DR ADRIAN BEVAN

MULTIVARIATE ANALYSIS AND ITS USE IN HIGH ENERGY PHYSICS

1) CUT BASED ANALYSES AND LINEAR DISCRIMINANTS

Lectures given at the department of Physics at CINVESTAV, Instituto Politécnico Nacional, Mexico City
28th August - 3rd Sept 2018

LECTURE PLAN

- These lectures can be found online at:
  - https://pprc.qmul.ac.uk/~bevan/teaching/MLinHEP.html
- My teaching notes (linked from the above) include other courses on machine learning that may be of interest.

LECTURE PLAN

- Introduction
- Cut based analyses
  - Figures of merit for optimisation
- Linear discriminants
  - Fisher's linear discriminant
- Examples
- Summary
- Suggested reading

INTRODUCTION

- In High Energy Physics we analyse data for many different reasons:
  - Event selection at the trigger level;
  - Calibration of systems;
  - Particle identification;
  - Object reconstruction;
  - Offline event selection for analysis (cut based analysis);
  - Background suppression (candidate selection);
    - As a part of a cut-based selection;
    - As an input to a fit; either as one discriminating variable, or as the single discriminating variable.
- Multivariate analysis (MVA) is a tool that helps us in all of these areas.

INTRODUCTION

- These lectures cover a number of different topics:
  - Cut based analysis and linear discriminants
  - Decision Trees
  - Neural Networks
  - Optimisation
  - Deep Learning
  - Support Vector Machines
- All of the topics are used in the field; the most common techniques are cut based analysis and the use of decision trees; but modern deep learning methods are gaining popularity.
- The concepts that underpin optimisation are also relevant for fit based approaches; although some of the methods used are tailored to machine learning (ML) methods.

INTRODUCTION

- These lectures will not cover coding examples, although if you are interested in coding examples there are several tools we can discuss outside of the lectures.

- The TMVA toolkit in ROOT has good documentation and would be a natural place to start working with ML in HEP as it has an interface to data via ROOT objects.

- The following set of lecture notes includes examples in TensorFlow, and includes a HEP analysis problem (the Kaggle Higgs challenge data).
  - https://pprc.qmul.ac.uk/~bevan/teaching/PML.html

CUT BASED ANALYSES

- A typical analysis workflow is schematically represented below:

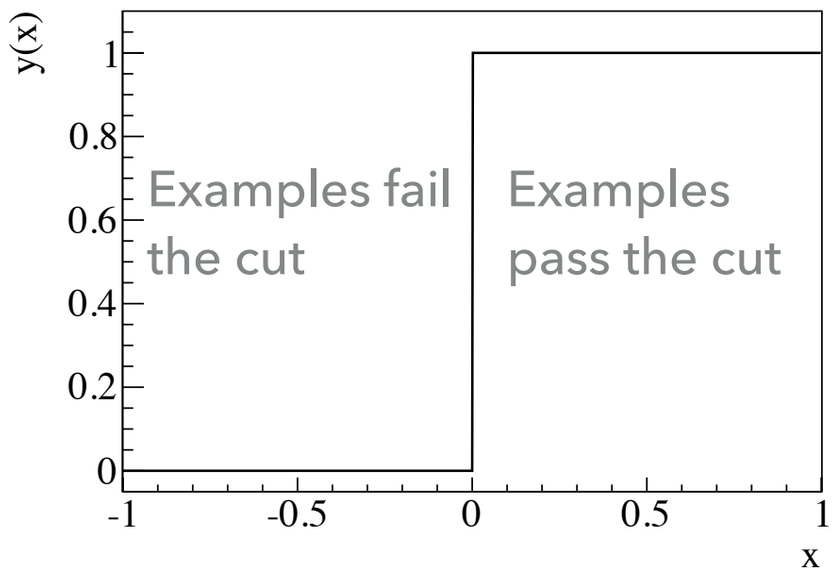
- Cut-based selection (sometimes called a cut and count analysis) is the method of applying a set of cuts (usually rectangular) to the data in order to determine how many events of different species remain.
  - The number of events left after applying cuts is then converted into a more interesting quantity such as a branching fraction, limit, or cross section.
  - How do we determine the cuts?

CUT BASED ANALYSES: FIGURES OF MERIT FOR OPTIMISATION

- Formally when we apply some cut we are imposing a threshold defined by some condition.
  - If the condition is satisfied then we retain the example for subsequent use.
  - If the condition is not satisfied the example is discarded.

- The condition can be thought of as some decision boundary:
  - this description will play an important role when we discuss ML algorithms.

- This is equivalent to applying a heavyside step function $H(x) = \frac{1}{2}(1 + sign(x))$.



We can map *x* to *x–b* if we want to place a cut at *x=b* in order to apply a non-trivial cut on the data.

We can invert the cut by mapping *x* to *–x* in the LHS of *H(x)*.

CUT BASED ANALYSES: FIGURES OF MERIT FOR OPTIMISATION

- Cut values (cuts) are either determined in an ad-hoc way driven by external constraints such as minimising resource:
  - This can done when as a pre-selection step when it is "obvious" that we are throwing away more background than signal; and a final selection optimisation will then be made.
  - If the preselection is imposed by some other resource driven constraint such as reducing the number of data to analyse in a fit.
- Or the cuts are determined by maximising some figure of merit (FOM) motivated by the problem being addressed.

CUT BASED ANALYSES: FIGURES OF MERIT FOR OPTIMISATION

- Popular FOMs include:

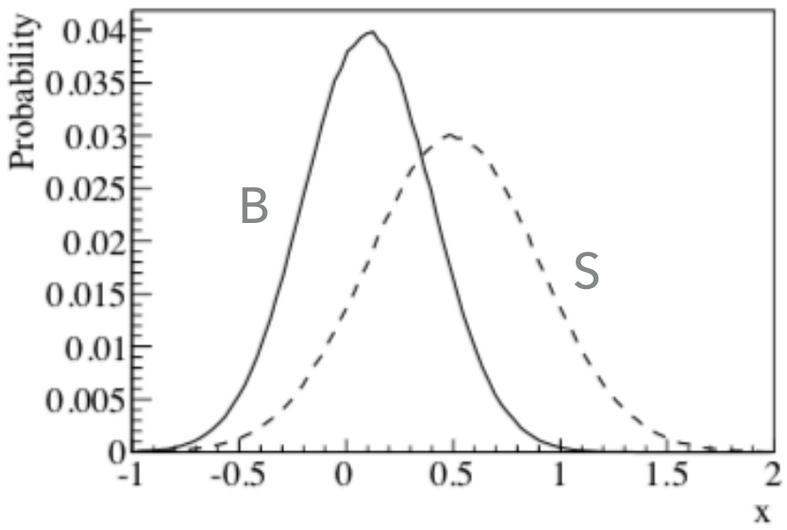
S/B Signal to noise: For example this is a standard FOM used when testing detectors in order to control false positive rates. If a S/N of 10 is achieved for signal pulses coming out of a silicon detector in the laboratory then the chance of using spurious noise in track reconstruction is small.

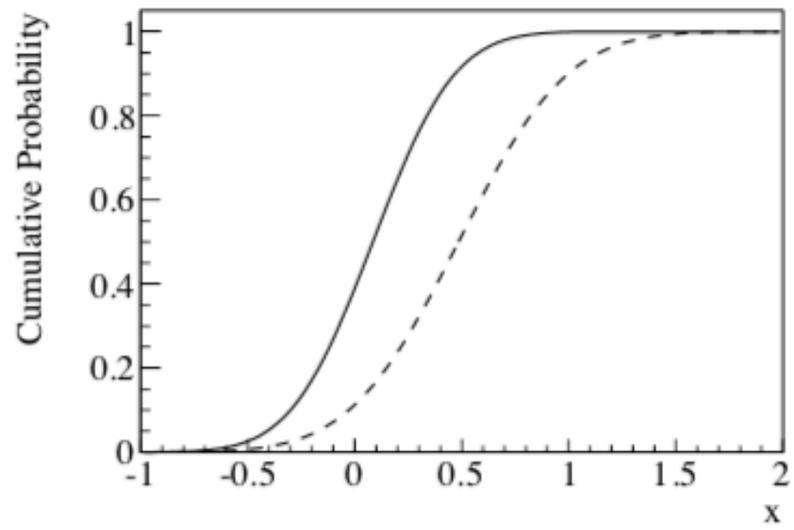
$S/\sqrt{S+B}$

Signal significance: This is one form of the signal significance metric. If the signal and background are both large then the denominator is equivalent to the uncertainty on the count rate for S+B. This FOM optimises the ratio of S to this uncertainty estimate.

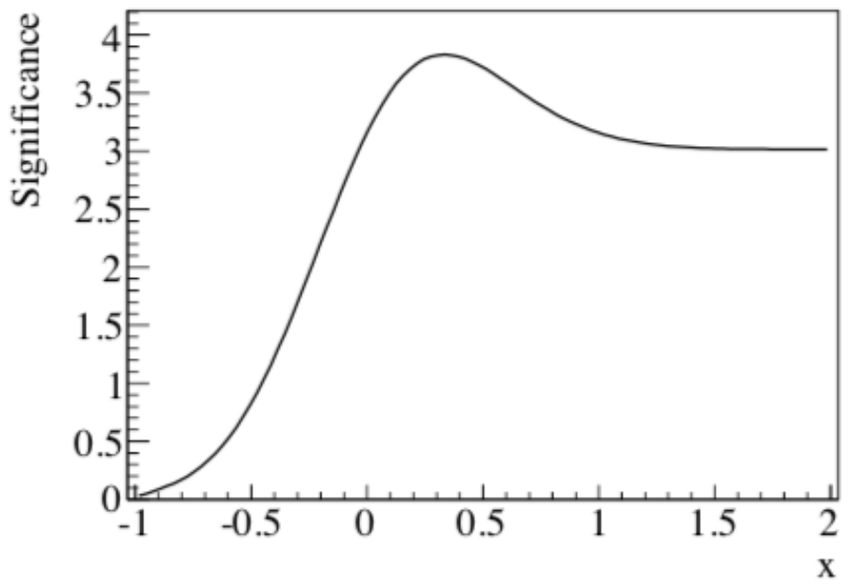
CUT BASED ANALYSES: FIGURES OF MERIT FOR OPTIMISATION

- Consider the following toy example: two types of event described by Gaussian distributions:





- Maximise $S/\sqrt{S+B}$ to find an "optimal" cut.



*While it is not often done, this same logic can be used when applying ML to our analysis. The problem with that approach is that it often requires a significant amount of computing resource to implement such an approach.

CUT BASED ANALYSES: FIGURES OF MERIT FOR OPTIMISATION

- If we have the ability to do so it is better to optimise the MVA on a figure of merit related to a measurement we are making*. For example:
  - Expected measurement precision
  - Expected limit for a search
  - Maximise purity (here we define purity as S/(S+B), and so we would maximise that as the FOM.

*While it is not often done, this same logic can be used when applying ML to our analysis. The problem with that approach is that it often requires a significant amount of computing resource to implement such an approach.

CUT BASED ANALYSES: FIGURES OF MERIT FOR OPTIMISATION

- e.g. Consider a counting experiment that is a search for a signal where there is a large known background, b.
- We can use the discovery significance $Z_0=\sqrt{q_0}$ as the figure of merit to optimise on: $q_0 = -2\ln(\mathscr{L}(s=0)/\mathscr{L})$
- Where the likelihood is Poisson: $\mathscr{L} = \frac{(s+b)^n e^{(-s+b)}}{n!}$
- From this we can compute

$$q_0 = -2\ln\left(\frac{b^n e^{-b}}{(s+b)^n e^{-(s+b)}}\right)$$

$$Z_0 = \sqrt{q_0} = \sqrt{2(n\ln(n/b)+b-n)}$$

*Here we invoke Wilks' theorem to relate the likelihood ratio to a X^2 distribution in order to be able to relate the significance to the likelihood ratio given by the null hypothesis of no signal as an outcome. e.g. see Cowan et al., Eur.Phys.J.C71:1554,2011.

CUT BASED ANALYSES: FIGURES OF MERIT FOR OPTIMISATION

- Most cut based analyses consider cuts independently; so optimisation has to be done either:
  - Simultaneously for all cuts that are being made on an analysis. This approach results in M^N iterations for the optimisation, where N is the number of cuts being made and M is the number of samples tested on each cut.
    - This approach suffers from the curse of dimensionality.
  - One at a time iteratively until such time as cut values do not change significantly [a pragmatic solution to overcome the M^N scaling].
    - Scales like some (usually small) number times MxN.
- How can we extract more information from our data?

LINEAR DISCRIMINANTS

- We can construct combinations of input features (dimensions that we want to cut on) to see if we are able to utilise the way that examples (events or candidates) cluster in the hyperspace that is formed by all of the features.
  - Our ability to separate examples in this feature space will generally be better than considering lower dimensional projections of the problem and recursively iterating through until we have considered the whole space.
- The simplest approach is to construct a linear combination of features.
- In HEP the Fisher discriminant has been widely used on a number of experiments.

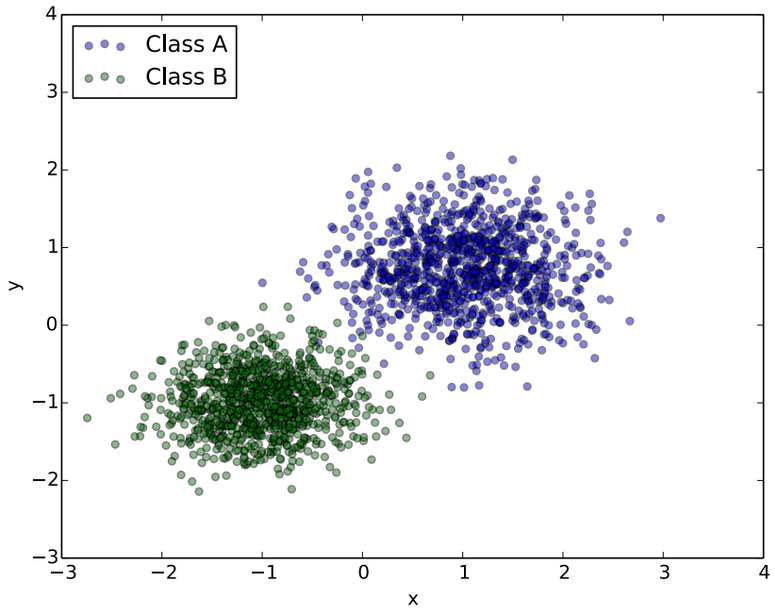
LINEAR DISCRIMINANTS: FISHER'S LINEAR DISCRIMINANT

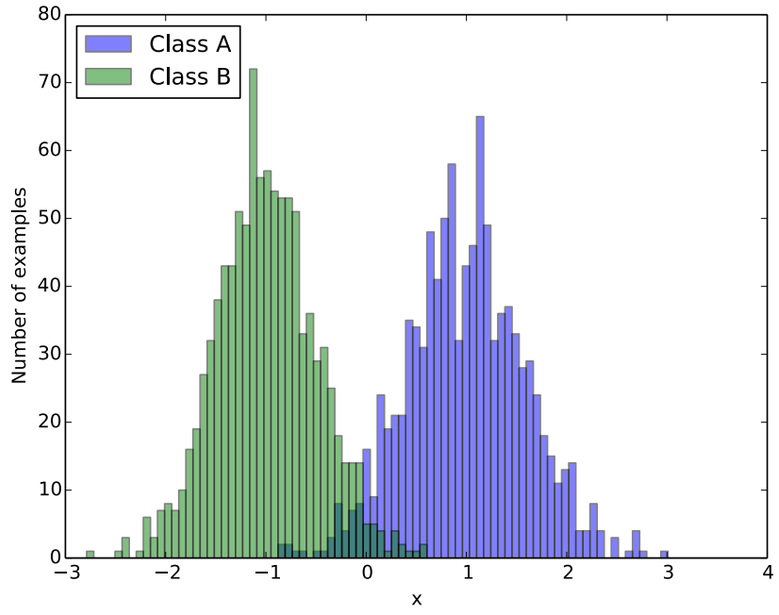
- This is an analytic algorithm that was inspired by the classification problem for species of iris in the 1930's[1].
- Starting point is the assumption that data are distributed according to a multi-Gaussian probability (e.g. random sampling of data), and that one wishes to maximise the separation between different classes (types) of iris.
- Maximise the separation of the mean distributions.
- Minimise the sum of the covariances.

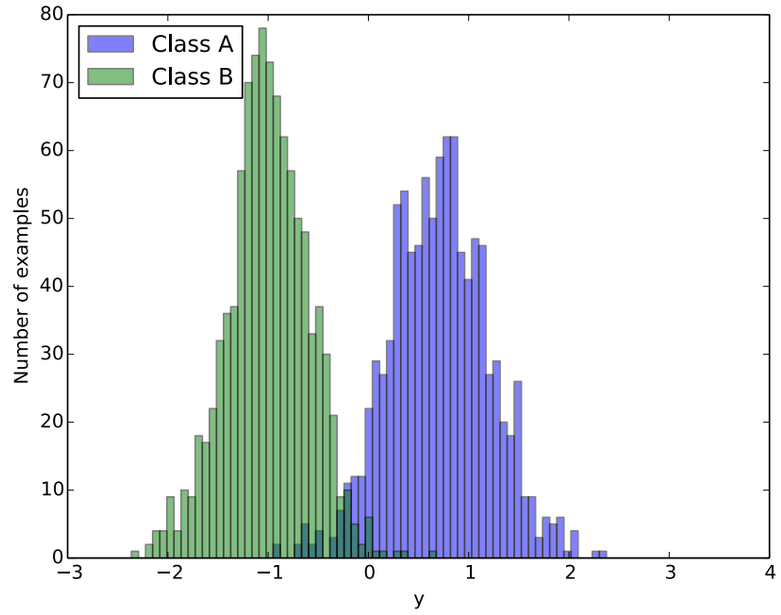
[1] R. A. Fisher, Ann. Eug., 7, 179188 (1936).

LINEAR DISCRIMINANTS: FISHER'S LINEAR DISCRIMINANT

- Consider the problem where we have two classes of event. Some events of type A (signal) and some events of type B (background).
- These events are described by a 2D feature space, consisting of the dimensions x and y.
- We want to compute $\mathcal{F}$ in order for us to be able to distinguish between types A and B.







[1] R. A. Fisher, Ann. Eug., 7, 179188 (1936).

LINEAR DISCRIMINANTS: FISHER'S LINEAR DISCRIMINANT

- The Fisher discriminant is given by

$$F = \alpha^T x + \beta$$

$$\alpha = W^{-1}(\mu_A - \mu_B)$$

  - α: a vector of weight parameters
  - x: data with the dimension of the input feature space
  - W: sum of covariance matrices for classes A and B
  - $\mu_{A,B}$: mean value of class A or B

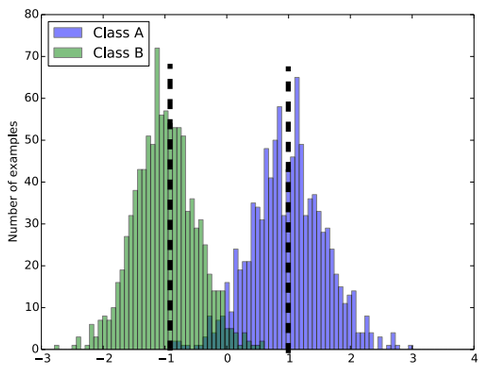
NOTE: I like this algorithm as an example as it can help us understand the data and how to separate classes before looking in more detail at a neural network as the underlying equations appear later in the course.

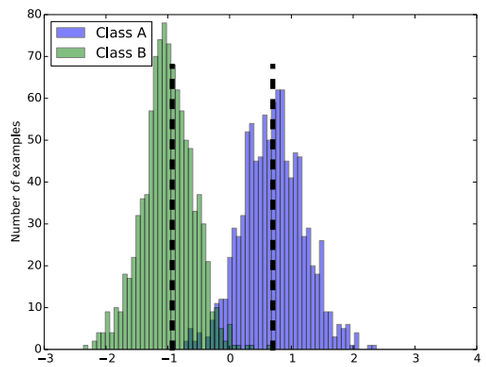
It is also a simple algorithm that can be used as a benchmark to check that a more sophisticated algorithm performs at least as well as this one. Such a sanity check can be useful in low dimensional physics problems and it may or may not be useful for you machine learning work in the future.

[1] R. A. Fisher, Ann. Eug., 7, 179188 (1936).

LINEAR DISCRIMINANTS: FISHER'S LINEAR DISCRIMINANT

- Consider this 2D problem:





$\mu_{A,B}$ are the mean values of A and B, respectively, given by

$$\mu_{A,B;u} = \frac{1}{N}\sum_{i=1}^{N} u_i, \quad u = x, y$$

i.e. $(\mu_{B,y})^T=(\mu_{B,x}, \mu_{B,y})$ and $(\mu_{A,y})^T=(\mu_{A,x}, \mu_{A,y})$.

- The means (μ) and standard deviations (σ) describe the distribution of data; where $\sigma_{A,B}$ are 2D covariance matrices.

- We can compute the mean (M) and variance (Σ) of the Fisher distribution using

$$M_{A,B} = \alpha^T \mu_{A,B} = \sum_i \alpha_i \mu_{A,B},$$

$$\Sigma^2_{A,B} = \alpha^T \sigma^2_{A,B} \alpha = \sum_i \sum_j \alpha_i \sigma_{ij\ A,B} \alpha_j$$

[1] R. A. Fisher, Ann. Eug., 7, 179188 (1936).

LINEAR DISCRIMINANTS: FISHER'S LINEAR DISCRIMINANT

- Optimise J, where

$$J(\alpha) = \frac{[M_A - M_B]^2}{\Sigma_A^2 + \Sigma_B^2}$$

$$\begin{aligned}
[M_A - M_B]^2 &= \left[\sum_{i=1}^{n} \alpha_i (\mu_A - \mu_B)_i\right]\left[\sum_{j=1}^{n} \alpha_j (\mu_A - \mu_B)_j\right] \\
&= \sum_{i,j=1}^{n} \alpha_i (\mu_A - \mu_B)_i (\mu_A - \mu_B)_j \alpha_j, \\
&= \alpha^T B \alpha,
\end{aligned}$$

$$\begin{aligned}
\Sigma_A^2 + \Sigma_B^2 &= \alpha^T \sigma_A^2 \alpha + \alpha^T \sigma_B^2 \alpha, \\
&= \alpha^T W \alpha,
\end{aligned}$$

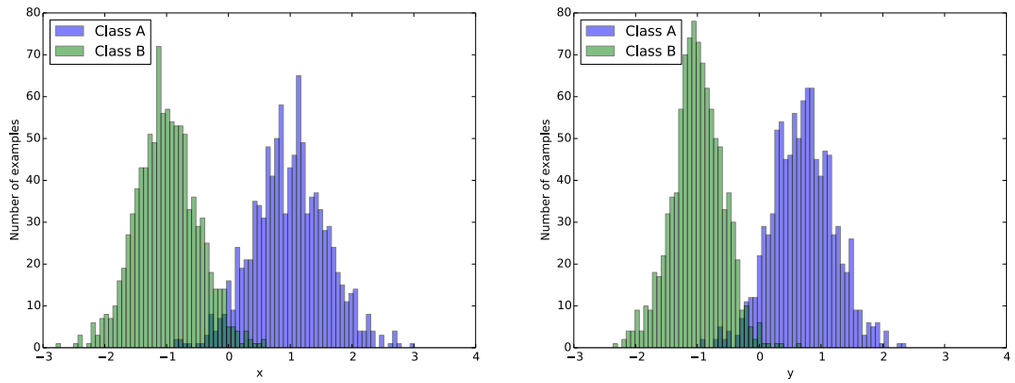
- Thus:

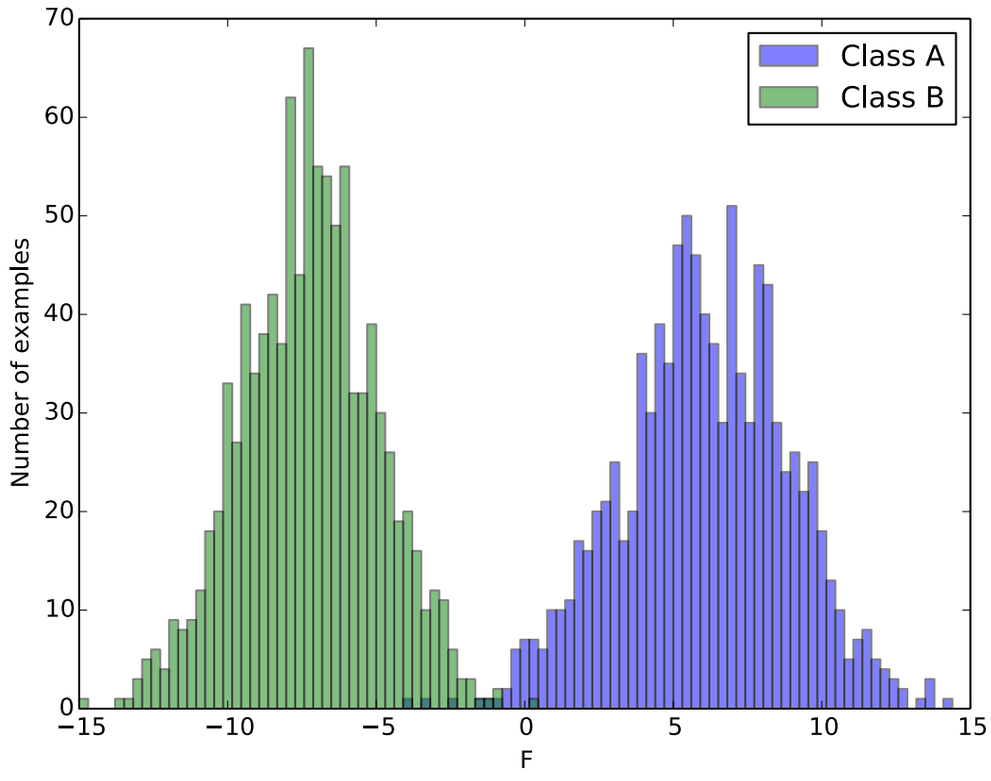
$$J(\alpha) = \frac{\alpha^T B \alpha}{\alpha^T W \alpha} \quad \text{which is optimised for} \quad \frac{\partial J(\alpha)}{\partial \alpha} = 0$$

- resulting in: $\alpha \propto W^{-1}(\underline{\mu}_A - \underline{\mu}_B)$.
  - The α are given up to an arbitrary scale factor.
  - The mean value of $\mathcal{F}$ can be offset by β arbitrarily.

[1] R. A. Fisher, Ann. Eug., 7, 179188 (1936).

LINEAR DISCRIMINANTS: FISHER'S LINEAR DISCRIMINANT





- Coming back to our example problem we can compute $\mathcal{F}$ using

$$F = \alpha^T x + \beta$$

- $\mathcal{F}$ is a linear combination of x and y (like a rotated axis in the (x, y) plane) along which we can project the data in order to obtain a smaller overlap than in either of the individual features (x or y).
- The offset β is arbitrary and is ignored here as it just changes the position of an example on the $\mathcal{F}$ axis without changing the separation between the classes.
- Separation in $\mathcal{F}$ is better than in either x or y.
- Separation in $\mathcal{F}$ is also better than that using rectangular cuts.

[1] R. A. Fisher, Ann. Eug., 7, 179188 (1936).

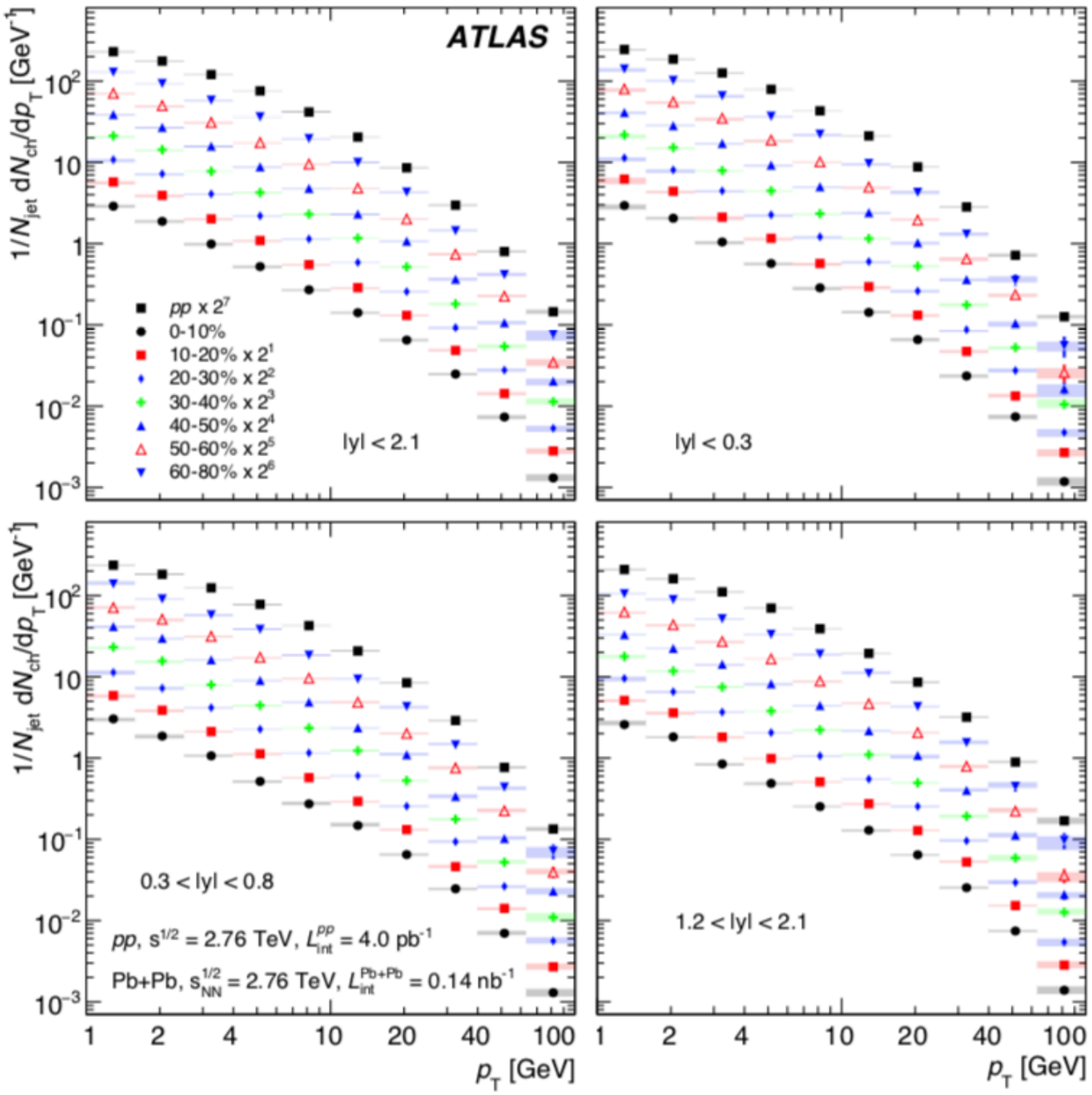
EXAMPLES: CUT AND COUNT

- Measurement of jet fragmentation in Pb+Pb and pp collisions at ATLAS (2.76 TeV).
  - Use triggers to isolate potentially interesting events.
  - Use event selection to identify jets; and from these compute the number of jets in a given p_T and rapidity* range.
- This allows the computation of cross sections.

*The rapidity $y = 0.5 \ln [(E + p_z) / (E - p_z)]$, where E and p_z are the energy and the component of the momentum along the beam direction, respectively.

ATLAS Collaboration, Eur. Phys. J. C **77** (2017) 379

EXAMPLES: CUT AND COUNT



- These plots are unfolded charged particle distributions of the pT.
- Error bars indicate statistical uncertainties and the shaded bands indicate systematic uncertainties.
- The paper has a number of other results that use the same cut-and-count approach in order to extract information of interest.

ATLAS Collaboration, Eur. Phys. J. C **77** (2017) 379

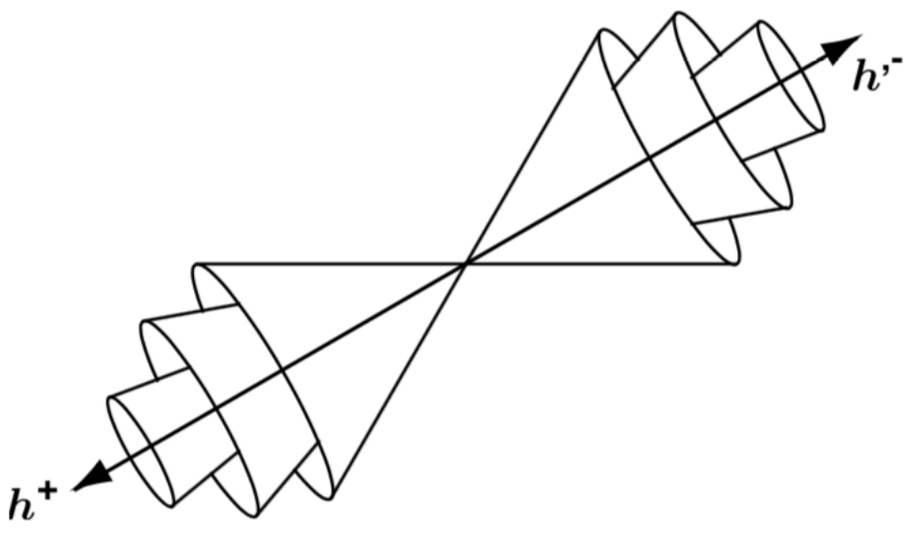
EXAMPLES: FISHER DISCRIMINANT

$$F = \alpha^T x + \beta$$

- Using a Fisher discriminant to suppress light quark background in order to study B mesons. Produced via

$$e^+e^- \to B\overline{B}$$

- Light quarks are fast in the CM frame, whereas the B mesons are slow and decay isotropically.



- BaBar used a fisher discriminant based on work by the CLEO collaboration. This divided the data up into 9 concentric cones to parameterise the energy flow for BB vs qq-bar (q=u, d, s, c) events.

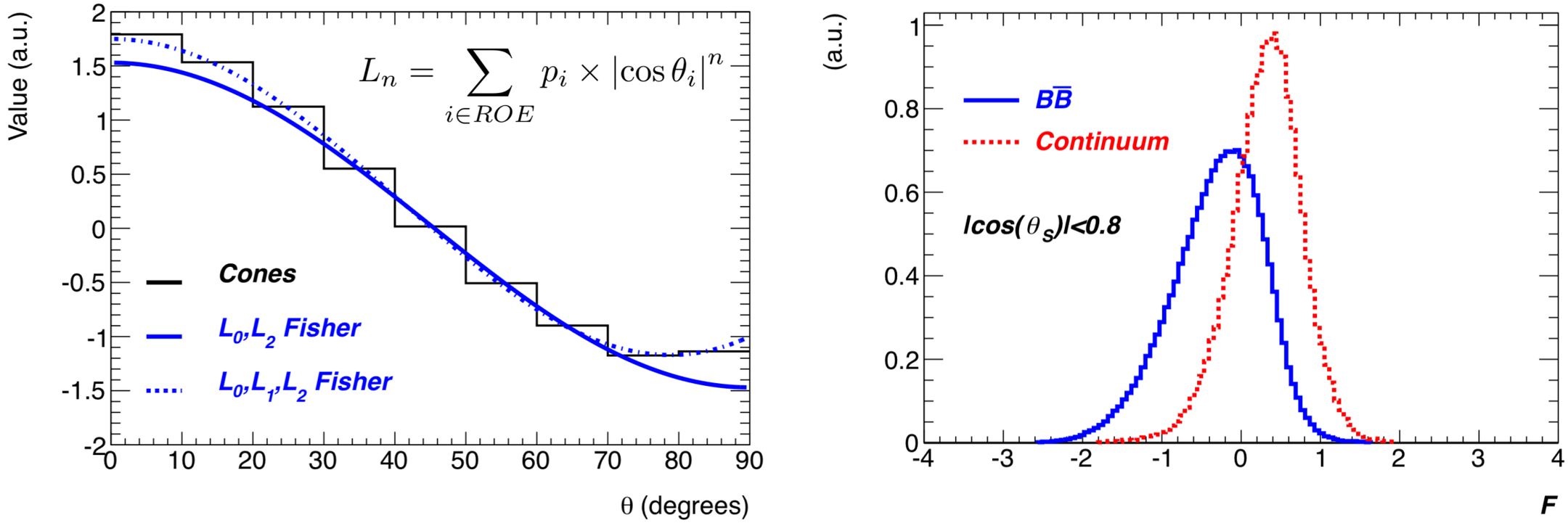
- During the course of data taking new energy-flow variables were introduced that improved upon the “CLEO Fisher”.

A. Bevan et al., Eur. Phys. J. C**74** (2014) 3026

EXAMPLES: FISHER DISCRIMINANT

$$F = \alpha^T x + \beta$$

- A typical Fisher used by BaBar is illustrated here:



- The discriminating variables are the sums L_0, L_1 and L_2; typically a Fisher constructed from L_0 and L_2 is used.
- Usually F was used as a discriminating variable in a maximum likelihood fit. The aim of the fit would be to extract a signal yield or some other parameter (e.g. CP violation asymmetry).

A. Bevan et al., Eur. Phys. J. C**74** (2014) 3026

SUMMARY

- We've discussed the concept of cut-and count analyses and their roles in our field.
- Linear discriminants are a simple extension of a cut based analysis:
  - Dimensional reduction is achieved by combining information from many features into a single dimension.
- Examples from past publications have been discussed.

SUGGESTED TOOLS

- TMVA has implementations of cut based optimisation and Fisher discriminants implemented in it.
- scikit learn has linear (and quadratic) discriminants implemented.
- These are simple algorithms to implement if you want to try them out yourselves; you will obtain a deeper understanding of the process if you choose to follow that route.

SUGGESTED READING (HEP RELATED)

- There are numerous papers that have been written in the HEP community that use the cut and count approach.
  - Essentially all experimental results start off with this approach, and some will then apply more complicated methods in order to extract the measurements of interest from data.
- Fisher discriminants have been widely used in some experiments, particularly the B Factories.
- You will be able to find many examples of using these methods on the archive and in journals.
- Chapters 4 and 9 of A. Bevan et al., Eur. Phys. J. C**74** (2014) 3026 may also be of use to you with understanding these methods.

SUGGESTED READING (NON-HEP)

- There are many non-HEP related source of information on this type of analysis. Below are a few suggested starting points:
  - R. A. Fisher, Ann. Eug., **7**, 179188 (1936).
  - Vidal et al., Generalized Principal Component Analysis (2016), Springer.
  - Cowan, Statistical Data Anslysis, (1998) Oxford University Press.