Eur. Phys. J. C 56, 153–159 (2008)
DOI 10.1140/epjc/s10052-008-0637-z

THE EUROPEAN
PHYSICAL JOURNAL C

Special Article – Scientific Note

Background and signal estimation for a low mass Higgs boson at the LHC

S. Paganis[a], D.R. Tovey

The University of Sheffield, Hicks Building, Hounsfield Road, S3 7RH Sheffield, U.K.

Received: 28 December 2007 / Revised version: 30 April 2008 /
Published online: 3 June 2008 – © Springer-Verlag / Società Italiana di Fisica 2008

Abstract. The discovery of the Higgs boson(s) is the major goal of the LHC which will start taking data in 2008. In this work a data driven extraction of the background and statistical signal significance in the $H \to ZZ \to 4\ell$ decay channel is presented. The background for Higgs masses as low as 130 GeV can be extracted with an error of 20%, using a sideband measurement from a single 30 fb^{-1} experiment. The predicted background distribution is best described by a double asymmetric Gaussian. An analytic formula is introduced which provides an accurate p-value that a Higgs discovery claim is consistent with a background fluctuation. The formula can be used in a single real measurement at LHC using as input the measured background and the profile likelihood asymmetric errors of this measurement. The method presented here can be applied to the general case of extrapolating from a signal-free data region to a candidate signal region. This is the case of supersymmetry searches at the LHC.

PACS. 14.80.Bn; 06.20.Dk

1 Introduction

The discovery of the standard model (SM) Higgs boson is the major goal of the Large Hadron Collider (LHC). The first proton–proton LHC data at 14 TeV center of mass energy are expected in 2008. The Higgs boson mass is a free parameter in the SM, however there is strong expectation motivated by precision electroweak data [1] and direct searches [2] that a low mass Higgs (114.4–199 GeV, 95% confidence level) should be discovered at the LHC.

The experimentally cleanest signature for the discovery of the Higgs is its "golden" decay to four leptons (electrons and muons): $H \to ZZ \to 4\ell$. The excellent energy resolution and linearity of the reconstructed electrons and muons leads to a narrow 4-lepton invariant mass peak on top of a smooth background. The expected signal to background ratio after all experimental cuts depends on the Higgs mass itself. The major component of the background consists of the irreducible $pp \to ZZ \to 4\ell$ decays. The most challenging mass region is between 120–150 GeV where one of the Z-bosons is off-shell giving low transverse momentum leptons. In this region backgrounds from $pp \to Zb\bar{b} \to 4\ell$ and $pp \to t\bar{t} \to 4\ell$ are significant requiring tight lepton isolation cuts which reduce these backgrounds to levels well below the $pp \to ZZ$ background. As an example, in the 130 GeV mass region a S/B $\simeq 3/1$ is expected with 5–7 events expectation on the background and 12–15 events expectation on the signal after 30 fb^{-1} of integrated luminosity which corresponds to 3 years of LHC running. For lower masses closer to the LEP limit (114.4 GeV) both S/B and signal yield decrease even further, thus minimizing the potential of the 4-lepton channel in that region.

From these S/B expectations it is clear that in order to maximize the Higgs discovery potential using the golden channel, the most accurate achievable knowledge of the background and its associated error is essential. The main characteristic of the 4ℓ final state at the LHC is the presence of a background continuum dominated by $pp \to ZZ \to 4\ell$ decays. This background can be estimated by (i) theoretical predictions, (ii) by a combination of theoretical predictions and subsequent constraints using experimental LHC data, or (iii) by performing a sideband measurement and subsequently extrapolating to the signal region (SR). As discussed in [3], the first two methods suffer from theoretical uncertainties, the luminosity measurement uncertainty (5%–10% only for method (i)) and systematic uncertainties such as lepton reconstruction efficiency. In addition, for low Higgs mass, the $Zb\bar{b}$ contribution is significant with a theoretical error ranging from 20%–50% [4] (ideally it should be experimentally controlled with early LHC data). For these reasons a data-driven background extraction using the sideband measurement (iii) may provide the most accurate determination of the background. This method is not as sensitive to theoretical and luminosity uncertainties, and errors due to lepton and isolation efficiencies, but is limited by the statistics in the sideband.

A background prediction based on a subsidiary measurement is best motivated in searches where an alternate hypothesis is lacking. This is the example of supersymmetry (SUSY) searches at the LHC. In the case of the Higgs

[a] e-mail: paganis@mail.cern.ch

search discussed here, there is a clear alternate hypothesis, the prediction of a standard model Higgs. Inclusion of the signal hypothesis in the case of a low mass $H \to 4\ell$ requires special attention. While a measurement of the $M4\ell$ background is possible from the data itself, the signal must be modelled and uncertainties in its distribution must be considered. In the low mass $M4\ell$ case, we have two flavours of leptons (electrons and muons) coming from on-shell and off-shell Z-boson decays. While the on-shell Z can be measured with data, the off-shell Z is reconstructed by lower energy leptons with large uncertainties in the linearity, energy scale and resolution. The effects of these uncertainties in the $M4\ell$ modelling for the signal must be considered in the calculation of confidence levels, but given the very low expected statistics for a SM Higgs, it is presently unclear if this inclusion brings any gains. For these reasons, we chose to study the problem first with the background-only hypothesis, and the extension of the method presented here to include the signal hypothesis is the subject of future work.

The importance of the discovery of the Higgs boson necessitates a rigorous study of the background extraction and the calculation of the significance of a candidate signal observation. In this work we report on a data driven method for reliably predicting the Higgs background and accurately calculating the confidence level that an observation is consistent with a background fluctuation. The main results of the method are: (i) the predicted background using the sideband is practically unbiased and can be well described by a double asymmetric Gaussian. The uncertainty on the background prediction can be extracted using the error obtained by the profile likelihood method; (ii) an analytic formula (9) which provides the correct probability that an observation is consistent with a background fluctuation. This formula uses the value of the predicted background and its associated asymmetric uncertainty from a sideband fit; (iii) the formula can be applied during a real single experiment by using the profile likelihood errors from a fit to the sideband. Equation (9) can be easily extended to include arbitrary systematic uncertainties.

The method presented here can be applied on ATLAS and CMS analyses to provide an accurate estimate of the significance of an observation. It can also be used in the general case of extrapolating from a signal-free data region to a signal region as in the case of SUSY searches.

2 Background extraction

The low mass $H \to 4\ell$ search at the LHC focuses on narrow Gaussian-like bumps on top of a smooth background dominated by the $pp \to ZZ \to 4\ell$ continuum. The width of such a bump at low masses is dominated by the detector energy linearity and resolution. Measurement of the non-negligible background in the signal-free region and prediction of the background in a candidate signal region is a method that has been considered by both ATLAS [5] and CMS [3].

In this section we study the performance of a background extraction method which exploits the presence of a signal-free sideband-region. The two quantities of interest are the bias (systematic uncertainty) and the uncertainty on the mean from the sideband measurement. At this point it is important to define the above uncertainties. First the shape of the background is not *a priori* known and hence an assumption must be made. We define the assumed shape as $\mathcal{L}(B;\boldsymbol{\delta})$, where B is the number of background events in the signal region and $\boldsymbol{\delta}$ a number of parameters which define the shape of the background. For a particular background shape choice, predictions of B and $\boldsymbol{\delta}$ can be made by fitting the data in the sideband. The result of such a fit will produce:

- B_{pred}: the predicted background in the SR;
- $\sigma_{B_{\text{pred}}}$: the fit error in the predicted background;
- $S_{B_{\text{pred}}}$: the syst. uncertainty (shift) on B_{pred}.

While the predicted background B_{pred} is a measurement given a single experiment, the uncertainties are unknown. However there are procedures based on Monte Carlo (MC) pseudo-experiments which allow for estimation of the above uncertainties. These procedures typically involve studies of all different assumed background shapes which are consistent with the data.

To explain what these uncertainties involve we will take two ideal limiting cases: (i) in the case of infinite statistics in the sideband and knowledge of the true shape of the background we have:

$$B_{\text{pred}} = B_{\text{true}}\,, \quad \sigma_{B_{\text{pred}}} = 0\,, \quad S_{B_{\text{pred}}} = 0, \qquad (1)$$

which means that the uncertainties vanish, and (ii) in the case of finite statistics in the sideband and knowledge of the true shape of the background, the $\sigma_{B_{\text{pred}}}$ is the statistical uncertainty from the sideband fit. The $S_{B_{\text{pred}}} = 0$ since the background shape is known. In a real experiment the background shape is unknown and the statistics in the sideband is finite. Thus, in general $S_{B_{\text{pred}}}$ does not vanish and $\sigma_{B_{\text{pred}}}$ should also receive contributions from uncertainties due to the unknown shape. The experimental challenge is how to estimate these uncertainties given a single experiment and this will be the subject of the rest of this section.

Several existing LHC studies of the Higgs discovery potential assume arbitrary systematic uncertainties [6]. Here we will determine the expected level of bias in the background estimation and we will argue that the uncertainty $\sigma_{B_{\text{pred}}}$ can be estimated using the profile likelihood error of the fit. For the case of $H \to 4\ell$ this uncertainty is dominated by the statistics in the sideband. The systematic bias $S_{B_{\text{pred}}}$ can only be studied for a particular background shape through running MC pseudo-experiments.

2.1 Fitting procedure

In this section we study the distribution of B_{pred} as extracted from sideband fits of toy MC experiments. Each experiment corresponds to an LHC integrated luminosity of 30 fb^{-1} (about 3 years of running).

For our studies a distribution for the $H \to 4\ell$ background is assumed based on Monte Carlo. The simulation includes all three major background components $pp \to ZZ, Zbb, t\bar{t}$, and has gone through certain analysis cuts

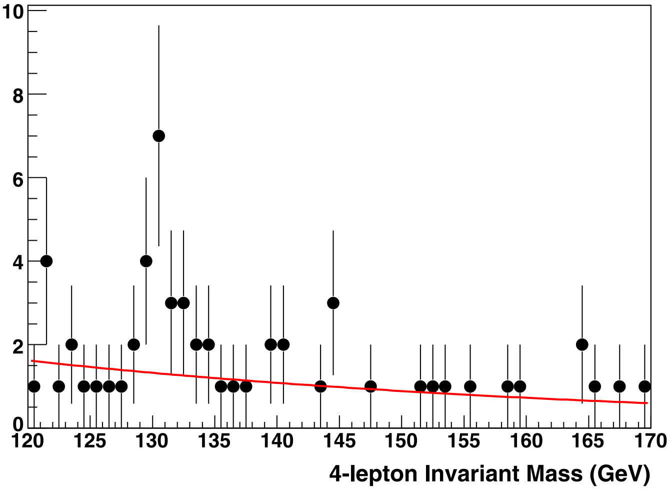

Fig. 1. The 4-lepton invariant mass for a 30 fb^{-1} pseudo-experiment at the LHC. A 130 GeV Higgs has been added. The sideband fit uses an unbinned extended likelihood method and does not include the signal region 127–133 GeV

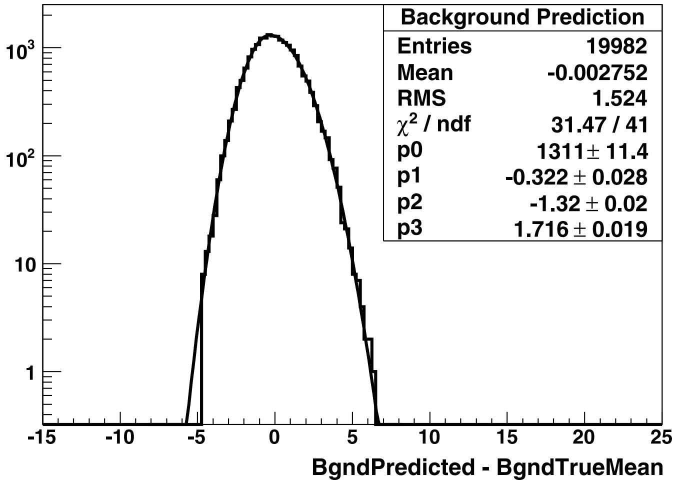

Fig. 2. The distribution of the predicted background from the sideband fit with respect to the true mean background. The mean of the distribution is unbiased and the RMS = 1.524 events. A double asymmetric Gaussian fit provides a good description of the distribution

using a simple fast simulation algorithm. The resulting distribution which corresponds to one 3-year LHC experiment contains on average $B_{\text{true}} = 7$ events in a chosen candidate signal region 127–133 GeV. The background extraction method is evaluated over a large number of such background-only experiments. For each experiment we perform an unbinned extended likelihood fit of the sideband from 120 to 170 GeV, excluding the signal region. All fits were performed using MINUIT [7]. The result of the fit for each experiment is B_{pred} the predicted background in the sideband, and its associated fitting errors. For each experiment we also perform a measurement of the events in the signal region, N_{obs}, which follows the Poisson distribution: $\mathcal{P}(N_{\text{obs}}; B_{\text{true}})$.

Biases coming from uncertainties in the shape distribution of the background, are taken into account because the model parameters are allowed to float in these fits. In order to check the effects of the parametrization itself, a series of background shapes (and normalizations) were assumed, ranging from the true shape distributed as $(M - \delta_1)\exp(-\delta_0\sqrt{M})$ to such shapes as a straight line and $\exp(-\delta_0 M)$ which although inconsistent with the true shape, they would still provide a good fit to the sideband data. An example of such a sideband fit in the presense of an average signal of 15 Higgs events is shown in Fig. 1.

In Fig. 2 the predicted background B_{pred} in the SR is shown for the $B_{\text{true}} = 7$ case. The deviation from the true mean is very small (~ 0.003 events) i.e. B_{pred} is practically unbiased and the RMS = 1.52 events. The distribution is best described by a double asymmetric Gaussian which gives the correct probability content in the tails.

2.2 Background uncertainty from profile likelihood

The sideband fits we performed in the previous section led to a distribution of B_{pred}, best described by a double asymmetric Gaussian. We expect that this distribution must be closely related to the error obtained from the fit. The aim of this section is to study this connection. Since during a real single experiment we will perform a single final fit to the data, we would like to investigate to what extend the resulting error from the fit can be used to reproduce the distribution of B_{pred} in Fig. 2. As we will see the asymmetric fitting error on B_{pred} as obtained using the profile likelihood method, provides a powerful prior to the distribution of B_{pred}.

In this section we use the profile likelihood method to obtain estimates of the distribution of our background prediction in the signal region. The background likelihood function is defined as $\mathcal{L}(B; \boldsymbol{\delta})$ where B is the parameter of interest (the background events in the signal region) and $\boldsymbol{\delta}$ the rest of the parameters of the problem (the parameters that describe the background function). If the maximum likelihood is $\mathcal{L}_{\max} = \mathcal{L}(B^*; \boldsymbol{\delta}^*)$, then the profile likelihood function is the reduced likelihood obtained by maximizing over the parameter $\boldsymbol{\delta}$ at δ_0^*: $\mathcal{L}_{\text{prof}}(B) = \max_{\boldsymbol{\delta}} \mathcal{L}(B; \boldsymbol{\delta})$. The profile likelihood error is found by determining the values of $B = B_{\pm 1\sigma}$ which shift the profile (log) likelihood by a half-unit:

$$2(\ln \mathcal{L}_{\text{prof}}(B) - \ln \mathcal{L}_{\text{prof}}(B^*)) = 1\,. \qquad (2)$$

The profile likelihood error on B_{pred} from a single fit is calculated numerically for each experiment using the MINOS algorithm of the MINUIT program [7]. The distribution of these asymmetric errors for our pseudo-experiments are shown in Figs. 3 and 4. The immediate observation is that the mean of the profile errors

$$\frac{\text{ErrorHigh} + \text{ErrorLow}}{2} = \frac{1.407 + 1.625}{2} = 1.516\,, \qquad (3)$$

is very close to the true error given by the RMS spread of B_{pred} shown in Fig. 2. The spread of the profile errors is quite small, of order 10%. Assuming an exponential background shape $\exp(-\delta_0 M)$ to fit the true $(M - \delta_1) \times$

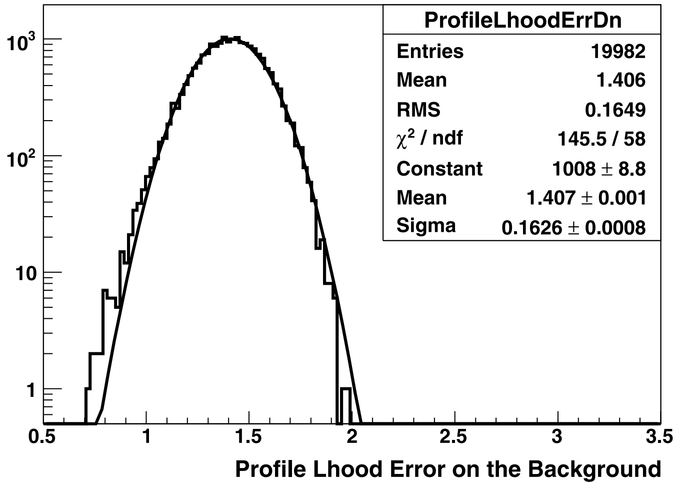

Fig. 3. The distribution of the lower profile likelihood error $\sigma^{\text{low}}_{B_{\text{pred}}}$ on the background from the sideband fit. A Gaussian fit gives a mean error of 1.407 which in combination with the upper error from Fig. 4 is very close to the RMS = 1.52 shown in Fig. 2

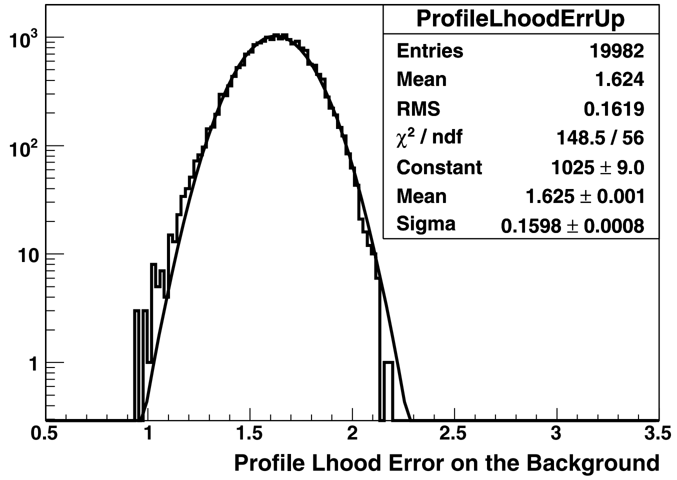

Fig. 4. The distribution of the upper profile likelihood error $\sigma^{\text{up}}_{B_{\text{pred}}}$ on the background from the sideband fit. A Gaussian fit gives a mean error of 1.625 which in combination with the lower error from Fig. 3 is very close to the RMS = 1.52 shown in Fig. 2

$\exp(-\delta_0\sqrt{M})$ shape, leads to practically the same error in the background. However, in that case a small bias appears of $S_{B_{\text{pred}}} = 0.15$, due to the difference in the shape. In conclusion the impact of background shape differences relative to the effect of the background error coming from the statistics in the sideband, are small and they do not affect the results presented here.

3 Calculation of the p-value

Accurate calculation of the p-value and the significance of an observation in the signal region, requires knowledge of the background prediction distribution shown in Fig. 2. However during a single experiment and a single measurement from a low statistics sideband as in Fig. 1, the background prediction distribution is unknown. The asymmetric uncertainty estimates $\sigma^{\text{up}}_{B_{\text{pred}}}$ and $\sigma^{\text{low}}_{B_{\text{pred}}}$ obtained in the previous section are also not known, however a single experiment will provide an estimate of them with a 10% accuracy as shown in Figs. 3 and 4. Here we will use these asymmetric uncertainty estimates $\sigma^{\text{up}}_{B_{\text{pred}}}$ and $\sigma^{\text{low}}_{B_{\text{pred}}}$ on the predicted background B_{pred} to evaluate certain $H \to 4\ell$ discovery criteria. After some necessary definitions, we introduce an analytic expression to calculate the p-value in the presence of the background uncertainties extracted in the previous section. Then we discuss how to apply the method in a real experiment given a single measurement. Finally we study the performance of the analytical calculation and the Cousins–Highland method [8] in obtaining the true p-value (i.e. we study the coverage of these methods).

3.1 Definitions

The chance that we claim a discovery in the presence of background only is called the rate of type I error, α. In the case of a Poisson fluctuating background and in the absence of any uncertainties on the background, α is given by:

$$\alpha = \sum_{N=N_{\text{crit}}}^{\infty} \mathcal{P}(N; B_{\text{pred}})\,, \tag{4}$$

where N_{crit} is the critical event number above which we would claim a discovery. While (4) defines a critical N_{crit} that corresponds to a discovery, in a real experiment where a N_{obs} number of events is measured, we define the p-value as the probability, under the background only assumption, to measure N_{obs} events or more:

$$p = \sum_{N=N_{\text{obs}}}^{\infty} \mathcal{P}(N; B_{\text{pred}})\,. \tag{5}$$

This p-value can be turned to a significance of an observation (number of σ, $S = n\sigma$) as follows:

$$p = \int_{S}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\, \mathrm{d}x\,. \tag{6}$$

In this paper we are concerned with the calculation of the p-value p in the presence of uncertainties in the background prediction from a low statistics subsidiary measurement (the sideband fit).

In the presence of an uncertainty on the background (systematic and statistical due to the sideband), both N_{obs} and B_{pred} in (5) are distributed and the problem becomes 2-dimensional. Since the distributions depend on the unknown background parameters B_{true} and $\boldsymbol{\delta}_{\text{true}}$, we define a function $L(N_{\text{obs}}, B_{\text{pred}}|B_{\text{true}}, \boldsymbol{\delta}_{\text{true}})$ [9,10]. In this case the criterion for discovery given by the type I error (and consequently the p-value) is modified:

$$\alpha' = \int_W L(N_{\mathrm{obs}}, B_{\mathrm{pred}} | B_{\mathrm{true}}, \boldsymbol{\delta}_{\mathrm{true}})\, \mathrm{d}N_{\mathrm{obs}}\, \mathrm{d}B_{\mathrm{pred}}\,, \qquad (7)$$

where the region W on the $(N_{\mathrm{obs}}, B_{\mathrm{pred}})$ plane is defined by cuting beyond some critical line

$$N_{\mathrm{obs}} \geq N_{\mathrm{crit}} = N_{\mathrm{crit}}(B_{\mathrm{pred}} | B_{\mathrm{true}}, \boldsymbol{\delta}_{\mathrm{true}})\,, \qquad (8)$$

which in general depends on B_{pred} given B_{true} and the true background shape. Equation (7) defines an α' for every set of the unknown (nuisance) parameters B_{true} and $\boldsymbol{\delta}_{\mathrm{true}}$. The challenge is to define a method to calculate the correct confidence levels and significance of an observation in the presence of these parameters [9].

3.2 Analytical calculation

We will now introduce, and examine the validity of, an analytic expression to calculate the p-value in the presence of background uncertainties. This approach requires as input the distribution of the predicted background B_{pred} and its goal is to return an accurate p-value. The basic idea is that the presence of an uncertainty in the predicted background always reduces the significance. For a certain measured number of events N_{obs} in a signal region, downward fluctuations on B_{pred} would lead to an apparent larger excess of signal events, thus leading to an overestimate of the significance. Upward fluctuations have the opposite effect. The resulting p-value, p' given all possible fluctuations is always greater than p given by (5), leading to an overestimate of the significance (undercoverage). This is demonstrated in Fig. 5 where the B_{pred} vs. N_{obs} plane is shown for an assumed mean predicted background of 7 events: $\overline{B}_{\mathrm{pred}} = B_{\mathrm{true}} = 7$. The red vertical line $N_{\mathrm{crit}} = 25$ in Fig. 5 corresponds to an approximate 5σ true statistical fluctuation. The underestimate of the p-value is shown in green and occurs for background predictions below the $\overline{B}_{\mathrm{pred}} = 7$. The overestimate is shown in red. The correct p-value can be recovered by adding/subtracting the corresponding probabilities of any possible fluctuation of the predicted background: in Fig. 5 we would add the green portions and subtract the red. This is analytically possible only when the distribution of the predicted background events B_{pred} is known.

The analytic formula which calculates the p-value in the presence of a background B_{pred} that is distributed as a double Gaussian with $\sigma_{B^{\mathrm{up}}_{\mathrm{pred}}}$, $\sigma_{B^{\mathrm{low}}_{\mathrm{pred}}}$ is given by

$$\begin{aligned} p' = 1 &- \sum_{N=0}^{N_{\mathrm{obs}}-1} \mathcal{P}(N; \overline{B}_{\mathrm{pred}}) \\ &- \sum_{i=1}^{\infty} G_i^{\mathrm{up}} \left(\sum_{N=N_{\mathrm{obs}}}^{B^{\mathrm{high}}_{i,\mathrm{max}}-1} \mathcal{P}(N; \overline{B}_{\mathrm{pred}} + i) \right) \\ &+ \sum_{i=1}^{\overline{B}_{\mathrm{pred}}} G_i^{\mathrm{low}} \left(\sum_{N=B^{\mathrm{low}}_{i,\mathrm{max}}}^{N_{\mathrm{obs}}} \mathcal{P}(N; \overline{B}_{\mathrm{pred}} - i) \right)\,, \end{aligned} \qquad (9)$$

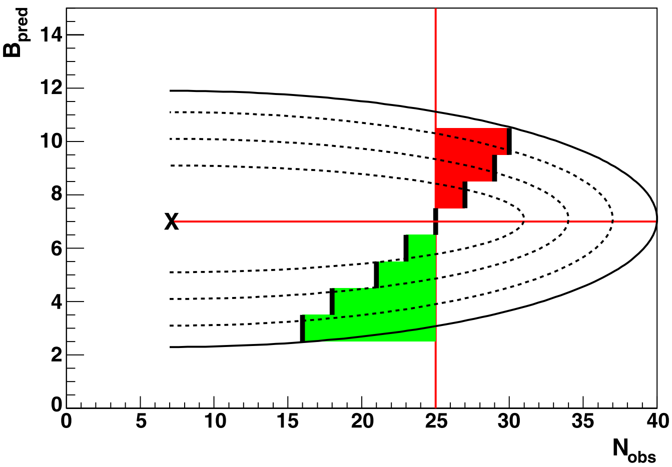

Fig. 5. The distribution of the predicted background from the sideband fit with respect to the observed events in the signal region, for a mean $\overline{B}_{\mathrm{pred}} = 7$. The *vertical red line* denotes the critical number of events $N_{\mathrm{crit}} = 25$ corresponding to a true 5σ (Gaussian equivalent) fluctuation. The *ellipses* represent contours of $L(N_{\mathrm{obs}}, B_{\mathrm{pred}} | B_{\mathrm{true}}, \boldsymbol{\delta}_{\mathrm{true}})$ with eccentricity $\epsilon = \sigma_{B_{\mathrm{pred}}}/\sigma_{N_{\mathrm{obs}}}$ (7). For example the solid ellipse corresponds to the contour corresponding to $\sigma_{B_{\mathrm{pred}}} = 5$ and $\sigma_{N_{\mathrm{obs}}} = 33$ events. For a certain observation of $N_{\mathrm{obs}} = 25$ events, downward fluctuations on B_{pred} would give an apparent extra excess of signal events, thus leading to an overestimate of the significance (an underestimate of the p-value). Upward fluctuations have the opposite effect. The *coloured regions* show the level of overestimation of signal for downward fluctuations (*green*) and the level of underestimation for upward fluctuations (*red*). Taking into account all possible fluctuations of B_{pred} leads to an overestimation of the significance

where

$$G_N^{\mathrm{low}} = \int_{N/2}^{(N+1)/2} G(\overline{B}_{\mathrm{pred}}, \sigma_{B^{\mathrm{low}}_{\mathrm{pred}}})\, \mathrm{d}B_{\mathrm{pred}}\,,$$

$$G_N^{\mathrm{up}} = \int_{N/2}^{(N+1)/2} G(\overline{B}_{\mathrm{pred}}, \sigma_{B^{\mathrm{up}}_{\mathrm{pred}}})\, \mathrm{d}B_{\mathrm{pred}}\,,$$

are the Gaussian probability weights for each possible value $N = B_{\mathrm{pred}}$ of the predicted background. These weights are different for $N > \overline{B}_{\mathrm{pred}}$ (up) and $N < \overline{B}_{\mathrm{pred}}$ (low). Equation (9) gives the p-value taking into account the distribution of the predicted background B_{pred}. It simply adds/subtracts to the p-value of (5) terms corresponding to the different possible values of the prediction B_{pred}. When $B_{\mathrm{pred}} = \overline{B}_{\mathrm{pred}}$ we get the 0th term; when $B_{\mathrm{pred}} = \overline{B}_{\mathrm{pred}} + 1$ then the excess of signal events is underpredicted and the Poisson probabilities corresponding to this underprediction must be subtracted (the 1st term). This subtraction continues until a critical $B^{\mathrm{high}}_{\mathrm{max}}$ that corresponds to the same $n\sigma$ test as the N_{obs} given $B_{\mathrm{pred}} = \overline{B}_{\mathrm{pred}}$ (the black vertical lines in Fig. 5). Similarly when $B_{\mathrm{pred}} = \overline{B}_{\mathrm{pred}} - 1$ the excess of signal events is overpredicted and their probabilities are added. Every term must be weighted by the distribution of the background prediction which is very close to a double asymmetric Gaussian (see Fig. 2).

Equation (9) can be easily generalized to include the systematic bias uncertainty which is relatively small over a wide choice of background parametrizations tried for the $H \to 4\ell$ case.

3.3 Application to a real experiment

Although (9) provides an accurate analytic way of calculating the p-value, it assumes the knowledge of the predicted background distribution. The crucial argument that applies in the case of the $H \to 4\ell$ channel is that we can still use (9) by replacing $\overline{B}_{\rm pred}$ by the measured background from the sideband $B_{\rm Meas}$ and the systematic uncertainty $\sigma^{\rm up}_{B_{\rm pred}}$ by $\sigma^{\rm up}_{B_{\rm Meas}}$, and $\sigma^{\rm low}_{B_{\rm pred}}$ by $\sigma^{\rm low}_{B_{\rm Meas}}$ without a significant loss of accuracy. This is a central result which may allow use of this method at the LHC.

For a single experiment at the LHC, a single fit on the sideband will yield a prediction of the background in the signal region, $B_{\rm Meas}$, and two asymmetric MINOS errors $\sigma^{\rm up}_{B_{\rm Meas}}$ and $\sigma^{\rm low}_{B_{\rm Meas}}$. Equation (9) can then be used to obtain the p-value given a measurement of the events in the signal region. Although systematic effects from the shape of the background are already taken into account (the model parameters are floating in the fit), a family of different parametrizations that can still describe the background can be used to study the stability of the p-value on such shape effects.

3.4 Results

A widely used approach to calculate confidence intervals in the presence of systematic uncertainties on the background is the Cousins–Highland (CH) method [8], in which the pure statistical fluctuations of the background are convoluted with systematic uncertainties. Here for comparison we also use the CH method by considering the profile likelihood error on the background and throwing MC pseudo-experiments. The performance of (9) in providing the correct p-value is summarized in Table 1. The results of the CH method and (9) are given in the last two columns in terms of the significance. The two methods are tested for several values of the true mean background (first column) for various $N\sigma$ tests (fourth column). The results should be compared with the true p-value (eighth column). An immediate observation is a dramatic drop in the significance due to the presence of the background uncertainties. The last line shows that a 5σ observation in the signal region, actually corresponds to a 3.91σ significance. This reduction can be reasonably accurately estimated by plugging the profile likelihood errors from the fit (columns 6 and 7) to the CH method (upper error) or (9) (both errors). This allows for a single experiment at the LHC to accurately estimate the significance using the measured $B_{\rm Meas}$ and $\sigma^{\rm up}_{B_{\rm Meas}}$ and $\sigma^{\rm low}_{B_{\rm Meas}}$ from a single fit to the data. The proposed method performs better than the CH method for which we have used the upper MINOS error. The clear advantage of (9) is that no MC experiments have to be thrown. This allows studies of more general situations where the unknown

Table 1. For a mean number of background events in the signal region $B_{\rm true}$ (column 1) and given critical number of events $N_{\rm crit}$ (column 2) which corresponds to a Poisson probability given in column 3 and an equivalent Gaussian $N\sigma$ significance shown in column 4, the table gives the true p-value for claiming a discovery for background-only experiments (column 8). The extracted significance using the Cousins–Highland method and (9) are shown in columns 9 and 10 respectively

$B_{\rm true}$	B_0 [1]	Prob[2]	Approximate Gaussian $N\sigma$ [3]	$\overline{B}_{\rm pred}$	Sideband errors[4]		True $N\sigma$ (MC)[5]	CH $N\sigma$ (MC)[6]	Calculated $N\sigma$ [7]
					lower	upper			
4	12	9.20×10^{-4}	3	4	1.036	1.253	2.41	2.32	2.44
5	14	7.00×10^{-4}	3	5	1.164	1.382	2.43	2.37	2.48
6	16	5.10×10^{-4}	3	6	1.299	1.517	2.48	2.48	2.52
7	17	9.60×10^{-4}	3	7	1.407	1.625	2.53	2.49	2.56
8	19	6.50×10^{-4}	3	8	1.505	1.723	2.57	2.53	2.60
9	20	10.0×10^{-4}	3	9	1.614	1.832	2.59	2.54	2.61
10	22	7.00×10^{-4}	3	10	1.700	1.918	2.59	2.56	2.61
11	23	10.4×10^{-4}	3	11	1.798	2.016	2.58	2.55	2.61
5	17	1.97×10^{-5}	4	5	1.164	1.382	3.32	3.07	3.24
7	21	1.45×10^{-5}	4	7	1.407	1.625	3.26	3.17	3.26
7	25	1.07×10^{-7}	5	7	1.407	1.625	3.91	3.71	3.88

Explanation of table columns:

[1] Critical number of events in the signal region beyond which a discovery is claimed ($N_{\rm obs} \geq N_{\rm crit}$).

[2] Corresponding Poisson probability that the background will fluctuate above the $N_{\rm obs}$ events in column 2.

[3] Corresponding approximate Gaussian $N\sigma$'s equivalent used in the coverage study.

[4] Upper and lower profile likelihood MINOS errors in the background from the sideband measurement (mean values).

[5] True $N\sigma$ significance in the presence of the background error extracted by throwing pseudo-experiments.

[6] Cousins–Highland $N\sigma$ using the sideband profile likelihood MINOS error from the fit as the error on the background.

[7] Calculated $N\sigma$ significance using the sideband error as the error on the background, based on (9).

systematic bias in the background has to be included and a whole range for this bias has to be scanned.

4 Summary and conclusions

A data driven method for background extraction and p-value estimation in the search for the Higgs boson in the $H \rightarrow ZZ \rightarrow 4\ell$ channel at the LHC was presented. For Higgs masses as low as 130 GeV and integrated luminosity of $30\,\text{fb}^{-1}$, the background in the signal region is best described by a double asymmetric Gaussian and can be extracted with a very small bias. Given a single experiment, the background uncertainty, which is dominated by the statistics in the sideband, is about 20%. The presence of this uncertainty on the background reduces the statistical significance as shown in Table 1. This reduced significance can be calculated by using the analytic formula given by (9) with the upper and lower profile likelihood (MINOS) errors of the fit as a background uncertainty. The method presented here reproduces very well the true signal significance which means that it has excellent coverage properties. It can be applied in the general case of extrapolating from a signal-free region to a candidate signal region. It can also be generalized for the case where the systematic uncertainty (bias from the fit) $S_{B_{\text{pred}}}$ on the background is significant, by shifting the common mean of the Gaussians in (9). The effect of not *a priori* knowing the position of the signal peak (look elsewhere effect) to the significance can be readily extracted.

A couple of final remarks: although a range of different background parametrizations were studied without significantly affecting the results presented here, a formal study of their effects on the p-value is still needed. One could expect large effects in the case where the signal is close to a turning point of the background and the two sidebands are significantly different. In this case one must calculate the p-value by arbitrarily varying the parameters $\boldsymbol{\delta}$ of the fit, and the parametrization itself. As stated in the introduction a background prediction based on a subsidiary measurement is best motivated in searches as SUSY where an alternate hypothesis is lacking. The method presented here can be extended to include an alternate hypothesis (e.g. standard model Higgs signal). However since our focus is on the treatment of systematic effects on the estimation of the p-value, we postpone the inclusion of an alternate hypothesis for a future paper.

In statistical terms the method is frequentistic (B and $\boldsymbol{\delta}$ are treated as nuisance parameters), but with a Bayesian flavor (first we fit the data and then we calculate the significance in a separate step).

Acknowledgements. The authors would like to thank Kyle Cranmer and Markus Schumacher for their valuable input. This work was supported in part by the EU FP6 Marie Curie network Artemis grant MRTN-CT-2006-035657.

References

1. ALEPH, DELPHI, L3 and OPAL Collaborations, arXiv: hep-ex/0612036
2. ALEPH, DELPHI, L3 and OPAL Collaborations, Phys. Lett. B **565**, 61 (2003)
3. CMS Physics Technical Design Report, Vol. 2, CERN/LHC 2006-021
4. J. Campbell et. al., Phys. Rev. D **73**, 054 007 (2006)
5. ATLAS Collaboration, ATLAS Technical Design Report, Vol. 2, CERN/LHC 99-15
6. S. Asai et. al., Eur. Phys. J. C **32**, 19 (2004)
7. F. James, MINUIT, D507, CERN (1978)
8. R.D. Cousins, V.L. Highland, Nucl. Instrum. Methods A **320**, 331 (1992)
9. K. Cranmer, Phystat2003, arXiv:physics/0310108
10. K. Cranmer, Statistical Problems in Particle Physics, Astrophysics and Cosmology, Proceedings PHYSTAT 05, ed. by L. Lyons, M.K. Unel, Imperial College Press