Multivariate Discriminants
A Thumbnail Sketch

Harrison B. Prosper
Florida State University

DESY, Hamburg, Germany

April 3, 2014

Outline

- Introduction
- Classification
  - In Theory
  - In Practice
- Illustrative Example
- Summary

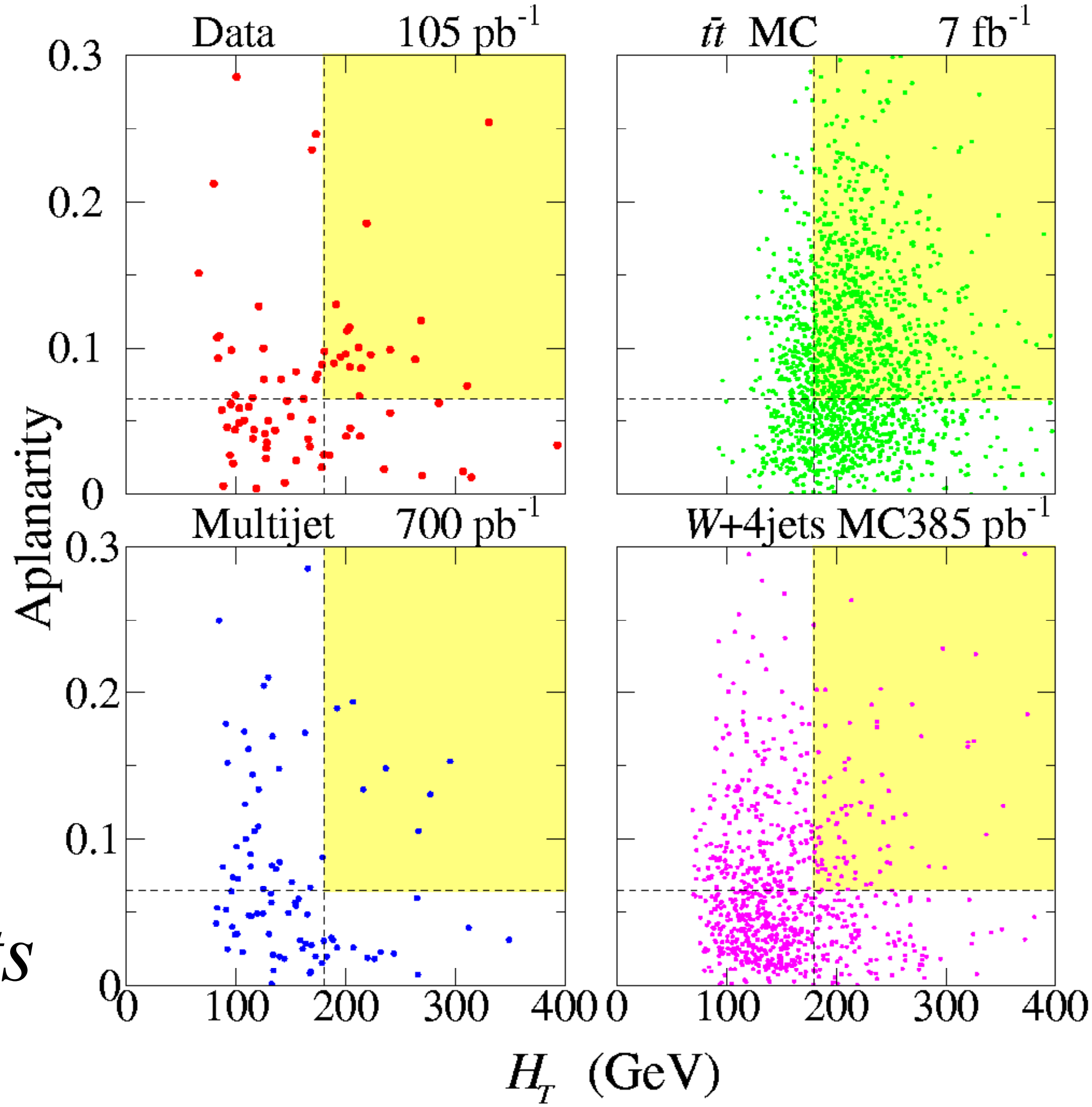
INTRODUCTION

Introduction – Multivariate Data

DØ 1995
Top quark discovery

$x = (A, H_{\mathrm{T}})$

$$p\bar{p} \rightarrow t\bar{t} \rightarrow l + jets$$



Introduction – General Approaches

Two general approaches:

Machine Learning

Given training data $\boldsymbol{T} = (y, \boldsymbol{x}) = (y, x)_1, \ldots (y, x)_N$, a function space $\{f\}$, and a constraint on these functions, teach a machine to learn the mapping $y = f(x)$.

Bayesian Learning

Given training data $\boldsymbol{T}$, a function space $\{f\}$, the likelihood of the training data, and a prior defined on the space of functions, infer the mapping $y = f(x)$.

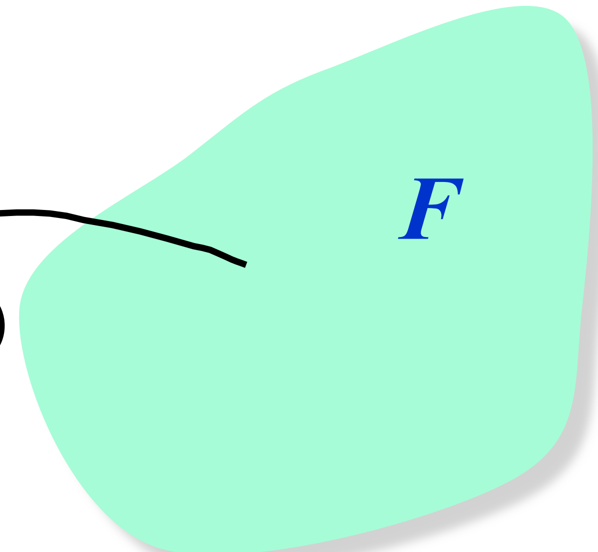
Machine Learning

Choose

Function space $\boldsymbol{F} = \{ f(x, \boldsymbol{w}) \}$

Constraint $\boldsymbol{C}$

Loss function* $\boldsymbol{L}$



$C(\boldsymbol{w})$

$f(x, \boldsymbol{w}^*)$

Method

Find $f(x)$ by minimizing the empirical risk $R(w)$

$$R[f_w] = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i, w))$$

subject to the constraint $C(\boldsymbol{w})$

*The loss function measures the cost of choosing badly

Machine Learning

Many methods (e.g., neural networks, boosted decision trees, rule-based systems, random forests,…) use the

quadratic loss

$$L(y, f(x, w)) = [y - f(x, w)]^2$$

and choose $f(x, w^*)$ by minimizing the

constrained mean square empirical risk

$$R[f_w] = \frac{1}{N}\sum_{i=1}^{N}[y_i - f(x_i, w)]^2 + C(w)$$

Bayesian Learning

Choose

Function space $F = \{ f(x, \boldsymbol{w}) \}$

Likelihood $p(\boldsymbol{T} \mid \boldsymbol{w})$, $\boldsymbol{T} = (\boldsymbol{y}, \boldsymbol{x})$

Loss function L

Prior $p(\boldsymbol{w})$

Method

Use Bayes' theorem to assign a probability (density)

$$p(\boldsymbol{w} \mid \boldsymbol{T}) = p(\boldsymbol{T} \mid \boldsymbol{w})\, p(\boldsymbol{w}) \,/\, p(\boldsymbol{T})$$

$$= p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\, p(\boldsymbol{x} \mid \boldsymbol{w})\, p(\boldsymbol{w}) \,/\, p(\boldsymbol{y} \mid \boldsymbol{x})\, p(\boldsymbol{x})$$

$$\sim p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{w})\, p(\boldsymbol{w}) \qquad \text{(assuming } p(\boldsymbol{x} \mid \boldsymbol{w}) = p(\boldsymbol{x})\text{)}$$

to *every* function in the function space.

Bayesian Learning

Given, the posterior density $p(\mathbf{w} \mid T)$, and *new* data x one computes the (predictive) distribution

$$p(y \mid x, T) = \int p(y \mid x, w)\, p(w \mid T)\, dw$$

If a definite value for y is needed for every x, this can be obtained by minimizing the (risk) function,

$$R[f_w] = \int L(y, f)\, p(y \mid x, T)\, dy$$

which for $L = (y - f)^2$ approximates $f(x)$ by the average

$$f(x) \simeq \bar{y}(x, T) \equiv \int y\, p(y \mid x, T)\, dy$$

Bayesian Learning

Suppose that ***y*** has only two values **0** and **1**, for every x, then

$$f(x) = \int y\, p(y \mid x, T)\, dy$$

reduces to

$$f(x) = p(1 \mid x, T)$$

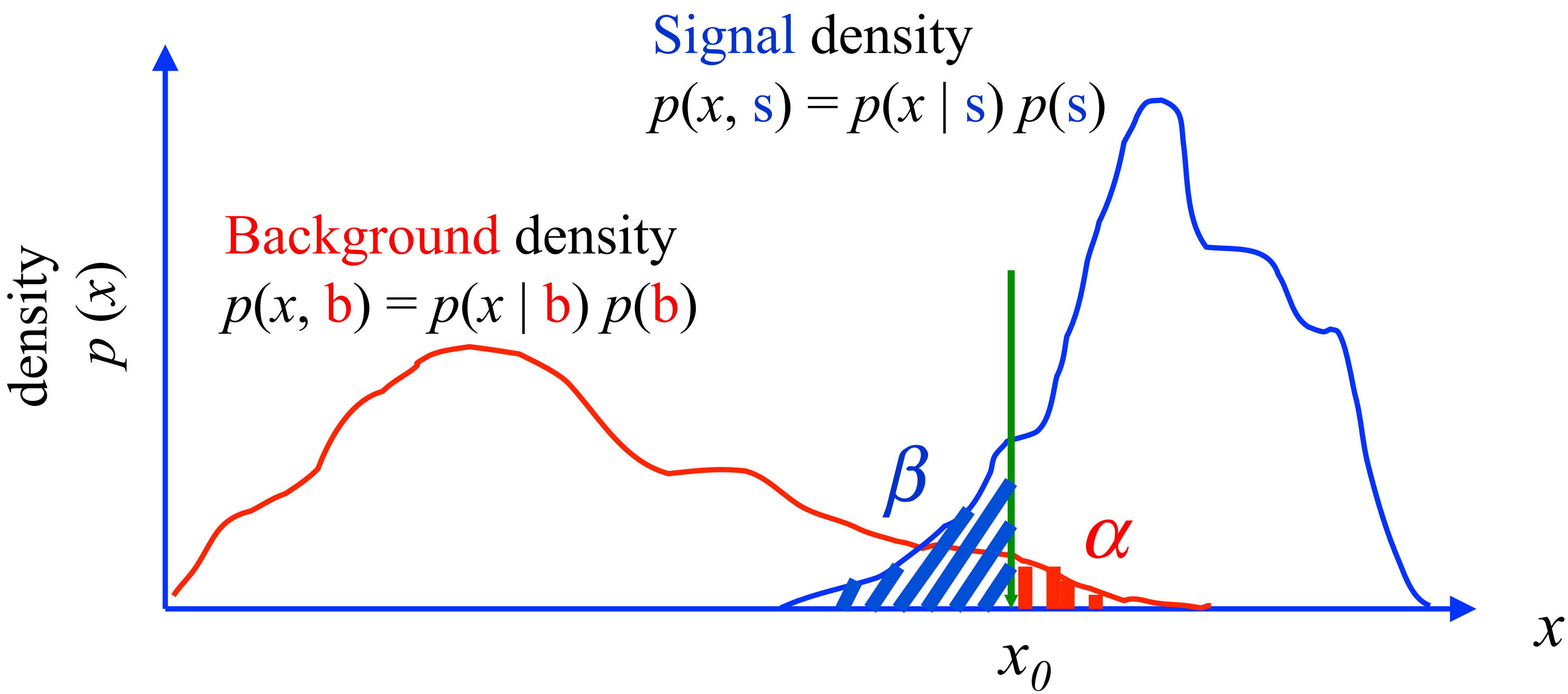
where y = **1** is associated with objects to be kept and y = **0** with objects to be discarded. For example, in an e-mail filter, we can reject junk e-mail using the (*complement of the*) rule

$$\boxed{\text{if } p(1 \mid x, T) > q \text{ accept } x}$$

which is called the **Bayes classifier**

CLASSIFICATION
IN THEORY

Classification: Theory



Optimality criterion: minimize the error rate, $\alpha + \beta$

Classification: Theory

The total loss L arising from classification errors is given by

$$L = L_b \int H(f)\, p(x,b)\, dx + L_s \int [1 - H(f)]\, p(x,s)\, dx$$

Cost of background misclassification

Cost of signal misclassification

where $f(x) = 0$ defines a decision boundary
such that $f(x) > 0$ defines the acceptance region

$H(f)$ is the Heaviside step function:

$H(f) = 1$ if f > 0, 0 otherwise

Classification: Theory

1-D example

$$L = L_b \int H(x - x_0)\, p(x,b)\, dx + L_s \int [1 - H(x - x_0)]\, p(x,s)\, dx$$

Minimizing the total loss L with respect to the boundary x_0

leads to the result:

$$\boxed{\frac{L_b}{L_s} = \frac{p(x_0, s)}{p(x_0, b)} = \left[\frac{p(x_0 \mid s)}{p(x_0 \mid b)}\right] \frac{p(s)}{p(b)}}$$

The quantity in brackets is just the **likelihood ratio**. The result, in the context of hypothesis testing (with $p(s) = p(b)$), is called the **Neyman-Pearson lemma** (1933)

Classification: Theory

The ratio

$$\frac{p(x,s)}{p(x,b)} = \frac{p(s\,|\,x)}{p(b\,|\,x)} \equiv B(x), \quad p(s\,|\,x) = p(x,s)\,/\,p(x)$$

$$p(b\,|\,x) = p(x,b)\,/\,p(x)$$

is called the **Bayes discriminant** because of its close connection to Bayes' theorem:

$$\frac{B(x)}{1+B(x)} = p(s\,|\,x) = \frac{p(x\,|\,s)\,p(s)}{p(x\,|\,s)\,p(s) + p(x\,|\,b)\,p(b)}$$

Classification: The Bayes Connection

Consider the mean squared risk in the limit $N \rightarrow$ infinity,

$$
\begin{aligned}
R[f] &= \frac{1}{N}\sum_{i=1}^{N}[y_i - f(x_i, w)]^2 + C(w) \\
&\rightarrow \int dx \int dy [y - f(x, w)]^2 \, p(y, x) \\
&= \int dx \, p(x) \left[\int dy \, (y - f)^2 \, p(y \mid x) \right]
\end{aligned}
$$

where we have written $p(y \mid x) = p(y, x) \,/\, p(x)$ and where we have assumed that the effect of the constraint (in this limit) is negligible.

Classification: The Bayes Connection

Now minimize the functional $R[f]$ with respect to f. If the function f is sufficiently flexible, then $R[f]$ will reach its absolute minimum. Then for any small change δf in f

$$\delta R[f] = 2\int dx\, p(x)\,\delta f \left[\int dy\,(y - f)\,p(y \mid x) \right] = 0$$

If we require the above to hold for all variations δf, for all x, then the term in brackets must be zero.

$$\int dy\,(y - f)\,p(y \mid x) = 0$$

Classification: The Bayes Connection

Since for the signal class s, $y = 1$, while for the background, b, $y = 0$, we obtain the important result:

$$\boxed{f = \int y\, p(y \mid x)\, dy = p(1 \mid x) \equiv p(s \mid x)}$$

See, Ruck et al., *IEEE Trans. Neural Networks* 4, 296-298 (1990);
Wan, *IEEE Trans. Neural Networks* 4, 303-305 (1990);
Richard and Lippmann, *Neural Computation*. 3, 461-483 (1991)

In summary:

1. Given sufficient training data T and
2. a sufficiently flexible function $f(x, w)$, then $f(x, w)$ will approximate $p(s \mid x)$, if $y = 1$ is assigned to objects of class s and $y = 0$ is assigned to objects of class b

Classification: The Discriminant

In practice, we typically do not use $p(\mathbf{s} \mid x)$ directly, but rather the **discriminant**

$$D(x) = \frac{p(x \mid s)}{p(x \mid s) + p(x \mid b)} = \frac{\exp(\lambda)}{1 + \exp(\lambda)},$$

$$\text{where } \lambda(x) \equiv \ln[p(x \mid s) / p(x \mid b)]$$

This is fine because $p(\mathbf{s} \mid x)$ is a *one-to-one* function of $D(x)$ and therefore both have the same discrimination power

$$p(s \mid x) = \frac{D(x)}{D(x) + [1 - D(x)] / a}, \quad a = p(s) / p(b)$$

CLASSIFICATION IN PRACTICE

Classification: In Practice

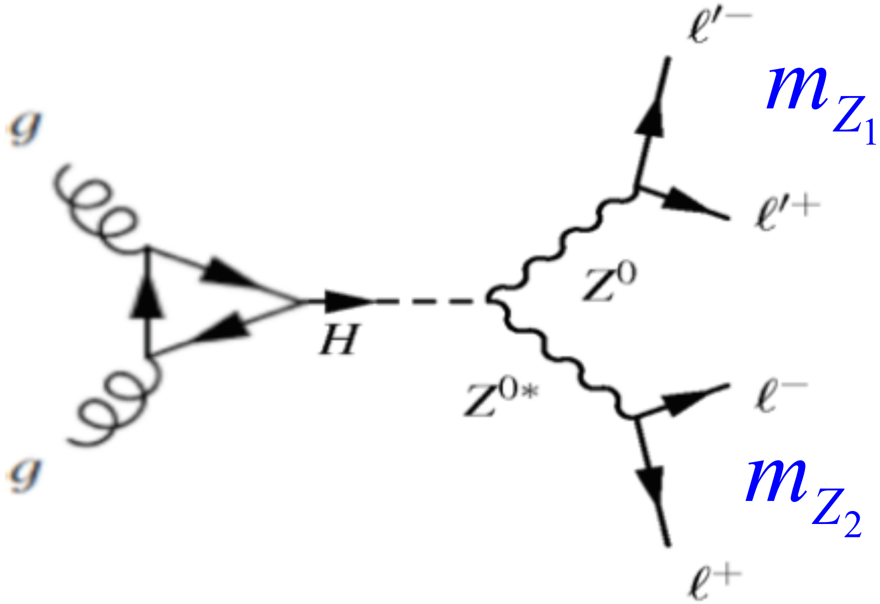
Here is a *short* list of multivariate (MVA) methods that can be used for classification:

- Random Grid Search
- Fisher Discriminant
- Quadratic Discriminant
- Naïve Bayes (Likelihood Discriminant)
- Kernel Density Estimation
- Support Vector Machines
- Binary Decision Trees
- Neural Networks
- Bayesian Neural Networks
- RuleFit
- Random Forests

ILLUSTRATIVE EXAMPLE

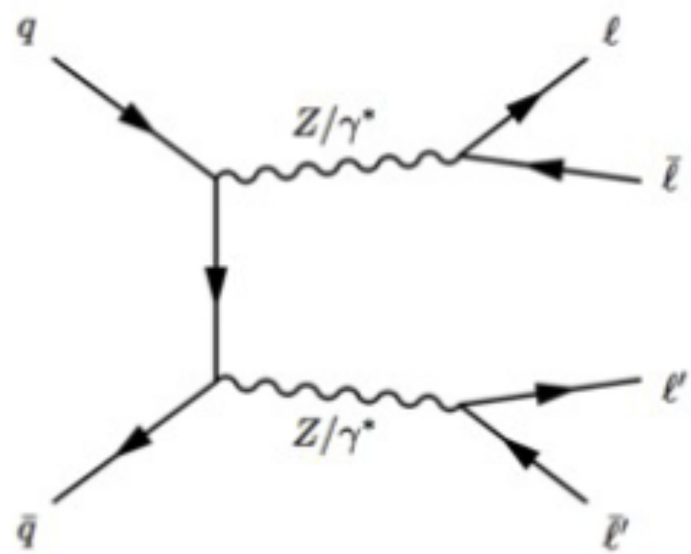
Example – H to ZZ to 4 Leptons

Signal



$$pp \to H \to ZZ \to \ell^+\ell^-\ell'^+\ell'^-$$

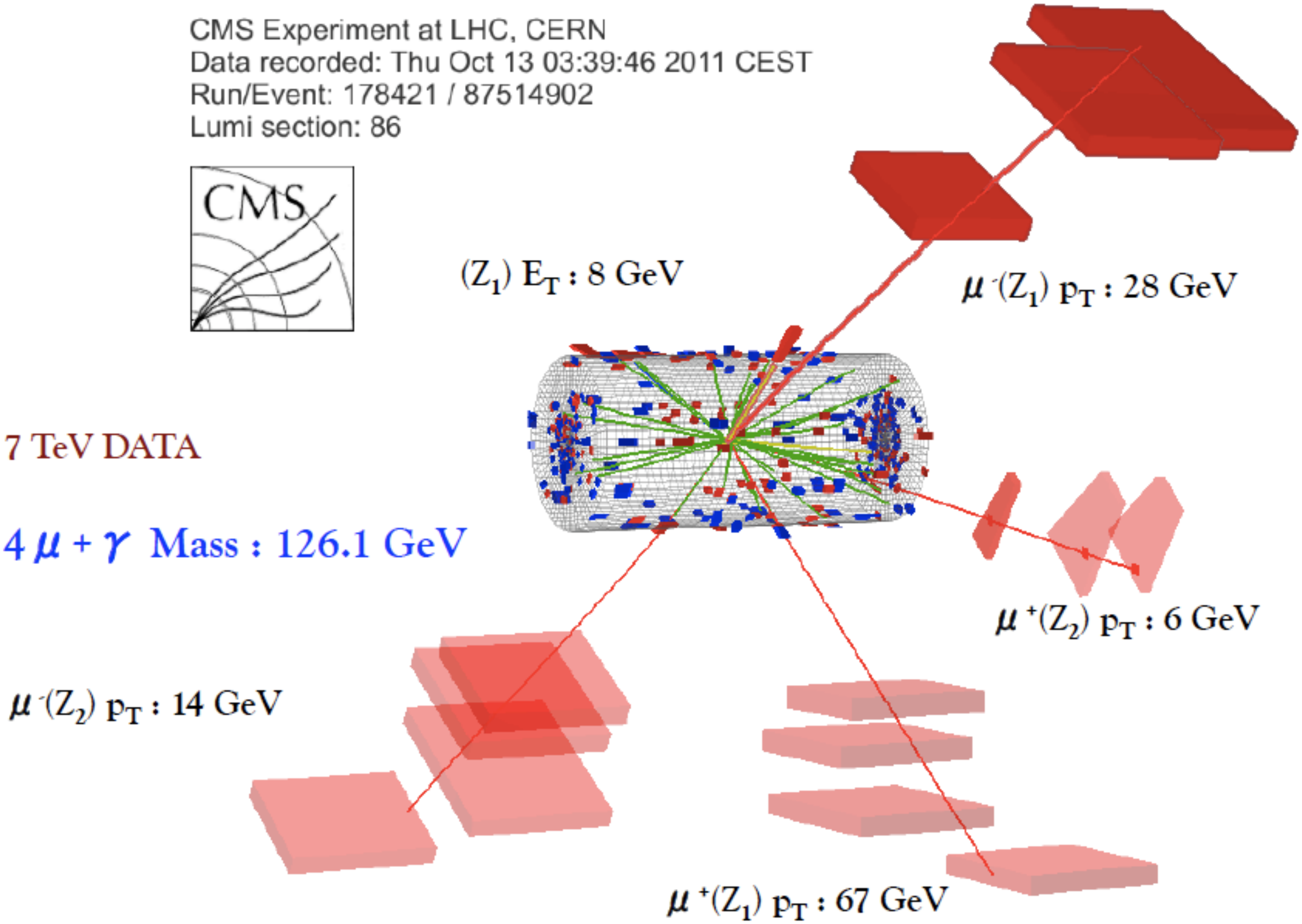
Background



$$pp \to ZZ \to \ell^+\ell^-\ell'^+\ell'^-$$

We shall use this example to illustrate a few of the methods.
We start with $p(s) / p(b) \sim 1 / 20$ and use $x = (m_{Z1}, m_{Z2})$

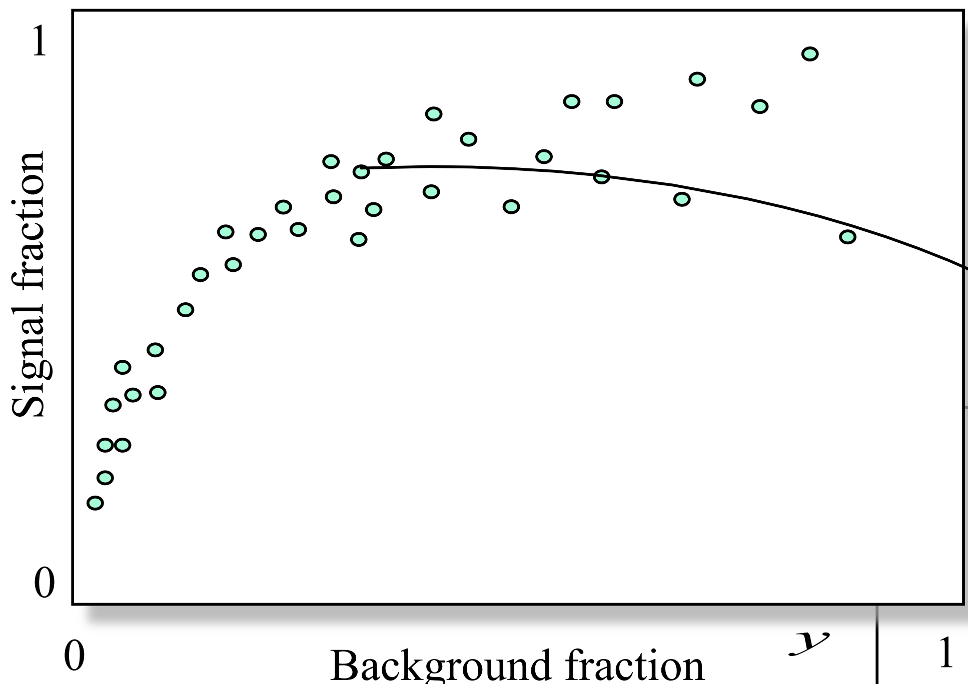
A 4-Lepton Event from CMS



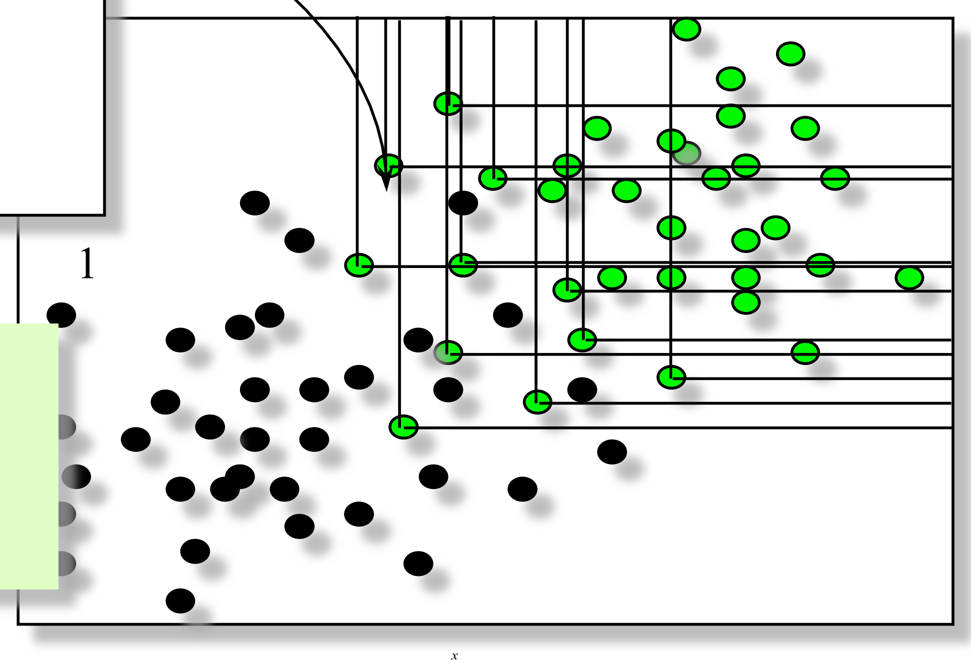
Random Grid Search

Random Grid Search (RGS)

Take each point of the signal class as a **cut-point**



$$x > x_i, \quad y > y_i$$



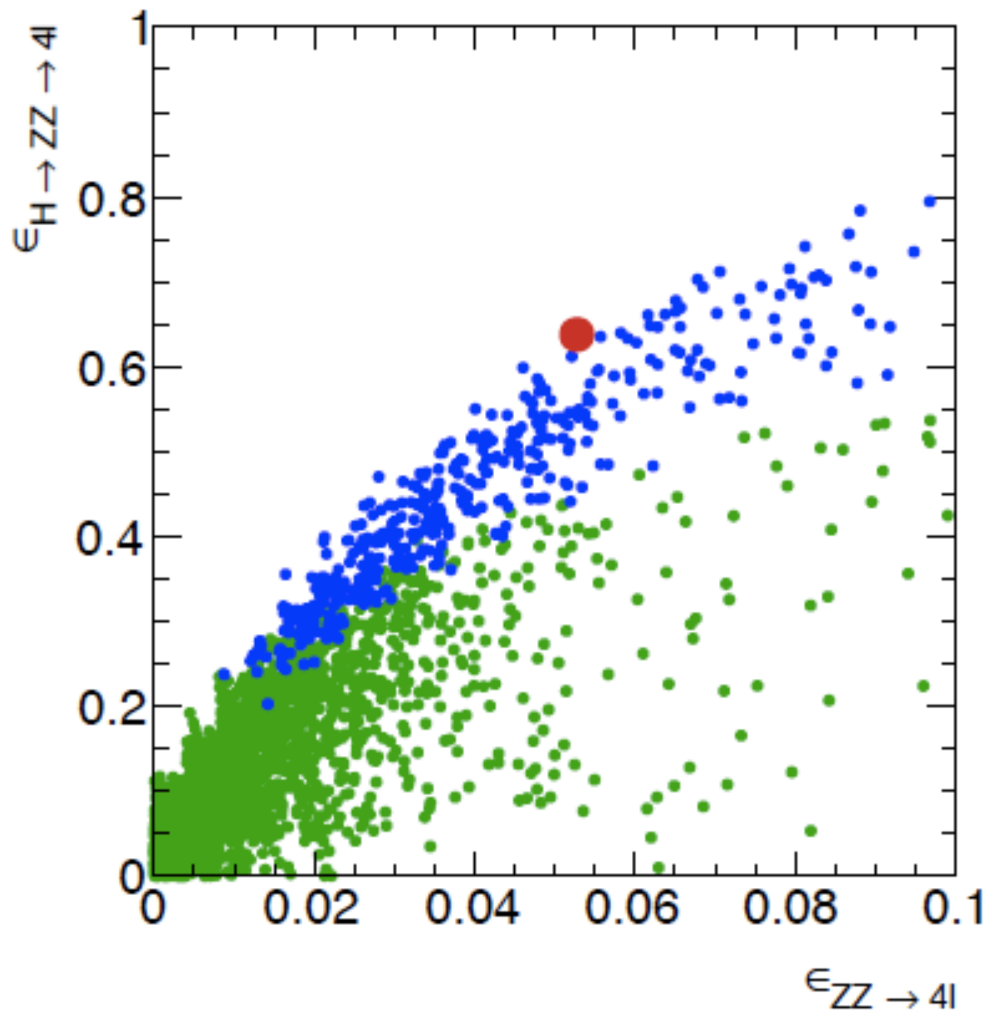
N_{tot} = # events before cuts

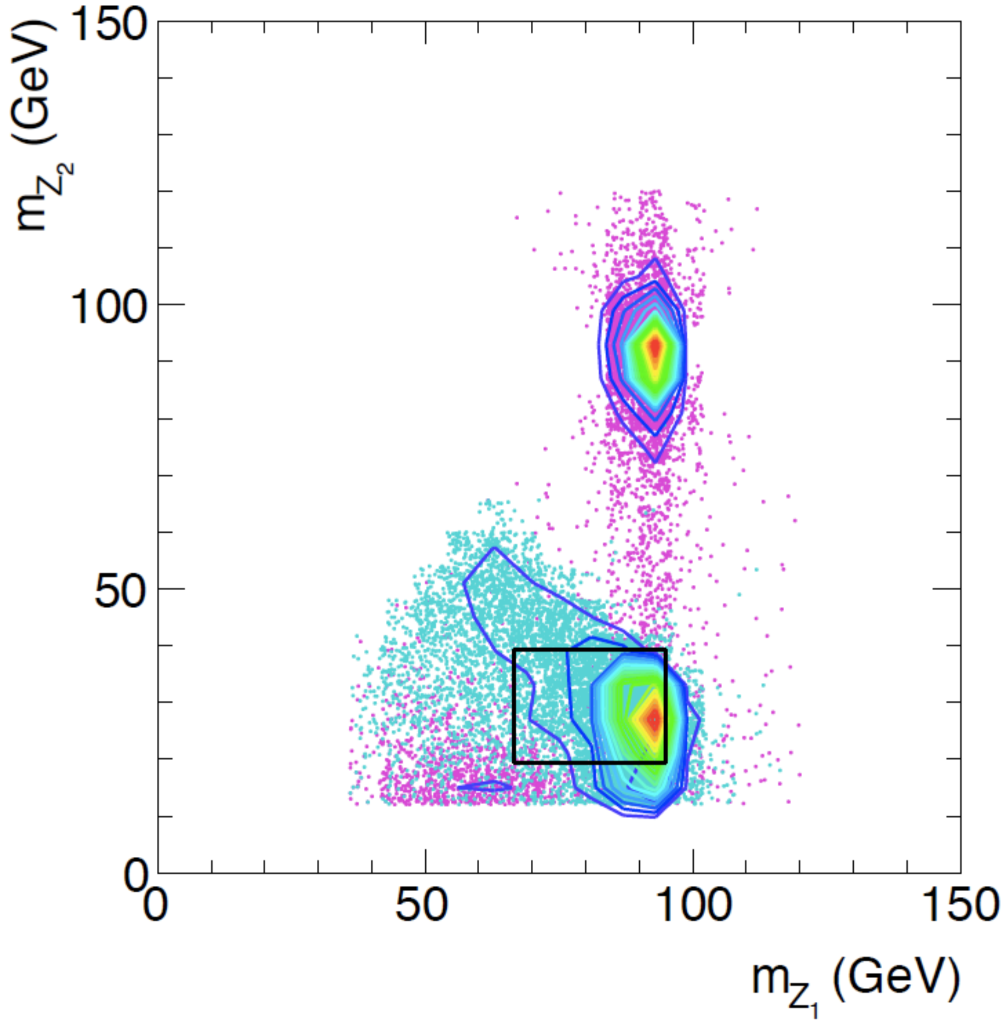
N_{cut} = # events after cuts

Fraction = N_{cut}/N_{tot}

H.B.P. et al., Proceedings, CHEP 1995

Example – H to ZZ to 4Leptons





The red point gives $p(s \mid x) / p(b \mid x) \sim 1 / 1$

Linear & Quadratic Discriminants

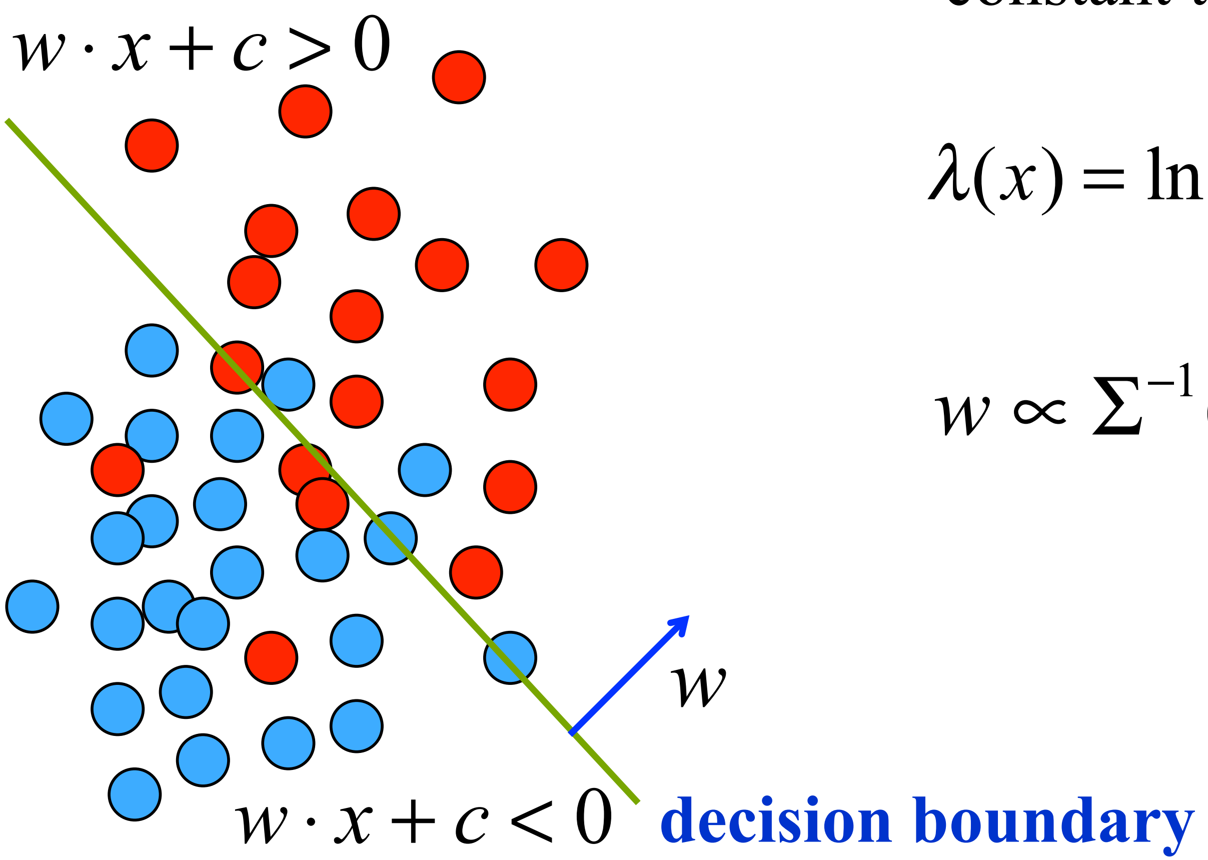
Fisher (Linear) Discriminant

$$B(x) = \frac{p(x \mid s)\, p(s)}{p(x \mid b)\, p(b)}$$

Take $p(x \mid s)$ and $p(x \mid b)$ to be Gaussian (and dropping the constant term) yields

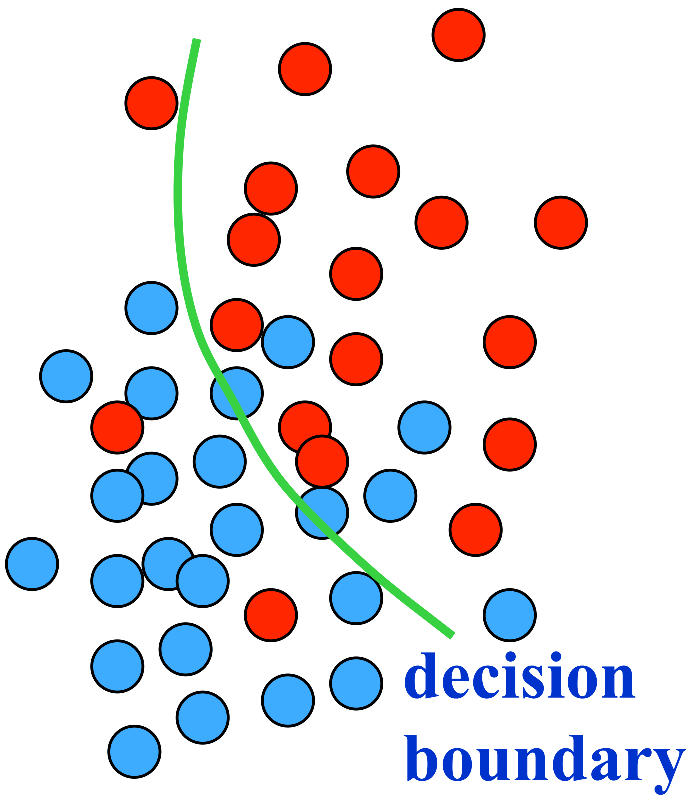
$$\lambda(x) = \ln \frac{G\left(x \mid \mu_s, \Sigma\right)}{G\left(x \mid \mu_b, \Sigma\right)} \rightarrow w \cdot x + c$$

$$w \propto \Sigma^{-1}\left(\mu_s - \mu_b\right)$$



Quadratic Discriminant

If we use *different* covariance matrices for the signal and the background densities, we obtain the **quadratic discriminant**:

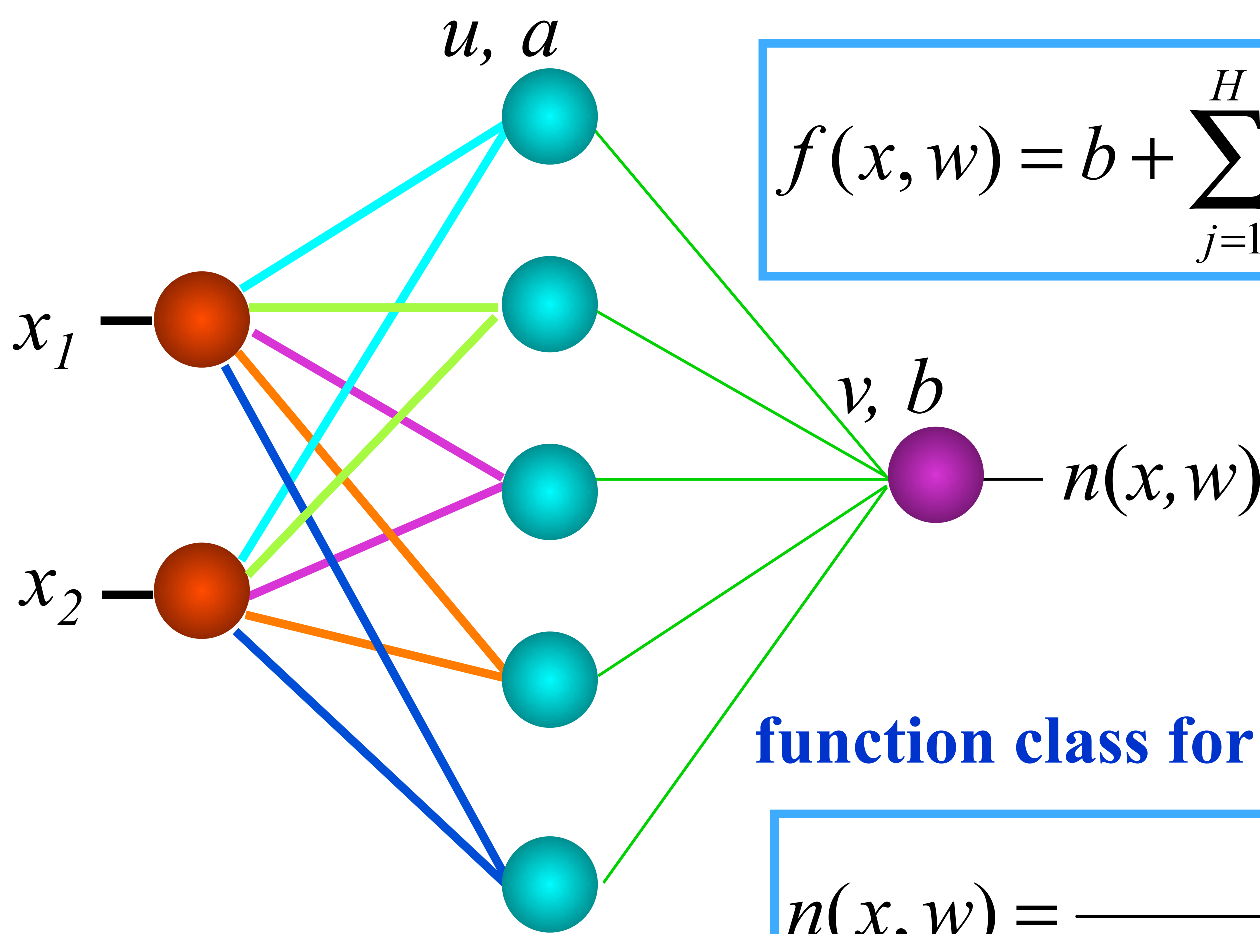


$$\lambda(x) = (x - \mu_b)^T \Sigma_b^{-1} (x - \mu_b) - (x - \mu_s)^T \Sigma_s^{-1} (x - \mu_s)$$

a fixed value of which defines a curved surface that partitions the space $\{x\}$ into signal-rich and background-rich regions

Neural Networks

Neural Networks



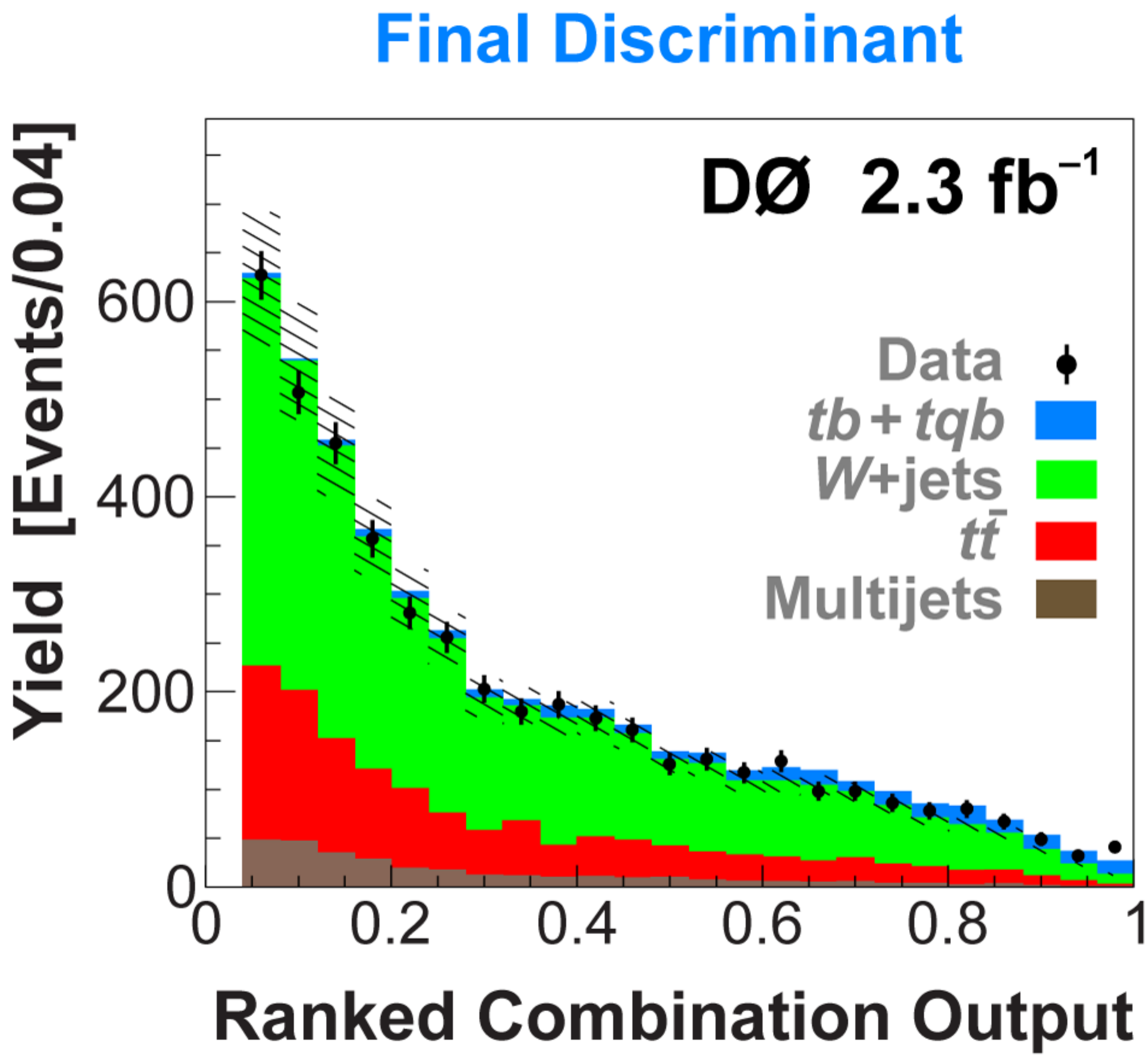
function class for regression

$$f(x,w) = b + \sum_{j=1}^{H} v_j \tanh\left[a_j + \sum_{i=1}^{P} u_{ij} x_i \right]$$

function class for classification

$$n(x,w) = \frac{1}{1+\exp[-f(x,w)]}$$

D0 Single Top Discovery, 2009



Kernel Density Estimation

Kernel Density Estimation

Basic Idea

Place a kernel function at each point and adjust their widths to obtain the best approximation

Parzen Estimation (1960s)

$$p(x) = \frac{1}{N}\sum_n \varphi\left(\frac{x - x_n}{h}\right) \qquad 1 \le n \le N$$

Mixtures

$$p(x) = \sum_j \varphi(x, j)\, q(j) \qquad j << N$$

Kernel Density Estimation

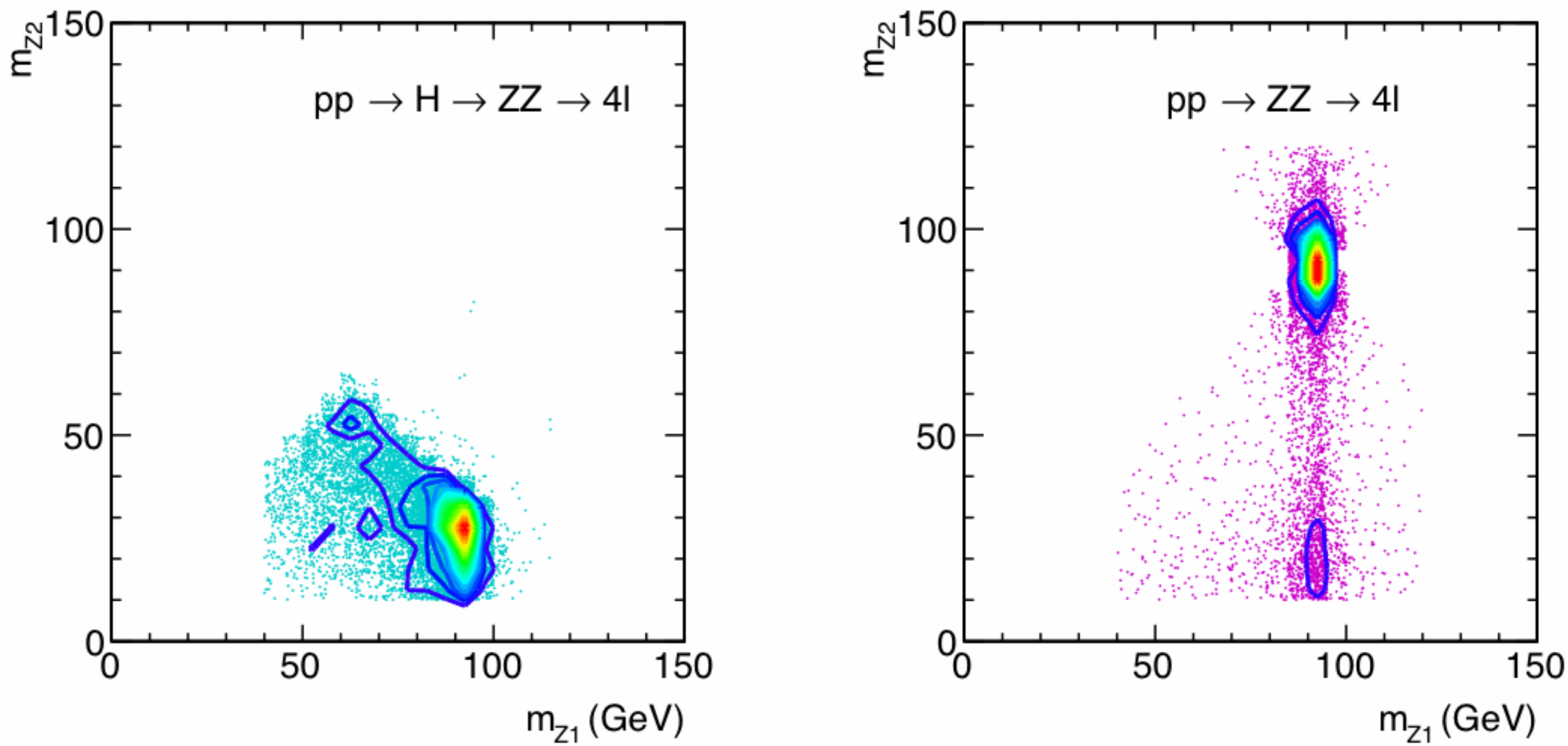
Why does it work? In the limit N goes to infinity

$$p(x) = \frac{1}{N}\sum_{n=1}^{N}\varphi\left(\frac{x-x_n}{h}\right) \to \int \varphi\left(\frac{x-z}{h}\right)p(z)dz$$

the true density $p(x)$ will be recovered provided that the kernel converges to a d-dimensional δ-function:

$$\varphi\left(\frac{x-x_n}{h}\right) \to \delta^d(x-z)$$

KDE of Signal and Background



3000 points / KDE

Summary

- Multivariate methods can be applied to many aspects of data analysis. In this talk, we considered classification in which the classification error rate is minimized.

- It is found that the Bayes discriminant, or any function thereof, is the function that minimizes the error rate.

- There are many ways to approximate this function. But, since no one method is guaranteed to be the best in all circumstances, it is good practice to experiment with a few of them.