M. Papucci and K. Trabelsi, Phys. Rev. D **89** (2014) no.3, 033016 [arXiv:1309.2293 [hep-ph]].

[22] J. Charles, O. Deschamps, S. Descotes-Genon, H. Lacker, A. Menzel, S. Monteil, V. Niess and J. Ocariz *et al.*, Phys. Rev. D **91** (2015) 7, 073007 [arXiv:1501.05013 [hep-ph]].

[23] B. Aubert *et al.* [BaBar Collaboration], Phys. Rev. D **79** (2009) 072009 doi:10.1103/PhysRevD.79.072009 [arXiv:0902.1708 [hep-ex]].

[24] I. Adachi *et al.*, Phys. Rev. Lett. **108** (2012) 171802 doi:10.1103/PhysRevLett.108.171802 [arXiv:1201.4643 [hep-ex]].

[25] R. Aaij *et al.* [LHCb Collaboration], Phys. Rev. Lett. **115** (2015) no.3, 031601 doi:10.1103/PhysRevLett.115.031601 [arXiv:1503.07089 [hep-ex]].

[26] I. I. Y. Bigi, V. A. Khoze, N. G. Uraltsev and A. I. Sanda, Adv. Ser. Direct. High Energy Phys. **3** (1989) 175. doi:10.1142/9789814503280_0004

[27] J. Neyman and E. S. Pearson, Phil. Trans. R. Soc. Lond. A 1933 231 289-337;

[28] S. S. Wilks, Ann. Math. Statist. 9 (1938), no. 1, 60–62.

[29] F. C. Porter, arXiv:0806.0530 [physics.data-an].

[30] S. Fichet and G. Moreau, Nucl. Phys. B **905** (2016) 391 [arXiv:1509.00472 [hep-ph]].

[31] S. Fichet, Nucl. Phys. B **911** (2016) 623 [arXiv:1603.03061 [hep-ph]].

[32] G. P. Dubois-Felsmann, D. G. Hitlin, F. C. Porter and G. Eigen, hep-ph/0308262.

[33] G. Eigen, G. Dubois-Felsmann, D. G. Hitlin and F. C. Porter, Phys. Rev. D **89** (2014) 3, 033004 [arXiv:1301.5867 [hep-ex]].

[34] K. A. Olive *et al.* [Particle Data Group Collaboration], Chin. Phys. C **38** (2014) 090001.

[35] J. Charles *et al*, work in progress.

[36] Y. Amhis *et al.* [Heavy Flavor Averaging Group (HFAG) Collaboration], arXiv:1412.7515 [hep-ex].

[37] S. Aoki *et al.*, arXiv:1607.00299 [hep-lat].

[38] M. Schmelling, Phys. Scripta **51** (1995) 676.

[39] H. Ruben, Ann. Math. Statist. Volume 33, Number 2 (1962), 542-570

[40] A. Castaño-Martnez and F. López-Blázquez, TEST (2005) 14:397.

[41] M. Constantinou *et al.* [ETM Collaboration], Phys. Rev. D **83** (2011) 014505 [arXiv:1009.5606 [hep-lat]].

[42] J. Laiho and R. S. Van de Water, PoS LATTICE **2011** (2011) 293 [arXiv:1112.4861 [hep-lat]].

[43] S. Durr, Z. Fodor, C. Hoelbling, S. D. Katz, S. Krieg, T. Kurth, L. Lellouch and T. Lippert *et al.*, Phys. Lett. B **705** (2011) 477 [arXiv:1106.3230 [hep-lat]].

[44] R. Arthur *et al.* [RBC and UKQCD Collaborations], Phys. Rev. D **87** (2013) 094514 [arXiv:1208.4412 [hep-lat]].

[45] T. Bae *et al.* [SWME Collaboration], arXiv:1402.0048 [hep-lat].

[46] B. Blossier *et al.*, JHEP **0907** (2009) 043 [arXiv:0904.0954 [hep-lat]].

[47] C. T. H. Davies, C. McNeile, E. Follana, G. P. Lepage, H. Na and J. Shigemitsu, Phys. Rev. D **82** (2010) 114504 [arXiv:1008.4018 [hep-lat]].

[48] A. Bazavov *et al.* [Fermilab Lattice and MILC Collaboration], Phys. Rev. D **85** (2012) 114506 [arXiv:1112.3051 [hep-lat]].

[49] A. Bazavov *et al.* [Fermilab Lattice and MILC Collaborations], Phys. Rev. D **90** (2014) 7, 074509 [arXiv:1407.3772 [hep-lat]].

[50] N. Carrasco, P. Dimopoulos, R. Frezzotti, P. Lami, V. Lubicz, F. Nazzaro, E. Picca and L. Riggio *et al.*, Phys. Rev. D **91** (2015) 5, 054507 [arXiv:1411.7908 [hep-lat]].

[51] D. d'Enterria and P. Z. Skands, arXiv:1512.05194 [hep-ph].

[52] G. Dissertori, A. Gehrmann-De Ridder, T. Gehrmann, E. W. N. Glover, G. Heinrich, G. Luisoni and H. Stenzel, JHEP **0908**, 036 (2009) [arXiv:0906.3436 [hep-ph]].

[53] G. Abbiendi *et al.* [OPAL Collaboration], Eur. Phys. J. C **71**, 1733 (2011) [arXiv:1101.1470 [hep-ex]].

[54] S. Bethke *et al.* [JADE Collaboration], Eur. Phys. J. C **64**, 351 (2009) [arXiv:0810.1389 [hep-ex]].

[55] G. Dissertori, A. Gehrmann-De Ridder, T. Gehrmann, E. W. N. Glover, G. Heinrich and H. Stenzel, Phys. Rev. Lett. **104**, 072002 (2010) [arXiv:0910.4283 [hep-ph]].

[56] J. Schieck *et al.* [JADE Collaboration], Eur. Phys. J. C **73**, no. 3, 2332 (2013) [arXiv:1205.3714 [hep-ex]].

[57] T. Becher and M. D. Schwartz, JHEP **0807**, 034 (2008) [arXiv:0803.0342 [hep-ph]].

[58] R. A. Davison and B. R. Webber, Eur. Phys. J. C **59**, 13 (2009) [arXiv:0809.3326 [hep-ph]].

[59] R. Abbate, M. Fickinger, A. H. Hoang, V. Mateu and I. W. Stewart, Phys. Rev. D **83**, 074021 (2011) [arXiv:1006.3080 [hep-ph]].

[60] T. Gehrmann, G. Luisoni and P. F. Monni, Eur. Phys. J. C **73**, no. 1, 2265 (2013) [arXiv:1210.6945 [hep-ph]].

[61] A. H. Hoang, D. W. Kolodrubetz, V. Mateu and I. W. Stewart, Phys. Rev. D **91** (2015) 9, 094018 [arXiv:1501.04111 [hep-ph]].

[62] S. Bethke, G. Dissertori and G. P. Salam, EPJ Web Conf. **120** (2016) 07005. doi:10.1051/epjconf/201612007005

[63] M. Beneke, Phys. Lett. B **620** (2005) 143 [hep-ph/0505075].

[64] B. O. Lange, M. Neubert and G. Paz, Phys. Rev. D **72** (2005) 073006 [hep-ph/0504071].