Approximating Decomposed Likelihood Ratios using Machine Learning

Juan G. Pavez S.
Universidad Técnica Federico Santa María
In collaboration with Kyle Cranmer (NYU)

ACAT 2016-Valparaíso

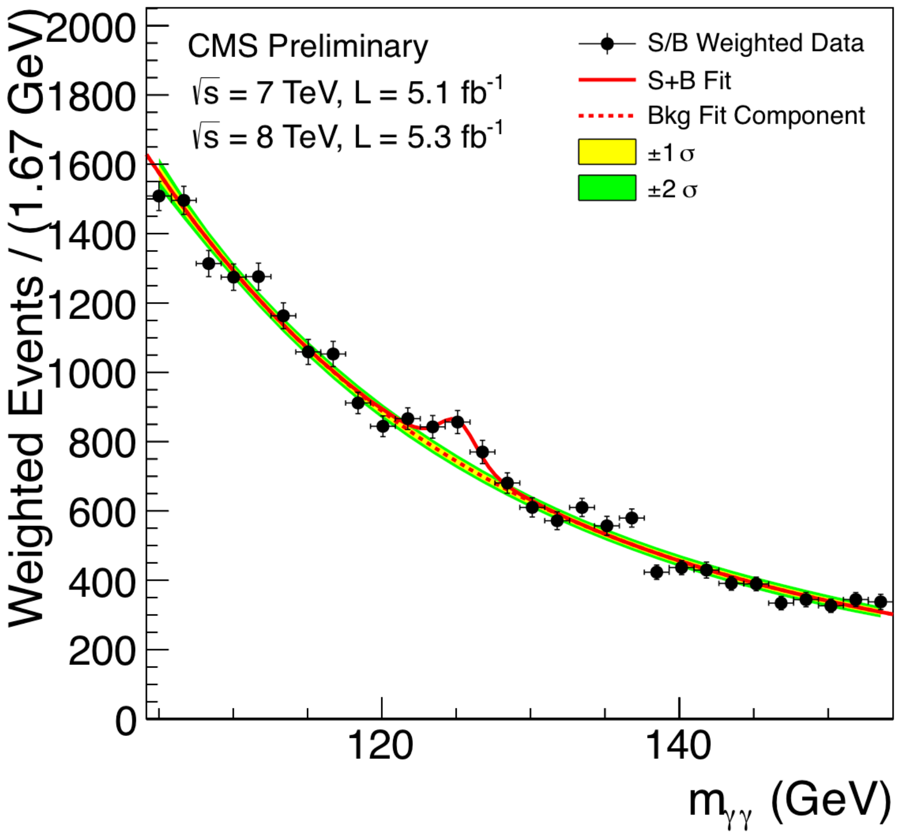
Outline

- **The problem**: How can we make likelihood based inference when the likelihood function is unknown?.
- **Solution:** Approximating likelihood ratios using machine learning and decomposing likelihood ratios.
- **Applications:** Signal vs. Background hypothesis testing, maximum likelihood estimation of parameters.
- **EFT Morphing:** Estimating coupling parameters for Higgs EFT Morphing.

The problem

Likelihood based inference when the likelihood function is unknown.

Statistics for Discovery



Higgs signal on pair of photon mass.

- **Likelihood ratios** are one of the main tools used in **HEP** when reporting results from a experiment.
- They also allow the incorporation of **systematics** effects (using the **profile likelihood ratio**).

5 SIGMA -> DISCOVERY!

Statistics for Discovery

The **Neyman-Pearson** Lemma:

$$\varphi(x) = \frac{p(x \mid \theta_0)}{p(x \mid \theta_1)} < k_\alpha$$

Given a null hypothesis θ_0 and an alternative hypothesis θ_1, this test is the most powerful test.

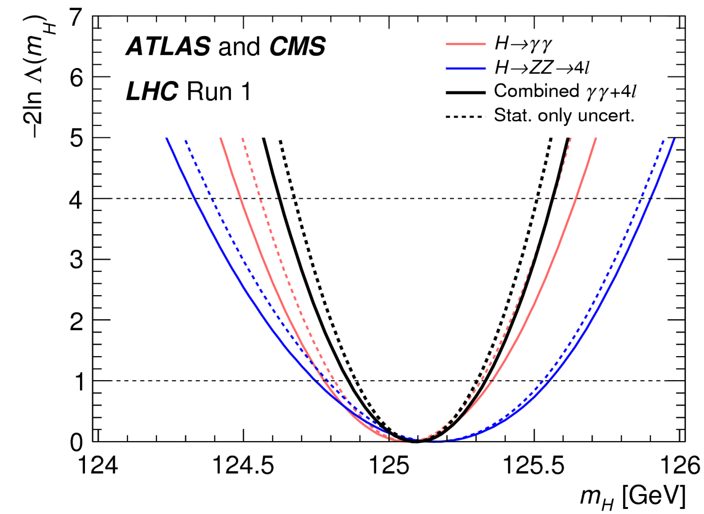
We can define the **Likelihood ratio** test as:

$$\Lambda(x) = \frac{L(\theta_0 \mid x)}{L(\theta_1 \mid x)}$$

Where $-2\ln\Lambda(x)$ asymptotically follows a χ_p^2 distribution with degrees of freedoms equals to the difference on size of θ_1 and θ_0.

Statistics for Discovery

- Likelihood ratios are used extendedly on HEP:
  - Search and discovery (Hypothesis testing).
  - Parameter estimation (Maximum Likelihood).
  - Limits (Confidence Intervals).
- **The problem:**
  - **Most of the times the Likelihood function is unknown.**
  - **We work with complex simulated multidimensional data where estimation is computationally intractable.**

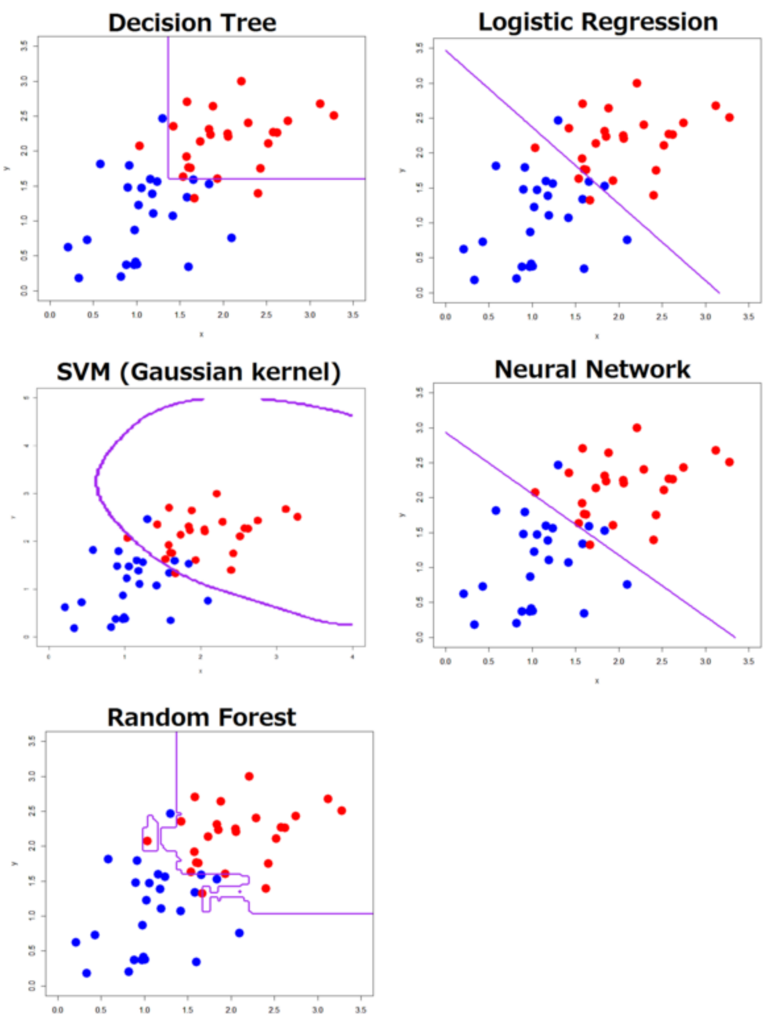


Parameter estimation of the Higgs mass using log-likelihood ratios.

Solution

Approximating decomposed likelihood ratios using machine learning.

Machine Learning in HEP



How different classifiers classify the same data.

- Classification algorithms have become a standard tool on HEP.
- The goal: Classify Signal events vs. Background events.
- **TMVA** is the common choice when applying machine learning in HEP.
- In recent years the collaboration between both fields (**ML** and **HEP**) has increased a lot:
  - Higgs Boson Challenge on Kaggle.
  - Flavours of Physics (charged lepton flavour violation) on Kaggle.
  - ALEPH workshop at NIPS.

Approximating Likelihood Ratios using Machine Learning

- Noticeable we can use the all the power of machine learning to approximate Likelihood Ratios (**Kyle Cranmer, 2015**):
- Given a classifier score $s(x;\theta_0,\theta_1)$ trained to classify between signal and background data.
- Let $p(s(x;\theta_0,\theta_1)|\theta)$ the probability distribution of the score variable conditioned by θ.
  - This distribution can be estimated using histograms or any estimation technique.

$$X = [x_1, x_2, \ldots] \longrightarrow \boxed{\textbf{Classifier}} \longrightarrow s(X)$$

Dimensionality Reduction

Approximating Likelihood Ratios using Machine Learning

- Then it can be proved that the **likelihood ratio**:

$$T(D) = \prod_{i=1}^{N} \frac{p(x_i \mid \theta_0)}{p(x_i \mid \theta_1)}$$

- Is **equivalent** to the ratio:

$$T(D) = \prod_{i=1}^{N} \frac{p(s(x;\theta_0,\theta_1) \mid \theta_0)}{p(s(x;\theta_0,\theta_1) \mid \theta_1)}$$

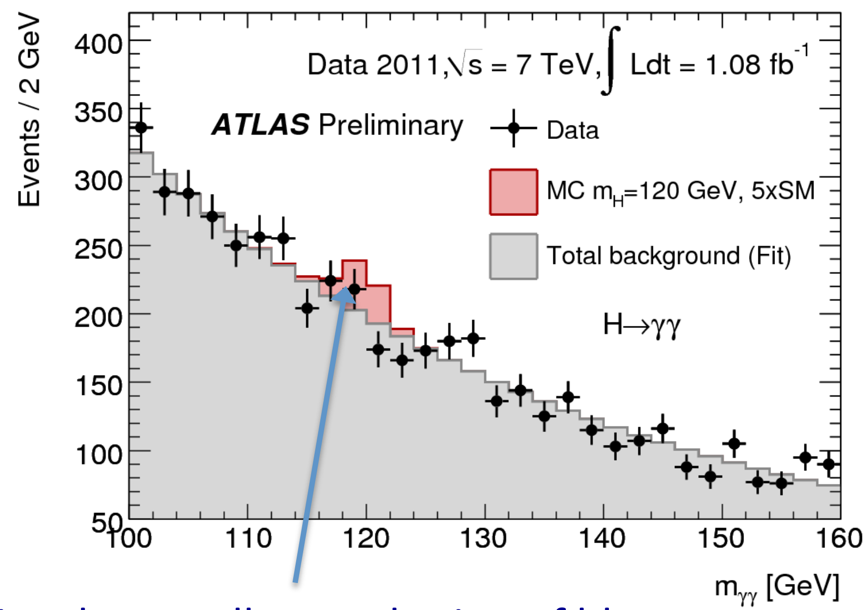
- **If the function $s(x;\theta_0,\theta_1)$ is monotonic with the ratio.**

$$s(x) \approx monotonic\left(\frac{p(x \mid \theta_1)}{p(x \mid \theta_0)}\right)$$

Most of the commonly used classifiers approximate a monotonic function of the ratio!.

Decomposing Likelihood Ratios

- Using this result we can derive another result very useful in many applications.
- Often want to separate a signal from various backgrounds, where the signal is only a **small perturbation** of the only backgrounds hypothesis.
- Another application is to test for
  - Null: SM Higgs + Bkg.
  - Alternate: BSM Higgs + Bkg.



Signal as small perturbation of bkg.

Decomposing Likelihood Ratios

- Formally we can define a **mixture model** as:

$$p(x|\theta) = \sum_i w_i(\theta) p_i(x|\theta)$$

- Where $w_i(\theta)$ define the contribution of each distribution to the full model.
- Then the Likelihood ratio between two mixture models is:

$$\frac{p(x|\theta_0)}{p(x|\theta_1)} = \frac{\sum_i w_i(\theta_0) p_i(x|\theta_0)}{\sum_j w_j(\theta_1) p_j(x|\theta_1)}$$

- Which is equivalent to (**Cranmer, 2015**):

$$\frac{p(x|\theta_0)}{p(x|\theta_1)} = \sum_i \left[\sum_j \frac{w_j(\theta_1)}{w_i(\theta_0)} \frac{p_j(s_{i,j}(x;\theta_0,\theta_1)|\theta_1)}{p_i(s_{i,j}(x;\theta_0,\theta_1)|\theta_0)} \right]^{-1}$$

Decomposing Likelihood Ratios

- Now, the likelihood ratio is decomposed into distributions of **pairwise trained classifiers**.
- Moreover, in the common case that the pairwise distributions $p(s_{i,j}(x;\theta_1,\theta_0)\,|\,\theta)$ are independent of θ the only free parameters are $w_i(\theta)$.
- It is possible to estimate using **maximum likelihood** the signal or background contributions. Keeping θ_0 fixed:

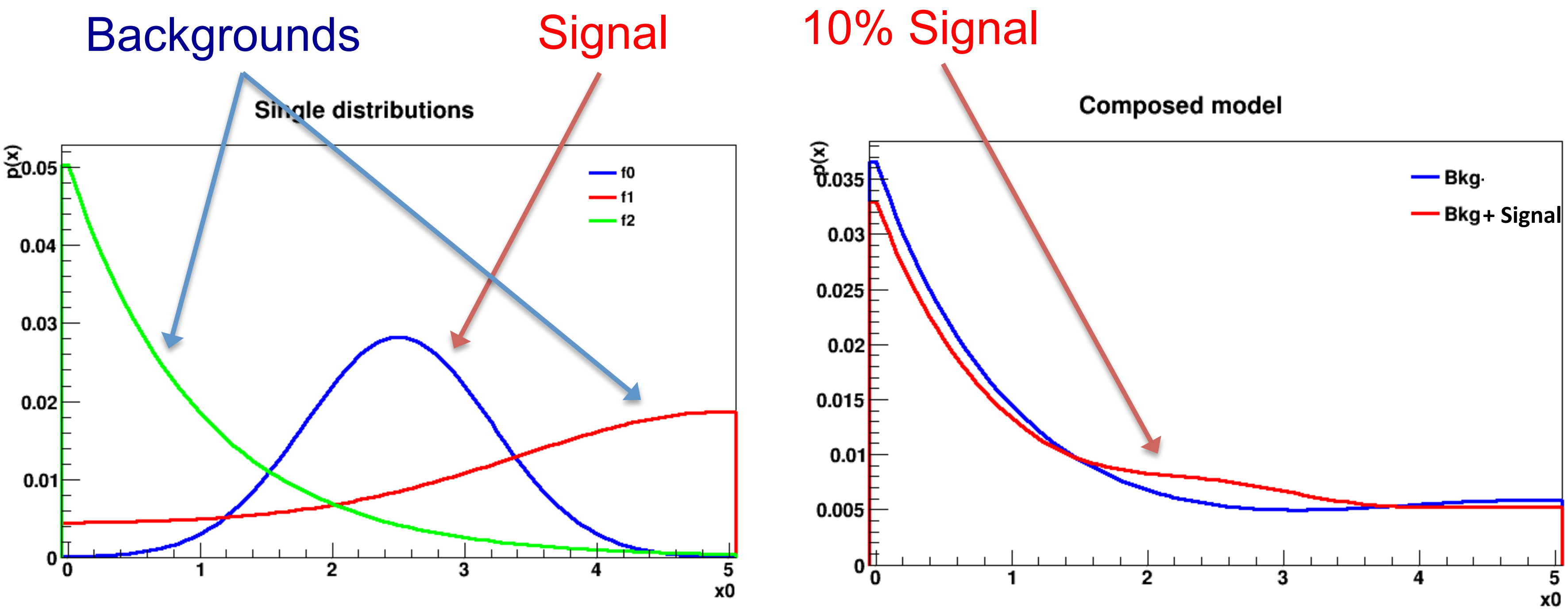
$$\hat{\theta}_1 = \underset{\theta_1}{\operatorname{argmax}} \prod_{e=1}^{n} \frac{p\ (x_e\,|\,\theta_1)}{p\ (x_e\,|\,\theta_0)}$$

Applications

Signal vs. Background hypothesis testing, maximum likelihood estimation of parameters.

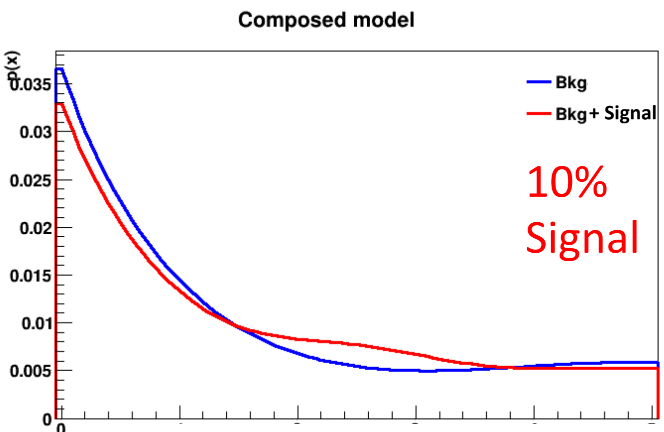
Applications

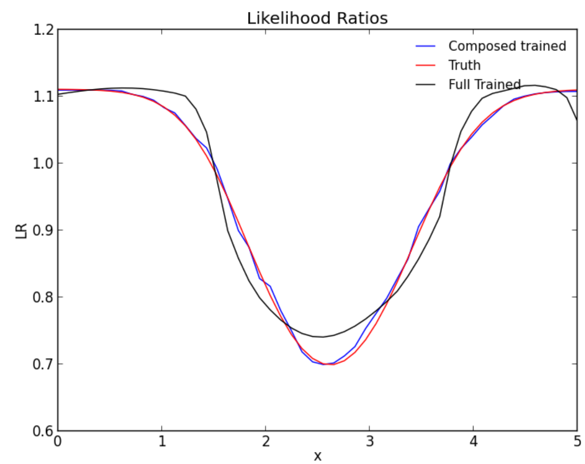
- First, consider a **simple mixture model** of 1-dim distributions.
- One of the distributions correspond to signal while the others are background.

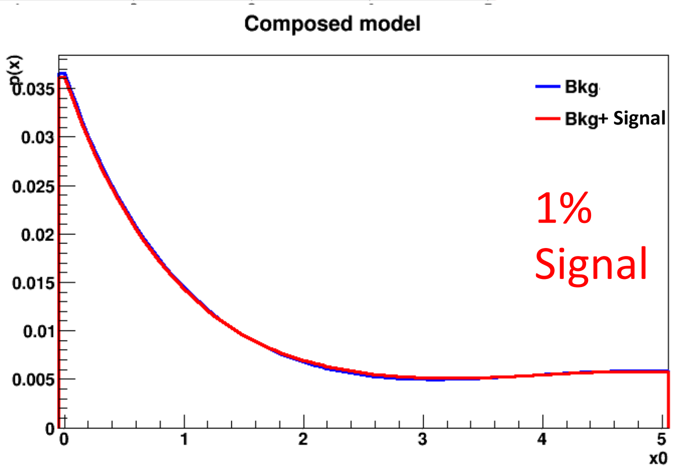


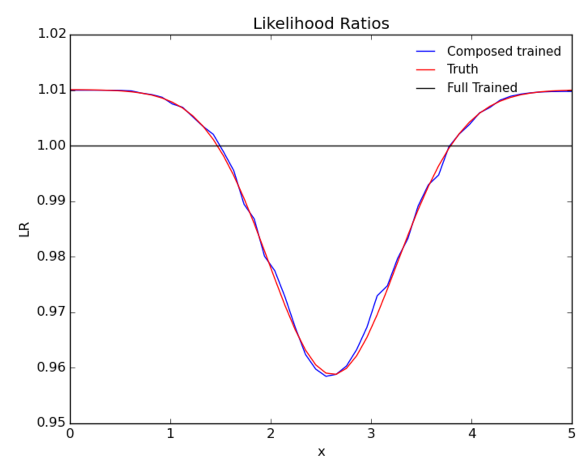
Applications

- We fit the **pairwise classifiers** and a **single classifier** (both MLPs) and then compute the approximated likelihood ratios and compare it to the **true ratio**.



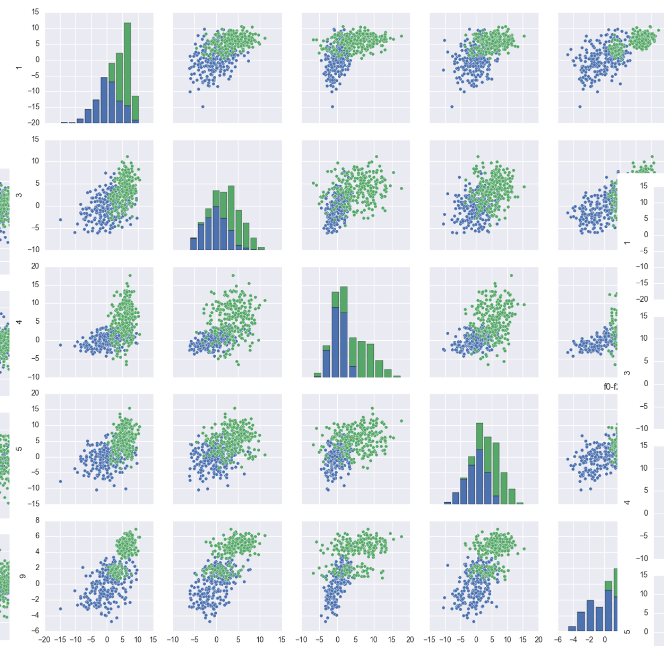




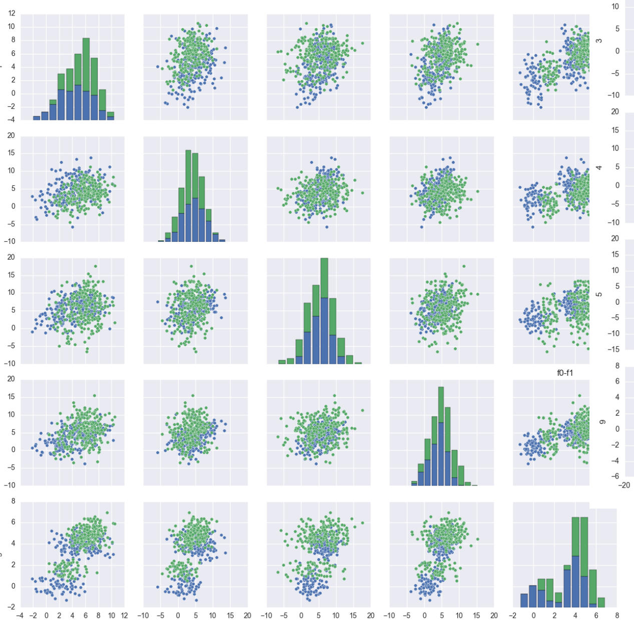


Applications

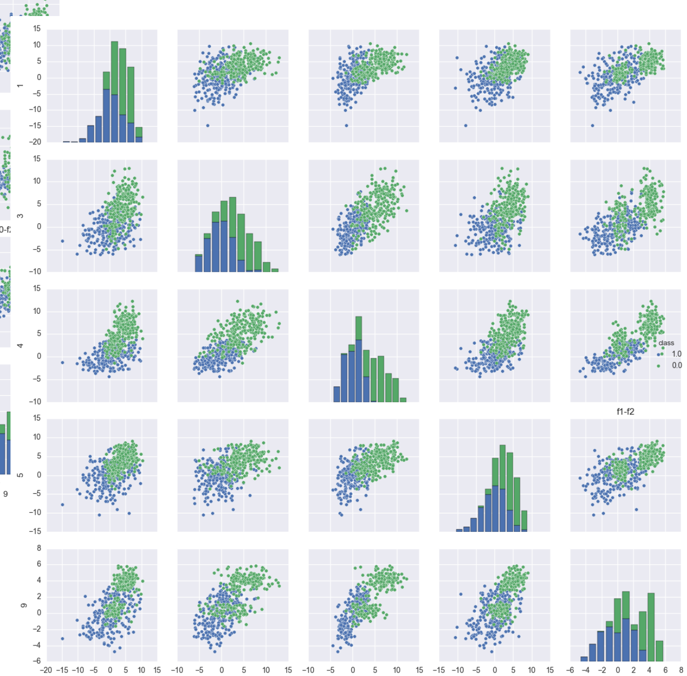
- We do the same but now with a **much harder** model: Three 10-dim distributions, each one is a mixture of Gaussians.



f0/f2



f0/f1

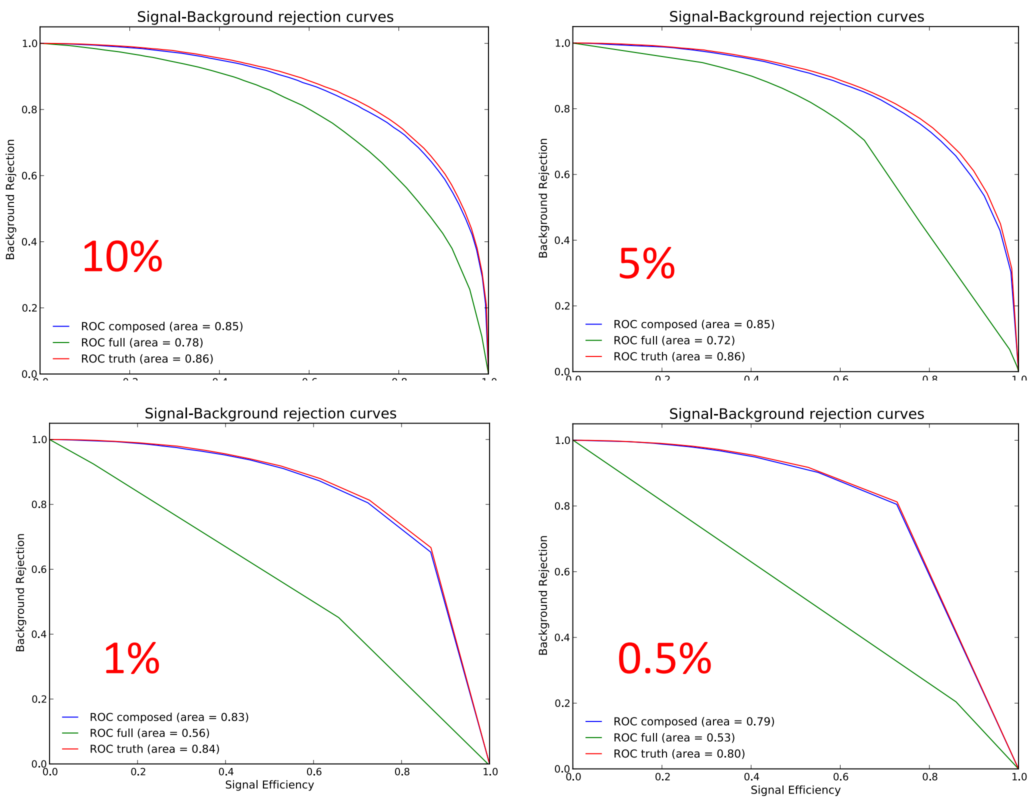


f1/f2

Applications

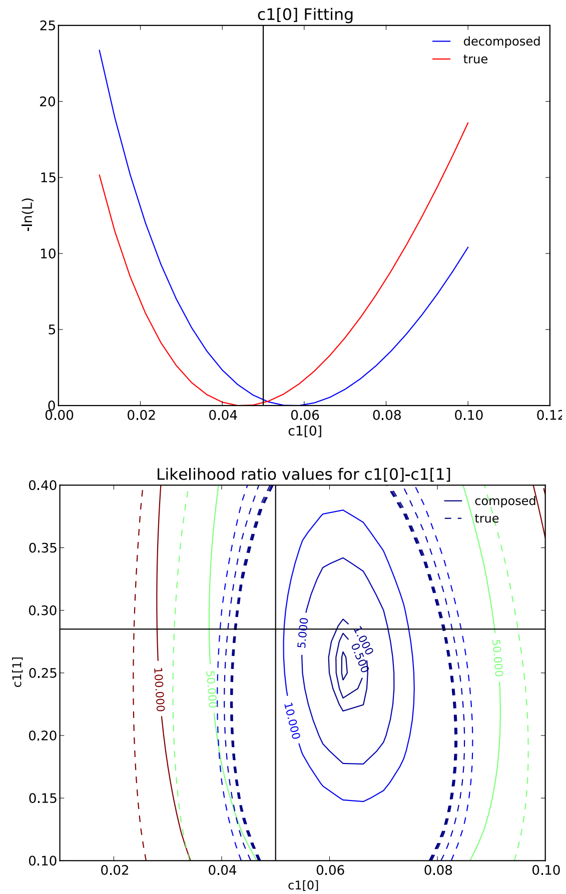
- Again we can vary the signal contribution and observe that the **decomposed model** has optimal results even for very small signal (0.5%).

AUC	Truth	Dec.	F0/F1
10%	0.86	0.85	0.78
5%	0.86	0.85	0.72
1%	0.84	0.83	0.56
0.5%	0.80	0.79	0.53

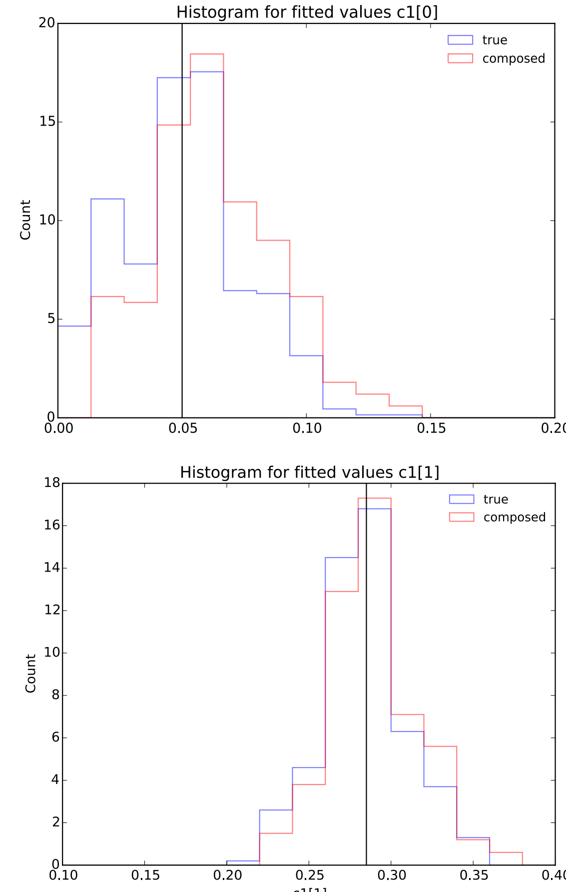


Applications

- It is possible to fit the signal contribution or background contributions (or both) values using **maximum likelihood** on the approximated likelihood ratios.



Fit for one single pseudo-experiment for:
A) Signal contribution.
B) Signal and Bkg.

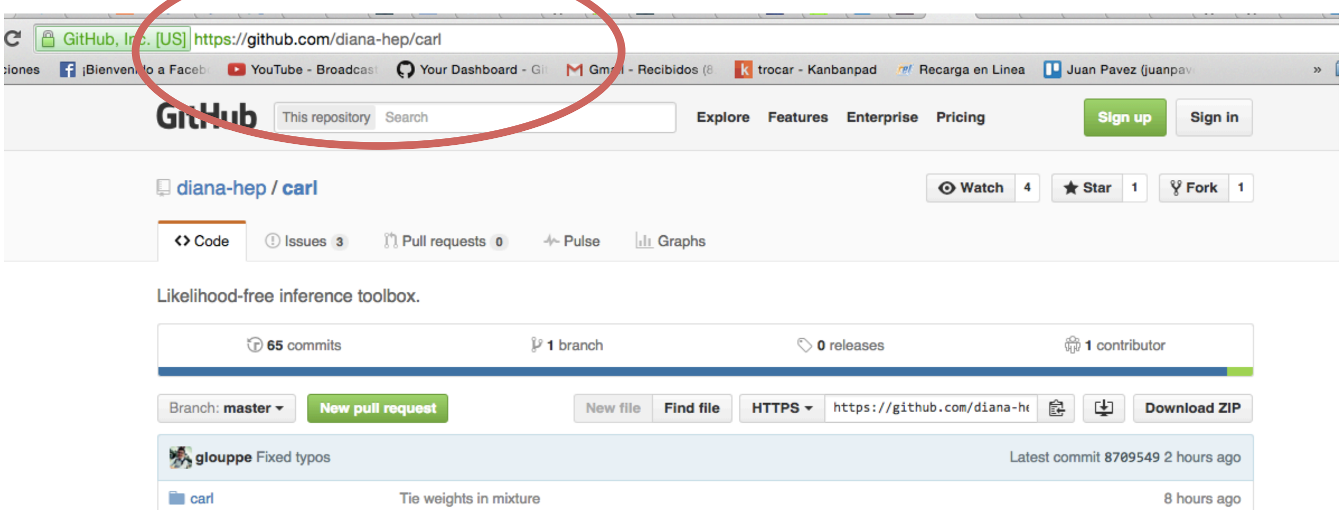


Histogram of fit of many pseudo-experiments for:
A) Signal contribution.
B) Bkg contribution.

Applications

- An open source **Python** package (**Carl**) has been implemented by **Gilles Louppe** allowing to easily use approximated likelihood ratio inference.

github.com/diana-hep/carl



Higgs EFT Morphing

Estimating coupling parameters for Higgs EFT Morphing.

Higgs EFT Morphing

- It is possible to expand the **Standard Model Lagrangian** adding **non-SM** couplings of the Higgs boson to SM particles using an ***effective field theory (EFT)*** approach.

- This allows to search for deviation from **SM** predictions for Higgs boson properties.

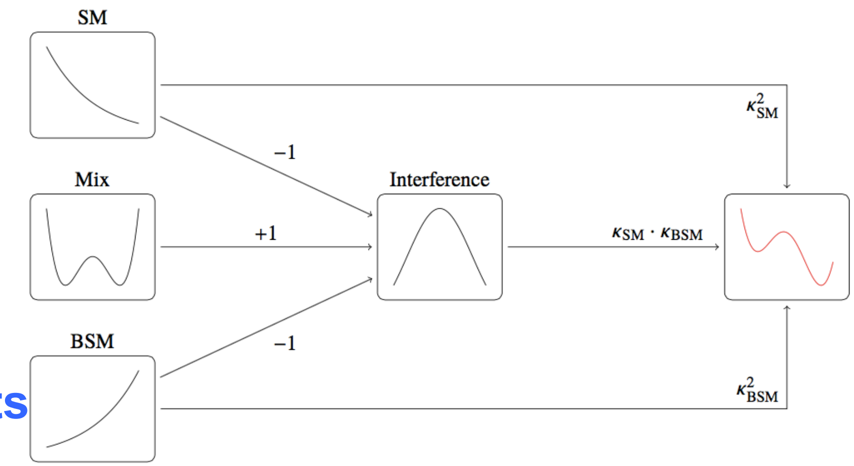
- While in **Run 1** a few number of coupling parameters was considered is expected than in **Run 2** many more **BSM** parameters will be considered.

$$
\begin{aligned}
\mathcal{L}_0^V = \Big\{ & c_\alpha \kappa_{SM} \left[\frac{1}{2} \tilde{g}_{HZZ} Z_\mu Z^\mu + \tilde{g}_{HWW} W_\mu^+ W^{-\mu} \right] \\
& - \frac{1}{4} \left[c_\alpha \kappa_{H\gamma\gamma} \tilde{g}_{H\gamma\gamma} A_{\mu\nu} A^{\mu\nu} + s_\alpha \kappa_{A\gamma\gamma} \tilde{g}_{A\gamma\gamma} A_{\mu\nu} \tilde{A}^{\mu\nu} \right] \\
& - \frac{1}{2} \left[c_\alpha \kappa_{HZ\gamma} \tilde{g}_{HZ\gamma} Z_{\mu\nu} A^{\mu\nu} + s_\alpha \kappa_{AZ\gamma} \tilde{g}_{AZ\gamma} Z_{\mu\nu} \tilde{A}^{\mu\nu} \right] \\
& - \frac{1}{4} \left[c_\alpha \kappa_{Hgg} \tilde{g}_{Hgg} G_{\mu\nu}^a G^{a,\mu\nu} + s_\alpha \kappa_{Agg} \tilde{g}_{Agg} G_{\mu\nu}^a \tilde{G}^{a,\mu\nu} \right] \\
& - \frac{1}{4} \frac{1}{\Lambda} \left[c_\alpha \kappa_{HZZ} Z_{\mu\nu} Z^{\mu\nu} + s_\alpha \kappa_{AZZ} Z_{\mu\nu} \tilde{Z}^{\mu\nu} \right] \\
& - \frac{1}{2} \frac{1}{\Lambda} \left[c_\alpha \kappa_{HWW} W_{\mu\nu}^+ W^{-\mu\nu} + s_\alpha \kappa_{AWW} W_{\mu\nu}^+ \tilde{W}^{-\mu\nu} \right] \\
& - \frac{1}{\Lambda} c_\alpha \left[\kappa_{H\partial\gamma} Z_\nu \partial_\mu A^{\mu\nu} + \kappa_{H\partial Z} Z_\nu \partial_\mu Z^{\mu\nu} + \kappa_{H\partial W} \left(W_\nu^+ \partial_\mu W^{-\mu\nu} + h.c. \right) \right] \Big\} X_0.
\end{aligned}
$$

Effective Lagrangian of spin-0 particle to gauge bosons.
A framework for Higgs characterisation. http://arxiv.org/abs/1306.6464

Higgs EFT Morphing

- **Problem**: Distributions of observables is dependant on the parameters of the **EFT**.
- **Morphing Method:** Provides description of any observable as a linear combination of a minimal set of orthogonal base samples.



Coupling constants

$$\boxed{S(g_1, g_2, \ldots) = \sum_{i=1}^{N} w^{(i)}(g1, g2, \ldots) S(g_1^{(i)}, g_2^{(i)}, \ldots)}$$

Sample of Interest   Weights   Base samples

Simple morphing procedure for one BSM coupling in decay or production.
A Morphing Technique for signal modelling in a Multidimensional space of non-SM coupling parameters.

cds.cern.ch/record/2066980

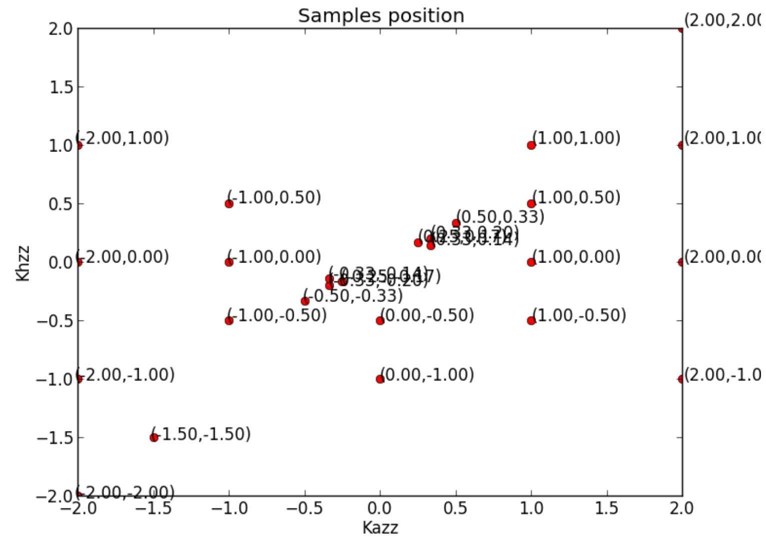
Higgs EFT Morphing

- The team working on **Morphing** has provide us samples for **VBF** production $H \to WW^* \to e\nu\mu\nu$ and $H \to WW^* \to 4l$.
- **15 samples and 5 samples** are needed in the base.
- **Problem:** In areas of the **coupling space** not covered by the base of samples statistical fluctuations increase a lot.

  ➔ This affect the **fitting** procedure
- **Solution:** Choose a pair of **orthogonal** sets of bases and sum as:

$$\boxed{B_{full} = \alpha_1(g_1, g_2, g_3)B_1 + \alpha_2(g_1, g_2, g_3)B_2}$$

- Good coverture of the coupling space while allowing a **smooth** change of bases when fitting.



Available samples on coupling space for 2BSM VBF.

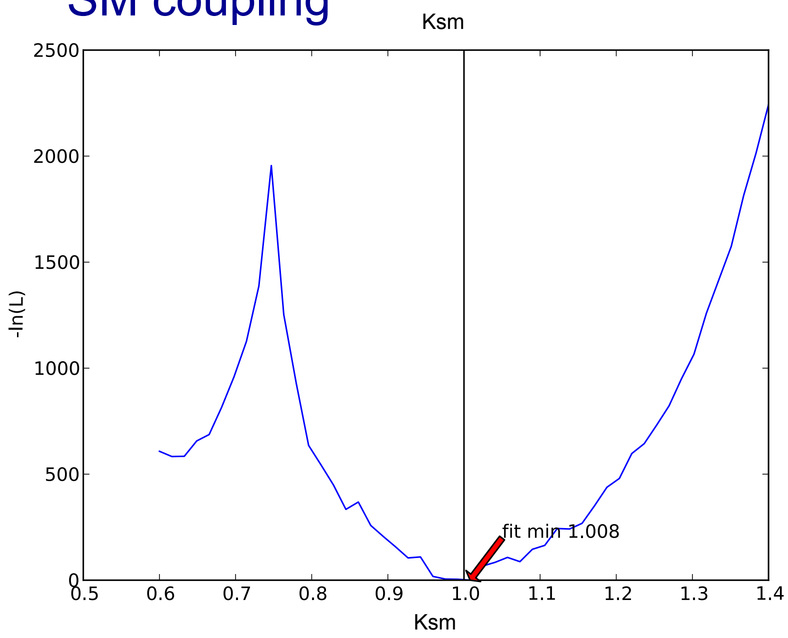
Higgs EFT Morphing **Preliminary Results**

$H \rightarrow WW^* \rightarrow 4l$

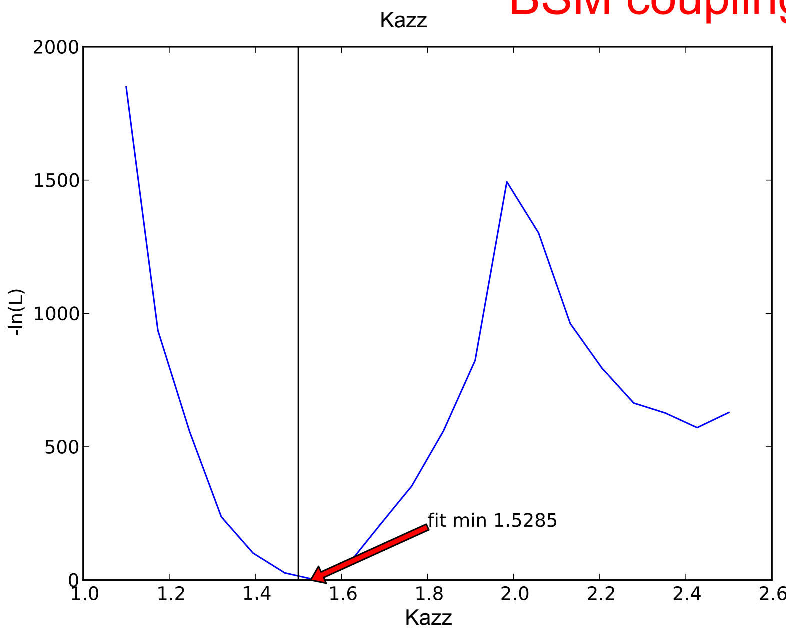
1 BSM

- Likelihood fit for coupling parameters using decomposed approximated likelihood ratios.
- Fitting sample S(1.,1.5). Good agreement with real values.

SM coupling



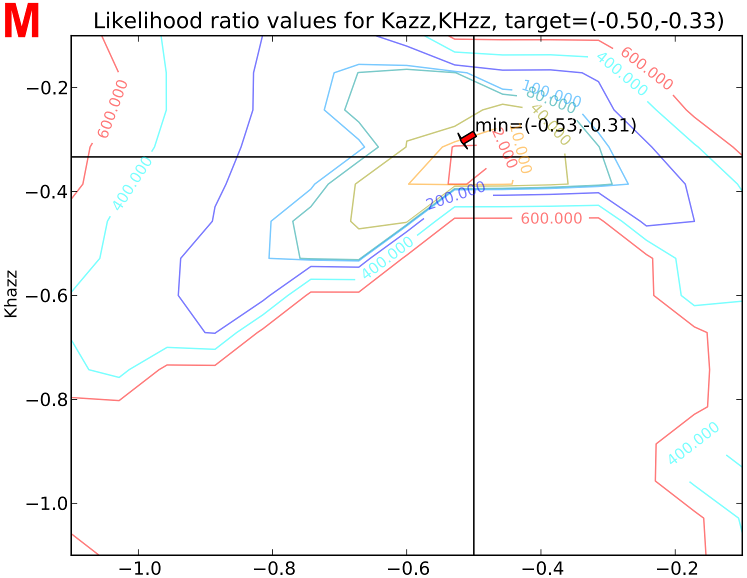
BSM coupling

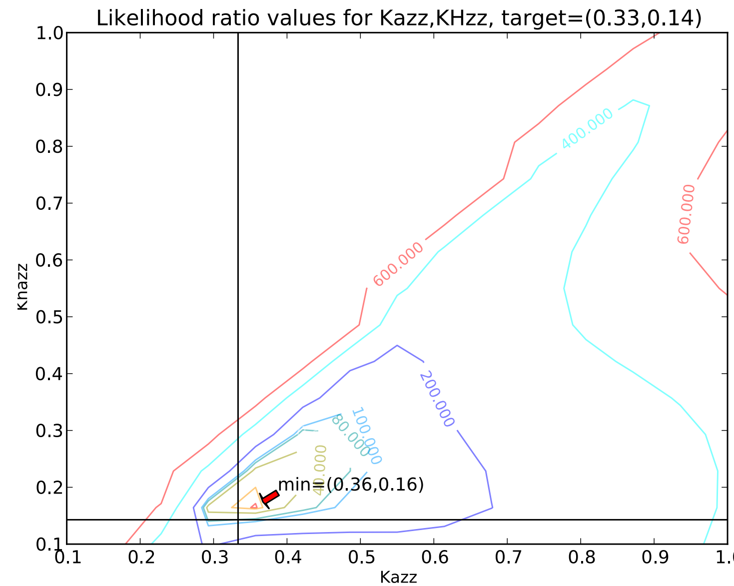


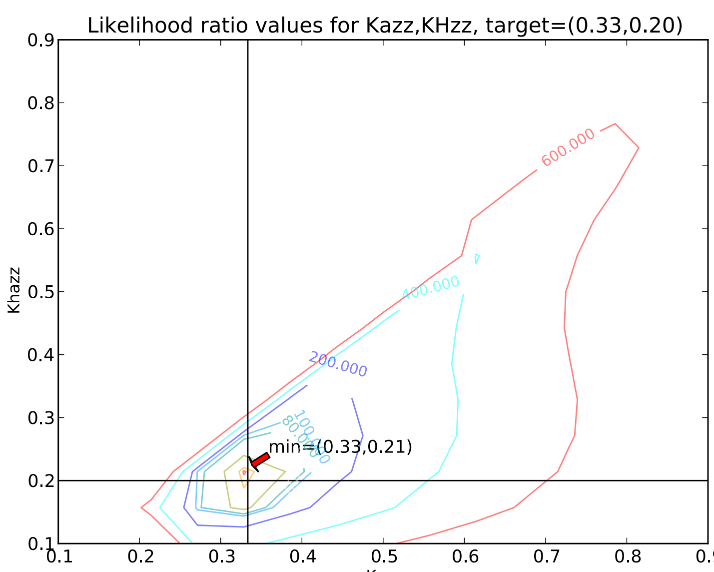
Higgs EFT Morphing **Preliminary Results**

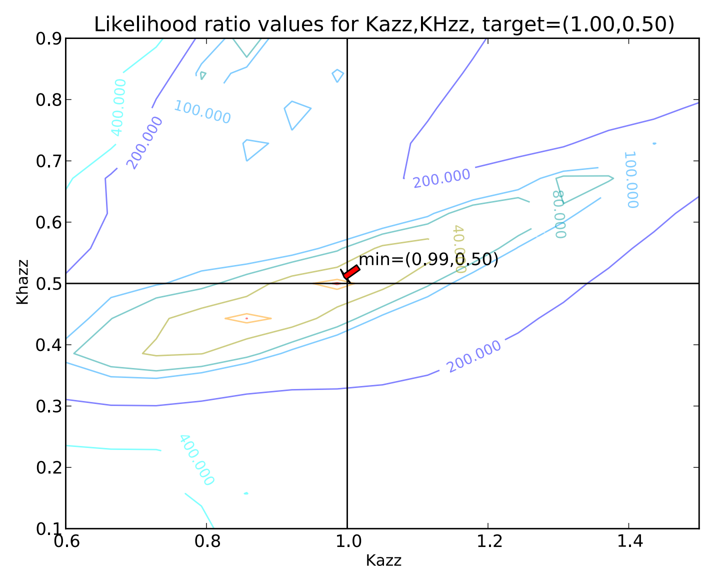
$H \to WW^* \to e\nu\mu\nu$

2 BSM









Conclusions

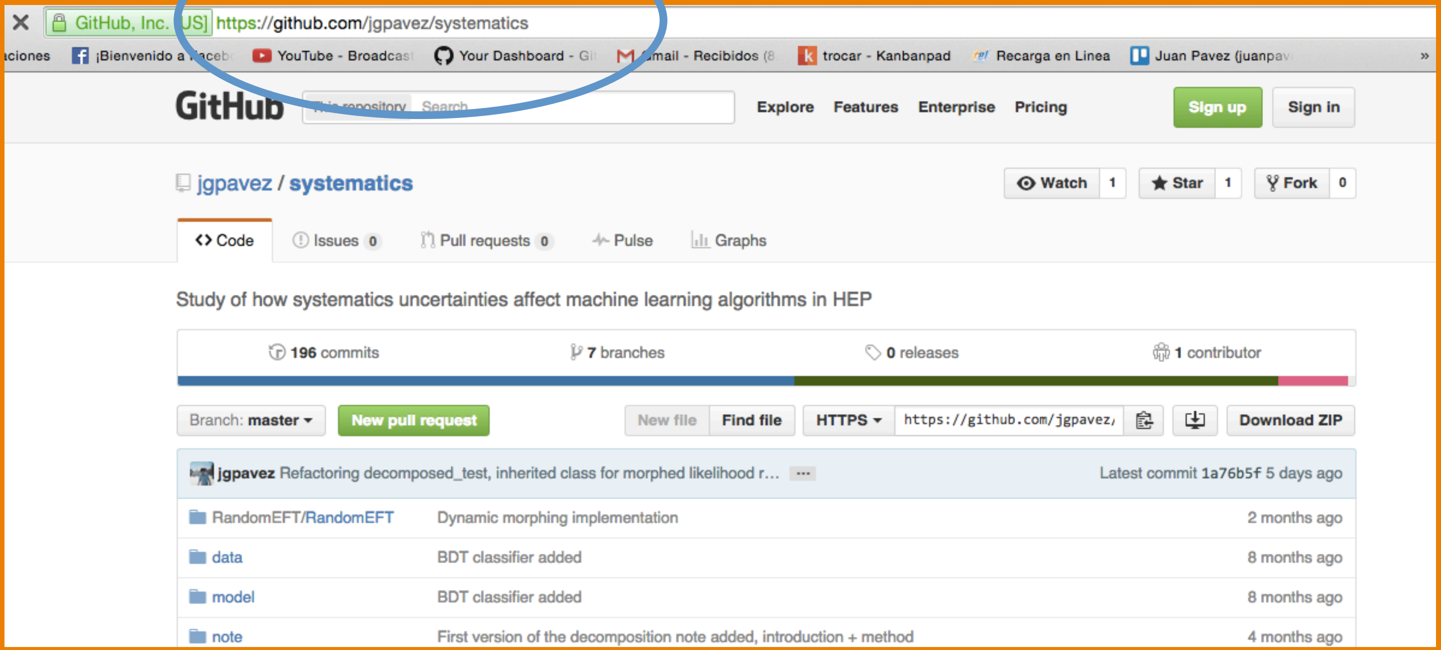
Conclusions & Future Work

- Machine Learning approximation of Likelihood ratios is a great alternative for Likelihood estimation methods, such as **Approximate Bayesian Computation (ABC)**.
  - This method allow to use all last advances on Machine Learning (e.g. Deep Learning, Tree based methods, SVMs).
- The technique is also an alternative to **MEM (Matrix Element Method)**, but faster since this approach is not event-based (but need many simulated samples).

- We need to understand how **errors** (e.g. training error, poison fluctuations from histogram estimation) affects the **final approximation**.
- Integration with common tools used in **HEP** might be needed (we have been working mainly with python frameworks).

Thank You!

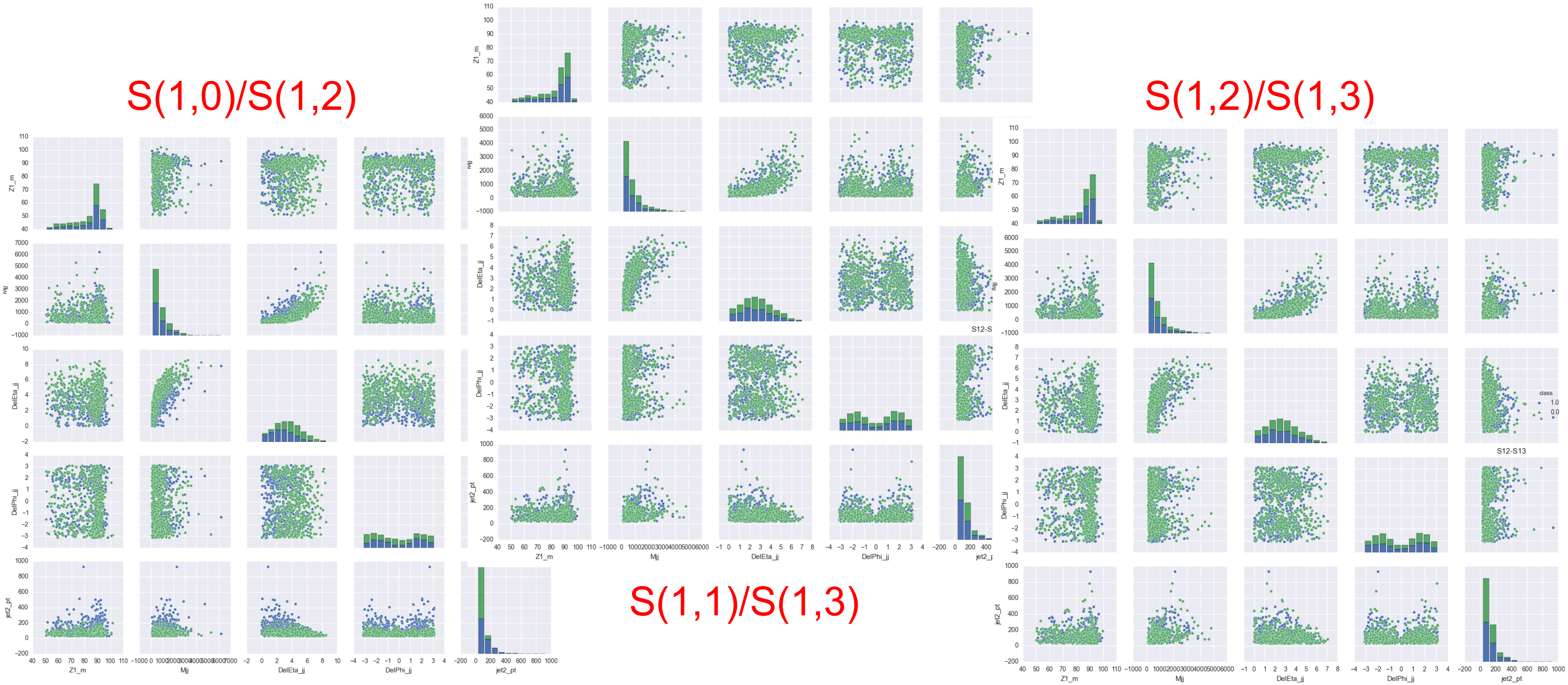
All the code and plots on this presentation can be found at:
github.com/jgpavez/systematics



Backup

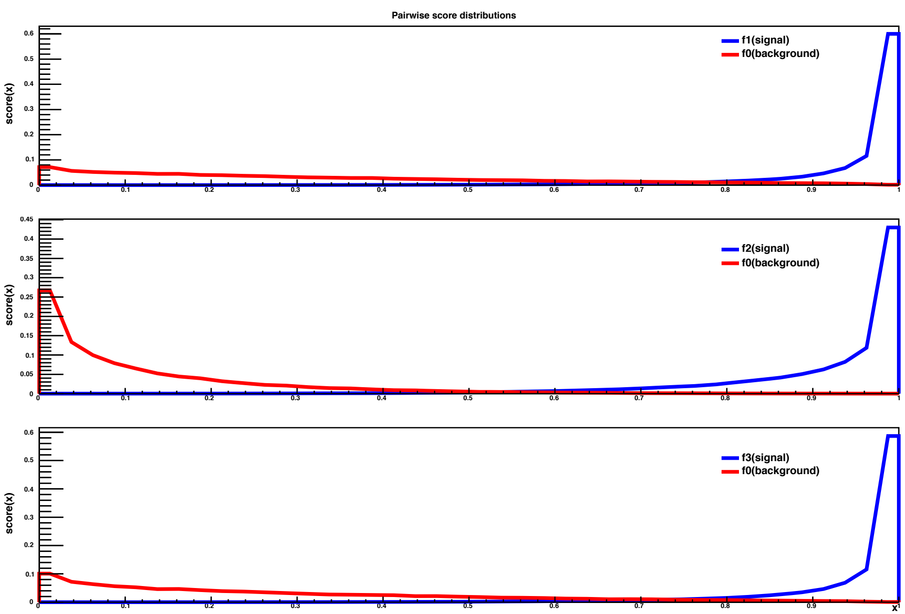
Backup

- Datasets for VBF $H \rightarrow WW^* \rightarrow 4l$ with one BSM coupling.

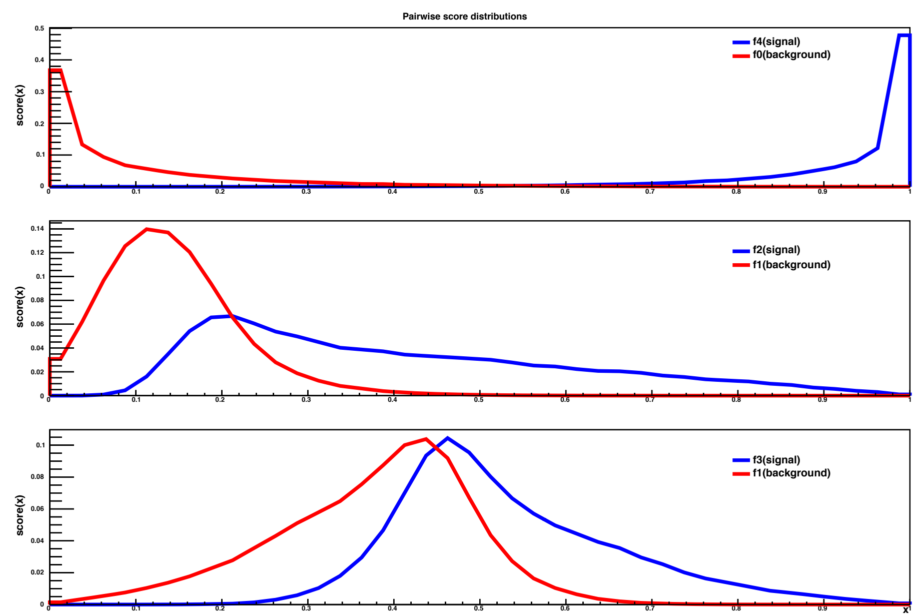


Backup

- Score distributions for pairwise trained classifiers (BDTs) for **VBF** $H \rightarrow WW^* \rightarrow 4l$ datasets.



S(1,0)/S(1,2),S(1,0)/S(1,1),S(1,0)/S(1,3)

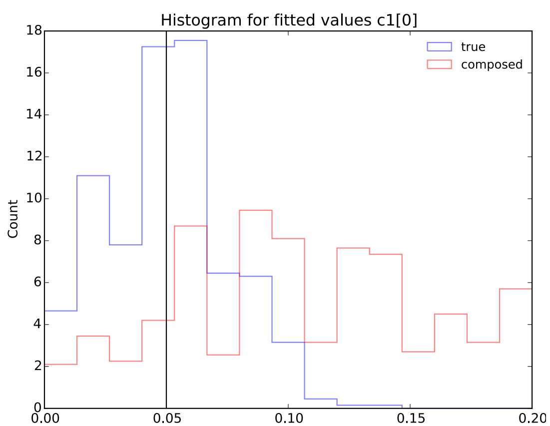


S(1,0)-S(0,1),S(1,2)-S(1,1),S(1,2)-S(1,3)

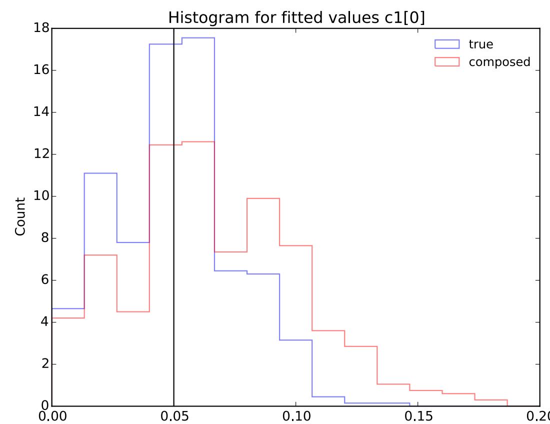
Backup

- Studies on how the quality of the classifier training affect the final approximation.
- Expected : **More training data -> Better Approximation** (Better fits).
- Results for 10-dim harder model.

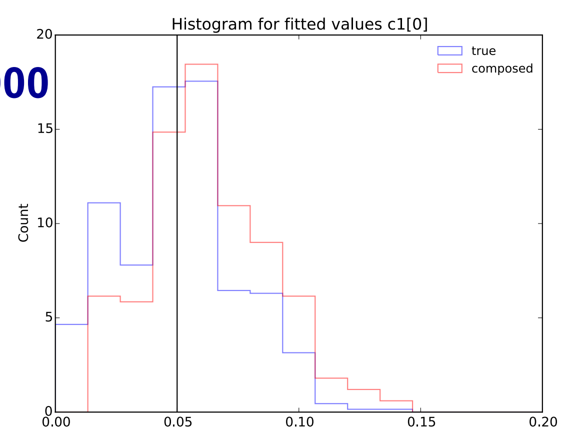
Training data: 100



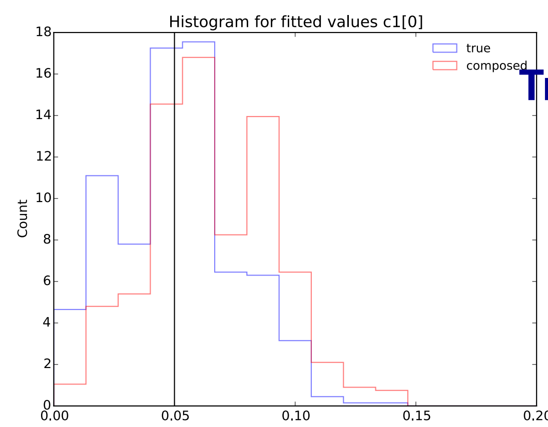
Training data: 1000



Training data: 10000



Training data: 100000



Backup

Loss Function	$L[y, f(x)]$	Minimizing Function
Binomial Deviance	$\log[1 + e^{-yf(x)}]$	$f(x) = \log \frac{\Pr(Y = +1\|x)}{\Pr(Y = -1\|x)}$
SVM Hinge Loss	$[1 - yf(x)]_+$	$f(x) = \text{sign}[\Pr(Y = +1\|x) - \frac{1}{2}]$
Squared Error	$[y - f(x)]^2 = [1 - yf(x)]^2$	$f(x) = 2\Pr(Y = +1\|x) - 1$
"Huberised" Square Hinge Loss	$-4yf(x)$, $\quad yf(x) < -1$ $[1 - yf(x)]_+^2$ $\quad$ otherwise	$f(x) = 2\Pr(Y = +1\|x) - 1$

Classifiers minimizing functions. The Elements of Statistical Learning, Hastie et al.

Bibliography

- Cranmer, K. (2015). Approximating Likelihood Ratios with Calibrated Discriminative Classifiers. arXiv preprint arXiv:1506.02169.

- ATLAS Collaboration (Belyaev, K. et al.). A morphing technique for signal modelling in a multidimensional space of non-SM coupling parameters. cds.cern.ch/record/2066980.

- Artoisenet, P., De Aquino, P., Demartin, F., Frederix, R., Frixione, S., Maltoni, F., ... & Seth, S. (2013). A framework for Higgs characterisation. *Journal of High Energy Physics*, *2013*(11), 1-38.