PROCEEDINGS OF SCIENCE

Absorbing systematic effects to obtain a better background model in a search for new physics

S. Caron[a], G. Cowan[b], E. Gross[c], S. Horner*[a] and J. E. Sundermann[a]

[a] *Physikalisches Institut,*
Albert-Ludwigs-Universität Freiburg,
Hermann-Herder-Straße 3, 79104 Freiburg i. Br., Germany
[b] *Royal Holloway,*
University of London,
Egham, Surrey TW20 0EX, UK.
[c] *Weizmann Institute of Science,*
Rehovot 76100, Israel

E-mail: `stephan.horner@gmail.com`

This contribution presents a novel approach to estimate the Standard Model backgrounds based on modifying Monte Carlo predictions within their systematic uncertainties. The improved background model is obtained by altering the original predictions with successively more complex correction functions in signal-free control selections. Statistical tests indicate when sufficient compatibility with data is reached. In this way, systematic effects are absorbed into the new background model. The same correction is then applied on the Monte Carlo prediction in the signal region. Comparing this method to other background estimation techniques shows improvements with respect to statistical and systematic uncertainties. The proposed method can also be applied in other fields beyond high energy physics.

13th International Workshop on Advanced Computing and Analysis Techniques in Physics Research, ACAT2010
February 22-27, 2010
Jaipur, India

*Speaker.

1. Introduction

The way to discover new physics beyond the Standard Model (SM) is to measure a significant deviation from the SM prediction in a signal region, that is a region of phase space where new physics is expected to appear. It is therefore essential that one can have utmost confidence in an estimate on this SM prediction in order to avoid false discoveries and overlooked signals. State-of-the-art Monte Carlo (MC) generators yield such estimates by modelling the relevant physics processes. However, systematic effects due to an imperfect detector and shortcomings in the underlying models of the MC generators lead to an insufficient description of the data. A way to verify and improve the validity of the MC prediction is to compare it with data in a signal-free control region in phase space.

2. Concept of the method

The proposal of this contribution, first introduced in detail in [1], is to reweight the MC estimate by multiplying it with an appropriate correction function. The correction function depends on a set of adjustable parameters which are determined by fitting the modified estimate to the data in the control region. Then the same function is to be applied on the corresponding template in the signal region.

To illustrate the proposed method consider two possible scenarios for measurements in a control region shown in Fig. 1. Both measurements are compatible with the uncertainty of their respective predictions, which is obtained by varying known systematic effects. In Fig. 1, left, however, the data strongly deviate from the central prediction which hints at substantial systematic effects being present in that scenario. This assumption is further supported by a *p*-value, which quantifies the agreement between prediction and data (see [1]), of only about 0.3%. In Fig. 1, right, the deviations seem compatible with statistical fluctations as is reflected by a *p*-value of about 56%. In both cases, the data shall now be used in an attempt to obtain a better background model following the procedure outlined above.

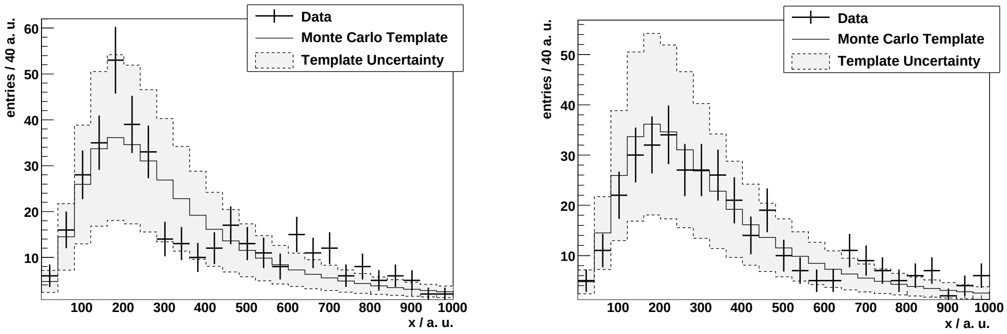

Figure 1: Two scenarios for measurements in a control region having the same Monte Carlo prediction.

Correction function	*p*-value
none (fixed to unity)	0.0027
Constant	0.0033
Linear	0.0038
Quadratic	0.018
Cubic	0.052
4th degree	0.33
5th degree	0.46
6th degree	0.46
7th degree	0.69
8th degree	0.63
9th degree	0.68
10th degree	0.60

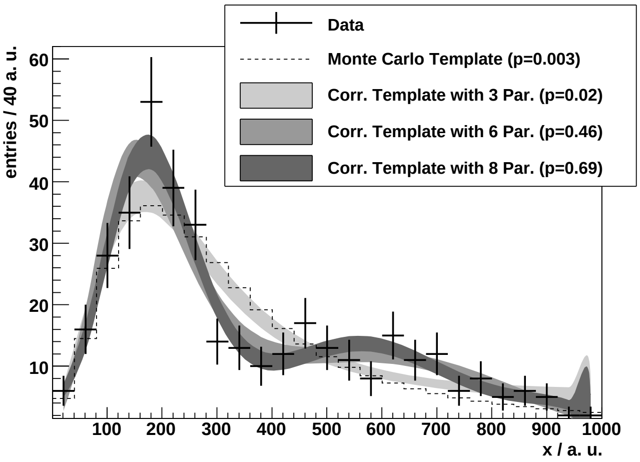

Figure 2: The Monte Carlo estimate for the data distribution in Fig. 1, left, is modified with correction functions of an increasing number of parameters until a satisfactory goodness-of-fit is reached, expressed by the *p*-value.

2.1 First scenario: Large systematic effects

Choosing the best correction. For this example ordinary polynomials are taken as correction functions. Starting with the first scenario, the central prediction ("zeroth-order model") is modified with polynomials of order 2, 5, and 7, displayed in increasing shades of grey in Fig. 2, right. The width of the bands corresponds to the respective statistical uncertainty. The table on the left shows the *p*-values for functions up to order 10. Using a correction function of degree 5 yields the first model with an acceptable goodness-of-fit of 0.46, which can be improved even further by including more parameters.

The compatibility peaks at a value of 0.69 when using 8 adjustable parameters. The decline for more complex models results from increasing the number of free parameters while not gaining a substantial improvement in terms of data description. The model with the highest *p*-value is taken as the new improved background estimate.

Alternative starting templates. Apart from the statistical error of the fit, an additional uncertainty arises from the choice of the starting template. To investigate the dependency of the corrected model on a particular shape of the original hypothesis, additional starting templates are selected from within the systematic uncertainty of the MC prediction, as shown in Fig. 3, left. As was mentioned above, in the case of real data one would vary the MC prediction according to known systematic effects, thereby obtaining a set of possible starting templates. All those templates are corrected separately with the polynomial yielding the highest goodness-of-fit, shown in Fig. 3, right. In addition, the true model from which the data were generated is displayed as the black solid line. After correction the new models nicely converge to the true model, almost regardless of the shape of the starting template. Comparing the true model with the original MC prediction (thick red line on the left) reveals the rather extreme systematic effects, which call for a correction using several parameters.

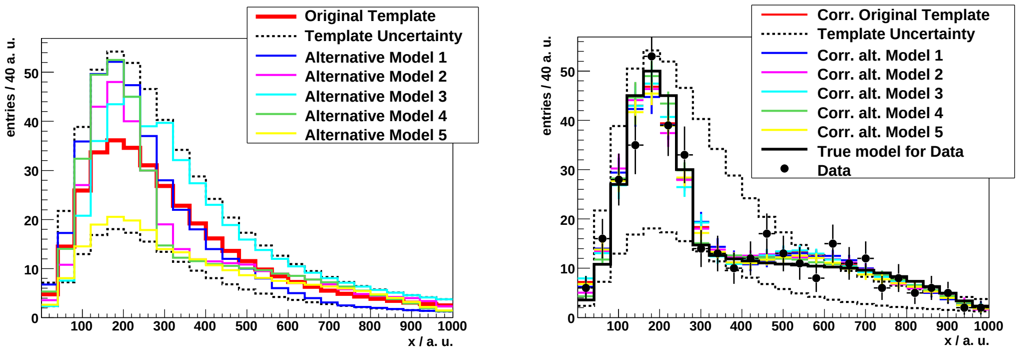

Figure 3: Selecting different templates within the systematic uncertainty of the original Monte Carlo prediction to estimate their influence on the corrected model.

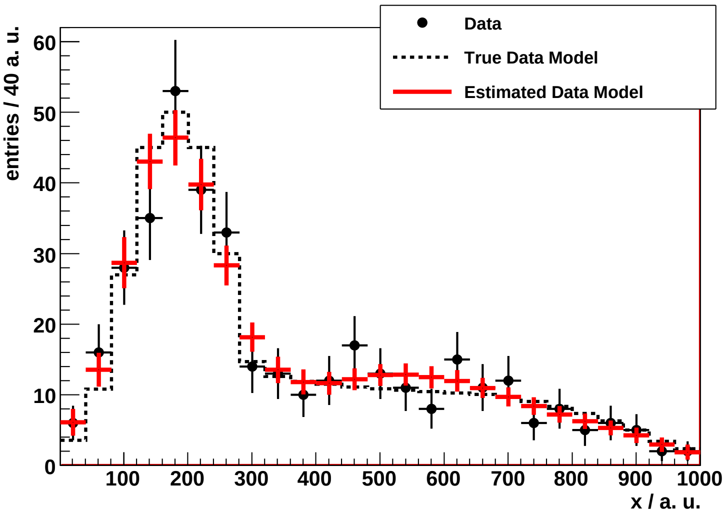

Figure 4: The estimated and the true model for the data agree well within the indicated uncertainty.

New background model and its uncertainty. A priori, none of the different starting templates in Fig. 3, left, can be favored over the others. Thus, the best estimated model is finally taken as the mean value of all corrected templates. Its total uncertainty is calculated by generating 2000 toy data sets from this estimate and applying the proposed method on every one of them.

The bin-wise RMS of the corrected models' distribution together with the inter-bin correlation is then taken as an estimate for the statistical error. Fig. 4 contrasts the best estimated model with the true model for the data. The true model is nicely reproduced.

2.2 Second scenario: No systematic effects

The second scenario shown in Fig. 1, right, shall illustrate the usefulness of the proposed method when there is apparently only little or no systematic deviation present. As before, the different MC templates of Fig. 3, left, are modified with the correction function giving the respective highest *p*-value.

As a limiting case, this scenario was generated without systematic effects. Hence the central prediction would constitute the best model. Still, the proposed method has the various templates converge to the true model - see Fig. 5, left. Fig. 5, right, shows the estimated and the true model for the data which agree well within the indicated uncertainty, which could be substantially reduced. The offset at higher x-values results from a bias introduced by the data.

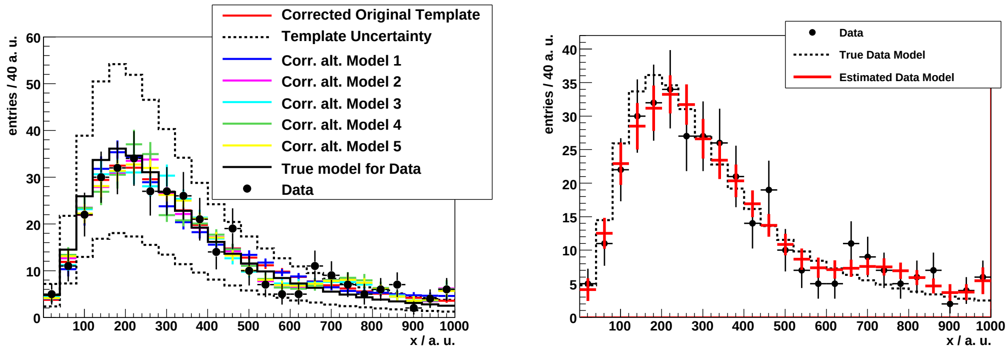

Figure 5: Correcting the templates in Fig. 3, left, for the second scenario yields nice agreement with the true model (left). On the right, the averaged estimated model with its uncertainty is shown. .

2.3 Extrapolation to signal region

Once the improved background model has been determined following the procedure described above the same correction is to be applied on the Monte Carlo prediction for the background in the signal region. In addition, systematic effects associated with the transfer of the correction from control to signal region need to be considered, such as the different influence on the shapes of the distributions by certain systematic sources. These have to be treated on a case-by-case basis and lie beyond the scope of this contribution. An advantage of this method is that the templates in control and signal region need not have identical shapes, only the systematic effects have to influence them in a similar way.

3. Performance against using control region data as a model

If a control region can be defined such that the shapes of the relevant background processes are practically identical to the ones in the signal region a simple scaling of the data can be used to get a model for the background in the signal region. In a first approximation, the uncertainties of such a model are simply the square root values of the data. For simplicity, assume the efficiency of signal to control region to be unity. In this case the model determined in the control region can be taken as-is for the signal region. Consider again the second scenario discussed in subsection 2.2, depicted in Fig. 1, right.

3.1 Background estimates for a new physics search

In a search for new physics one is often interested in the high mass tails of distributions for being the most sensitive regions to discover new phenomena. Suppose this region to include all

Table 1: Number of expected events for $x > 600$ a.u. predicted by different models. The error of the corrected model is the same as the one from the data in this case, in general it is smaller (see Fig. 6). The MC template is identical to the true data model since no systematic effects were introduced in this scenario.

Model	Number of expected events	Relative error
Original prediction (MC template)	43.9 ± 21.9	50%
Corrected model	59.9 ± 7.6	12.7%
Data as model	62.0 ± 7.9	12.7%

x-values greater than 600 a.u. of Fig. 1. Table 1 summarizes the expected number of events and its uncertainty for the original prediction, the corrected model and when taking the data as the model. Both the data model and the corrected model have a comparable and much smaller uncertainty than the original prediction. The inter-bin correlation boosts the error of the corrected model to the level of the data uncertainty in this example.

In order to obtain a general statement on how the error of the proposed method compares with the one from the data 10000 pseudo data sets are created from the true data model. In principle, a smaller uncertainty than when using the data as model can be achieved if the number of parameters used is smaller than the number of data bins. Fig. 6 shows the distribution of events for x-values greater than 600 a.u. Taking the data as the model, it produces an unbiased prediction of 43.92 events for the mean value with an error of 6.68 events, as expected in agreement with the true values of 43.89 and 6.63 within the statistical limitation of the sample. Applying the proposed method yields on average a value of 44.14 and a reduced error of 6.26. The mean value is slightly positively biased but only by about 4% of the quoted uncertainty.

More knowledge about the true shape of the distribution can reduce the uncertainty of the method even further since fewer parameters will be needed for the adjustment of appropriate starting templates. As a limiting case, five templates which only differ in their normalization with respect to the true model are employed. The resulting distribution of expected events is also displayed in Fig. 6 as the dashed red line, demonstrating a further decrease of the error to 5.92.

3.2 Significance of a possible signal

In order to investigate how the different errors affect the discovery potential, two toy measurements for the two regions $x > 600$ a.u. and $x > 800$ a.u. are assumed to be 99 and 52 events respectively as shown in table 2. High energy physics folklore considers a measurements to be a discovery if the probability, assuming only known physics, of observing data as or less likely is smaller than $2.9 * 10^{-7}$, which corresponds to the integrated tail of a Gaussian distribution beyond five standard deviations ("5σ discovery”).

Using the data from the control region as the background model one would claim a discovery since the significance, which is calculated by convoluting the Poisson probability for the data with the Gaussian prior function representing the systematic uncertainty of the background (see e.g. [2] and [3]), surpasses the 5σ threshold. Taking instead the predicted mean value and error of the proposed method using the different starting templates the significance grows to 5.12 and 5.29 for the two regions. It can be even further raised to 5.25 and 5.38 when using the set of "same shape

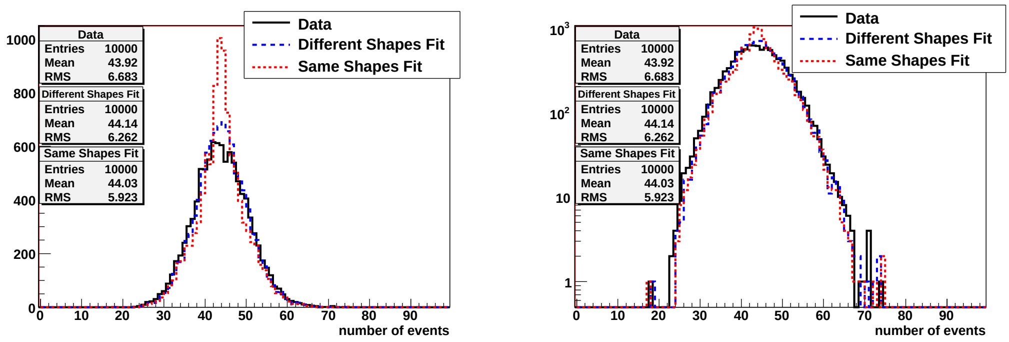

Figure 6: Distributions of number of expected events in the region $x > 600$ a.u. predicted by different models with linear and logarithmic ordinate. The proposed method using the "different shapes" from Fig. 3 yields a smaller RMS than using the data as a model. The uncertainty can be further reduced by using templates more similar to the true model, in this case only differing in the scale but having the "same shape".

Table 2: The significance of a discovery can be increased by using the proposed method instead of data from the control region as a background model. The measurements of 99 and 52 events for the two regions were chosen to allow for a 5σ discovery when using the data. The increase in significance is equivalent to a saving in luminosity as described in the text.

	$x > 600$ a.u.: 99 events		$x > 800$ a.u.: 52 events	
	Background:	Significance:	Background:	Significance:
Data	43.92 ± 6.68	5.01	15.62 ± 3.93	5.10
Different Shapes	44.14 ± 6.26	5.12	15.56 ± 3.60	5.29
Same Shapes	44.03 ± 5.92	5.25	15.53 ± 3.45	5.38

templates". The jump in significance is equivalent to an increase in luminosity of 4% and 12% for the two regions respectively when using the same shape fit results instead of the data. Thus, by using the proposed method the required integrated luminosity for a discovery is reduced. This effect gets bigger the smaller the inspected tail region compared to the region in x which was used to determine the background model.

4. Summary and conclusion

The underlying idea of the method presented in this contribution is to correct the Monte Carlo background estimates for systematic deviations. To that end, they are multiplied with successively more complex correction functions until a statistical test reports good compatibility with data in a control region. The correction determined that way is then applied on the corresponding templates in the signal region yielding an improved background model to search for new physics.

While systematic effects are absorbed by the correction functions, the total uncertainty of the model can be reduced compared to other common methods. In order to avoid absorbing a possible signal in the fit carried out in the control region the Monte Carlo estimates can be varied according to known systematic effects, thereby obtaining constraints on the maximal acceptable modification of the templates.

The usefulness of the proposed method is not restricted to high energy physics. It can be applied in other scientific fields where one uses data from control regions to estimate the background in a signal region and is confronted with large systematic uncertainties.

Acknowledgments

S. Caron, S. Horner and J. E. Sundermann gratefully acknowledge the support by the BMBF, the Landesstiftung Baden-Württemberg and the Graduiertenkolleg ”Physik an Hadron-Beschleunigern” of the DFG. E. Gross is obliged to the Benoziyo center for High Energy Physics, to the the Israeli Science Foundation (ISF), the Minerva Gesellschaft and the German Israeli Foundation (GIF) for supporting this work.

References

[1] S. Caron, G. Cowan, E. Gross, S. Horner and J.E. Sundermann *Absorbing systematic effects to obtain a better background model in a search for new physics, JINST* 4 P10009 (2009), [arXiv:0909.3718].

[2] J. T. Linnemann, *Measures of significance in HEP and astrophysics*, in the proceedings of *PhyStat2003: Statistical Problems in Particle Physics, Astrophysics, and Cosmology*, September 8-11, SLAC, Stanford, California U.S.A. (2003) physics/0312059, see online at http://www.slac.stanford.edu/econf/C030908/papers/MOBT001.pdf.

[3] ATLAS collaboration, G. Aad et al., *Expected performance of the ATLAS experiment, detector, trigger and physics* , [arXiv:0901.0512].