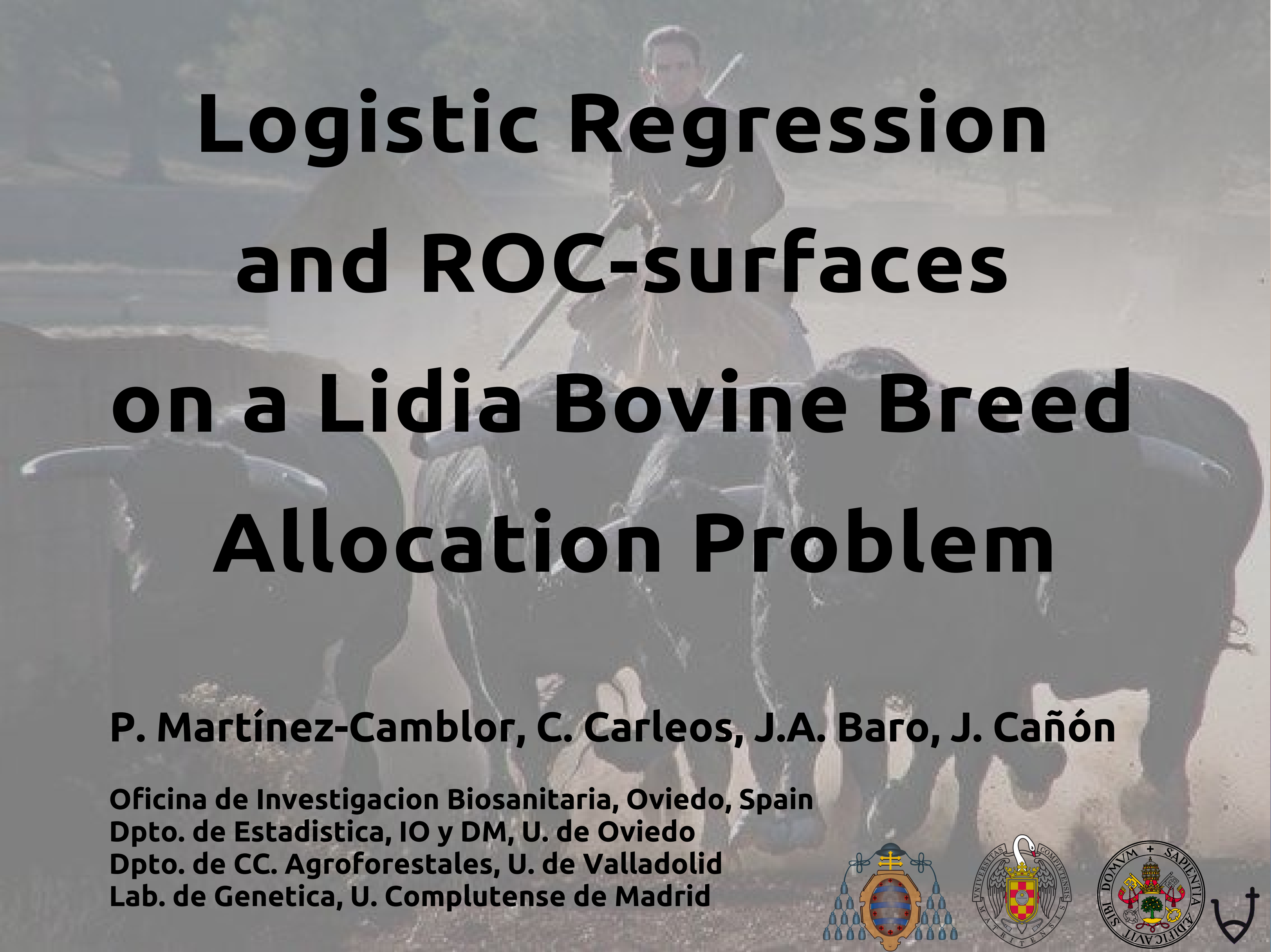


Logistic Regression and ROC-surfaces on a Lidia Bovine Breed Allocation Problem

P. Martínez-Camblor, C. Carleos, J.A. Baro, J. Cañón

Oficina de Investigacion Biosanitaria, Oviedo, Spain
Dpto. de Estadistica, IO y DM, U. de Oviedo
Dpto. de CC. Agroforestales, U. de Valladolid
Lab. de Genetica, U. Complutense de Madrid

Intro

The lidia or fighting bull breed

- a rare case of selection for bovine behaviour
- selection records kept for >500 yrs
- fragmented into lines called *encastes*
- different levels of gene flow among them

37

Objectives

markers for individual identity & breed assignment

- binomial logistic regression applied on each line
- capability to separate: Area Under ROC Curve AUC
- identify microsatellite loci related with each line
- competitive for animal allocation

Material

- blood samples from 1,811 males and females
- same-generation, random individuals from 70 lines
- sample size within line ranged from 7 to 59
- genotyped 24 microsatellite loci, most chromosomes
- alleles per locus ranged from 7 to 22

Material

- what we already knew:
  - STRUCTURE: optimum number of clusters (K) ~ 60

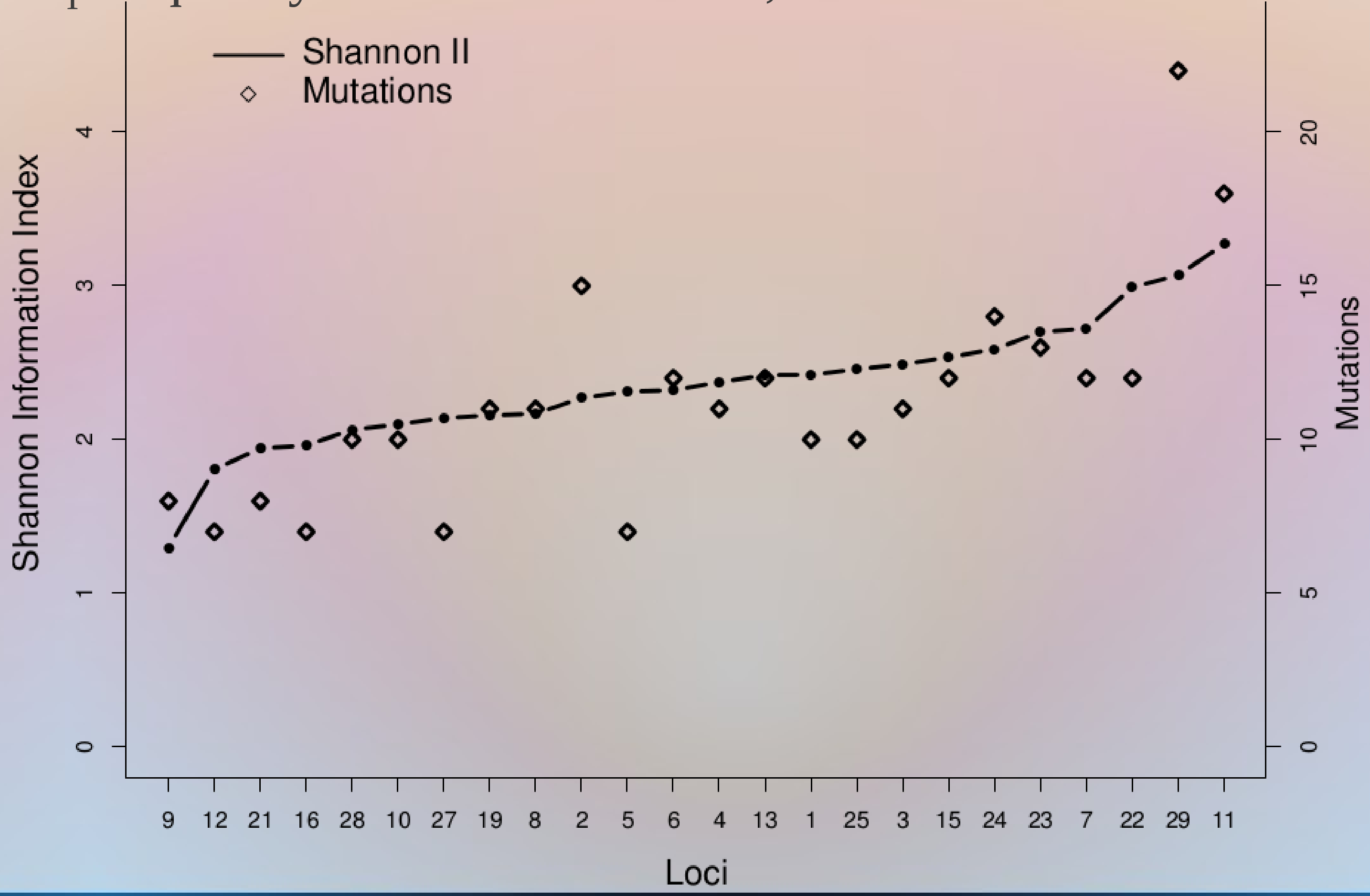
    with average FST 0.46, ranging from 0.27 to 0.86
  - FST values per K are drift rates since coalescence,

    given independence and not-admixture

Material

- other info:
  - Shannon information index $SII = -\sum_{i=1}^{n} [f_i \log_2(f_i)]$

f_i frequency of the i-th mutation, n no. of mutations observed)



Problem

- we have N subjects
  - n_i with the studied characteristic (line/encaste)
  - N – n_i without the characteristic
- we have k variables
  - genotypes
  - from which to classify the individuals
  - with or without the characteristic

Methods

logistic regression model for the conditional mean:

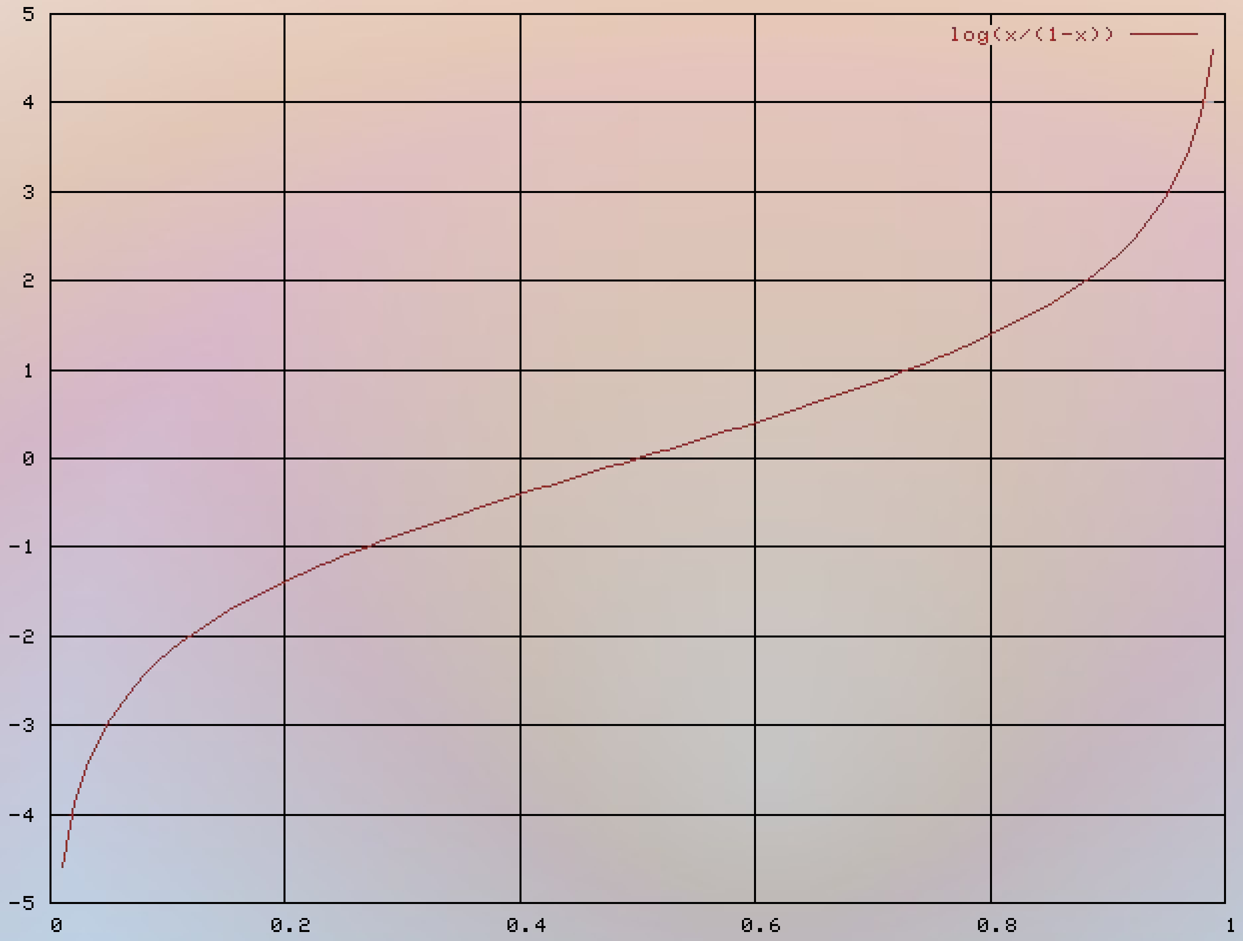
$$\pi(x)=[\mathrm{e}^{x\beta^{t}}][1+\mathrm{e}^{x\beta^{t}}]^{-1}$$

logit-transformed:

$$g(x)=\log\left(\frac{\pi(x)}{1-\pi(x)}\right)=x\beta'$$

Logit as link function

the logit of a probability transforms it in odds



Logistic regression

ML coefficient estimation:

- for each variable in the model,

e^{β_j} (1 ≤ j ≤ k) is the associated Odd Ratio

$g(x) = x\beta'$ how likely the individual belongs

- R package
- code:

```
LR<- glm(belong ~ genotype[,I],family=binomial(link = "logit"))
```

further details: Hosmer & Lemeshow (2000)

Dichotomous

Genotype coded as 2 dichotomous covariates

for each allele, two different covariates:

  - at least one copy?,   and
  - two copies?

selection of covariates if significant on ROC-surface:

- 40 out of 540 candidate allele covariates
- on least p-value at allele-line independence test

Model

- line: fit k (70 lines) logistic regressions
- assign to line with highest score *g*
- if $g_l (x_i) = \max \{g_1(x_i),\ldots,g_k(x_i)\}$ classification correct, wrong otherwise

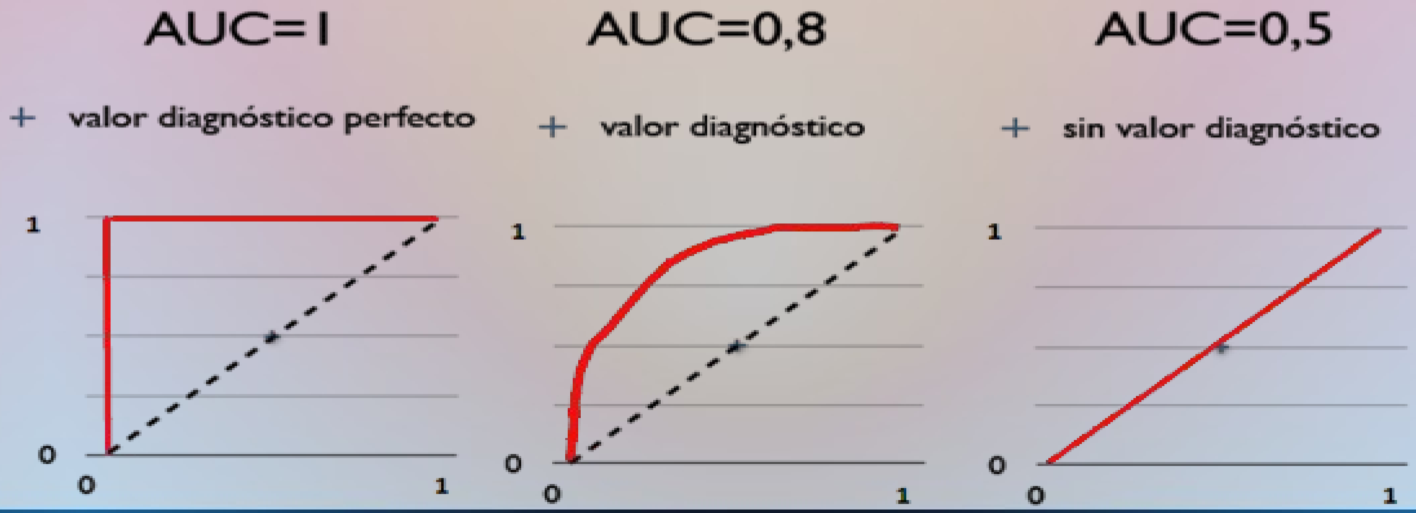
ROC curves & AUC

ROC plots describe intrisic accuracy of classification models

- sensitivity S_E → ability of the model to asign

  vs.
- complement of specificity $(1 - S_P)$ → inability to recognize as different

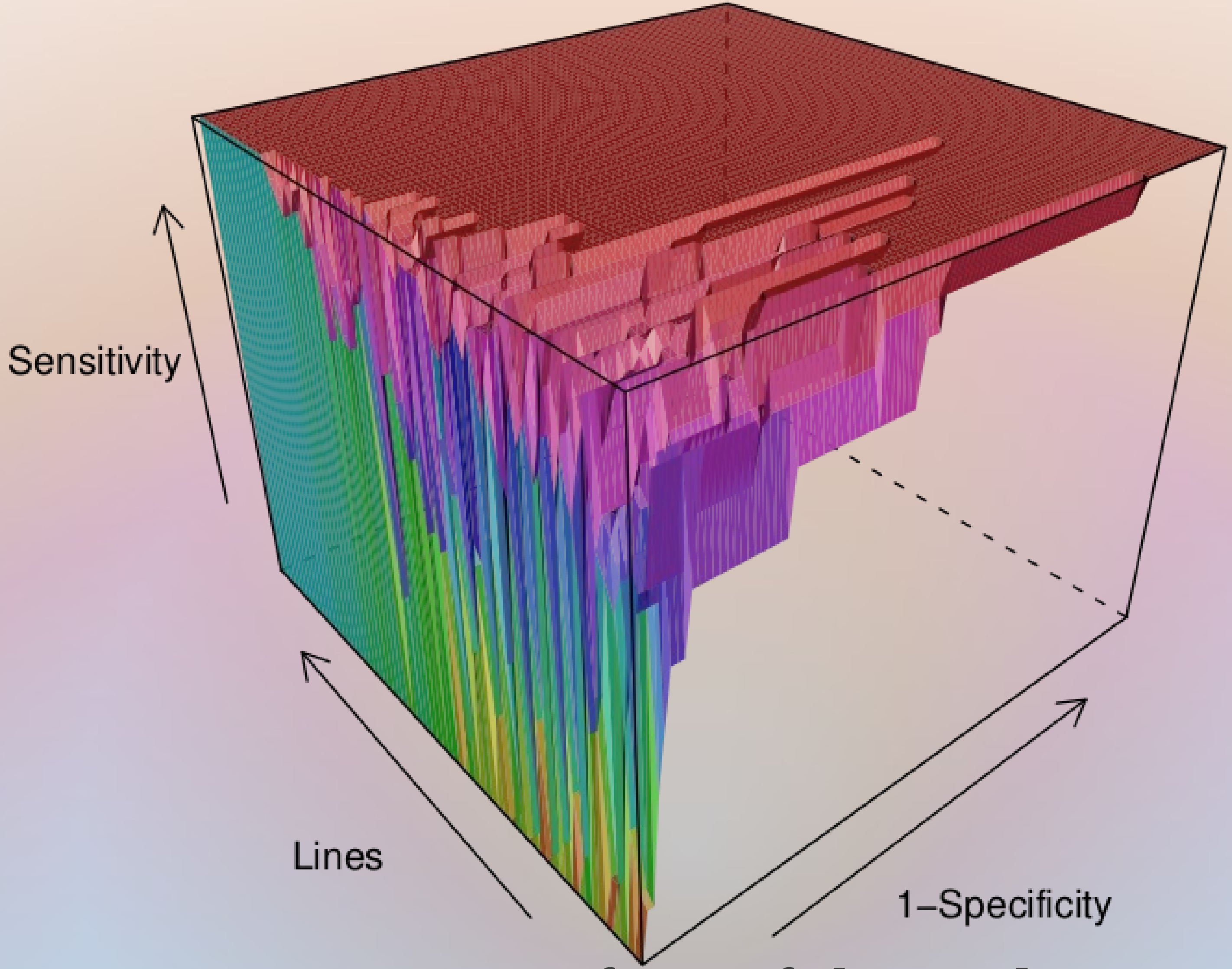
AUC = $\int_0^1$ ROC(t) dt measures quality of binary classification



Results

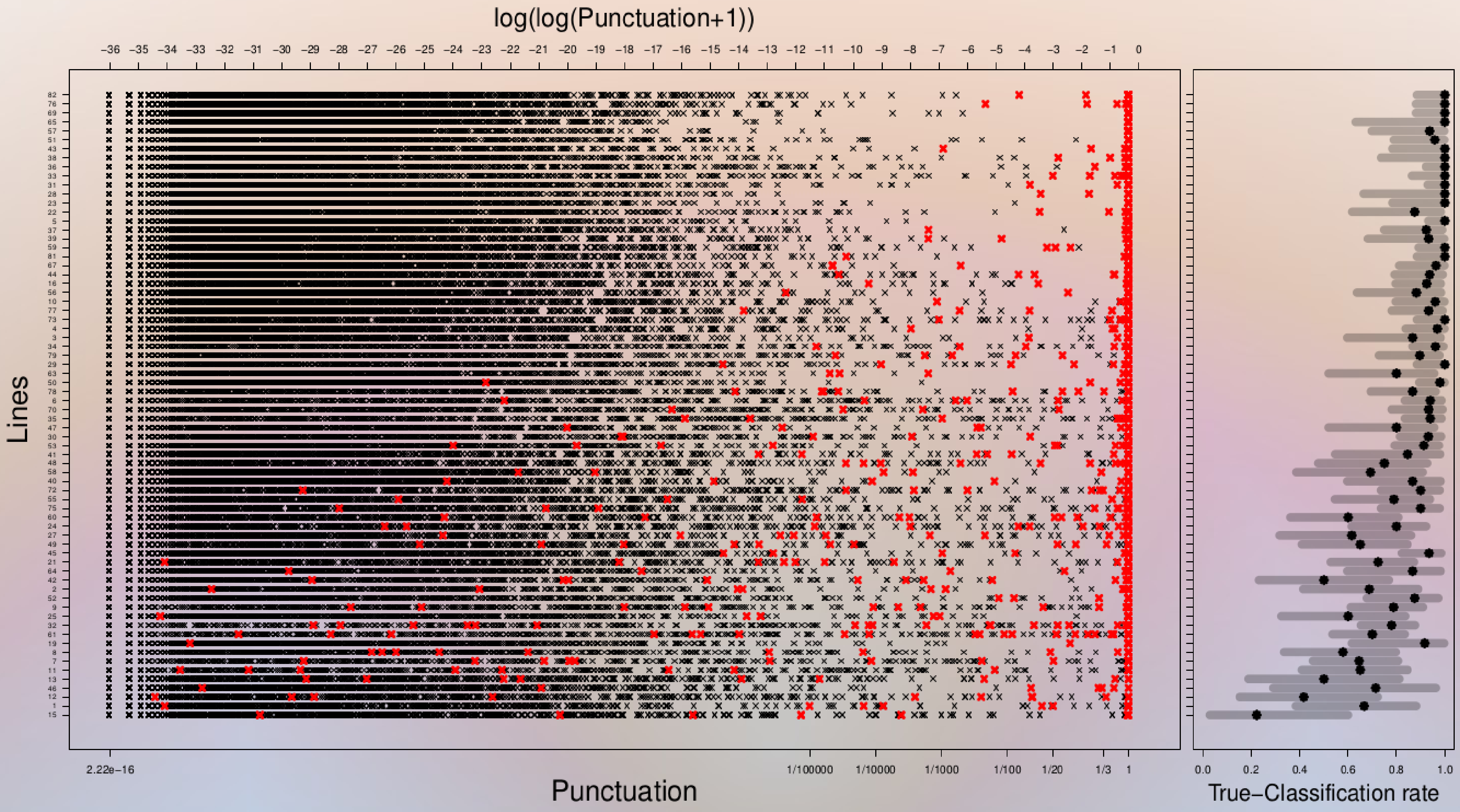
- true-classification-rate: 0.879
- worse than maximum-likelihood methods: 0.910
- BUT: better true-classification rates for some lines
- Logistic Regression results similar to Data-mining/Machine Learning (Guinand et al., 2002)

Results



ROC-surface of the 70 lines

Results



Score of each bull at each line (red at true line)

Conclusions

- multivariate models for line allocation with microsatellite genotypes
- covariates as binarily coded one- or two-copies
- AUC as a quality measure of the classification
- ROC-surface as a criteria for covariate selection
- multivariate model accounts for *between* and *within* loci dependencies
- possible improvement of results when used on SNP

THANK YOU

for your attention

- financed by grants from:
  - INIA (RTA2011-00060-C02-O2)
  - Ministerio de Ciencia e Innovación, Spain

    (AGL2010-15903, MTM2008-01519 & MTM2011-23204)