Available on CMS information server

CMS AN -2010/261

The Compact Muon Solenoid Experiment

Analysis Note

The content of this note is intended for CMS internal use and distribution only

08 September 2010

Computing the contamination from fakes in leptonic final states

H. Bakhshian, A. Fahim (IPM, Tehran, Iran), A. Calderon, P. Lobelle, J. Marco, P. Martinez, T. Rodrigo, R. Vilar (IFCA, CSIC-Universidad de Cantabria, Santander, Spain), G. Cappello, M. Chiorboli, A. Tricomi (INFN, Sezione di Catania, Italy), J. Cuevas, J. Fernandez, L. LLoret (Universidad de Oviedo, Spain), G. Dissertori, P. Milenovic, F. Moortgat, P. Nef, L. Pape, F. Ronga, L. Sala, B. Stieger, K. Theofilatos, D. Treille and M. Weber (ETH Zurich, Zurich, Switzerland).

Abstract

This note presents general formulae for the calculation of backgrounds caused by fake leptons, based on the so-called fakeable object method. Although it is applied here only to leptons, the method is more general and could also be of interest to other applications. We generalize the well known approach by including the tight-to-loose efficiency for prompt leptons, by allowing for an arbitrary number of leptons, and for events containing both electrons and muons. Formulae are also derived for the numbers of events which include given numbers of fake leptons in the final selection and provide constraints on specific types of backgrounds. The statistical and systematic uncertainties on the estimates are derived and applied to examples taken from top pair production and supersymmetry. An alternative method using recursive formulae, which may have advantages in certain conditions, is also discussed.

CMS Draft Analysis Note

The content of this note is intended for CMS internal use and distribution only

2010/09/08
Head Id: 13082
Archive Id: 13294M
Archive Date: 2010/07/05
Archive Tag: trunk

Computing the contamination from fakes in leptonic final states

Hamed Bakhshian[1], Alicia Calderon[2], Gigi Cappello[3], Massimiliano Chiorboli[3], Javier Cuevas[4], Guenther Dissertori[5], Ali Fahim[1], Javier Fernandez[4], Lara LLoret[4], Patricia Lobelle[2], Jesus Marco[2], Pablo Martinez[2], Predrag Milenovic[5], Filip Moortgat[5], Pascal Nef[5], Luc Pape[5], Teresa Rodrigo[2], Frederic Ronga[5], Leonardo Sala[5], Benjamin Stieger[5], Konstantinos Theofilatos[5], Daniel Treille[5], Alessia Tricomi[3], Rocio Vilar[2], and Matthias Weber[5]

[1] IPM, Tehran, Iran
[2] Santander, IFCA, CSIC-Universidad de Cantabria, Spain
[3] Catania, INFN Sezione di Catania, Italy
[4] Oviedo, Universidad de Oviedo, Spain
[5] ETH Zurich, Zurich, Switzerland

Abstract

This note presents general formulae for the calculation of backgrounds caused by fake leptons, based on the so-called fakeable object method. Although it is applied here only to leptons, the method is more general and could also be of interest to other applications. We generalize the well known approach by including the tight-to-loose efficiency for prompt leptons, by allowing for an arbitrary number of leptons, and for events containing both electrons and muons. Formulae are also derived for the numbers of events which include given numbers of fake leptons in the final selection and provide constraints on specific types of backgrounds. The statistical and systematic uncertainties on the estimates are derived and applied to examples taken from $t\bar{t}$ and supersymmetry. An alternative method using recursive formulae, which may have advantages in certain conditions, is also discussed.

This box is only visible in draft mode. Please make sure the values below make sense.

PDFAuthor:	Luc Pape
PDFTitle:	Computing the contamination from fakes in leptonic final states
PDFSubject:	CMS
PDFKeywords:	CMS, physics, SUSY, top

Please also verify that the abstract does not use any user defined symbols

Contents

1 Introduction

In this note, we derive the formulae to estimate the background caused by fake leptons. It should be noted, however, that this formulation could in principle also apply to different problems, like the b-tagging or the identification of hadronically decaying taus. The estimation of the background contribution due to fake leptons is important, for instance, for the Opposite Sign (OS) dileptonic $t\bar{t}$, contaminated by $W + jets$, and for Same Sign (SS) dileptonic SUSY, contaminated by $t\bar{t}$.

In this context, "fakes" are taken in a broad sense. They may be truly fakes, e.g. jets taken as electrons. But they may also be real leptons from a heavy flavour decay or electrons from γ conversions. The contamination from such leptons can be controlled by their isolation (Iso) and identification (ID). In contrast, "prompt" leptons are the leptons characterizing the signal, like leptons from W decay in $t\bar{t}$ or from SS chargino decay in SUSY. In addition, real leptons, as from Drell-Yan in $t\bar{t}$ or from SS $W^{\pm}W^{\pm}$ in SUSY, will be counted as prompt, as their isolation and identification properties are nearly indistinguishable from the ones of signal leptons. They are supposed to be estimated and subtracted by other methods.

The estimation of the fake lepton contribution is based on the "fakeable objet" method, see e.g. [1, 2] for recent presentations. It looks for leptons, satisfying loose Iso+ID criteria, which pass or fail tight criteria, i.e. the criteria for signal selection. All other criteria for the loose selection (like lepton p_T cuts, etc) are supposed to be the same as for the tight selection. The ratio of fake leptons passing the tight criteria over fake leptons passing the loose criteria is called the "fake ratio", f. This ratio can be determined from the abundantly produced QCD or other events as a function of the lepton kinematics. It is assumed that, once the lepton kinematics is taken into account, the same "universal" fake ratios can be applied to the other backgrounds to extract the signal.

The fakeable objet method is not new and has been presented in several occasions and applied to single lepton or dilepton events. In this note, the method is generalized in two ways. First, we include the "prompt ratio", p, of tight to loose leptons also for prompt leptons, which is usually ignored and assumed to be 1. This ratio could be measured in $Z^0 + jets$ events by a tag and probe method (or perhaps in $W + 0jets$ events, but only in a limited kinematical range). Second, we derive the formulae for an arbitrary number of leptons, including the case of both e and μ (with different f and p) in the same event. We also derive the formulae for the estimation of statistical and systematic uncertainties and correlations.

In addition, recursive formulae are derived which allow different fake ratios, directly determined from the different backgrounds, to be used for the estimate of signal and background contributions. This may avoid the difficulty of estimating a really universal fake ratio, which may depend on many event characteristics.

Here, we only work out the algebra for the estimation of the number of events with prompt leptons, but do not attempt to apply it to Monte Carlo simulation (MC) nor real data. We plan to apply these different methods and compare their results.

2 Fake rates for single lepton events

To start with, let us consider the simplest case of single leptons. This is relevant for the $W + jets$ and for the semi-leptonic $t\bar{t}$ channels, contaminated by fake leptons mainly from QCD. The same is true for SUSY single lepton channels.

Suppose that the total number, N_l, of leptons passing the loose criteria is made of N_p prompt and N_f fake leptons. The numbers N_p and N_f are not directly measurable. However, they are related to the numbers of events where no lepton, N_{t0}, or 1 lepton, N_{t1}, pass the tight criteria by

$$\begin{aligned} N_l &= N_p + N_f = N_{t0} + N_{t1} \\ N_{t0} &= (1-p)N_p + (1-f)N_f \\ N_{t1} &= pN_p + fN_f \end{aligned} \tag{1}$$

In these expressions, f is the ”fake ratio”, i.e. the probability for a fake lepton passing the loose criteria to also pass the tight ones, and p, or ”prompt ratio”, the probability for a prompt lepton passing the loose criteria to also pass the tight ones. The ratios f and p depend on the lepton kinematics (p_T and η). The factors can be interpreted as averages over the lepton spectra, specifically over the fake leptons for f and over the prompt leptons for p. Equivalently, they can be used as event by event weights over the fake and prompt leptons. But, we do not know which leptons are fake and which are prompt. What can be measured, instead, is N_{t1} and N_{t0}.

These relations are easily inverted to obtain N_p and N_f, from which the number of events with a prompt lepton and its contamination at the level of the loose selection are derived:

$$\begin{aligned} N_p &= \frac{1}{p-f}\left[(1-f)N_{t1} - fN_{t0}\right] \\ N_f &= \frac{1}{p-f}\left[pN_{t0} - (1-p)N_{t1}\right] \end{aligned} \tag{2}$$

Then, the number of selected prompt (signal) events is given by $N_{signal} = N_p^{pass} = pN_p$ and the number of fakes passing the tight cuts by $N_{contam} = N_f^{pass} = fN_f$.

In the factor $(p-f)$ in front of the square brackets, p is usually assumed to be 1. The working point for the loose selection is usually taken such that $p \geq 90\%$ and $f \simeq 10-20\%$, hence this factor is indeed expected to be close to $(1-f)$. But here we want to keep p explicitely in the expressions, as at some stage of precision it will matter.

We can now give a more precise definition of the factors in equation (2). As p and f depend on the lepton kinematics, the factors can be used as weights applied to the individual leptons. The overall factor $1/(p-f)$ should be applied to both passing and failing leptons. Additionally, the weight for the estimate of prompt leptons should be $(1-f)$ for passing leptons, f for failing leptons. For the estimate of fake leptons, it should be $(1-p)$ for passing leptons and p for failing leptons. The same convention for the weights will also be encountered below for multi-lepton estimates. It is also seen that the number of events with a prompt lepton is given by the number where the lepton passes the tight criteria from which the number of events failing, both multiplied by a function of the fake ratio f, has to be subtracted. On the other hand, the number of events with a fake lepton is obtained from the number where the lepton fails the tight criteria from which the number of events passing, both multiplied by a function of the prompt ratio p, has to be subtracted.

It should be emphasized that not only N_{signal}, but also N_{contam} is important to estimate, as the latter provides a handle on the QCD background for the examples mentioned above.

An alternative notation could be based on using small numbers inside the square brackets:

$$\epsilon = \frac{f}{1-f} \ , \ \eta = \frac{1-p}{p} \tag{3}$$

The interpretation of these quantities is as follows. The number of prompt leptons will be derived from N_{t1}, but this also contains a contribution from fakes. The number of fake leptons which fail the tight cuts is $N_f^{fail} = (1-f)N_f$. The number which pass the tight cuts and give rise to contamination is then

$$N_f^{pass} = fN_f = \frac{f}{1-f}N_f^{fail} = \epsilon N_f^{fail}$$

It shows that the fake leptons passing the tight cuts is expressed as a function of the ones which fail, weighted by ϵ. Similarly, for N_f which will be derived from N_{t0} from which the contribution of prompt leptons should be subtracted. Using an obvious notation, we have:

$$N_p^{fail} = \frac{1-p}{p}N_p^{pass} = \eta N_p^{pass}$$

This notation also facilitates the comparison with [1] . The factor multiplying the square bracket has in common

$$p - f = p(1-\frac{f}{p}) = p(1-f-\frac{f}{p}+f) = p(1-f)(1-\epsilon\eta) \tag{4}$$

The nicest expressions are obtained if, instead of computing N_p and N_f, we take the events which pass the tight cuts, namely N_p^{pass} and N_f^{pass}. Then, factoring out $(1-f)$ or p from the square bracket in (2)

$$\begin{aligned} N_p^{pass} &= p\frac{1-f}{p-f}\left[N_{t1} - \frac{f}{1-f}N_{t0}\right] = \frac{1}{1-\epsilon\eta}\left[N_{t1} - \epsilon N_{t0}\right] \\ N_f^{pass} &= f\frac{p}{p-f}\left[N_{t0} - \frac{1-p}{p}N_{t1}\right] = \frac{\epsilon}{1-\epsilon\eta}\left[N_{t0} - \eta N_{t1}\right] \end{aligned} \tag{5}$$

It is seen that the number of events with a prompt passing lepton is given by the total number with leptons passing the tight criteria from which the number of events failing, multiplied by ϵ, has to be subtracted. The factor in front of the square brackets is usually ignored, as p is assumed to be 1 and, as seen above, this factor is indeed expected to be close to unity. More precisely, these factors are to be used as weights applied to the leptons, with the factor ϵ weighting the prompt lepton which fails the tight criteria.

The number of fake passing leptons is given by the total number of events with no leptons passing the tight cut multiplied by ϵ, from which we subtract the number of events with 1 lepton passing, multiplied by $\epsilon\eta$.

In conclusion, all leptons (prompt or fake) which fail the tight cuts have to be weighted by ϵ, fake leptons which pass by $\epsilon\eta$ and prompt leptons which pass by 1.

3 Fake rates for same flavour dilepton events

Fake rates are important for the background estimate in $t\bar{t}$ dileptons, and in SUSY for the SS dileptons. For Same Flavour (SF) dilepton events, the total number of leptons passing the loose cuts consists of N_{pp} events with both prompt leptons, N_{fp} events with one lepton prompt and one fake and N_{ff} events where both leptons are fake. These numbers can be related to the measurable numbers of these events N_{tx} $(x = 0, 1, 2)$ with 0, 1, or 2 leptons passing the tight cuts, the remaining ones failing these cuts.

$$
\begin{aligned}
N_l &= N_{pp} + N_{fp} + N_{ff} = N_{t2} + N_{t1} + N_{t0} \\
N_{t0} &= (1-p)^2 N_{pp} + (1-p)(1-f) N_{fp} + (1-f)^2 N_{ff} \\
N_{t1} &= 2p(1-p) N_{pp} + [f(1-p) + p(1-f)]\, N_{fp} + 2f(1-f) N_{ff} \\
N_{t2} &= p^2 N_{pp} + pf N_{fp} + f^2 N_{ff}
\end{aligned}
\tag{6}
$$

These equations assume that the prompt and the fake ratios for different leptons are independent of each other. It is seen that the factors p and $(1-p)$ are again weighting (or are averaged over) the distribution of prompt leptons and f and $(1-f)$ the distributions of fake leptons.

The matrix of these linear equations can be inverted, yielding:

$$
\begin{aligned}
\begin{pmatrix} N_{pp} \\ N_{fp} \\ N_{ff} \end{pmatrix} &= \frac{f-p}{det} \begin{pmatrix} f^2 & -f(1-f) & (1-f)^2 \\ -2fp & p(1-f) + f(1-p) & -2(1-p)(1-f) \\ p^2 & -p(1-p) & (1-p)^2 \end{pmatrix} \begin{pmatrix} N_{t0} \\ N_{t1} \\ N_{t2} \end{pmatrix} \\
det &= -(p-f)^3
\end{aligned}
\tag{7}
$$

From this, the number of events with prompt leptons is derived:

$$
N_{pp} = \frac{1}{(p-f)^2} \left[(1-f)^2 N_{t2} - f(1-f) N_{t1} + f^2 N_{t0} \right] \tag{8}
$$

with the number of signal events being given by $N_{signal} = p^2 N_{pp}$.

This formula confirms clearly the recipe to apply weights to the leptons, which are summed (with the correct signs) over all events in order to extract the number of events with prompt leptons only. The factor in front of the square brackets is again near $1/(1-f)^2$ and is a weight affecting all leptons. Inside the brackets, every lepton passing the tight cuts is weighted by $(1-f)$, the ones failing are weighted by f, as seen in the single lepton case.

The second term has a minus sign, as expected for the subtraction of the contamination. A subtelty is, however, that the third term is added, instead of subtracted. This is due to the fact that N_{t1} has a contribution from N_{fp}, but also from N_{ff} and the latter is double counted because it has two fake leptons. Hence, the N_{t0} contribution has to be subtracted from the contamination, which corresponds to the third term with a positive sign.

In addition to the number of prompt events, we also obtain from the matrix inversion the number of events with 1 prompt and 1 fake lepton and the number of events with both fake leptons:

$$
\begin{aligned}
N_{fp} &= \frac{1}{(p-f)^2} \left[-2fp N_{t0} + [f(1-p) + p(1-f)]\, N_{t1} - 2(1-p)(1-f) N_{t2} \right] \\
N_{ff} &= \frac{1}{(p-f)^2} \left[p^2 N_{t0} - p(1-p) N_{t1} + (1-p)^2 N_{t2} \right]
\end{aligned}
\tag{9}
$$

For the mixed case of N_{fp}, the procedure is more complicated. A factor $1/(p-f)$ has to be applied to both leptons, hence its square. Failing leptons are weighted by f or p and passing leptons by $(1-f)$ or $(1-p)$, as before. As we do not know for N_{t0} which lepton was prompt and which one fake, we weight alternatively lepton 1 with f and 2 with p and lepton 1 with p and 2 with f, adding the weights (hence the factor 2). The same is done with $(1-f)$ and

$(1-p)$ for N_{t2}. For N_{t1}, the failing lepton gets alternatively a weight f and p and the passing lepton simultaneously $(1-p)$ and $(1-f)$.

The corresponding backgrounds remaining in the selected (tight) sample are then respectively pfN_{fp} and f^2N_{ff}. For the dileptonic $t\bar{t}$ analysis, the first one measures mainly the $W+jets$ and the second the QCD backgrounds. For SS SUSY dileptons, the first one corresponds to the dominant $t\bar{t}$ background, where the prompt lepton comes from a top decay and the fake lepton from the leptonic b decay of the other top. It might also include a contribution from $W+jets$, where the W decays into τ and the τ into ρ, the latter faking an electron. The second expression then gives a handle on the remaining QCD contribution.

The above notation can again be changed to the one using small numbers, for easier comparison with [1].

$$N_{pp}^{pass} = \frac{1}{(1-\epsilon\eta)^2}\left[N_{t2} - \epsilon N_{t1} + \epsilon^2 N_{t0}\right] \tag{10}$$

The factor in front of the square brackets is again near unity and $1/(1-\epsilon\eta)$ is a weight affecting all leptons. Inside the brackets, every prompt lepton failing the tight cuts is weighted by ϵ. The justification of this weight was commented on in the single lepton section. The third term is added, rather than subtracted, for the same reason of double counting the fake leptons as explained in relation with equation (8).

The weights in this formula agree with the method presented in [1] and generalize it to the case with $p \neq 1$. But it seems to depart from [1] as far as the signs are concerned, as it is said that "the sum of the weights over the selected events is the background prediction", which would imply a negative sign in front of the third term (maybe corrected in later versions).

For the mixed case of N_{fp} and for N_{ff}, we can rewrite the above formulae as

$$\begin{aligned} N_{fp}^{pass} &= \frac{\epsilon}{(1-\epsilon\eta)^2}\left[-2\epsilon N_{t0} + \left[1+\epsilon\eta\right]N_{t1} - 2\eta N_{t2}\right] \\ N_{ff}^{pass} &= \frac{\epsilon^2}{(1-\epsilon\eta)^2}\left[N_{t0} - \eta N_{t1} + \eta^2 N_{t2}\right] \end{aligned} \tag{11}$$

The weighting rules for the background contributions are an extension of the ones described above for equation (10). As in the single lepton case, failing leptons get a weight ϵ, fake passing leptons $\epsilon\eta$ and prompt passing leptons 1. For the mixed case of N_{fp}, in case both leptons pass, each of them is weighted alternatively one by η the other by 1, hence the factor 2. When both fail, the same is done, weighting by ϵ and 1.

These formulae can also be compared to the ones given in [3], by taking $\eta = 0$ ($p = 1$). The estimate of the QCD contribution in this reference is identical to our N_{ff}^{pass}, namely $N_{QCD} = \epsilon^2 N_{t0}$. But in the estimate of the $W+jets$ background, our N_{fp}^{pass}, the formula of [3] corresponds to $N_{Wj} = \epsilon N_{t1} - N_{QCD}$ and is missing a factor 2 in front of N_{QCD}. The need for this factor 2 is indeed not obvious when the formula is written in this form. It is, however, clearly needed when looking at formula (9). It is due to the fact that one lepton is prompt, the other fake.

4 Fake rates for same flavour trilepton events

To get more insight into the general structure of the formula for the estimation of the prompt lepton contribution, and in particular the signs of the weights, we also derived the formula for

3 SF leptons in the final state. This may also be useful for the estimation of fakes from, e.g., $Z^0 + jets$ in the $W + Z^0$ channel. Another case of interest is the $\tilde{\chi}^0\tilde{\chi}^\pm$ trilepton final state in SUSY. With an obvious extension of the notation of previous Section, we have:

$$
\begin{aligned}
N_l &= N_{ppp} + N_{fpp} + N_{ffp} + N_{fff} = N_{t3} + N_{t2} + N_{t1} + N_{t0} \\
N_{t0} &= (1-p)^3 N_{ppp} + (1-p)^2(1-f)N_{fpp} + (1-p)(1-f)^2 N_{ffp} + (1-f)^3 N_{fff} \\
N_{t1} &= 3p(1-p)^2 N_{ppp} + \left[2p(1-p)(1-f) + f(1-p)^2\right] N_{fpp} \\
&+ \left[2f(1-f)(1-p) + p(1-f)^2\right] N_{ffp} + 3f(1-f)^2 N_{fff} \\
N_{t2} &= 3p^2(1-p)N_{ppp} + \left[2pf(1-p) + p^2(1-f)\right] N_{fpp} \\
&+ \left[2pf(1-f) + (1-p)f^2\right] N_{ffp} + 3f^2(1-f)N_{fff} \\
N_{t3} &= p^3 N_{ppp} + p^2 f N_{fpp} + p f^2 N_{ffp} + f^3 N_{fff}
\end{aligned}
\tag{12}
$$

The inversion of the matrix of these linear equations gives:

$$
A = \begin{pmatrix}
-f^3 & f^2(1-f) & -f(1-f)^2 & (1-f)^3 \\
3f^2p & -f(f+2p-3pf) & (1-f)(p+2f-3pf) & -3(1-f)^2(1-p) \\
-3fp^2 & p(p+2f-3pf) & -(1-p)(f+2p-3pf) & 3(1-f)(1-p)^2 \\
p^3 & -(1-p)p^2 & (1-p)^2 p & -(1-p)^3
\end{pmatrix}
$$

$$
\begin{pmatrix} N_{ppp} \\ N_{fpp} \\ N_{ffp} \\ N_{fff} \end{pmatrix} = \frac{A}{(p-f)^3} \begin{pmatrix} N_{t0} \\ N_{t1} \\ N_{t2} \\ N_{t3} \end{pmatrix}
\tag{13}
$$

From this, the number of events with prompt leptons is derived:

$$
N_{ppp} = \frac{1}{(p-f)^3}\left[(1-f)^3 N_{t3} - f(1-f)^2 N_{t2} + f^2(1-f)N_{t1} - f^3 N_{t0}\right]
\tag{14}
$$

with the number of signal events being given by $N_{signal} = p^3 N_{ppp}$. It is seen that the weights can be computed by the same method as presented for the dileptons. It also shows that the sign is positive for the number of events where all leptons pass the tight criteria and flips between positive and negative for all successive terms.

The individual background contributions are now:

$$
\begin{aligned}
N_{fpp} &= \frac{1}{(p-f)^3}\left[3pf^2 N_{t0} - [f^2(1-p) + 2pf(1-f)]N_{t1}\right. \\
&\left. + [2f(1-p)(1-f) + p(1-f)^2]N_{t2} - 3(1-p)(1-f)^2 N_{t3}\right] \\
N_{ffp} &= \frac{1}{(p-f)^3}\left[-3p^2 f N_{t0} + [2pf(1-p) + p^2(1-f)]N_{t1}\right. \\
&\left. - [f(1-p)^2 + 2p(1-p)(1-f)]N_{t2} + 3(1-p)^2(1-f)N_{t3}\right] \\
N_{fff} &= \frac{1}{(p-f)^3}\left[p^3 N_{t0} - p^2(1-p)N_{t1} + p(1-p)^2 N_{t2} - (1-p)^3 N_{t3}\right]
\end{aligned}
\tag{15}
$$

The weighting method is a straightforward application of the rules presented in Section 3. In particular, the multiplicative constants are obtained by summing over the alternative assignments to the leptons. Also, the rule for the signs can be clarified. The term for which the number of leptons passing the tight cuts is equal to the number of prompt leptons always has a positive sign. For the other terms, it flips successively between positive and negative as we move away from it (left and right).

A typical SM reaction leading to trileptons is the WZ^0 production. The background contributions inside the tight cuts are $p^2 f N_{fpp}$ dominantly from $Z^0 + jets$, with a jet faking a lepton from a W. The others are $pf^2 N_{ffp}$ from $W + jets$ with two jets faking leptons and the last one, $f^3 N_{fff}$, would be QCD with three fake leptons, but these contributions is expected to be small as two fake leptons must produce a Z^0 peak.

In the notation with small numbers, the above equations read

$$
\begin{aligned}
N_{ppp}^{pass} &= \frac{1}{(1-\epsilon\eta)^3}\left[N_{t3} - \epsilon N_{t2} + \epsilon^2 N_{t1} - \epsilon^3 N_{t0}\right] \\
N_{fpp}^{pass} &= \frac{\epsilon}{(1-\epsilon\eta)^3}\left[3\epsilon^2 N_{t0} - \epsilon[\epsilon\eta + 2]N_{t1} + [1 + 2\epsilon\eta]N_{t2} - 3\eta N_{t3}\right] \\
N_{ffp}^{pass} &= \frac{\epsilon^2}{(1-\epsilon\eta)^3}\left[-3\epsilon N_{t0} + [1 + 2\epsilon\eta]N_{t1} - \eta[\epsilon\eta + 2]N_{t2} + 3\eta^2 N_{t3}\right] \\
N_{fff}^{pass} &= \frac{\epsilon^3}{(1-\epsilon\eta)^3}\left[N_{t0} - \eta N_{t1} + \eta^2 N_{t2} - \eta^3 N_{t3}\right]
\end{aligned}
\tag{16}
$$

5 Fake rates for e μ events

The expression for the number of prompt leptons from SF dileptons in Section 3 can be modified to include the case where different leptons, e and μ, are part of the same configuration. The notation is now N_{pp}, N_{pf}, N_{fp} and N_{ff}, where the first label refers to e, the second to μ. The numbers passing and failing the tight cuts are written as N_{txy}, with x giving the numbers of electrons passing the tight cuts and y the number for muons. The basic equations are then:

$$
\begin{aligned}
N_l &= N_{pp} + N_{fp} + N_{ff} = N_{t11} + N_{t10} + N_{t01} + N_{t00} \\
N_{t00} &= (1-p_e)(1-p_\mu)N_{pp} + (1-p_e)(1-f_\mu)N_{pf} + (1-f_e)(1-p_\mu)N_{fp} + (1-f_e)(1-f_\mu)N_{ff} \\
N_{t10} &= p_e(1-p_\mu)N_{pp} + p_e(1-f_\mu)N_{pf} + f_e(1-p_\mu)N_{fp} + f_e(1-f_\mu)N_{ff} \\
N_{t01} &= (1-p_e)p_\mu N_{pp} + (1-p_e)f_\mu N_{pf} + (1-f_e)p_\mu N_{fp} + (1-f_e)f_\mu N_{ff} \\
N_{t11} &= p_e p_\mu N_{pp} + p_e f_\mu N_{pf} + f_e p_\mu N_{fp} + f_e f_\mu N_{ff}
\end{aligned}
\tag{17}
$$

The inverted matrix is:

$$
A = \begin{pmatrix}
f_e f_\mu & -(1-f_e)f_\mu & -f_e(1-f_\mu) & (1-f_e)(1-f_\mu) \\
-f_e p_\mu & (1-f_e)p_\mu & f_e(1-p_\mu) & -(1-f_e)(1-p_\mu) \\
-p_e f_\mu & (1-p_e)f_\mu & p_e(1-f_\mu) & -(1-p_e)(1-f_\mu) \\
p_e p_\mu & -(1-p_e)p_\mu & -p_e(1-p_\mu) & (1-p_e)(1-p_\mu)
\end{pmatrix}
$$

$$
\begin{pmatrix} N_{pp} \\ N_{pf} \\ N_{fp} \\ N_{ff} \end{pmatrix} = \frac{A}{(p_e - f_e)(p_\mu - f_\mu)} \begin{pmatrix} N_{t00} \\ N_{t10} \\ N_{t01} \\ N_{t11} \end{pmatrix}
\tag{18}
$$

The number of events with two prompt leptons is then:

$$N_{pp} = \frac{1}{p_e - f_e}\frac{1}{p_\mu - f_\mu} \times \left[(1-f_e)(1-f_\mu)N_{t11} - (1-f_e)f_\mu N_{t10} - f_e(1-f_\mu)N_{t01} + f_e f_\mu N_{t00}\right] \tag{19}$$

with the number of signal events being given by $N_{signal} = p_e p_\mu N_{pp}$.

The weighting of the leptons is seen to be the straightforward extension of the method discussed in Section 3 where now the electrons and the muons are distinguished. Also, the second and third term reproduce correctly the second term in equation (9), as N_{t1} was the sum of cases where the first or the second lepton passed the tight cuts. This formula would apply to the $(e\mu)$ configurations in the dileptonic $t\bar{t}$ or in the SS dileptons of SUSY.

Also the individual backgrounds can be derived

$$\begin{aligned}
N_{pf} &= \frac{1}{p_e - f_e}\frac{1}{p_\mu - f_\mu}\left[-f_e p_\mu N_{t00} + (1-f_e)p_\mu N_{t10} + f_e(1-p_\mu)N_{t01} - (1-f_e)(1-p_\mu)N_{t11}\right] \\
N_{fp} &= \frac{1}{p_e - f_e}\frac{1}{p_\mu - f_\mu}\left[-p_e f_\mu N_{t00} + (1-p_e)f_\mu N_{t10} + p_e(1-f_\mu)N_{t01} - (1-p_e)(1-f_\mu)N_{t11}\right] \\
N_{ff} &= \frac{1}{p_e - f_e}\frac{1}{p_\mu - f_\mu}\left[p_e p_\mu N_{t00} - (1-p_e)p_\mu N_{t10} - p_e(1-p_\mu)N_{t01} + (1-p_e)(1-p_\mu)N_{t11}\right]
\end{aligned} \tag{20}$$

The corresponding backgrounds remaining in the selected (tight) sample of $t\bar{t}$ events are then respectively $(p_e f_\mu N_{pf} + p_\mu f_e N_{fp})$ from $W + jets$ contamination and $f_e f_\mu N_{ff}$ from QCD. In the SS SUSY dileptons events, the dominant backgrounds remaining in N_{pf} and N_{fp} would be from $t\bar{t}$ with the prompt lepton from W decay and the fake one from the leptonic decay of the b-quark from the other top. The N_{ff} contribution would again be a measure of the QCD background.

Expressed as a function of the small numbers for passing leptons, the formulae become:

$$\begin{aligned}
N_{pp}^{pass} &= \frac{1}{(1-\epsilon_e\eta_e))(1-\epsilon_\mu\eta_\mu)}\left[N_{t11} - \epsilon_\mu N_{t10} - \epsilon_e N_{t01} + \epsilon_e\epsilon_\mu N_{t00}\right] \\
N_{pf}^{pass} &= \frac{\epsilon_\mu}{(1-\epsilon_e\eta_e))(1-\epsilon_\mu\eta_\mu)}\left[-\epsilon_e N_{t00} + N_{t10} + \epsilon_e\eta_\mu N_{t01} - \eta_\mu N_{t11}\right] \\
N_{fp}^{pass} &= \frac{\epsilon_e}{(1-\epsilon_e\eta_e))(1-\epsilon_\mu\eta_\mu)}\left[-\epsilon_\mu N_{t00} + \eta_e\epsilon_\mu N_{t10} + N_{t01} - \eta_e N_{t11}\right] \\
N_{ff}^{pass} &= \frac{\epsilon_\mu\epsilon_e}{(1-\epsilon_e\eta_e))(1-\epsilon_\mu\eta_\mu)}\left[N_{t00} - \eta_e N_{t10} - \eta_\mu N_{t01} + \eta_e\eta_\mu N_{t11}\right]
\end{aligned} \tag{21}$$

6 Recursive formulae for dilepton events

The formulae from Section 3 can be put in a different form which exhibits the recursivity between the N_{ff}, N_{fp} and N_{pp}. To understand the approach, let us first take the case with $p = 1$. Then, from the last formula in (9)

$$N_{ff} = \frac{1}{(1-f)^2}N_{t0}$$

in other words the number of events where no leptons pass the tight cuts (N_{t0}) is given by the number of events with both fake leptons at the level of the loose cuts multiplied by the probability for both leptons to fail the tight cuts.

The number of events with one prompt and one fake lepton, the fake lepton failing the tight cuts, is to first approximation (small f) determined by N_{t1}. From this value, we have to subtract the contribution from N_{ff} with one of the fake leptons passing the tight cuts (2 combinations)

$$N_{fp} = \frac{1}{(1-f)}\left[N_{t1} - 2fN_{ff}\right]$$

The number of events with two prompt leptons is in first approximation N_{t2}. From this, we need to subtract N_{fp} with the fake lepton passing and N_{ff} with both fake leptons passing the tight cuts

$$N_{pp} = \left[N_{t2} - fN_{fp} - f^2N_{ff}\right]$$

It is easily verified that this reproduces correctly formulae (8) and (9) with $p = 1$.

In the more general case with $p \neq 0$ it is seen from the last equation in (6) that

$$N_{pp} = \frac{1}{p^2}\left[N_{t2} - pfN_{fp} - f^2N_{ff}\right] \tag{22}$$

As next step, from the last but one equation in (6), we get for the N_{fp} events where one lepton passes and one fails the tight cuts

$$\left[f(1-p) + p(1-f)\right]N_{fp} = N_{t1} - 2f(1-f)N_{ff} - 2p(1-p)N_{pp}$$

in which N_{pp} can be replaced to make the formula recursive:

$$\left[f(1-p) + p(1-f)\right]N_{fp} = N_{t1} - 2\frac{f}{p}[p(1-f) - f(1-p)]N_{ff} + 2f(1-p)N_{fp} - 2\frac{1-p}{p}N_{t2}$$

and thus

$$N_{fp} = \frac{1}{p(p-f)}\left[pN_{t1} - 2(p-f)fN_{ff} - 2(1-p)N_{t2}\right] \tag{23}$$

Starting from the equation for N_{ff} given in (9), the N_{fp} and N_{pp} can be computed recursively.

An interest of these recursive formulae is that N_{ff} and N_{fp} correspond to specific backgrounds. Given the different kinematical properties of the different backgrounds, the fake ratios may depend on the type of background. Calling the ratios f_{ff} and f_{fp}, it is easily verified that the recursive formulae take the form

$$\begin{aligned}
N_{pp} &= \frac{1}{p^2}\left[N_{t2} - pf_{fp}N_{fp} - f_{ff}^2N_{ff}\right] \\
N_{fp} &= \frac{1}{p(p-f_{fp})}\left[pN_{t1} - 2(p-f_{ff})f_{ff}N_{ff} - 2(1-p)N_{t2}\right] \\
N_{ff} &= \frac{1}{(p-f_{ff})^2}\left[p^2N_{t0} - p(1-p)N_{t1} + (1-p)^2N_{t2}\right]
\end{aligned} \tag{24}$$

Hence, they allow different fake ratios to be used for different backgrounds. Rather than trying to determine a universal fake ratio as a function of the event properties, which may be difficult in practice, they open the possibility to determine individual fake ratios directly from the contributing backgrounds and apply them to compute the number of signal events and the backgrounds in the selected sample.

The recursive formulae can also be derived for the $e-\mu$ events. Starting from the last equation in (17), N_{pp} can be expressed as a function of N_{pf}, N_{fp} and N_{ff}. Then, using the two previous equations and replacing the expression for N_{pp}, yields the recursive relations for N_{pf} and N_{fp}. With an obvious notation for the fake ratios, the equations are

$$\begin{aligned}
N_{pp} &= \frac{1}{p_e p_\mu}\left[N_{t11} - p_e f_{\mu,pf} N_{pf} - f_{e,fp} p_\mu N_{fp} - f_{e,ff} f_{\mu,ff} N_{ff}\right] \\
N_{pf} &= \frac{1}{p_e(p_\mu - f_{\mu,pf})}\left[p_\mu N_{t10} - f_{e,ff}(p_\mu - f_{\mu,ff})N_{ff} - (1-p_\mu)N_{t11}\right] \\
N_{fp} &= \frac{1}{p_\mu(p_e - f_{e,fp})}\left[p_e N_{t01} - f_{\mu,ff}(p_e - f_{e,ff})N_{ff} - (1-p_e)N_{t11}\right] \\
N_{ff} &= \frac{1}{p_e - f_{e,ff}}\frac{1}{p_\mu - f_{\mu,ff}}\left[p_e p_\mu N_{t00} - (1-p_e)p_\mu N_{t10} - p_e(1-p_\mu)N_{t01} + (1-p_e)(1-p_\mu)N_{t11}\right]
\end{aligned} \tag{25}$$

A generalization of these equations to an arbitrary number of SF leptons will be presented in the Appendix.

7 Weighting rules

From the preceding sections, it is now possible to infer the general rules for the weighting in the various cases.

First, the overall factor will be of the form $1/(p-f)^n$. This will multiply a sum which runs over terms made of the product of $n = n_p + n_f$ weights, one weight per lepton. The rules for the weighting can be derived from the formula in Section 2 as follows:

- for a prompt lepton estimate, f for leptons failing the tight cuts
- for a prompt lepton estimate, $(1-f)$ for leptons passing the tight cuts
- for a fake lepton estimate, p for leptons failing the tight cuts
- for a fake lepton estimate, $(1-p)$ for leptons passing the tight cuts

We can count the number of leptons in each of the four categories above. We choose a fixed number, i, of failing leptons, hence a number $n-i$ of passing leptons (i.e. for $N_{t,n-i}$). The total number of fake and prompt leptons is n_f and n_p, respectively. What remains undefined is the repartition of the fakes and prompt leptons among the passing and failing ones. Let us take the number of fake leptons failing as j. Then we have

- number of fake failing leptons = j
- number of fake passing leptons = $n_f - j$
- number of prompt failing leptons = $i-j$
- number of prompt passing leptons = $n-i-n_f+j = n_p - i + j$

As the numbers of leptons in each category must be zero or positive, the values of j are restricted. This implies that $j \le n_f$ and $i-j \le n_p$, giving together $i - n_p \le j \le n_f$. In addition, $j \ge 0$ and $j \le i$. Together, we have $max(i-n_p, 0) \le j \le min(n_f, i)$. To each of the allowed values of j corresponds a valid configuration of prompt and fake leptons contributing to a fixed topology $N_{t,n-i}$.

The contribution to the weights then consists of one or several terms of the form

$$p^j f^{i-j}(1-p)^{n_f-j}(1-f)^{n_p-i+j} \tag{26}$$

The sum of the above terms will run over $max(i-n_p, 0) \leq j \leq min(n_f, i)$.

Next, we compute the constant counting the number of possible assignments of (26) to the leptons. For a given term $N_{t,n-i}$ the numbers of passing and failing leptons are fixed and the individual leptons are identified as passing or failing. What is not known is how they are distributed among fake and prompt leptons. Therefore, permutations can only occur among the passing or among the failing leptons. We can take, for example, the i failing leptons. If all of them would be distinguishable, we would have $i!$ permutations. But as the j fake leptons and the $(i-j)$ prompt leptons are undistinguishable, this number should be divided by $j!(i-j)!$. For the $(n-i)$ passing leptons, the total permutations are $(n-i)!$ with (n_f-j) fake and (n_p-i+j) prompt indistinguishable leptons. The number of possible assignments of (26) is thus

$$C_{ij} = \frac{i!}{j!(i-j)!} \cdot \frac{(n-i)!}{(n_f-j)!(n_p-i+j)!}$$

Finally, we have to determine the sign of the contribution of $N_{t,n-i}$. We have seen in Section 4 that the term for which the number of passing leptons is equal to the number of prompt leptons always has a positive sign. The other terms have successively a flip of sign as i moves away from it. Hence, the sign is $(-1)^{n-i-n_p} = (-1)^{n_f-i}$.

It is easily verified that these rules give indeed the constant factors and the signs in the formulae of the preceding sections. The generalization of the formulae to an arbitrary number of leptons are derived in Appendix B.

The rules can also be defined for the case where small numbers are used. The overall factor in front of the square brackets is $1/(1-\epsilon\eta)^n$, the other weights are:

- for a prompt or fake lepton estimate, ϵ for leptons failing the tight cuts
- for a prompt lepton estimate, 1 for leptons passing the tight cuts
- for a fake lepton estimate, $\epsilon\eta$ for leptons passing the tight cuts

In this case, the terms will be of the form

$$(1-\epsilon\eta)^{-n}\ \epsilon^i\ (\epsilon\eta)^{n_f-j}$$

The counting of possible assignments and the rule for the signs is the same as above.

8 Uncertainties for dileptons

The quantities of interest are the uncertainties on the numbers of events passing the tight cuts. We start from the error propagation for the equations written in terms of the small numbers. They can be rewritten as

$$N^{pass}_{n_f,n_p} = \sum_{i=0}^{n} (-1)^{n_f-i} N_{t,n-i} w_i \tag{27}$$

where

$$w_i = \sum_{j=jmin}^{jmax} w_{ij} = \sum_{j=jmin}^{jmax} C_{ij}(1-\epsilon\eta)^{-n}\ \epsilon^i\ (\epsilon\eta)^{n_f-j} \tag{28}$$

with C_{ij} being the ratio of factorials. As seen before, the dominant term is the one with $n - i = n_p$, or $i = n_f$. Then $max(n_f - n_p, 0) \leq j \leq n_f$. As $\epsilon\eta$ is a very small number, we can neglect these contributions, i.e. take $j = n_f$. It is easily verified that in this case $C_{n_f n_F} = 1$ and we have

$$N^{pass}_{n_f,n_p} \simeq N_{t,n_p} w_{n_f} \quad , \quad w_{n_f} \simeq (1 - \epsilon\eta)^{-n} \, \epsilon^{n_f} \tag{29}$$

As the $N_{t,n-i}$ are statistically uncorrelated and assumed to be Poisson distributed, the statistical uncertainty on $N^{pass}_{n_f,n_p}$ is

$$\Delta N^{pass}_{n_f,n_p}\Big|_{stat} = \sqrt{\sum_{i=0}^{n} w_i^2 N_{t,n-i}} \tag{30}$$

Keeping only the leading term, it becomes

$$\Delta N^{pass}_{n_f,n_p}\Big|_{stat} \simeq w_{n_f} \sqrt{N_{t,n_p}} = (1 - \epsilon\eta)^{-n} \, \epsilon^{n_f} \sqrt{N_{t,n_p}} \tag{31}$$

There are also systematic uncertainties originating from the uncertainties on p and f. Assuming that the uncertainties on p and f are uncorrelated (they are normally measured in different reactions), the uncertainty on the number of events passing is obtained as

$$\Delta N^{pass}_{n_f,n_p}\Big|_{syst} = \sqrt{\left(\frac{\partial N^{pass}_{n_f,n_p}}{\partial p}\Delta p\right)^2 + \left(\frac{\partial N^{pass}_{n_f,n_p}}{\partial f}\Delta f\right)^2} \tag{32}$$

The uncertainties on f and p are propagated to uncertainties on ϵ and η by

$$\begin{aligned} \frac{\partial\epsilon}{\partial p} &= 0 \ , \quad \frac{\partial\epsilon}{\partial f} = \left(\frac{1}{f} + \frac{1}{1-f}\right)\epsilon \\ \frac{\partial\eta}{\partial f} &= 0 \ , \quad \frac{\partial\eta}{\partial p} = -\left(\frac{1}{p} + \frac{1}{1-p}\right)\eta \end{aligned} \tag{33}$$

We can now compute the uncertainties on the numbers of events passing for the SF dilepton channels given in (10) and (11), the general formulae being derived in Appendix C. To simplify the notation, we define D_{pp}, D_{fp} and D_{ff} to be the expressions in square brackets in these formulae. The contributions from $\Delta\epsilon$ and $\Delta\eta$ are then given by:

$$\begin{aligned} \frac{\Delta_f N^{pass}_{pp}}{N^{pass}_{pp}} &= \left|\tfrac{2\eta}{1-\epsilon\eta} - \tfrac{N_{t1}-2\epsilon N_{t0}}{D_{pp}}\right| \Delta\epsilon \ , & \frac{\Delta_p N^{pass}_{pp}}{N^{pass}_{pp}} &= \left(\frac{2\epsilon}{1-\epsilon\eta}\right)\Delta\eta \\ \frac{\Delta_f N^{pass}_{fp}}{N^{pass}_{fp}} &= \left|\tfrac{1}{\epsilon} + \tfrac{2\eta}{1-\epsilon\eta} - \tfrac{2N_{t0}-\eta N_{t1}}{D_{fp}}\right| \Delta\epsilon \ , & \frac{\Delta_p N^{pass}_{fp}}{N^{pass}_{fp}} &= \left|\frac{2\epsilon}{1-\epsilon\eta} - \frac{2N_{t2}-\epsilon N_{t1}}{D_{fp}}\right|\Delta\eta \\ \frac{\Delta_f N^{pass}_{ff}}{N^{pass}_{ff}} &= \left(\tfrac{2}{\epsilon} + \tfrac{2\eta}{1-\epsilon\eta}\right)\Delta\epsilon \ , & \frac{\Delta_p N^{pass}_{ff}}{N^{pass}_{ff}} &= \left|\frac{2\epsilon}{1-\epsilon\eta} - \frac{N_{t1}-2\eta N_{t2}}{D_{ff}}\right|\Delta\eta \end{aligned} \tag{34}$$

The systematic uncertainty (32) can be rewritten in the simpler form

$$\Delta N^{pass}_{pp}\big|_{syst} = \sqrt{\left(B_{pp,\epsilon}\Delta\epsilon\right)^2 + \left(B_{pp,\eta}\Delta\eta\right)^2} \, N^{pass}_{pp} \tag{35}$$

where $B_{pp,\epsilon}$ and $B_{pp,\eta}$ stand for the factors multiplying $\Delta\epsilon$ and $\Delta\eta$, respectively in (34). Similar expressions are obtained for the fp and ff cases. Note that all the systematic uncertainties on the numbers of events increase linearly with the number of events, unlike the statistical uncertainties.

It should be noted that the estimate of the uncertainties on f and p should include the statistical errors of their measurement (which could be made small if sufficient luminosity is accumulated), but also other effects more difficult to account for. Examples are the approximate p_T and η dependence, or a dependence on the number of jets, which may differ considerably in the reaction where they are measured from the channel where they are applied. Such effects may potentially induce large uncertainties on f and/or p which could invalidate the linear propagation used here. But is is assumed that at some stage these effects have been thoroughly studied and are under control.

The different $N^{pass}_{n_f,n_p}$ for different values of n_f and n_p will be correlated. Let us choose another combination $N^{pass}_{m_f,m_p}$ such that $m_f + m_p = n_f + n_p = n$.

The statistical uncertainties are correlated, because they both depend on the same numbers $N_{t,n-i}$. Assuming that the various $N_{t,n-i}$ are uncorrelated and that they are Poisson distributed, the variance is

$$V_{stat}(N^{pass}_{n_f,n_p}, N^{pass}_{m_f,m_p}) = \sum_{i=0}^{n} (-1)^{n_f+m_f} w_i w'_i N_{t,n-i} \tag{36}$$

where w'_i is the weight in $N^{pass}_{m_f,m_p}$ for the same i.

Keeping only the leading term, the correlation coefficient defined by

$$C(N^{pass}_{n_f,n_p}, N^{pass}_{m_f,m_p}) = \frac{V(N^{pass}_{n_f,n_p}, N^{pass}_{m_f,m_p})}{\Delta N^{pass}_{n_f,n_p} \Delta N^{pass}_{m_f,m_p}} \tag{37}$$

becomes, using (31).

$$C_{stat}(N^{pass}_{n_f,n_p}, N^{pass}_{m_f,m_p}) \simeq \frac{w_{n_f} w'_{n_f} N_{t,n_p} + w_{m_f} w'_{m_f} N_{t,m_p}}{w_{n_f}\sqrt{N_{t,n_p}}\, w'_{m_f}\sqrt{N_{t,m_p}}} = \frac{w'_{n_f}}{w'_{m_f}}\sqrt{\frac{N_{t,n_p}}{N_{t,m_p}}} + \frac{w_{m_f}}{w_{n_f}}\sqrt{\frac{N_{t,m_p}}{N_{t,n_p}}} \tag{38}$$

As the coefficients w_{n_f} and w'_{m_f} are leading while w_{m_f} and w'_{n_f} are subleading, the correlation coefficient of the statistical uncertainties is not expected to be large.

For the systematic uncertainties, correlations arise because p and f appear in both expressions. The variance of the number of events between e.g. pp and fp is

$$V_{syst}(N^{pass}_{pp}, N^{pass}_{fp}) = \left(B_{pp,\epsilon} B_{fp,\epsilon} \Delta\epsilon^2 + B_{pp,\eta} B_{fp,\eta} \Delta\eta^2\right) N^{pass}_{pp} N^{pass}_{fp} \tag{39}$$

and the correlation coefficient is

$$C_{syst}(N^{pass}_{n_f,n_p}, N^{pass}_{m_f,m_p}) = \frac{B_{pp,\epsilon} B_{fp,\epsilon} \Delta\epsilon^2 + B_{pp,\eta} B_{fp,\eta} \Delta\eta^2}{\sqrt{\left(B_{pp,\epsilon}\Delta\epsilon\right)^2 + \left(B_{pp,\eta}\Delta\eta\right)^2}\sqrt{\left(B_{fp,\epsilon}\Delta\epsilon\right)^2 + \left(B_{fp,\eta}\Delta\eta\right)^2}} \tag{40}$$

independent of the number of events.

To complete the discussion of the uncertainties for the dileptons, we should also consider the $e-\mu$ events. In this case, f and p as well as their uncertainties may be different for electrons and muons. A difficult question is whether there exists a correlation between the uncertainties of e and μ. As they are measured in different event samples, they are statistically uncorrelated.

But, if there is a dependence on the number of jets, for example, this could induce some correlation in the systematic uncertainties. In the following, we will nevertheless assume that no correlations exist.

The expressions for the numbers of events passing were given in (21). The uncertainties from $\Delta\epsilon$ are:

$$
\begin{aligned}
\frac{\Delta_{fe} N_{pp}^{pass}}{N_{pp}^{pass}} &= \left| \tfrac{\eta_e}{1-\epsilon_e\eta_e} - \tfrac{N_{t01}-\epsilon_\mu N_{t00}}{D_{pp}} \right| \Delta\epsilon_e \ , &
\frac{\Delta_{f\mu} N_{pp}^{pass}}{N_{pp}^{pass}} &= \left| \frac{\eta_\mu}{1-\epsilon_\mu\eta_\mu} - \frac{N_{t10}-\epsilon_e N_{t00}}{D_{pp}} \right| \Delta\epsilon_\mu \\
\frac{\Delta_{fe} N_{pf}^{pass}}{N_{pf}^{pass}} &= \left| \tfrac{\eta_e}{1-\epsilon_e\eta_e} - \tfrac{N_{t00}-\eta_\mu N_{t01}}{D_{pf}} \right| \Delta\epsilon_e \ , &
\frac{\Delta_{f\mu} N_{pf}^{pass}}{N_{pf}^{pass}} &= \left(\frac{1}{\epsilon_\mu} + \frac{\eta_\mu}{1-\epsilon_\mu\eta_\mu} \right) \Delta\epsilon_\mu \\
\frac{\Delta_{fe} N_{fp}^{pass}}{N_{fp}^{pass}} &= \left(\tfrac{1}{\epsilon_e} + \tfrac{\eta_e}{1-\epsilon_e\eta_e} \right) \Delta\epsilon_e \ , &
\frac{\Delta_{f\mu} N_{fp}^{pass}}{N_{fp}^{pass}} &= \left| \frac{\eta_\mu}{1-\epsilon_\mu\eta_\mu} - \frac{N_{t00}-\eta_e N_{t10}}{D_{fp}} \right| \Delta\epsilon_\mu \\
\frac{\Delta_{fe} N_{ff}^{pass}}{N_{ff}^{pass}} &= \left(\tfrac{1}{\epsilon_e} + \tfrac{\eta_e}{1-\epsilon_e\eta_e} \right) \Delta\epsilon_e \ , &
\frac{\Delta_{f\mu} N_{ff}^{pass}}{N_{ff}^{pass}} &= \left(\frac{1}{\epsilon_\mu} + \frac{\eta_\mu}{1-\epsilon_\mu\eta_\mu} \right) \Delta\epsilon_\mu
\end{aligned}
\tag{41}
$$

and from $\Delta\eta$:

$$
\begin{aligned}
\frac{\Delta_{pe} N_{pp}^{pass}}{N_{pp}^{pass}} &= \left(\tfrac{\epsilon_e}{1-\epsilon_e\eta_e} \right) \Delta\eta_e \ , &
\frac{\Delta_{p\mu} N_{pp}^{pass}}{N_{pp}^{pass}} &= \left(\frac{\epsilon_\mu}{1-\epsilon_\mu\eta_\mu} \right) \Delta\eta_\mu \\
\frac{\Delta_{pe} N_{pf}^{pass}}{N_{pf}^{pass}} &= \left(\tfrac{\epsilon_e}{1-\epsilon_e\eta_e} \right) \Delta\eta_e \ , &
\frac{\Delta_{p\mu} N_{pf}^{pass}}{N_{pf}^{pass}} &= \left| \frac{\epsilon_\mu}{1-\epsilon_\mu\eta_\mu} - \frac{N_{t11}-\epsilon_e N_{t01}}{D_{pf}} \right| \Delta\eta_\mu \\
\frac{\Delta_{pe} N_{fp}^{pass}}{N_{fp}^{pass}} &= \left| \tfrac{\epsilon_e}{1-\epsilon_e\eta_e} - \tfrac{N_{t11}-\epsilon_\mu N_{t10}}{D_{fp}} \right| \Delta\eta_e \ , &
\frac{\Delta_{p\mu} N_{fp}^{pass}}{N_{fp}^{pass}} &= \left(\frac{\epsilon_\mu}{1-\epsilon_\mu\eta_\mu} \right) \Delta\eta_\mu \\
\frac{\Delta_{pe} N_{ff}^{pass}}{N_{ff}^{pass}} &= \left| \tfrac{\epsilon_e}{1-\epsilon_e\eta_e} - \tfrac{N_{t10}-\eta_\mu N_{t11}}{D_{ff}} \right| \Delta\eta_e \ , &
\frac{\Delta_{p\mu} N_{ff}^{pass}}{N_{ff}^{pass}} &= \left| \frac{\epsilon_\mu}{1-\epsilon_\mu\eta_\mu} - \frac{N_{t01}-\eta_e N_{t11}}{D_{ff}} \right| \Delta\eta_\mu
\end{aligned}
\tag{42}
$$

The systematic uncertainty on the numbers of events are, for example

$$
\Delta N_{pp}^{pass}\big|_{syst} = \sqrt{\left(B_{pp,\epsilon e}\Delta\epsilon_e\right)^2 + \left(B_{pp,\epsilon\mu}\Delta\epsilon_\mu\right)^2 + \left(B_{pp,\eta e}\Delta\eta_e\right)^2 + \left(B_{pp,\eta\mu}\Delta\eta_\mu\right)^2}\; N_{pp}^{pass} \tag{43}
$$

9 Example of $t\bar{t}$

To get a feeling for the orders of magnitude, let us consider the case of dileptons, where we are interested in the number of passing prompt leptons. From the simulation results of Oviedo/Santander [5] and from UCSB/UCSD/FNAL [3] it is found that the selected dileptons consist of events approximately in the proportion of 400 from $t\bar{t}$, about 8 from $W + jets$ and a small number of QCD that we take as 1. From the last equation of (6), we can extract $N_{pp} = 494$, $N_{fp} = 89$ and $N_{ff} = 100$. Then the observed numbers of events are $N_{t2} = 409.15$, $N_{t1} = 179.90$ and $N_{t0} = 93.95$. The prompt ratio used is $p = 0.9 \pm 0.1$ and the fake ratio $f = 0.1 \pm 0.05$, assumed independent of p_T and η. The 50% uncertainty on f is taken from [2]. For these values, $\epsilon = 0.111$ and $\eta = 0.111$, so that $\Delta\epsilon = 0.0617$ and $\Delta\eta = 0.123$.

We will first consider only the leading terms. The numbers of events and their statistical uncertainties are then

$$\begin{aligned} N_{pp}^{pass} &= 419.4 \pm 20.7_{stat} \\ N_{fp}^{pass} &= 20.7 \pm 1.55_{stat} \\ N_{ff}^{pass} &= 1.19 \pm 0.12_{stat} \end{aligned}$$

They are compatible with the true numbers above, except for N_{fp} which is incompatible with 8. The latter is due to the leading term approximation, as will be seen below.

The correlation coefficient of the statistical uncertainties, given by (38) is, for example

$$C_{stat}(N_{pp}^{pass}, N_{ff}^{pass}) = 0.026 + 0.006 = 0.032$$

hence, it is indeed small, as mentioned above.

For the systematic uncertainties, the result from the leading terms and from the exact calculation including all terms are summarized in the following table:

	t-tbar		
	leading term	full calculation	$p = 1$
N_{pp}^{pass}	$419.4 \pm 20.7_{stat} \pm 13.0_{syst}$	$400.1 \pm 20.8_{stat} \pm 12.0_{syst}$	$390.3 \pm 20.3_{stat} \pm 9.8_{syst}$
N_{fp}^{pass}	$20.7 \pm 1.55_{stat} \pm 12.0_{syst}$	$8.01 \pm 1.65_{stat} \pm 11.5_{syst}$	$17.7 \pm 1.51_{stat} \pm 8.53_{syst}$
N_{ff}^{pass}	$1.19 \pm 0.12_{stat} \pm 1.34_{syst}$	$1.00 \pm 0.12_{stat} \pm 1.13_{syst}$	$1.16 \pm 0.12_{stat} \pm 1.29_{syst}$

and the exact calculation reproduces exactly the original numbers. They show that the number of events for the N_{fp}^{pass} is not reliably estimated from the leading term, but all other event numbers and all the uncertainties are good approximations. It also shows that for these statistics, the $t\bar{t}$ result would be dominated by statistics whereas the $W + jets$ and QCD are dominated by systematics. The number of signal events would be known at the 5% level and the $W + jets$ and QCD background at the 100% level. The latter would, nevertheless, provide an interesting bound on the QCD contribution.

The last column in the table starts from the same numbers of observed events $N_{t,n-i}$, but does the calculation with $p = 1 \pm 0$, i.e. the usual assumption for the computation of fakes. It is seen that the shift of the values from the true value is at about the level of the systematic uncertainty, thanks to the 50% uncertainty on f. If the latter would be improved, the result might become incompatible with the true value, especially for N_{fp}^{pass}.

The variance between the numbers of events, keeping only the leading terms, is given by

$$V_{syst}(N_{pp}^{pass}, N_{ff}^{pass}) = 8.18$$

and the corresponding correlation coefficient becomes

$$C_{syst}(N_{pp}^{pass}, N_{ff}^{pass}) = \frac{8.18}{13.0 \times 1.34} = 0.47$$

Note that the sign is positive, as both leading terms have a positive sign.

For the full calculation, the correlation coefficients are:

	statistical	systematic
$C(N_{pp}^{pass}, N_{fp}^{pass})$	-0.380	-0.996
$C(N_{pp}^{pass}, N_{ff}^{pass})$	0.042	-0.465
$C(N_{fp}^{pass}, N_{ff}^{pass})$	-0.298	0.386

It shows that the value of the correlation coefficients between systematics computed from the leading terms is totally unreliable, even the sign may be wrong. The correlation between N_{pp}^{pass} and N_{fp}^{pass} is almost 100%, the others are smaller. This strong (anti-)correlation is due to the fact that there is little other background and that, therefore, their sum is constrained to be nearly equal to N_{t2}. More quantitatively, looking back at the equations (34), it appears that if the N_{ti} are of the same order of magnitude (as is the case for $t\bar{t}$), the dependence on the N_{ti} drops out from the $B_{pp,\epsilon}$ and $B_{fp,\epsilon}$ and the corresponding ratios are of order -1 and -2, respectively. The uncertainties from $\Delta\epsilon$ are dominated by the $1/\epsilon$ (positive) term in $B_{fp,\epsilon}$, the ones from $B_{pp,\epsilon}$ are approximately $\simeq -1$ and the uncertainties from $\Delta\eta$ are suppressed compared to the $\Delta\epsilon$ ones. The correlation coefficient is therefore $\simeq -1$. The table also shows that the correlation coefficient from systematics is an order of magnitude larger than for the statistical uncertainties.

10 Example of SUSY

As a second example with somewhat different proportions of signal and backgrounds, we can use the SUSY search in the same sign dilepton channel. The signal is from the test points LM0 and LM1. The main backgrounds consist of $t\bar{t}$ and $W + jets$, with 1 prompt and 1 fake lepton, and QCD with both fake leptons. Numbers are based on simulation results from [6], each for 100 pb^{-1} at 10 TeV. By summing their numbers, we get approximately for LM0 (LM1) 123 (31) events, for $t\bar{t}\ +\ (W + jets)$ 12 events and for QCD 0.001 event. From the last equation of (6), we get $N_{pp} = 151.85$ (43), $N_{fp} = 133.33$ and $N_{ff} = 0.1$. The observed topologies then contain $N_{t2} = 135(43)$, $N_{t1} = 136.68(116.24)$ and $N_{t0} = 13.43(12.46)$ events for LM0 (LM1). The same values and uncertainties for f and p are used as in the previous example.

The leading term approximation and the complete calculation then give the results:

	LM0		
	leading term	full calculation	p = 1
N_{pp}^{pass}	$138.4 \pm 11.9_{stat} \pm 4.30_{syst}$	$123.0 \pm 12.0_{stat} \pm 7.57_{syst}$	$120.0 \pm 11.7_{stat} \pm 8.25_{syst}$
N_{fp}^{pass}	$15.8 \pm 1.35_{stat} \pm 9.11_{syst}$	$12.0 \pm 1.38_{stat} \pm 7.49_{syst}$	$14.9 \pm 1.30_{stat} \pm 8.07_{syst}$
N_{ff}^{pass}	$0.17 \pm 0.046_{stat} \pm 0.191_{syst}$	$0.001 \pm 0.049_{stat} \pm 0.167_{syst}$	$0.166 \pm 0.045_{stat} \pm 0.184_{syst}$

	LM1		
	leading term	full calculation	p = 1
N_{pp}^{pass}	$44.08 \pm 6.72_{stat} \pm 1.37_{syst}$	$31.0 \pm 6.83_{stat} \pm 6.80_{syst}$	$30.2 \pm 6.67_{stat} \pm 7.00_{syst}$
N_{fp}^{pass}	$13.40 \pm 1.24_{stat} \pm 7.74_{syst}$	$12.0 \pm 1.26_{stat} \pm 6.78_{syst}$	$12.6 \pm 1.20_{stat} \pm 6.83_{syst}$
N_{ff}^{pass}	$0.16 \pm 0.045_{stat} \pm 0.177_{syst}$	$0.001 \pm 0.047_{stat} \pm 0.167_{syst}$	$0.154 \pm 0.044_{stat} \pm 0.171_{syst}$

Many features are the same as for the $t\bar{t}$ example, except that the number of events for the N_{fp}^{pass} from the leading term is now closer to the correct value. On the other hand, the systematic uncertainty on N_{pp}^{pass} estimated from the leading term is now too low. This illustrates that the precision with which the leading term approximation estimates the values and uncertainties is not easily predictable. It is, therefore, advisable to perform the full calculation.

The last column in the table starts again from the same numbers of observed events $N_{t,n-i}$, but does the calculation with $p = 1 \pm 0$, the usual assumption for the computation of fakes. It is seen that the departure from the true value is less significant in this case than for the $t\bar{t}$. A reduction of the uncertainty on f might nevertheless give rise to an incompatible value for N_{ff}^{pass}.

The correlation coefficients from the full calculation are:

	LM0		LM1	
	statistical	systematic	statistical	systematic
$C(N_{pp}^{pass}, N_{fp}^{pass})$	-0.320	-1.000	-0.308	-1.000
$C(N_{pp}^{pass}, N_{ff}^{pass})$	0.077	-0.445	0.085	-0.133
$C(N_{fp}^{pass}, N_{ff}^{pass})$	-0.396	0.427	-0.388	0.109

The systematic uncertainties on N_{pp}^{pass} and N_{fp}^{pass} are again fully correlated, the other correlation coefficients being smaller, especially in the case of LM1.

The background from N_{ff}^{pass} (i.e. QCD) is rather severely constrained. At LM0, the statistical uncertainty for N_{pp}^{pass} (the signal) is increased by the systematics to 14.2, which still leaves the significance (the signal dropping to zero) at the 8.7 s.d. level for about 300 pb^{-1}. For the signal at LM1 the systematic uncertainty, which is of the same order as the statistical one, increases the total uncertainty to 9.6, which leads to a significance of 3.2 s.d. only. A four-fold increase in statistics would give $124 \pm 13.67_{stat} \pm 27.22_{syst}$, leading to a significance of 4.1. This slow increase in significance is due to the linear increase of the systematic uncertainty with the number of events, as can be seen from equation (35). Hence the ratio $N_{pp}/\Delta N_{pp}$ to the systematic uncertainty remains constant with luminosity. The significance levels off, when the statistical uncertainty becomes negligible w.r.t. the systematics, to a value of 4.6. It shows that, with the chosen values of the uncertainties on f and p and of the ratio $S/B = 2.6$ at LM1, the 5 s.d. level of significance would not be reachable. If we would improve the S/B ratio to e.g. ~ 4, we would get a signal of $34.29 \pm 6.81_{stat} \pm 4.99_{syst}$, giving 6.9 as best possible significance. Hence, a discovery would still be possible in principle, but would require a different optimization than the one assumed in this example. But it should be emphasized that the real uncertainties on f and p are still unknown, as well as their evolution with increasing luminosity. In any case, this example demonstrates that a very tight control over the fake and prompt ratios will be needed if we want to avoid spoiling the discovery potential. It also shows that the optimization of the selection criteria has to take into account the effect of uncertainties on fake and prompt ratios.

11 Summary and conclusions

Formulae were derived for the application of the fakeable object method to various physics processes. They include both the "fake ratio", the probability for a fake lepton passing the loose criteria to also pass the tight criteria, and the "prompt ratio", the probability for a prompt lepton to pass the tight criteria. The latter has so far been neglected. The first sections consider simple lepton multiplicities which allow us to establish the rules from which to infer the formulae, rules which are summarized in Section 7. Appendix B presents the most general formulae, applicable to any number of leptons and with both electrons and muons in the final state.

An important point is the presence of different signs for the contributions with different numbers of leptons passing the tight criteria. The term for which the number of leptons passing the tight cuts is equal to the number of prompt leptons always has a positive sign. For the other terms, it flips successively between positive and negative as we move away from it (left and right).

Another important point is that we derived not only the formulae for the estimate of the number of events with prompt leptons within the tight criteria, but also the numbers of events including mixtures of prompt and fake leptons. These can be associated to specific types of backgrounds and allow constraints on such contributions to be put.

Finally, the uncertainties associated to the limited knowledge of the fake and prompt ratios are computed and they are applied to the examples of $t\bar{t}$ and same sign SUSY dileptons. The latter illustrates the fact that in the optimization of the selection criteria the systematic uncertainties from the fake ratios have to be taken into account. Else, they may well spoil the possibility to make a discovery.

The fakeable object method assumes that a universal fake ratio can be determined, usually extracted from QCD events. As the lepton kinematics may be very different between the various backgrounds and the signal, the fake ratio is determined as a function of the lepton kinematics, hoping that this fully encompasses the differences between the different processes. We also established recursive relations which allow the fake ratio for each specific background, directly derived from this background, to be used to estimate the signal and background contributions. This avoids the introduction of potential inaccuracies inherent in the universal fake ratio approach.

A Fake rates for same flavour quadrilepton events

As a final cross check of the structure of the formula for the estimation of the prompt lepton contribution, we derived the formula for the 4 SF lepton final state. The equations are:

$$
\begin{aligned}
N_l &= N_{pppp} + N_{fppp} + N_{ffpp} + N_{fffp} + N_{ffff} = N_{t4} + N_{t3} + N_{t2} + N_{t1} + N_{t0} \\
N_{t0} &= (1-p)^4 N_{pppp} + (1-p)^3(1-f) N_{fppp} + (1-p)^2(1-f)^2 N_{ffpp} \\
&+ (1-p)(1-f)^3 N_{fffp} + (1-f)^4 N_{ffff} \\
N_{t1} &= 4p(1-p)^3 N_{pppp} + \left[3p(1-p)^2(1-f) + f(1-p)^3\right] N_{fppp} \\
&+ \left[2(1-p)^2 f(1-f) + 2p(1-p)(1-f)^2\right] N_{ffpp} \\
&+ \left[3(1-p)f(1-f)^2 + p(1-f)^3\right] N_{fffp} + 4f(1-f)^3 N_{ffff} \\
N_{t2} &= 6p^2(1-p)^2 N_{pppp} + \left[3p^2(1-p)(1-f) + 3pf(1-p)^2\right] N_{fppp} \\
&+ \left[p^2(1-f)^2 + (1-p)^2 f^2 + 4p(1-p)f(1-f)\right] N_{ffpp} \\
&+ \left[3(1-p)f^2(1-f) + 3pf(1-f)^2\right] N_{fffp} + 6f^2(1-f)^2 N_{ffff} \\
N_{t3} &= 4p^3(1-p) N_{pppp} + \left[3p^2(1-p)f + p^3(1-f)\right] N_{fppp} \\
&+ \left[2p(1-p)f^2 + 2p^2 f(1-f)\right] N_{ffpp} \\
&+ \left[3pf^2(1-f) + (1-p)f^3\right] N_{fffp} + 4f^3(1-f) N_{ffff} \\
N_{t4} &= p^4 N_{pppp} + p^3 f N_{fppp} + p^2 f^2 N_{ffpp} + p f^3 N_{fffp} + f^4 N_{ffff}
\end{aligned}
\tag{44}
$$

The inverted matrix gives:

$$A = \begin{pmatrix} f^4 & -f^3(1-f) & (1-f)^2 f^2 & -(1-f)^3 f & (1-f)^4 \\ -4f^3 p & f^2(f+3p-4fp) & -2f(1-f)(p+f-2pf) & (1-f)^2(p+3f-4pf) & -4(1-f)^3(1-p) \\ 6f^2p^2 & -3fp(p+f-2pf) & \begin{array}{c} f(1-p)(2p+f-3fp) \\ +p(1-f)(2f+p-3fp) \end{array} & -3(1-f)(1-p)(p+f-2pf) & 6(1-f)^2(1-p)^2 \\ -4fp^3 & p^2(p+3f-4pf) & -2p(1-p)(p+f-2pf) & (1-p)^2(3p+f-4pf) & -4(1-f)(1-p)^3 \\ p^4 & -p^3(1-p) & p^2(1-p)^2 & -p(1-p)^3 & (1-p)^4 \end{pmatrix}$$

$$\begin{pmatrix} N_{pppp} \\ N_{fppp} \\ N_{ffpp} \\ N_{fffp} \\ N_{ffff} \end{pmatrix} = \frac{A}{(p-f)^4} \begin{pmatrix} N_{t0} \\ N_{t1} \\ N_{t2} \\ N_{t3} \\ N_{t4} \end{pmatrix} \tag{45}$$

From this, the number of events with prompt leptons is derived as:

$$N_{pppp} = \frac{1}{(p-f)^4}\left[(1-f)^4 N_{t4} - f(1-f)^3 N_{t3} + f^2(1-f)^2 N_{t2} - f^3(1-f)N_{t1} + f^4 N_{t0}\right] \tag{46}$$

It confirms the rule of flipping signs given in previous Section.

For the background contributions, we get

$$\begin{aligned}
N_{fppp} &= \frac{1}{(p-f)^4}\left[-4pf^3 N_{t0} + [f^3(1-p) + 3pf^2(1-f)]N_{t1} - 2[f^2(1-p)(1-f) + pf(1-f)^2]N_{t2}\right. \\
&\quad \left. + [3f(1-p)(1-f)^2 + p(1-f)^3]N_{t3} - 4(1-p)(1-f)^3 N_{t4}\right] \\
N_{ffpp} &= \frac{1}{(p-f)^4}\left\{6p^2f^2 N_{t0} - 3[pf^2(1-p) + p^2 f(1-f)]N_{t1}\right. \\
&\quad + [f^2(1-p)^2 + 4pf(1-p)(1-f) + p^2(1-f)^2]N_{t2} \\
&\quad \left. -3[f(1-p)^2(1-f) + p(1-p)(1-f)^2]N_{t3} + 6(1-p)^2(1-f)^2 N_{t4}\right\} \\
N_{fffp} &= \frac{1}{(p-f)^4}\left[-4p^3 f N_{t0} + [3p^2 f(1-p) + p^3(1-f)]N_{t1} - 2[pf(1-p)^2 + p^2(1-p)(1-f)]N_{t2}\right. \\
&\quad \left. + [f(1-p)^3 + 3p(1-p)^2(1-f)]N_{t3} - 4(1-p)^3(1-f)N_{t4}\right] \\
N_{ffff} &= \frac{1}{(p-f)^4}\left[p^4 N_{t0} - p^3(1-p)N_{t1} + p^2(1-p)^2 N_{t2} - p(1-p)^3 N_{t3} + (1-p)^4 N_{t4}\right]
\end{aligned} \tag{47}$$

All of these formulae confirm the rules laid out before.

B Fake rates for any lepton configuration

Given the rules for the weights and for the signs of the contributions presented in Sections 3 to A, we can now infer the formula for the number of events with prompt leptons in the general case of n SF leptons in the final state:

$$N_{prompt} = p^n \frac{1}{(p-f)^n} \sum_{i=0}^{n} (-1)^i f^i (1-f)^{n-i} N_{t,n-i} \tag{48}$$

Following the discussion in Section 5, this can be further generalized to the final states with n electrons and m muons:

$$N_{prompt} = p_e^n p_\mu^m \frac{1}{(p_e - f_e)^n} \frac{1}{(p_\mu - f_\mu)^m} \sum_{i=0}^{n} \sum_{j=0}^{m} (-1)^{i+j} f_e^i (1-f_e)^{n-i} f_\mu^j (1-f_\mu)^{m-j} N_{t,n-i,m-j} \quad (49)$$

We will now establish the fully general formula for the case with n_p prompt and n_f fake SF leptons, N_{n_f,n_p}, from which equation (48) is a special case with only prompt leptons. The general rules have been summarized in Section 7.

Putting it all together leads to the general formula:

$$\begin{aligned} N_{n_f,n_p} &= \frac{1}{(p-f)^n} \sum_{i=0}^{n} (-1)^{n_f - i} N_{t,n-i} \times \\ & \sum_{j=jmin}^{jmax} \frac{i!(n-i)!}{j!(i-j)!(n_f-j)!(n_p-i+j)!} p^j f^{i-j} (1-p)^{n_f-j} (1-f)^{n_p-i+j} \end{aligned} \quad (50)$$

with $jmin = max(i - n_p, 0)$ and $jmax = min(n_f, i)$.

After factorizing to get small coefficients, we obtain for the passing leptons

$$\begin{aligned} N_{n_f,n_p}^{pass} &= \frac{\epsilon^{n_f}}{(1-\epsilon\eta)^n} \sum_{i=0}^{n} (-1)^{n_f - i} N_{t,n-i} \times \\ & \sum_{j=jmin}^{jmax} \frac{i!(n-i)!}{j!(i-j)!(n_f-j)!(n_p-i+j)!} \epsilon^{i-j} \eta^{n_f-j} \end{aligned} \quad (51)$$

For the usual choices of p and f, all coefficients are expected to be small except when $i = j = n_f$, which corresponds to N_{t,n_p}. For this term, the ratio of factorials is equal to 1 and the sign is positive.

To illustrate the way to obtain the general recursive equations, let us start with the $p = 1$ case. Then, the number of events with n_f fake and n_p prompt leptons passing the tight cuts, N_{n_f,n_p}^{pass}, is

$$N_{n_f,n_p}^{pass} = (1-f)^{n_f} N_{n_f,n_p}$$

This is given by the number of observed events N_{t,n_p}, from which we have to subtract the sum (over i) of all events with n_p passing leptons from configurations N_{n_f+i,n_p-i}. The latter are given by

$$C_i^{n_f+i} (1-f)^{n_f} f^i N_{n_f+i,n_p-i}$$

where C_i^n is the combinatorial factor $C_i^n = \frac{n!}{(n-i)!i!}$. The expression for the general case $p \neq 1$ rewritten as recursive relations with a fake ratio f_k for the configuration with k prompt leptons is then

$$\begin{aligned} N_{n_f,n_p} &= \frac{1}{p^{n_p}} \frac{p^{n_f}}{(p - f_{n_p})^{n_f}} \left[N_{t,n_p} - \sum_{i=1}^{n_p} C_i^{n_f+i} \left(\frac{p - f_{n_p-i}}{p} \right)^{n_f} (f_{n_p-i})^i p^{n_p-i} N_{n_f+i,n_p-i} \right. \\ & \left. - \sum_{i=1}^{n_f} (-1)^{i+1} C_i^{n_p+i} \left(\frac{1-p}{p} \right)^i N_{t,n_p+i} \right] \end{aligned} \quad (52)$$

In the special case of dileptons, it yields the equations of Section 6.

The very last step is to generalize equation (50) to allow for both electrons and muons. By following the above reasoning separately to the electrons and to the muons, the following expression is obtained:

$$\begin{aligned} N_{n_{fe},n_{pe},n_{f\mu},n_{p\mu}} = & \frac{1}{(p_e-f_e)^n}\frac{1}{(p_\mu-f_\mu)^m}\sum_{i=0}^{n}\sum_{k=0}^{m}(-1)^{n_f-i-k}N_{t,n-i,m-k}\times \\ & \sum_{j=jmin}^{jmax}\sum_{l=lmin}^{lmax}\frac{i!(n-i)!}{j!(i-j)!(n_{fe}-j)!(n_{pe}-i+j)!}\frac{k!(m-k)!}{k!(k-l)!(n_{f\mu}-k)!(n_{p\mu}-k+l)!}\times \\ & p_e^j f_e^{i-j}(1-p_e)^{n_{fe}-j}(1-f_e)^{n_{pe}-i+j}p_\mu^l f_\mu^{k-l}(1-p_\mu)^{n_{f\mu}-l}(1-f_\mu)^{n_{p\mu}-k+l} \end{aligned} \tag{53}$$

where $n_f = n_{fe} + n_{f\mu}$; $jmin = max(i - n_{pe}, 0)$ and $jmax = min(n_{fe}, i)$; $lmin = max(k - n_{p\mu}, 0)$ and $lmax = min(n_{f\mu}, k)$.

If small coefficients are preferred,

$$\begin{aligned} N^{pass}_{n_{fe},n_{pe},n_{f\mu},n_{p\mu}} = & \frac{\epsilon_e^{n_{fe}}}{(1-\epsilon_e\eta_e)^n}\frac{\epsilon_\mu^{n_{f\mu}}}{(1-\epsilon_\mu\eta_\mu)^m}\sum_{i=0}^{n}\sum_{k=0}^{m}(-1)^{n_f-i-k}N_{t,n-i,m-k}\times \\ & \sum_{j=jmin}^{jmax}\sum_{l=lmin}^{lmax}\frac{i!(n-i)!}{j!(i-j)!(n_{fe}-j)!(n_{pe}-i+j)!}\frac{k!(m-k)!}{k!(k-l)!(n_{f\mu}-l)!(n_{p\mu}-k+l)!}\times \\ & \epsilon_e^{i-j}\eta_e^{n_{fe}-j}\epsilon_\mu^{k-l}\eta_\mu^{n_{f\mu}-k} \end{aligned} \tag{54}$$

C Generalization of the uncertainties

General expressions for the statistical uncertainties were provided in Section 8 and will not be repeated here.

The quantities of interest are the uncertainties on the numbers of events passing the tight cuts. To simplify the calculation, we start from the error propagation for equation (50). It can be rewritten as

$$N^{pass}_{n_f,n_p} = p^{n_p}f^{n_f}N_{n_f,n_p} = \sum_{i=0}^{n}(-1)^{n_f-i}N_{t,n-i}w_i \tag{55}$$

where

$$w_i = \sum_{j=jmin}^{jmax} w_{ij} = \sum_{j=jmin}^{jmax} C_{ij}(p-f)^{-n}p^{n_p+j}(1-p)^{n_f-j}f^{n_f+i-j}(1-f)^{n_p-i+j} \tag{56}$$

with C_{ij} being the ratio of factorials. As seen in relation with (50), the dominant term will be

$$N^{pass}_{n_f,n_p} \simeq N_{t,n_p}w_{n_f} \ , \ w_{n_f} = (p-f)^{-n}p^n f^{n_f}(1-f)^{n_p} \tag{57}$$

The systematic uncertainties originate from the uncertainties on p and f. They can be propagated from the w_{ij} in (56):

$$\begin{aligned} \frac{\partial w_{ij}}{\partial p} &= w_{ij}\left[-\frac{n}{p-f}+\frac{n_p+j}{p}-\frac{n_f-j}{1-p}\right] \equiv w_{ij}C_{p,ij} \\ \frac{\partial w_{ij}}{\partial f} &= w_{ij}\left[\frac{n}{p-f}+\frac{n_f+i-j}{f}-\frac{n_p-i+j}{1-f}\right] \equiv w_{ij}C_{f,ij} \end{aligned} \tag{58}$$

from which the derivatives of w_i are given as

$$\frac{\partial w_i}{\partial p} = \sum_{j=jmin}^{jmax} \frac{\partial w_{ij}}{\partial p} = \sum_{j=jmin}^{jmax} w_{ij} C_{p,ij} \tag{59}$$

and a similar expression for $\frac{\partial w_i}{\partial f}$. Assuming again that the uncertainties on p and f are uncorrelated the uncertainty on the number of events passing is obtained as

$$\begin{aligned} \Delta N^{pass}_{n_f,n_p}\Big|_{syst} &= \sqrt{\left(\frac{\partial N^{pass}_{n_f,n_p}}{\partial p}\Delta p\right)^2 + \left(\frac{\partial N^{pass}_{n_f,n_p}}{\partial f}\Delta f\right)^2} \\ &= \sqrt{\left(\sum_{i=0}^{n}(-1)^{n_f-i} N_{t,n-i}\frac{\partial w_i}{\partial p}\Delta p\right)^2 + \left(\sum_{i=0}^{n}(-1)^{n_f-i} N_{t,n-i}\frac{\partial w_i}{\partial f}\Delta f\right)^2} \end{aligned} \tag{60}$$

The different $N^{pass}_{n_f,n_p}$ for different values of n_f and n_p will be correlated. Let us choose another combination $N^{pass}_{m_f,m_p}$ such that $m_f + m_p = n_f + n_p = n$.

For the systematic uncertainties, correlations arise because p and f appear in both expressions. The variance of the number of events is

$$V_{syst}(N^{pass}_{n_f,n_p}, N^{pass}_{m_f,m_p}) = \sum_{i=0}^{n}(-1)^{n_f-i} N_{t,n-i} \sum_{k=0}^{n}(-1)^{m_f-k} N_{t,m-k} V(w_i, w'_k) \tag{61}$$

where

$$V(w_i, w'_k) = \sum_{j=jmin}^{jmax}\sum_{l=lmin}^{lmax} V(w_{ij}, w'_{kl}) = \sum_{j=jmin}^{jmax}\sum_{l=lmin}^{lmax} w_{ij} w'_{kl}\left[C_p C'_p \Delta p^2 + C_f C'_f \Delta f^2\right] \tag{62}$$

with C'_p and C'_f defined analogously to (58) as a function of m_p and m_f.

If we keep again only the leading terms, the variance becomes

$$V_{syst}(N^{pass}_{n_f,n_p}, N^{pass}_{m_f,m_p}) \simeq N_{t,n_p} N_{t,m_p} w_{n_f} w'_{m_f}\left[C_p C'_p \Delta p^2 + C_f C'_f \Delta f^2\right] \tag{63}$$

and the correlation coefficient is

$$\begin{aligned} C_{syst}(N^{pass}_{n_f,n_p}, N^{pass}_{m_f,m_p}) &\simeq \frac{N_{t,n_p} N_{t,m_p} w_{n_f} w'_{m_f}\left[C_p C'_p \Delta p^2 + C_f C'_f \Delta f^2\right]}{N_{t,n_p} w_{n_f}\sqrt{C_p^2\Delta p^2 + C_f^2 \Delta f^2}\, N_{t,m_p} w'_{m_f}\sqrt{C'^2_p\Delta p^2 + C'^2_f\Delta f^2}} \\ &= \frac{C_p C'_p \Delta p^2 + C_f C'_f \Delta f^2}{\sqrt{C_p^2\Delta p^2 + C_f^2\Delta f^2}\sqrt{C'^2_p\Delta p^2 + C'^2_f\Delta f^2}} \end{aligned} \tag{64}$$

References

[1] D.Barge et al., CMS AN-2010/164, 07 June 2010 (v2).

[2] W.Andrews et al., CMS AN-2009/041, 21 March 2009.

[3] D.Barge et al., CMS AN-10-193.

[4] G.Karapostoli et al., presentation at the 10/06/10 SUSY meeting.

[5] P.Lobelle Pardo, Top dilepton workgroup meeting, 20/01/10.

[6] Y.Pakhotin, presentation at the 18/06/09 leptonic SUSY meeting; M.Pioppi, presentation at the 18/06/09 leptonic SUSY meeting; S.Padhi, presentation at the 31/07/09 leptonic SUSY meeting;.