A few notes on the 2d sideband method

These are some simple notes on the 2d sideband method, basically summarizing the method proposed by Zhijun in the photon purity note and some variants proposed by Henso et al. First I review the standard method, then I describe how to take into account the signal leakage in the background control region, and then discuss correlations. The main equations are 8, 16, 30, 35, 41, 43, 45.

Basic method

In the classical 2D sideband method investigated by Zhijun, one considers candidates in a 2D plane (x, y) formed by a photon identification variable x and an isolation variable y. The plane is then divided into four regions by means of rectangular cuts. We define:

- N^A: number of candidates passing signal cuts in both x and y (signal region)
- N^B: number of candidates passing signal cut in x and failing cut in y (control region 1)
- M^A: number of candidates failing signal cut in x and passing cut in y (control region 2)
- M^B: number of candidates failing signal cuts in both x and y (control region 3)

We also define numbers N^A_{sigMC}, N^B_{sigMC}, M^A_{sigMC}, M^B_{sigMC} for true signal candidates in the four regions taken from Monte Carlo (MC), N^A_{bkgMC}, N^B_{bkgMC}, M^A_{bkgMC}, M^B_{bkgMC} for fake candidates in the same regions taken from MC, and N^A_{sig}, N^A_{bkg} (and others) for background and signal candidates in data.

One starts with two simplifying hypotheses:

1. the correlation between the x and y variables is negligible for the background.
2. the number of signal candidates is negligible compared to the number of fake candidates in the 3 control regions:

$$N^B_{\text{bkg}} \gg N^B_{\text{sig}} \tag{1}$$
$$M^A_{\text{bkg}} \gg M^A_{\text{sig}} \tag{2}$$
$$M^B_{\text{bkg}} \gg M^B_{\text{sig}} \tag{3}$$

As a consequence of the first assumption, one can assume that $N^A_{\text{bkg}}/N^B_{\text{bkg}} = M^A_{\text{bkg}}/M^B_{\text{bkg}}$. The second assumptions leads to

$$N^B = N^B_{\text{bkg}} \tag{4}$$
$$M^A = M^A_{\text{bkg}} \tag{5}$$
$$M^B = M^B_{\text{bkg}} \tag{6}$$

Therefore, combining the two hypotheses:

$$N^A_{\text{bkg}} = N^B_{\text{bkg}} \times M^A_{\text{bkg}}/M^B_{\text{bkg}} = N^B \times M^A/M^B \tag{7}$$

which provides us with a fully data-driven estimation of the background. The signal in data in region A is therefore:

$$N^A_{\text{sig}} = N^A - N^A_{\text{bkg}} = N^A - N^B \frac{M^A}{M^B} \tag{8}$$

and the purity is

$$P = N^A_{\rm sig}/N^A = 1 - \frac{N^B}{N^A}\frac{M^A}{M^B} \tag{9}$$

The statistical errors on $N^A_{\rm sig}$ and P are respectively:

$$\begin{aligned} \sigma\left(N^A_{\rm sig}\right) &= \sqrt{N^A + \left(N^B\frac{M^A}{M^B}\right)^2\left(\frac{1}{N^B}+\frac{1}{M^A}+\frac{1}{M^B}\right)} && (10) \\ \sigma(P) &= \frac{N^B}{N^A}\frac{M^A}{M^B}\sqrt{\frac{1}{N^A}+\frac{1}{N^B}+\frac{1}{M^A}+\frac{1}{M^B}} && (11) \end{aligned}$$

The method is fully data-driven, but requires that the two hypothesis mentioned above be satisfied

Taking into account non-negligible correlations in the background 2d distribution

We can try to take into account non-negligible correlations in the following way:

$$\begin{aligned} N^A_{\rm sig} &= N^A - N^A_{\rm bkg} = N^A - N^A_{\rm bkg}\frac{N^B_{\rm bkg}\times M^A_{\rm bkg}/M^B_{\rm bkg}}{N^B_{\rm bkg}\times M^A_{\rm bkg}/M^B_{\rm bkg}} && (12) \\ &= N^A - \left(N^B_{\rm bkg}\frac{M^A_{\rm bkg}}{M^B_{\rm bkg}}\right)\left(\frac{N^A_{\rm bkg}}{N^B_{\rm bkg}}\frac{M^B_{\rm bkg}}{M^A_{\rm bkg}}\right) && (13) \\ &\approx N^A - \left(N^B_{\rm bkg}\frac{M^A_{\rm bkg}}{M^B_{\rm bkg}}\right)\left(\frac{N^A_{\rm bkgMC}}{N^B_{\rm bkgMC}}\frac{M^B_{\rm bkgMC}}{M^A_{\rm bkgMC}}\right) && (14) \\ &\approx N^A - \left(N^B\frac{M^A}{M^B}\right)\left(\frac{N^A_{\rm bkgMC}}{N^B_{\rm bkgMC}}\frac{M^B_{\rm bkgMC}}{M^A_{\rm bkgMC}}\right) && (15) \\ &\approx N^A - \left(N^B\frac{M^A}{M^B}\right)R_{\rm MC} && (16) \end{aligned}$$

The purity is

$$P = N^A_{\rm sig}/N^A = 1 - \frac{N^B}{N^A}\frac{M^A}{M^B}R_{\rm MC} \tag{17}$$

The statistical errors on $N^A_{\rm sig}$ and P are respectively:

$$\begin{aligned} \sigma\left(N^A_{\rm sig}\right) &= \sqrt{N^A + \left(N^B\frac{M^A}{M^B}R_{\rm MC}\right)^2\left(\frac{1}{N^B}+\frac{1}{M^A}+\frac{1}{M^B}+\left(\frac{\sigma(R_{\rm MC})}{R_{\rm MC}}\right)^2\right)} && (18) \\ \sigma(P) &= \frac{N^B}{N^A}\frac{M^A}{M^B}R_{\rm MC}\sqrt{\frac{1}{N^A}+\frac{1}{N^B}+\frac{1}{M^A}+\frac{1}{M^B}+\left(\frac{\sigma(R_{\rm MC})}{R_{\rm MC}}\right)^2} && (19) \end{aligned}$$

where

$$\sigma(R_{\rm MC}) = R_{\rm MC}\sqrt{\frac{1}{N^A_{\rm bkgMC}}+\frac{1}{N^B_{\rm bkgMC}}+\frac{1}{M^A_{\rm bkgMC}}+\frac{1}{M^B_{\rm bkgMC}}} \tag{20}$$

One must use background Monte Carlo to compute the quantity $R_{\rm MC} = \frac{N^A_{\rm bkgMC}}{N^B_{\rm bkgMC}}\frac{M^B_{\rm bkgMC}}{M^A_{\rm bkgMC}}$, which is related to the correlation between x and y for the background ($R_{\rm MC} = 1$ in case the 2 variables are not correlated). The method is data-driven with the exception of the factor $R_{\rm MC}$, which must be determined from background Monte Carlo.

Taking into account signal leakage in background control regions

The signal yield, even if small, may be not negligible in the three background control regions. We can take the signal leakage into account in the following way, as proposed by Henso (we start with the case where we neglect the correlations). For each region, we can correct the relation between the observed number of candidates and the number of background candidates in the following way:

$$N^B = N^B_{\rm bkg} + N^B_{\rm sig} = N^B_{\rm bkg} + N^A_{\rm sig}\frac{N^B_{\rm sig}}{N^A_{\rm sig}} \tag{21}$$

$$M^A = M^A_{\rm bkg} + M^A_{\rm sig} = M^A_{\rm bkg} + N^A_{\rm sig}\frac{M^A_{\rm sig}}{N^A_{\rm sig}} \tag{22}$$

$$M^B = M^B_{\rm bkg} + M^B_{\rm sig} = M^B_{\rm bkg} + N^A_{\rm sig}\frac{M^B_{\rm sig}}{N^A_{\rm sig}} \tag{23}$$

and we use signal MC to determine the three ratios

$$\frac{N^B_{\rm sig}}{N^A_{\rm sig}} = c_1 \approx \frac{N^B_{\rm sigMC}}{N^A_{\rm sigMC}} \tag{24}$$

$$\frac{M^A_{\rm sig}}{N^A_{\rm sig}} = c_2 \approx \frac{M^A_{\rm sigMC}}{N^A_{\rm sigMC}} \tag{25}$$

$$\frac{M^B_{\rm sig}}{N^A_{\rm sig}} = c_3 \approx \frac{M^B_{\rm sigMC}}{N^A_{\rm sigMC}} \tag{26}$$

Therefore

$$N^A_{\rm sig} = N^A - N^B_{\rm bkg}\frac{M^A_{\rm bkg}}{M^B_{\rm bkg}} = N^A - (N^B - N^A_{\rm sig}c_1)\frac{M^A - N^A_{\rm sig}c_2}{M^B - N^A_{\rm sig}c_3} \tag{27}$$

Now some simple algebra:

$$N^A_{\rm sig}M^B - (N^A_{\rm sig})^2c_3 = N^AM^B - N^AN^A_{\rm sig}c_3 - N^BM^A + N^BN^A_{\rm sig}c_2 + N^A_{\rm sig}M^Ac_1 - (N^A_{\rm sig})^2c_1c_2 \tag{28}$$

$$(N^A_{\rm sig})^2(c_1c_2 - c_3) + N^A_{\rm sig}(M^B + N^Ac_3 - N^Bc_2 - M^Ac_1) + (N^BM^A - N^AM^B) = 0 \tag{29}$$

Choosing the physical solution for $N^A_{\rm sig}$:

$$\begin{aligned} N^A_{\rm sig} &= \frac{-(M^B + N^Ac_3 - N^Bc_2 - M^Ac_1) + \sqrt{(M^B + N^Ac_3 - N^Bc_2 - M^Ac_1)^2 + 4(c_1c_2 - c_3)(N^AM^B - N^BM^A)}}{2(c_1c_2 - c_3)} \\ &= \frac{(M^B + N^Ac_3 - N^Bc_2 - M^Ac_1)(-1 + \sqrt{1 + \frac{4(c_1c_2-c_3)(N^AM^B-N^BM^A)}{(M^B+N^Ac_3-N^Bc_2-M^Ac_1)^2}}}{2(c_1c_2 - c_3)} \end{aligned} \tag{30}$$

The formula can be simplified if we assume (to be checked!) that:

$$|c_1c_2 - c_3| \ll \frac{(M^B + N^Ac_3 - N^Bc_2 - M^Ac_1)^2}{4|N^AM^B - N^BM^A|} \tag{31}$$

In that case, doing a first-order Taylor expansion of the square root part, we get

$$N^A_{\rm sig} = \frac{(M^B + N^Ac_3 - N^Bc_2 - M^Ac_1)(-1 + 1 + \frac{2(c_1c_2-c_3)(N^AM^B-N^BM^A)}{(M^B+N^Ac_3-N^Bc_2-M^Ac_1)^2})}{2(c_1c_2 - c_3)} \tag{32}$$

$$
\begin{aligned}
&= \frac{2(c_1c_2 - c_3)(N^AM^B - N^BM^A)}{(M^B + N^Ac_3 - N^Bc_2 - M^Ac_1)(2(c_1c_2 - c_3))} && (33)\\
&= \frac{N^AM^B - N^BM^A}{M^B + N^Ac_3 - N^Bc_2 - M^Ac_1} && (34)\\
&= \left(N^A - N^B\frac{M^A}{M^B}\right)\frac{1}{1 + \frac{c_3N^A - c_2N^B - c_1M^A}{M^B}} && (35)
\end{aligned}
$$

Note that if $c_1 = c_2 = c_3 = 0$ we get back the original expression:

$$N^A_{\text{sig}} = \frac{N^AM^B - N^BM^A}{M^B} = N^A - N^B\frac{M^A}{M^B} \tag{36}$$

The method is data-driven with the exception that the coefficients c_1, c_2, c_3 must be determined from signal Monte Carlo.

Taking into account both signal leakage and correlations

If we want to take into account also the correlations, we have to start from equation 14:

$$N^A_{\text{sig}} \approx N^A - \left(N^B_{\text{bkg}}\frac{M^A_{\text{bkg}}}{M^B_{\text{bkg}}}\right)\left(\frac{N^A_{\text{bkgMC}}}{N^B_{\text{bkgMC}}}\frac{M^B_{\text{bkgMC}}}{M^A_{\text{bkgMC}}}\right) \tag{37}$$

and replace N^B_{bkg} with $N^B - N^A_{\text{sig}}c_1$ and so on:

$$N^A_{\text{sig}} \approx N^A - \left((N^B - N^A_{\text{sig}}c_1)\frac{M^A - N^A_{\text{sig}}c_2}{M^B - N^A_{\text{sig}}c_3}\right)R_{\text{MC}} \tag{38}$$

Now back to the algebra:

$$N^A_{\text{sig}}M^B - (N^A_{\text{sig}})^2c_3 = N^AM^B - N^AN^A_{\text{sig}}c_3 - N^BM^AR_{\text{MC}} + N^BN^A_{\text{sig}}c_2R_{\text{MC}} + N^A_{\text{sig}}M^Ac_1R_{\text{MC}} - (N^A_{\text{sig}})^2c_1c_2R_{\text{MC}} \tag{39}$$

$$(N^A_{\text{sig}})^2(c_1c_2R_{\text{MC}} - c_3) + N^A_{\text{sig}}(M^B + N^Ac_3 - N^Bc_2R_{\text{MC}} - M^Ac_1R_{\text{MC}}) + (N^BM^AR_{\text{MC}} - N^AM^B) = 0 \tag{40}$$

so

$$N^A_{\text{sig}} = \frac{(M^B + N^Ac_3 - N^Bc_2R_{\text{MC}} - M^Ac_1R_{\text{MC}})(-1 + \sqrt{1 + \frac{4(c_1c_2R_{\text{MC}} - c_3)(N^AM^B - N^BM^AR_{\text{MC}})}{(M^B + N^Ac_3 - N^Bc_2R_{\text{MC}} - M^Ac_1R_{\text{MC}})^2}}}{2(c_1c_2R_{\text{MC}} - c_3)} \tag{41}$$

and if

$$\left|\frac{4(c_1c_2R_{\text{MC}} - c_3)(N^AM^B - N^BM^AR_{\text{MC}})}{(M^B + N^Ac_3 - N^Bc_2R_{\text{MC}} - M^Ac_1R_{\text{MC}})^2}\right| \ll 1 \tag{42}$$

then a first-order Taylor expansion in $\lambda = \frac{4(c_1c_2R_{\text{MC}} - c_3)(N^AM^B - N^BM^AR_{\text{MC}})}{(M^B + N^Ac_3 - N^Bc_2R_{\text{MC}} - M^Ac_1R_{\text{MC}})^2}$ leads us to:

$$N^A_{\text{sig}} = \left(N^A - N^B\frac{M^A}{M^B}R_{\text{MC}}\right)\frac{1}{1 + \frac{c_3N^A - c_2N^BR_{\text{MC}} - c_1M^AR_{\text{MC}}}{M^B}} \tag{43}$$

The typical values of λ I find are below 1% so the approximation is really good.

Now, it is a bit more complicated to derive the expression for the error on N^A_{sig} since one has to compute some derivatives. I write

$$N^A_{\text{sig}} = \frac{N^A - N^B\frac{M^A}{M^B}R_{\text{MC}}}{f(N^A, N^B, M^A, M^B, R_{\text{MC}}, c_1, c_2, c_3)} \tag{44}$$

where $f(N^A, N^B, M^A, M^B, R_{\rm MC}, c_1, c_2, c_3) = 1 + \frac{c_3 N^A - c_2 N^B R_{\rm MC} - c_1 M^A R_{\rm MC}}{M^B}$. Then:

$$\begin{aligned}
\frac{\partial N^A_{\rm sig}}{\partial N^A} &= 1/f - \left(N^A - N^B \frac{M^A}{M^B} R_{\rm MC}\right)/f^2 \times \frac{c_3}{M^B} \\
&= 1/f \left(1 - N^A_{\rm sig} \frac{c_3}{M^B}\right) \\
\frac{\partial N^A_{\rm sig}}{\partial N^B} &= -1/f \frac{M^A}{M^B} R_{\rm MC} + \left(N^A - N^B \frac{M^A}{M^B} R_{\rm MC}\right)/f^2 \times \frac{c_2 R_{\rm MC}}{M^B} \\
&= -1/f \frac{M^A}{M^B} R_{\rm MC} \left(1 - N^A_{\rm sig} \frac{c_2}{M^A}\right) \\
\frac{\partial N^A_{\rm sig}}{\partial M^A} &= -1/f \frac{N^B}{M^B} R_{\rm MC} + \left(N^A - N^B \frac{M^A}{M^B} R_{\rm MC}\right)/f^2 \times \frac{c_1 R_{\rm MC}}{M^B} \\
&= -1/f \frac{N^B}{M^B} R_{\rm MC} \left(1 - N^A_{\rm sig} \frac{c_1}{N^B}\right) \\
\frac{\partial N^A_{\rm sig}}{\partial M^B} &= 1/f \frac{N^B M^A}{(M^B)^2} R_{\rm MC} + \left(N^A - N^B \frac{M^A}{M^B} R_{\rm MC}\right)/f^2 \times \frac{c_3 N^A - c_2 N^B R_{\rm MC} - c_1 M^A R_{\rm MC}}{(M^B)^2} \\
&= 1/f \frac{N^B M^A}{(M^B)^2} R_{\rm MC} \left(1 + N^A_{\rm sig} \frac{c_3 N^A - c_2 N^B R_{\rm MC} - c_1 M^A R_{\rm MC}}{N^B M^A R_{\rm MC}}\right) \\
\frac{\partial N^A_{\rm sig}}{\partial R_{\rm MC}} &= -1/f \frac{N^B M^A}{M^B} + \left(N^A - N^B \frac{M^A}{M^B} R_{\rm MC}\right)/f^2 \times \frac{c_2 N^B + c_1 M^A}{M^B} \\
&= -1/f \frac{N^B M^A}{M^B} \left(1 - N^A_{\rm sig} \frac{c_2 N^B + c_1 M^A}{N^B M^A}\right) \\
\frac{\partial N^A_{\rm sig}}{\partial c_1} &= 1/f N^A_{\rm sig} \frac{M^A R_{\rm MC}}{M^B} \\
\frac{\partial N^A_{\rm sig}}{\partial c_2} &= 1/f N^A_{\rm sig} \frac{N^B R_{\rm MC}}{M^B} \\
\frac{\partial N^A_{\rm sig}}{\partial c_3} &= -1/f N^A_{\rm sig} \frac{N^A}{M^B}
\end{aligned}$$

Then the uncertainty on $N^A_{\rm sig}$ is

$$\sigma(N^A_{\rm sig}) = \sqrt{\sum \left(\frac{\partial N^A_{\rm sig}}{\partial x_i}\right)^2 \sigma_i^2} \tag{45}$$

For the purity, since $P = N^A_{\rm sig}/N^A$, the derivatives are:

$$\frac{\partial P}{\partial N^A} = -\frac{N^A_{\rm sig}}{(N^A)^2} + \frac{1}{N^A} \frac{\partial N^A_{\rm sig}}{\partial N^A} = \frac{1}{N^A}(-P + \frac{\partial N^A_{\rm sig}}{\partial N^A}) \tag{46}$$

$$\frac{\partial P}{\partial x}(x \neq N^A) = \frac{1}{N^A} \frac{\partial N^A_{\rm sig}}{\partial x} \tag{47}$$

For N^A, N^B, M^A, M^B the uncertainty σ is just their square root. For c_1, c_2, c_3, $R_{\rm MC}$ one just needs to propagate the uncertainty on the signal and background MC yields.

I have done a Toy Monte Carlo study to compare the error obtained with this formula to the square root of the variance of the distribution of $N^A_{\rm sig}$ that I obtain when I generate an ensemble of 100k pseudoexperiments, in each one of which I generate randomly - according to Poisson statistics and to the mean values as obtained from our data and Monte Carlo - the number of events in the four regions for signal MC, bkg MC, and data, and compute N^A_{sig} according to equation 41.

Here are the numbers I get using the April reprocessed data and MC, corrected isolation vs tightisEM (with or without fracm+weta1): So the formula reproduces the RMS of the distribution from the ToyMC

p_T bin	[10,15) GeV	[15,20) GeV	$[20, +\inf)$ GeV
error from formula	490.5	130.6	49.7
RMS of distribution	536.5	148.6	52.0

to within approximately 10%, which is probably because in the configuration where one removes only fracm and weta1, one of the control regions is not very populated, and so the gaussian approximation for the errors is not too precise. If I release also deltaE and Eratio to improve the statistics in the control regions and redo the exercise I get an agreement at the level of 1%:

p_T bin	[10,15) GeV	[15,20) GeV	$[20, +\inf)$ GeV
error from formula	186.9	53.6	27.8
RMS of distribution	188.8	54.4	27.9

It would be interesting to compare these numbers to those obtained with Mathematica for a cross-validation of both approaches.