Unfolding at CMS

Matthias Weber on behalf of the CMS Collaboration
Institute for Particle Physics, ETH Zurich, 8093 Zurich, Switzerland

Abstract
The unfolding techniques used by the CMS collaboration on the 2010 data analyses are presented. Each method is discussed on the basis of an experimental measurement. The main focus is on studying the sensitivity to different models used in the determination of the response matrix and the propagation of statistical errors.

1 Unfolding of experimental distributions

In order to allow a *direct* comparison of experimental measurements with theoretical predictions, the measurements must be unfolded for detector effects. This also permits the direct comparison of distributions from different experiments without knowledge of the detector response for each experiment. Moreover, the automated tuning of Monte Carlo generators using multiple measurements is greatly facilitated using unfolded data. In principle the measurements could be presented without corrections for detector effects together with a detector response matrix. The smearing of the theory distribution with a published response matrix for each single measurement and each experiment is far more complicated in Monte Carlo tuning efforts.

The measurements discussed here are based on data collected in proton-proton collisions with the Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC). A detailed description of the CMS detector can be found elsewhere [1]. The detector response matrix R in the unfolding procedure is usually derived using simulated Monte Carlo (MC) samples. The CMS detector response is derived via simulation based on GEANT4 [2]. The inversion of the response matrix can lead to unacceptable solutions, since small statistical fluctuations can lead to large effects in the solution. Even negative entries in the unfolded distribution are possible. This oscillating behaviour can be reduced by imposing the requirement that the true distribution is smooth. This smoothing procedure is known as regularization. The regularized unfolding methods used and investigated by CMS are iterative Bayesian unfolding [3], the SVD method [4] and Tikhonov regularization [5].

2 Iterative Unfolding: Charged Particle Multiplicities

The iterative Bayesian unfolding method is used in the measurement of the charged particle multiplicities [6]. The observed charged multiplicity $O = (O_1, \ldots, O_N)$ will in general be different from the true multiplicity $T = (T_1, \ldots, T_M)$ due to track reconstruction inefficiencies, the presence of secondary particles and decay products of long-lived hadrons. In this method the inversion is done in a stepwise iterative procedure. The unfolded distribution for the iteration step k is:

$$T_j^{(k)} = \sum_i \frac{R_{ij} T_j^{(k-1)}}{\sum_s R_{il} T_s^{(k-1)}} \cdot O_i. \tag{1}$$

After each iteration k the χ^2 between $T^{(k)}$ and $T^{(k-1)}$ is calculated. The procedure is stopped once the χ^2 converges and a stable solution T is found. The response matrix R is derived from samples generated with PYTHIA6 [7] with the tune D6T. It is checked that the final solution T does not depend on initial ansatz for the true distribution, using in one case a flat distribution and in the other case the MC generator level distribution. In a second step the robustness of the unfolding procedure is tested, using the PYTHIA6 DW and a PHOJET [8] response matrix to unfold pseudodata generated by PYTHIA6 tune D6T.

The observed differences shown in Fig. 1 are small over a large range, for higher multiplicity bins they are dominated by statistical fluctuations.

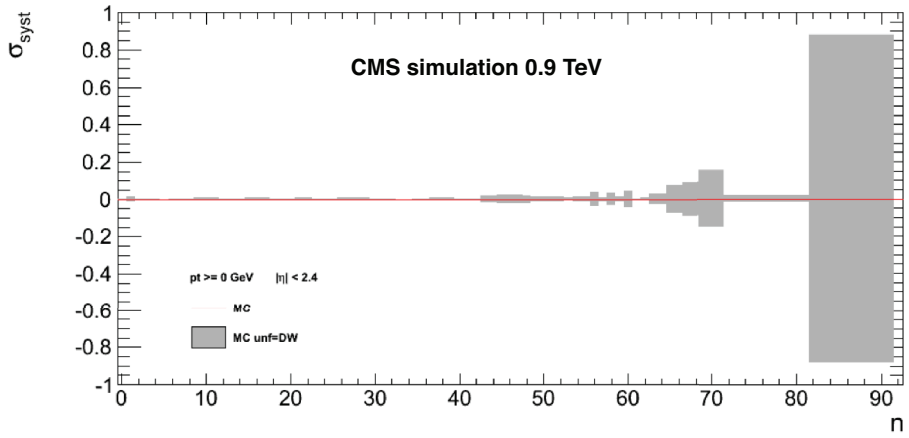

Fig. 1: The robustness of the unfolding procedure, unfolding PYTHIA6 D6T pseudodata distribution of the charged particle multiplicity n with the PYTHIA6 DW response matrix. σ_{syst} represents the relative difference between the unfolded result and MC thruth.

The covariance matrix of the unfolded spectrum is derived using a resampling technique. As shown in Fig. 2 the statistical errors are very dependent on the number of iterations. The errors increase as a function of the iterations, while the statistical bias decreases. For a small number of iterations the errors are smaller than $\sqrt{N}$, where N is the number of entries.

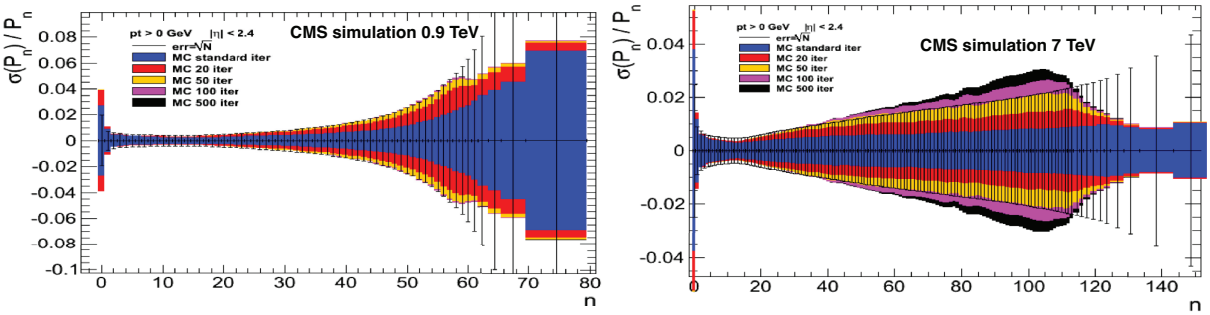

Fig. 2: The dependence of the relative statistical errors of the charged multiplicity spectrum on the number of iterations in the unfolding. The pseudodata is simulated with PYTHIA6 at $\sqrt{s} = 900$ GeV (left) and $\sqrt{s} = 7$ TeV (right).

3 SVD Unfolding: Hadronic Event Shapes

The SVD method is a special implementation of the Tikhonov regularization method, based on the singular value decomposition of the response matrix. The regularization method works as low-pass filter and suppresses small singular values, which would lead to oscillating behaviour. This method is used in the measurement of hadronic event-shape variables in 7 TeV proton-proton jet data [9]. The variable which we will use in the following is the central transverse thrust, which is defined as

$$\tau_{\perp,\mathcal{C}} \equiv 1 - \max_{\hat{n}_{\text{T}}} \frac{\sum_i |\vec{p}_{\perp,i} \cdot \hat{n}_{\text{T}}|}{\sum_i p_{\perp,i}}, \tag{2}$$

where $p_{\perp,i}$ are the jet transverse momenta. Well balanced dijet events have low thrust values close to 0, spherical multijet events have high values. The measurement is performed in several bins of the leading

jet transverse momentum $p_{\perp,1}$. The unfolded measured distributions are compared to predictions from the MC generators PYTHIA6, HERWIG++ [10], ALPGEN [11], MADGRAPH [12] and PYTHIA8 [13]. The measurement can be used in further tuning of MC generators.

The jet energy and jet position resolutions of the detector distort the event-shape distributions. Especially in the lower range of event shape values the off-diagonal elements of the response matrix are sizable (20-30%); only for higher event-shape values the diagonal element is around 90%. The response matrix is determined using simulated PYTHIA6 D6T QCD events. The consistency between the unfolded pseudodata distribution and the MC generator level distribution is checked for each generator using in all cases the PYTHIA6 D6T response matrix. Fig. 3 shows that for all other generators a good closure can be observed. The regularization parameter is chosen such that the χ^2 value between the unfolded and the generator level distribution is minimal (In this test we consider all generators but PYTHIA6). The full covariance matrix is used in the χ^2 calculation.

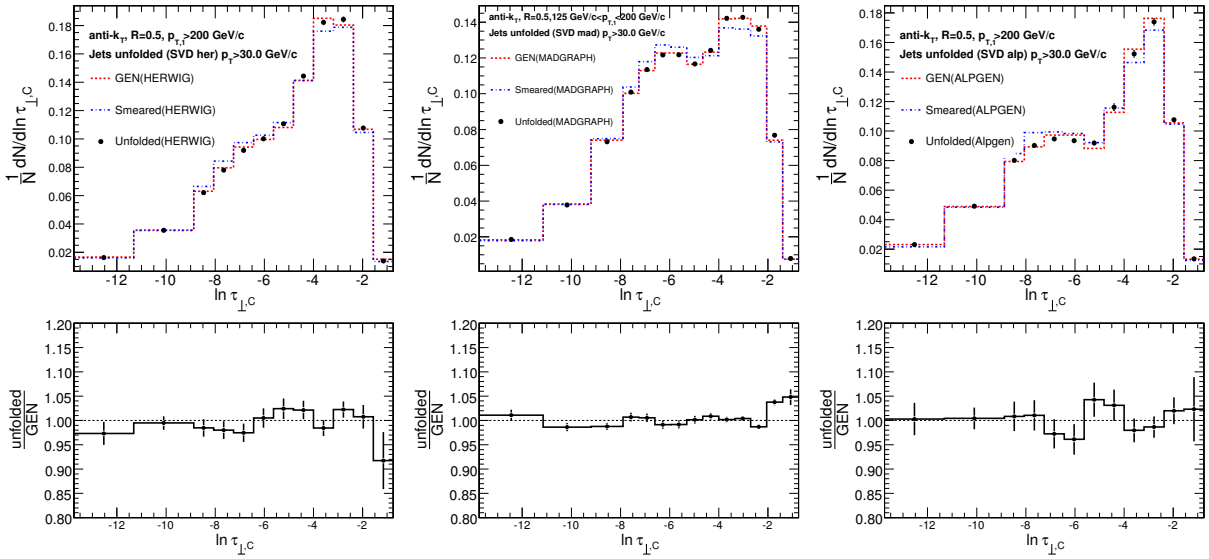

Fig. 3: The closure of the SVD unfolding procedure for the central transverse thrust distribution for HERWIG++ (left), MADGRAPH (middle) and ALPGEN (right) pseudodata. The ratios show the deviations from the generator level distribution.

Unfolding the data with a response matrix determined from MADGRAPH instead of PYTHIA6 gives consistent results. We check that no preference for one of the generators is introduced by the unfolding procedure. In a first step the χ^2 between the simulated pseudodata and the data distribution prior to unfolding is calculated. These χ^2 values are compared to χ^2 values between the generator level distributions and the data distributions after unfolding. The ordering is the same before and after unfolding and the values are similar. The iterative Bayesian unfolding is applied as a further cross-check. The resulting unfolded distribution agrees within 1% with the distribution of the SVD unfolding.

The covariance matrix of the SVD method is non diagonal with large bin-to-bin correlations in the errors (Fig. 4, left). The statistical error of a bin i is taken as the square root of corresponding diagonal element of the covariance matrix $\sqrt{C_{ii}}$. As illustrated in Fig. 4, right using PYTHIA6 pseudodata, the relative statistical errors prior to unfolding are for some regularization choices bigger than the relative statistical errors after the unfolding for almost all bins. With stronger regularization (small regularization parameter, e.g. Reg. 7), the errors can be smaller than for the raw data. The correlation between the bins of the unfolded distribution is bigger and the diagonal element of the covariance matrix smaller. In those

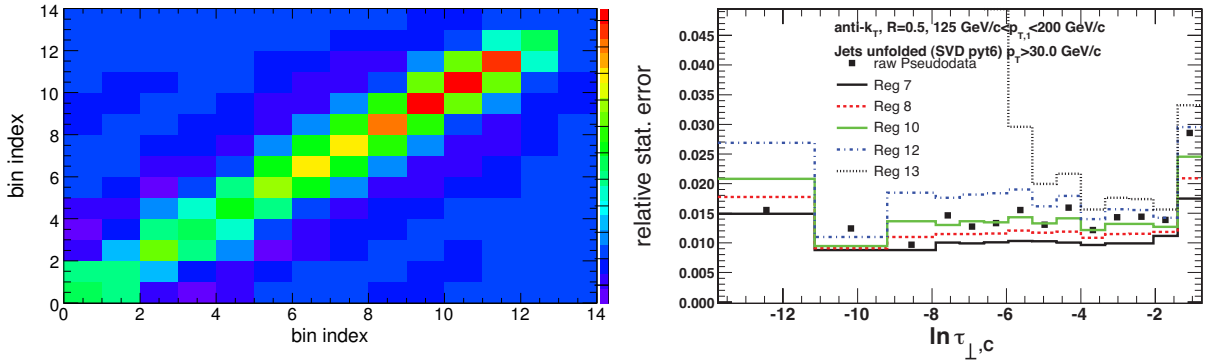

Fig. 4: The covariance matrix after unfolding using as regularization value 7 (left). Comparison of the relative statistical errors prior to unfolding and after the unfolding using several strengths of regularization (right). In both cases PYTHIA6 has been used to generate the pseudodata.

cases the statistical error is smaller than $\sqrt{N}$, where N is the bin entry in these bins after unfolding the distributions. For small regularization (higher regularization value. e.g Reg. 13) the errors approximate the errors of the matrix inversion and can be sizable.

The jet energy resolution uncertainty of 10% is treated in the measurement as uncertainty in the response matrix, i.e. the unfolding is repeated with a new response matrix determined from PYTHIA6 D6T with jet energies smeared by an additional 10%.

4 Tikhonov Regularization: Inclusive Jet Cross Section

The Tikhonov regularization method is investigated in the context of the inclusive jet cross section measurement. The true spectrum is distorted by the finite detector energy resolution. The inclusive jet transverse momentum spectrum is a steeply falling distribution covering a range of many orders of magnitude and thus more challenging than e.g. event shape distributions. The determination of the response matrix is difficult, since a huge phase space with dramatically different cross sections needs to be covered with sufficient statistics in all corners. A further complication is the fact that one particle level jet might be reconstructed as two jets in the detector or vice versa. The response matrix is thus calculated using a theory curve and smearing it with measured jet resolution functions. Cross-checks on Monte Carlo show that the solution of the Tikhonov method also depends on the choice of the number of bins for theory and data. The solution is found using the quasi-optimal approach [14]. In this approach the regularization parameter τ is varied in fine steps starting with larger parameters τ. For each step k the maximum deviation between the contents of all bins j of the unfolded histograms $\Delta(\tau) = \max_j |O_j(\tau_k) - O_j(\tau_{k-1})|$ is determined. Minima of Δ are stable solutions; the first minima depends on starting conditions of the iteration procedure and should be disregarded. In general the deepest minimum is preferred as solution, since it corresponds to the most stable solution as function of the regularization parameter τ.

The solution of the method shows a good closure using the deepest minimum. The effective input unfolding correction factor is correctly reproduced by this solution. The error propagation of the regularized solution by the matrix inversion can lead to large estimates of the statistical uncertainties as shown in Fig. 5. In data the inclusive jet spectrum uses several jet trigger streams, involving also heavily prescaled low p_{T} triggers. This can lead to artefacts in the unfolded distribution and the error propagation. The Tikhonov regularization is also used in the measurement of jet shapes.

5 Summary and Conclusion

Several unfolding methods used in CMS 2010 data analyses are presented: iterative Bayesian unfolding, SVD unfolding, Tikhonov regularization and matrix inversion. The usual tests performed in the unfolding

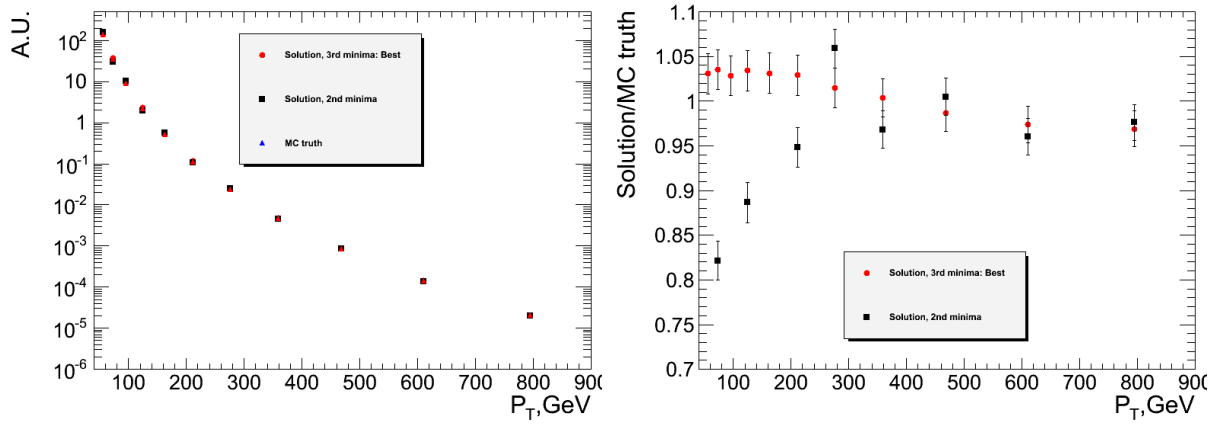

Fig. 5: The closure of the unfolding correction in the analysis of the inclusive jet cross section. On the left the MC truth spectrum and two unfolded spectra are shown. The ratio between the unfolded solutions and MC truth is shown on the right.

procedure involve closure tests and the model dependency. Uncertainties in the modelling of the response matrix are examined. The interpretation of the error propagation is discussed, especially the fact that errors are sometimes smaller than $\sqrt{N}$.

References

[1] CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST* **3** (2008) S08004.

[2] S. Agostinelli et al., "GEANT4: A simulation toolkit", *Nucl. Instrum. Meth.*, **A**506 (2003) 250.

[3] G. D'Agostini, "A Multidimensional unfolding method based on Bayes' theorem", *Nucl. Instrum. Meth.* **A362** (1995) 487.

[4] A. Höcker and V. Kartvelishvili, "SVD Approach to Data Unfolding", *Nucl. Instrum. Meth.* **A372** (1996) 46.

[5] A.N. Tikhonov, "On the solution of improperly posed problems and the method of regularization", *Sov. Math.* **5** (1964) 1035.

[6] CMS Collaboration, "Charged particle multiplicities in pp interactions at $\sqrt{s}$ =0.9, 2.36, and 7 TeV", *JHEP* **1** (2011) 79.

[7] T. Sjöstrand, S. Mrenna, and P. Z. Skands, "PYTHIA 6.4 Physics and Manual", *JHEP* **05** (2006) 026.

[8] R. Engel and J. Ranft, "Hadronic photon-photon interactions at high-energies", *Phys. Rev.* **D54** (1996) 4244, arXiv:hep-ph/9509373.

[9] CMS Collaboration, "Measurement of Hadronic Event Shapes in pp Collisions at $\sqrt{s} = 7$ TeV", *Phys. Lett. B* **699** (2011) 48.

[10] M. Bahr et al., "Herwig++ Physics and Manual", *Eur. Phys. J.* **C58** (2008) 639.

[11] M.L. Mangano et al., "ALPGEN, a generator for hard multiparton processes in hadronic collisions", *JHEP* **07** (2003) 001.

[12] J. Alwall et al., "MadGraph/MadEvent v4: The New Web Generation", *JHEP* **09** (2007) 028.

[13] T. Sjöstrand, S. Mrenna, and P. Z. Skands, "A Brief Introduction to PYTHIA8.1", *Comput. Phys. Commun.* **178** (2008) 852.

[14] V.B. Glasko, "Inverse problems of Mathematical Physics", *American Institute of Physics translation series* (1988). (2008) 852.