Mammograms, Machine Learning and Cumulus

Daniel F. Schmidt, Enes Makalic, Jennifer Stone and John L. Hopper

Centre for MEGA Epidemiology
The University of Melbourne

“Why Study Mammographic Density?”, 16–18th August, 2010

Outline

Outline

Background

- We are from computer science background
- Want to approach the problem from a machine learning perspective

Background

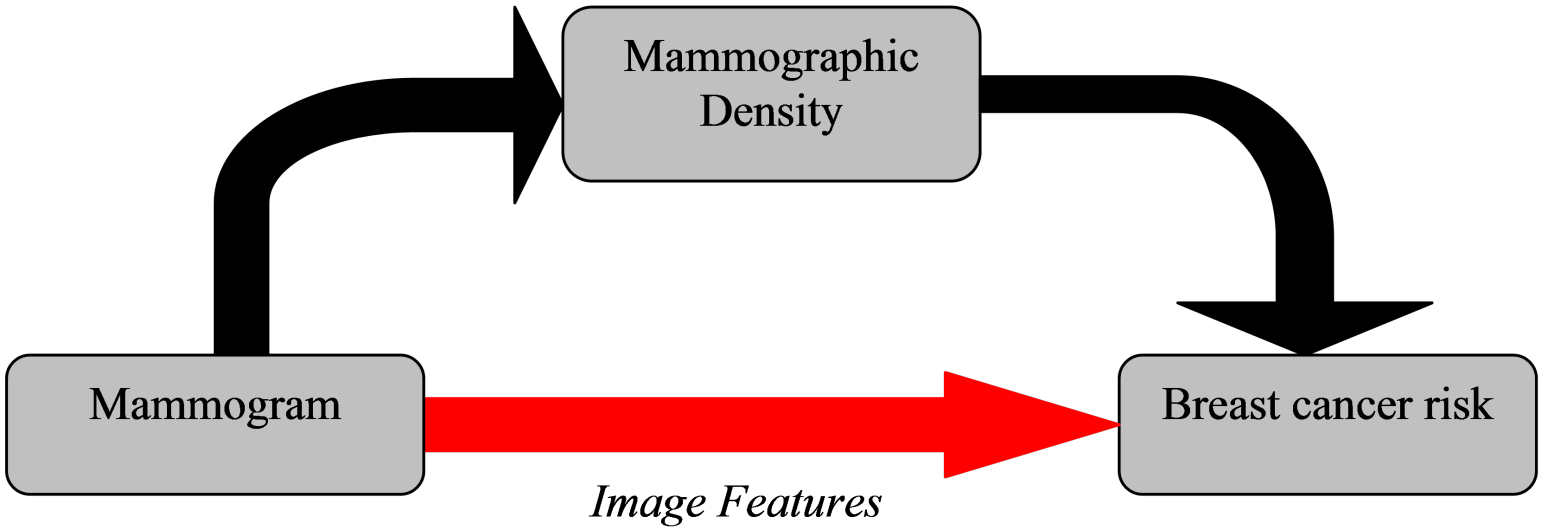
- We are from computer science background
- Want to approach the problem from a machine learning perspective

Background

- We are from computer science background
- Want to approach the problem from a machine learning perspective

Background

- We are from computer science background
- Want to approach the problem from a machine learning perspective



CUMULUS

- One the most important tools for measuring MD is CUMULUS
- Transforms an image (mammogram) into a single summary statistic
- If **Z** is a mammogram, then the CUMULUS measure is

$$\text{CUMULUS}(\mathbf{Z}) = \sum_{i,j} \text{I}(Z_{i,j} > t)$$

where the threshold t is chosen manually

- Though simple, CUMULUS appears to be the "gold standard" for breast cancer risk
  $\Rightarrow$ Would be of great use if CUMULUS could be fully automated

Outline

Training Dataset

- Aged matched film mammograms from Cambridge study
- Total people, $n = 343$; multiple mammograms per person
  - 161 cases
  - 182 controls
- Restricted to MLOs only
  $\Rightarrow$ MLOs showed stronger association on this dataset
- Multiple mammograms per person handled by averaging
- For each mammogram, we had
  - Age (AGE)
  - Percent Dense Area (PDA)
  - Square-root Dense Area (DA)
  - Square-root Non Dense Area (NDA)

Our Procedure (CIRRUSv1)

- Extract simple "features" from mammograms
  $\Rightarrow$ Total number of extracted features exceeded $2,500$
- Use features in a linear logistic regression model

$$p(y_i = \text{case}|\mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{x}_i\boldsymbol{\beta})}$$

  where $\mathbf{x}_i$ is the vector of features and y_i is the case-control status for individual i respectively
- Number of features is large; must use something more sophisticated than maximum likelihood/stepwise selection
  $\Rightarrow$ Bayesian regularised logistic regression employed

Our Procedure (CIRRUSv1)

- Extract simple "features" from mammograms
  $\Rightarrow$ Total number of extracted features exceeded $2,500$
- Use features in a linear logistic regression model

$$p(y_i = \text{case}|\mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{x}_i\boldsymbol{\beta})}$$

  where $\mathbf{x}_i$ is the vector of features and y_i is the case-control status for individual i respectively
- Number of features is large; must use something more sophisticated than maximum likelihood/stepwise selection
  $\Rightarrow$ Bayesian regularised logistic regression employed

Our Procedure (CIRRUSv1)

- Extract simple "features" from mammograms
  $\Rightarrow$ Total number of extracted features exceeded $2,500$
- Use features in a linear logistic regression model

$$p(y_i = \text{case}|\mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{x}_i\boldsymbol{\beta})}$$

  where $\mathbf{x}_i$ is the vector of features and y_i is the case-control status for individual i respectively
- Number of features is large; must use something more sophisticated than maximum likelihood/stepwise selection
  $\Rightarrow$ Bayesian regularised logistic regression employed

"Smoothing" CUMULUS Measurements

- A CUMULUS measurement of MD is of the form

$$\text{CUMULUS} = \text{MD} + \text{bias} + \text{noise}$$

  - MD is the true, underlying amount of MD (person dependent)
  - bias represents the limitations of the CUMULUS measurement scheme (person dependent)
  - noise is measurement noise due to various factors (measurement specific)

- Use PDA as "outcome", our image features as "exposures"
  $\Rightarrow$ Regularised linear regression

"Smoothing" CUMULUS Measurements

- A CUMULUS measurement of MD is of the form

$$\text{CUMULUS} = \text{MD} + \text{bias} + \text{noise}$$

  - MD is the true, underlying amount of MD (person dependent)
  - bias represents the limitations of the CUMULUS measurement scheme (person dependent)
  - noise is measurement noise due to various factors (measurement specific)

- Use PDA as "outcome", our image features as "exposures"
  $\Rightarrow$ Regularised linear regression

Outline

Procedure

- Based on the Cambridge dataset
- Pectoral muscle was manually removed
- Compare CIRRUSv1 against
  - CUMULUS: PDA adjusted for NDA and AGE
  - SmoothedCUMULUS: Smoothed PDA adjusted for AGE
- These are implemented in logistic regression models estimated using maximum likelihood
- Comparison metrics
  - Classification accuracy
  - Odds ratios
  - Correlation
- All metrics are computed *within-sample*

Classification

- Classifying with logistic regression
  $\Rightarrow$ if $p(y_i = \text{case}|\mathbf{x}_i) > 1/2$, classify individual i as CASE
- Compare methods on number of correct classifications (classification accuracy)
- Can produce a "confusion matrix"; example:

	CON	CAS
CON	TN	FP
CAS	FN	TP

Classification Results

- Baseline = 0.53
- CUMULUS = 0.6122,

	CON	CAS
CON	129	53
CAS	80	81

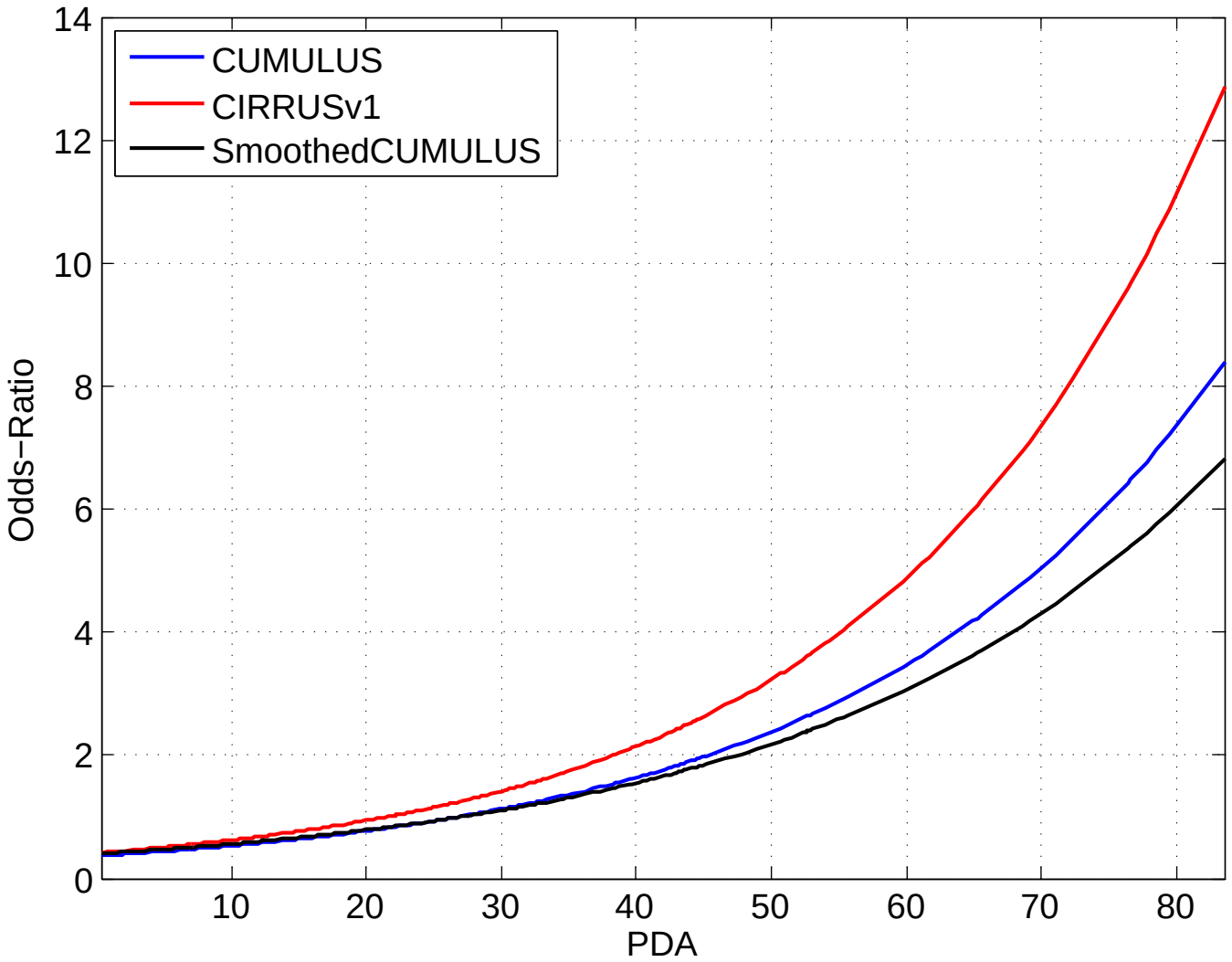
- CIRRUSv1 = 0.6735,

	CON	CAS
CON	161	21
CAS	91	70

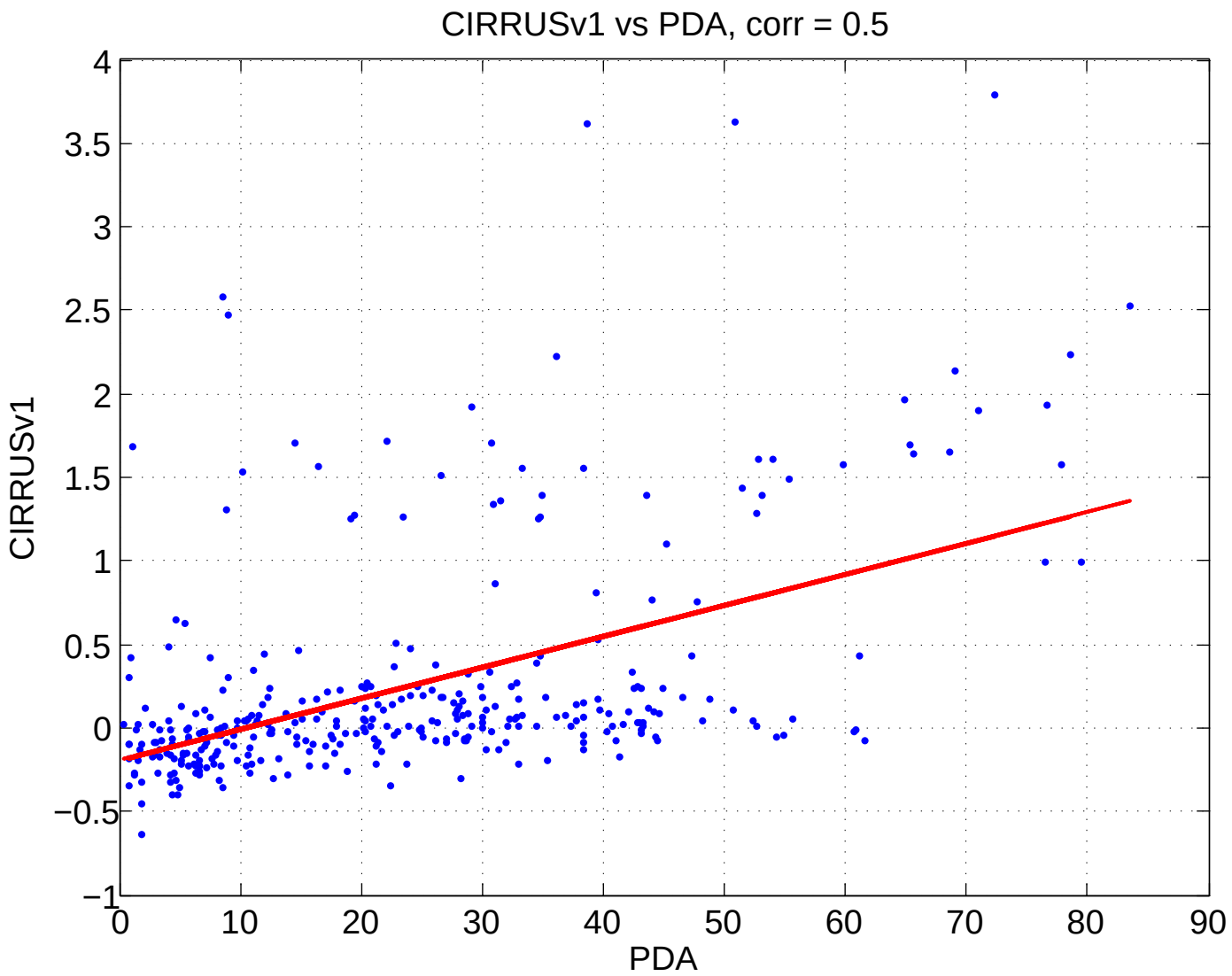
- SmoothedCUMULUS = 0.6501,

	CON	CAS
CON	136	46
CAS	74	87

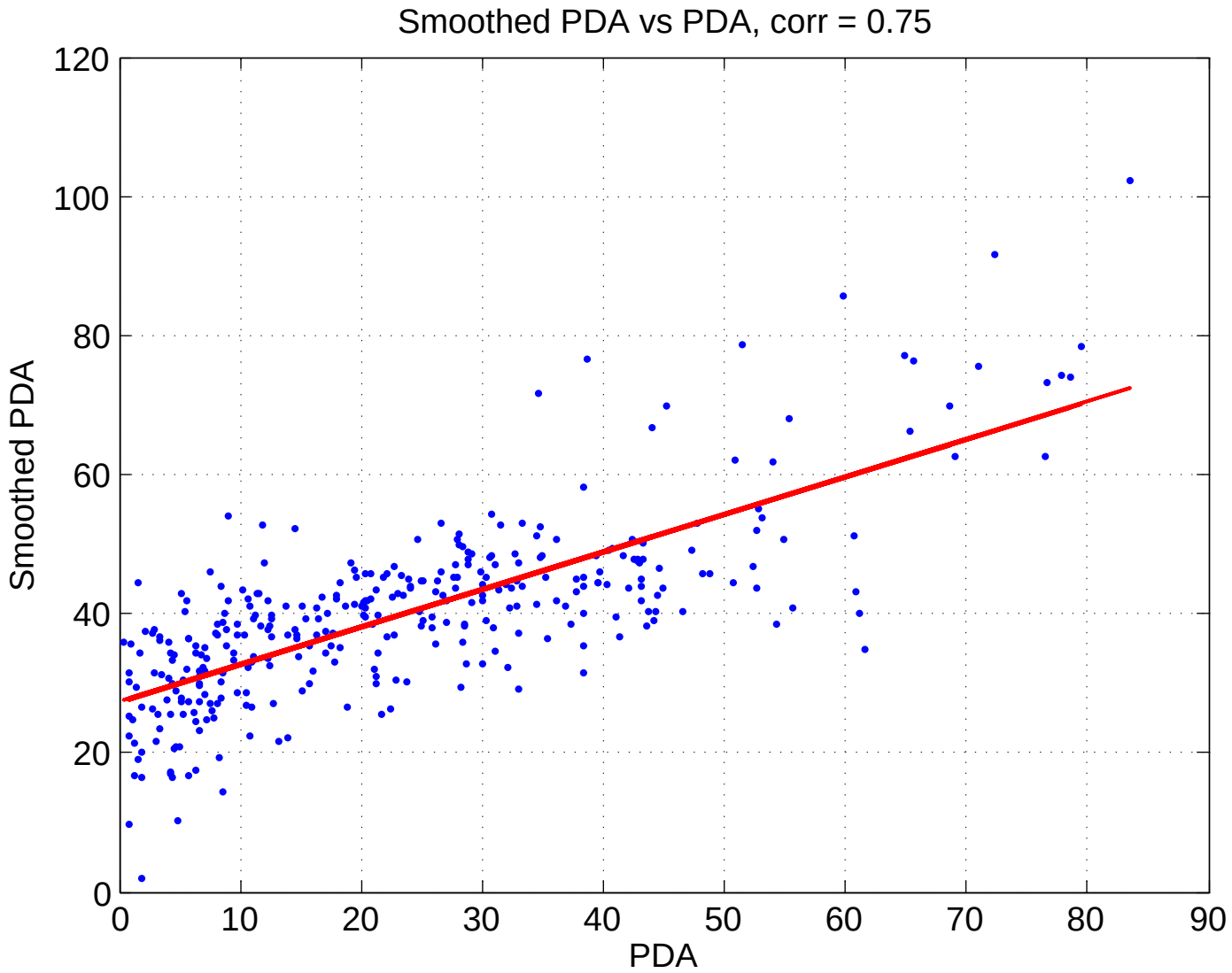
Odds-Ratios


CUMULUS
CIRRUSv1
SmoothedCUMULUS
Odds−Ratio
PDA
0
2
4
6
8
10
12
14
10
20
30
40
50
60
70
80

CUMULUS vs CIRRUSv1



CUMULUS vs SmoothedCUMULUS



Future Work and Conclusion

- More ideas to try; more advanced features
- Try on larger datasets/validation datasets
- Questions?