Mitigation of systematic errors induced biases in ML-based selection

Francisco Matorras, Pablo Martínez
IFCA
Instituto de Física de Cantabria (Santander, Spain)
Luis Crespo Ruiz
Universidad de Cantabria (Santander, Spain)

Foreword

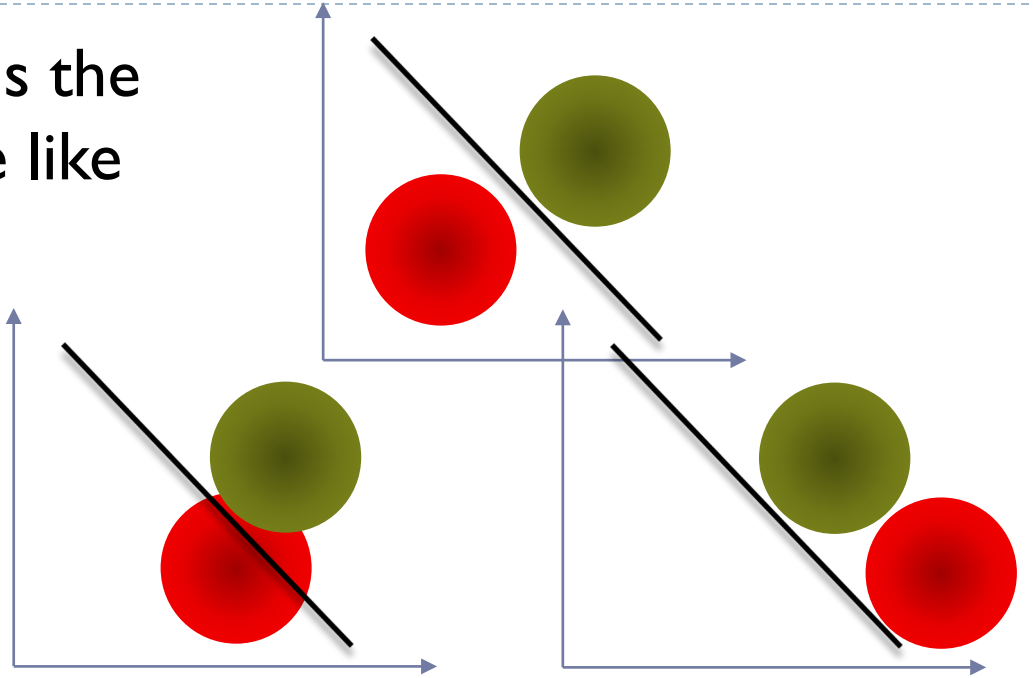
- This is an extract of the final project of Luis Crespo Ruiz to get his Degree in Physics
- **Application of multidimensional classification techniques to Particle Physics in the presence of systematic errors**
- https://repositorio.unican.es/xmlui/handle/10902/20598
- Not yet tested on full HEP analysis, but very promising results worth being shared

Motivation

- As you know, currently most particle physics analyses incorporate ML (or MVA) techniques.
- We all have heard many times the concern "what about systematics?", "this is a black-box", "what if you do not control the correlations..."
- A common practice to incorporate systematic uncertainties is to follow an equivalent procedure as for a classical cut analysis:
  1. I do my analysis based on **ideal samples**
  2. I estimate the effect of the systematics in the input variables
  3. I propagate the uncertainties through my selection (either cuts or MVA)
  4. I repeat my analysis on this distorted samples and estimate the effect on my final result: efficiency, cross section, significance...
- For ML based, this means **training in samples with ideal conditions**
- That's something, but
  - what if we rely on a variable that is poorly described and there are some other supposedly less discriminant but better in real life?

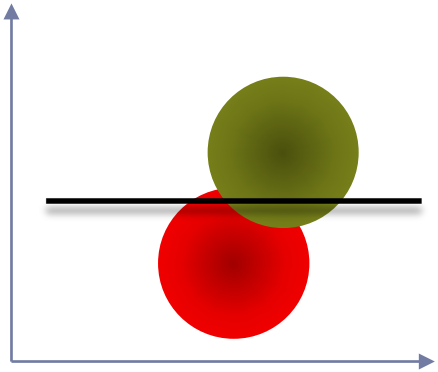
An example: separate green and red

- Your ideal sample tells the algorithm to separate like that



- But if you don't know very well how the x of your reds behaves...?

- It might be wiser to ignore the horizontal variable and cut differently



The method

- Can we make the ML algorithm learn the *weaknesses* of the variables in such cases?
- Propose to use the data augmentation technique
  - let's feed the machine with replicas with the weakness incorporated
- Relatively simple to implement in HEP
  - Given our MC (or data) original samples, replicate each event several times according to a law driven by the systematics
    - Basically, do the same you do to estimate the systematics
  - Train on these altered samples (no need to perform the costly MC simulation)

Testing the method

- As often happens, in ML difficult to demonstrate the general validity
- Run instead on an example
  - GEN + smearing-based example (true physics, simplified detector)
- Classification of the production of a dark matter candidate in association with a top pair (ttDM) versus the SM production of ttbar, for different masses of DM
  - Will show the extreme cases in DM mass, for low mass the two processes are basically indistinguishable while for high mass there is an easy separation
  - Few variables: 3-momentum for two jets and two leptons and MET, invariant masses...

Testing the method

- Study the classification for different algorithms
  - several shallow MLP, BDT, LD, SVM, Fisher
- Compare performance on systematic-modified samples after training under different conditions
  - Training on ideal samples (standard way of systematic evaluation)
  - Training with different choices of data augmented samples

Systematic on resolution

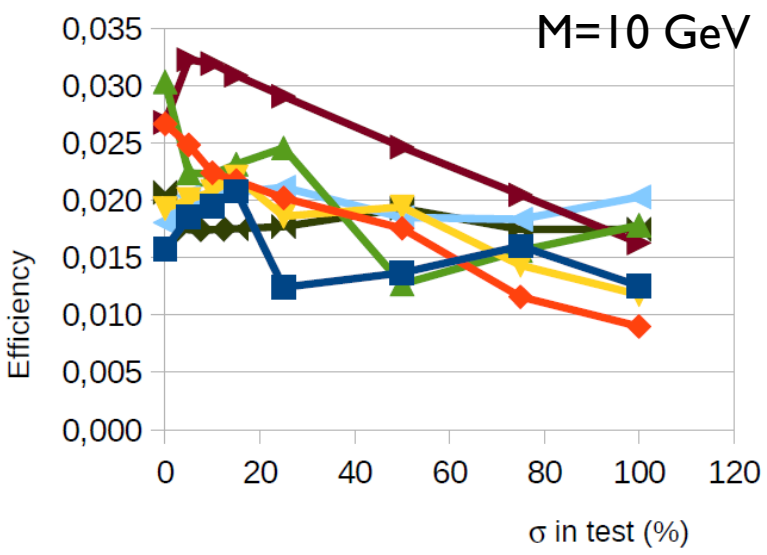
As an example, assume a systematic that implies worsen the resolution of some of the variables, with a random gaussian noise

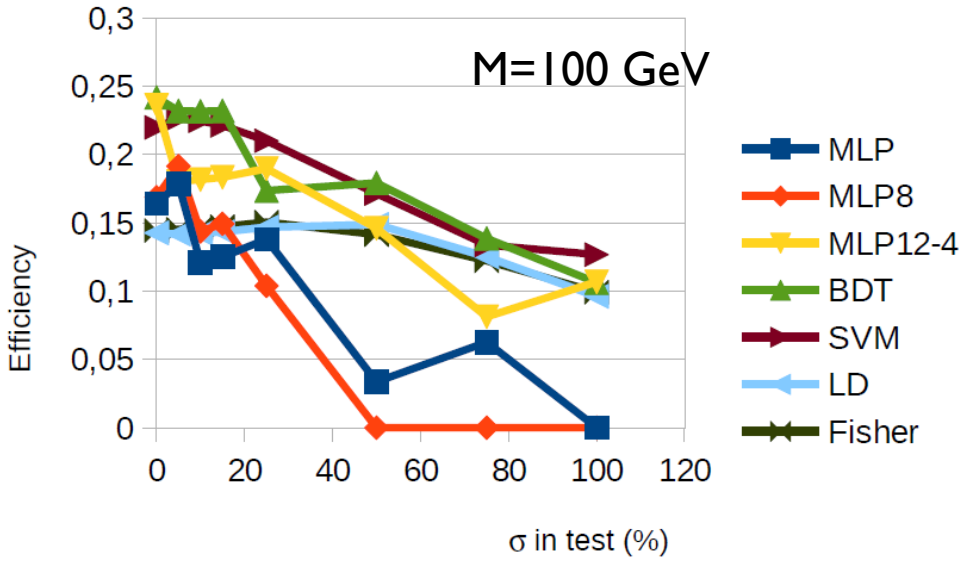
Jet energy resolution

- Check the effect of jet energy resolution
- Add an additional gaussian smearing to the jet energy (and propagate to derived quantities)
- Evaluate the signal efficiency at a different working points (1% efficiency for background in the plots shown)
- When trained on zero-systematic samples and tested on systematic-modified samples important degradation for "some" of the methods. Usual estimation of systematics

Jet energy resolution

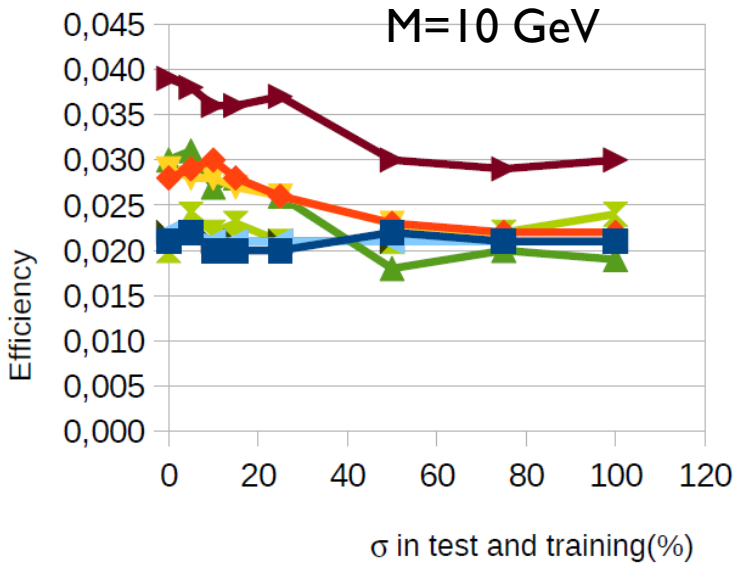
- Efficiency for signal as a function of the jet energy resolution uncertainty
- Some algorithms more robust than others, but variability on the particular training seen
- Some degradation in all cases

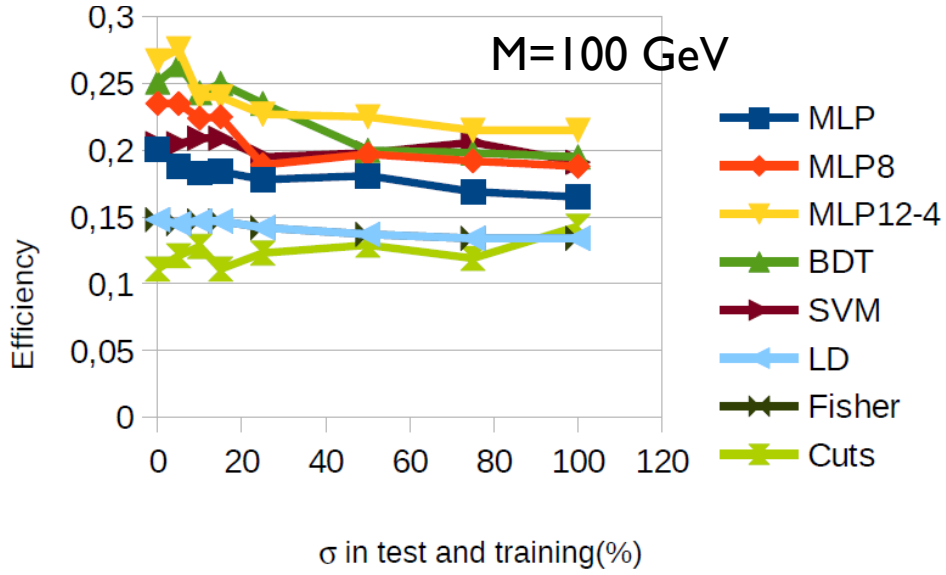




Training on data augmented

- Not surprisingly if we train with a smeared sample most of the effect is corrected
  - Here training on samples smeared with equal σ as test sample
- Systematic nearly cancelled even for very large effects
  - You might argue that this is obvious but still not always done...



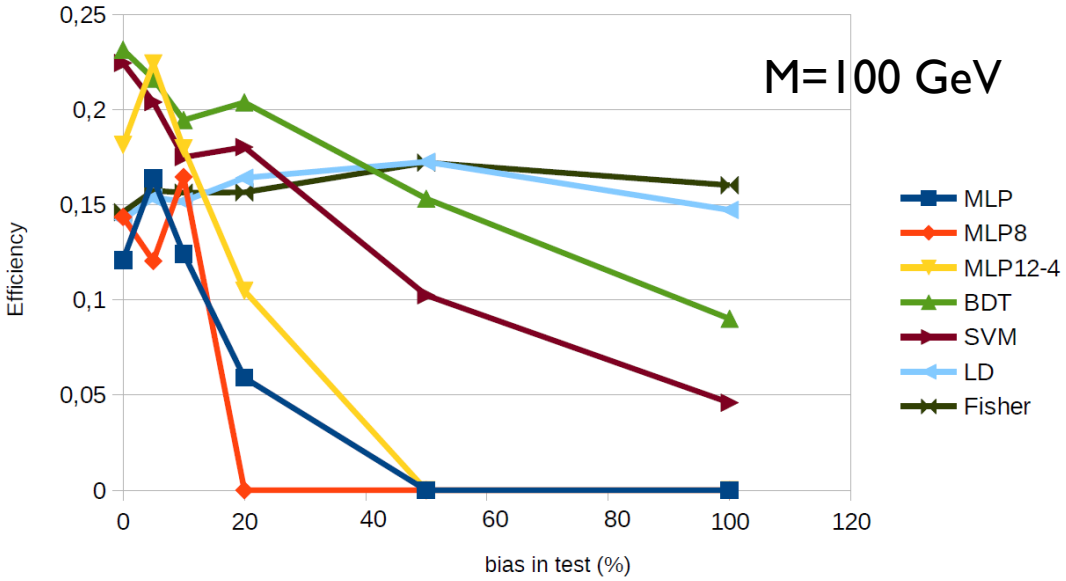


Systematic on scale/calibration

As another example, assume a systematic that implies a correlated bias in some of the variables

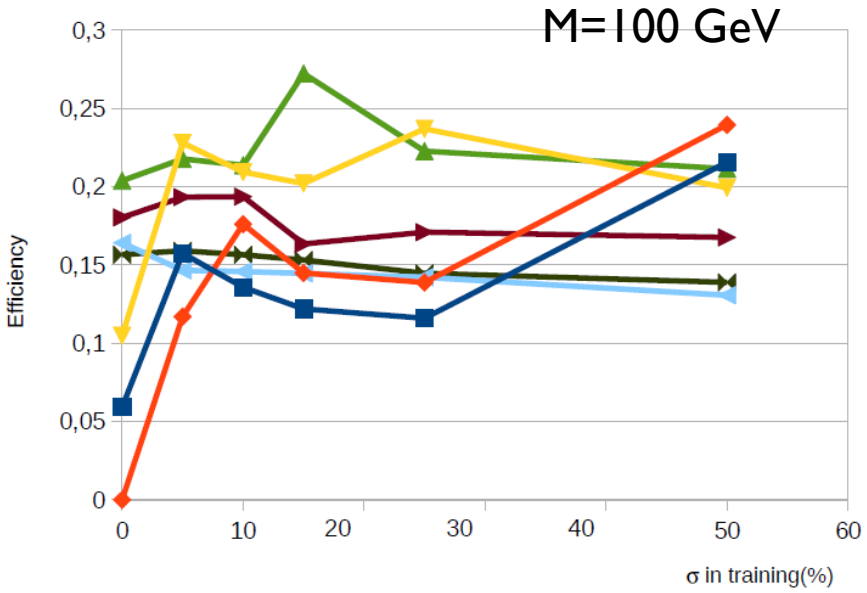
Jet energy scale

- Imagine instead a scale/calibration effect
  - The energy is wrong by a given fixed scale for all jets in all events
- Jet energy on test samples scaled by a constant term
- Jet energy in training samples is smeared
- In all cases derived variables are recalculated
- When trained on zero-systematic samples and tested on systematic samples, catastrophic degradation for "some" of the methods (NN).
  - Equivalent to a huge systematic uncertainty!



Training on data augmented

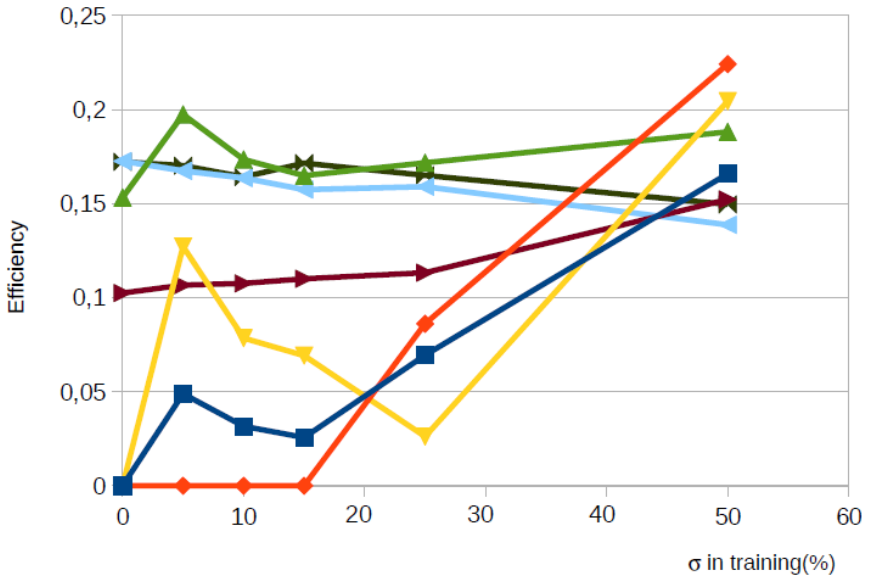
- Let's try to train with smeared samples
- Try for different size of the smearing
- Most of the effect for a 20% bias is cancelled when trainin with smeared sample with 5-20% sigma
- Similar result for a wide rang of smearing
- Large variability with algorithms but would reduce related systematics from 50-100% to a few %



Test with a 20% bias, M=100 GeV
Efficiency as a function of the smearing applied to the training samples

Training on data augmented

- Check with a larger (huge) bias of 50% (all energies scaled by 1.5)
- Again, response is mostly recovered (systematic uncertainty reduced to few %) when training with a smearing of similar size as the bias



Test with a 50% bias, M=100 GeV

Results

- Cannot draw general conclusions from this simplistic example but:
  - As it is very well known, the effect of the systematics is very strongly dependent on the type of algorithm, the working point and even the particular training.
  - Not so difficult to find examples where systematic uncertainties totally destroy the performance of ML algorithms.
  - Training on smeared samples cures most of the effect of resolution systematics, when the smearing is comparable to the systematic error
  - Training on smeared samples cures most of the effect of scale systematics, when the smear is comparable to the systematic error

Conclusions and outlook

- A very simple method based on data augmentation is proposed to mitigate the effect of systematic errors in ML-based analyses
  - Based on training on samples augmented from the original samples, which include the effect of the systematics
  - Don't need to resimulate events
- Easily implemented for most systematics, in a similar way as we usually calculate them
- It is implemented at the level of the variables, so it is valid for any ML algorithm.
- So far, only tested on simplified examples, but results promising
  - Can recover performance even for very large systematic uncertainties
  - ...but need to check on real physic examples