THE TWO CULTURES: A DISCUSSION

Katrin Newger
Supervisor: Christoph Jansen M.Sc. and Dipl.-Math. Georg Schollmeyer

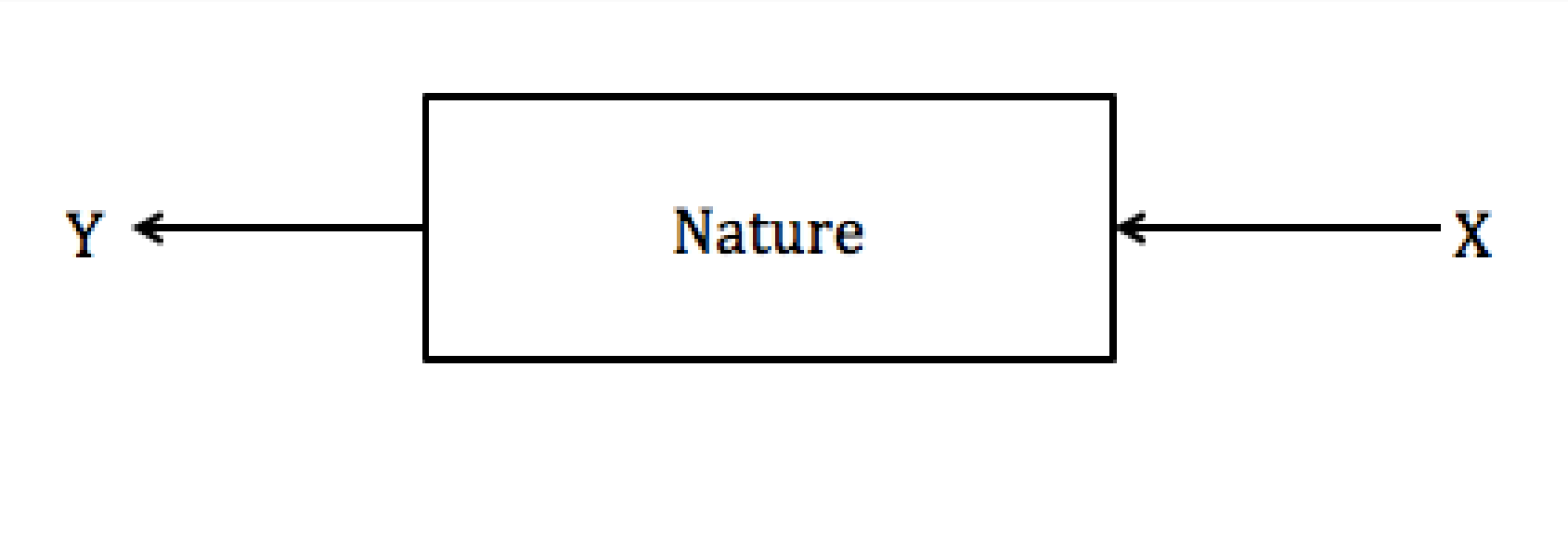
June 27, 2015

Department of Statistics, LMU Munich

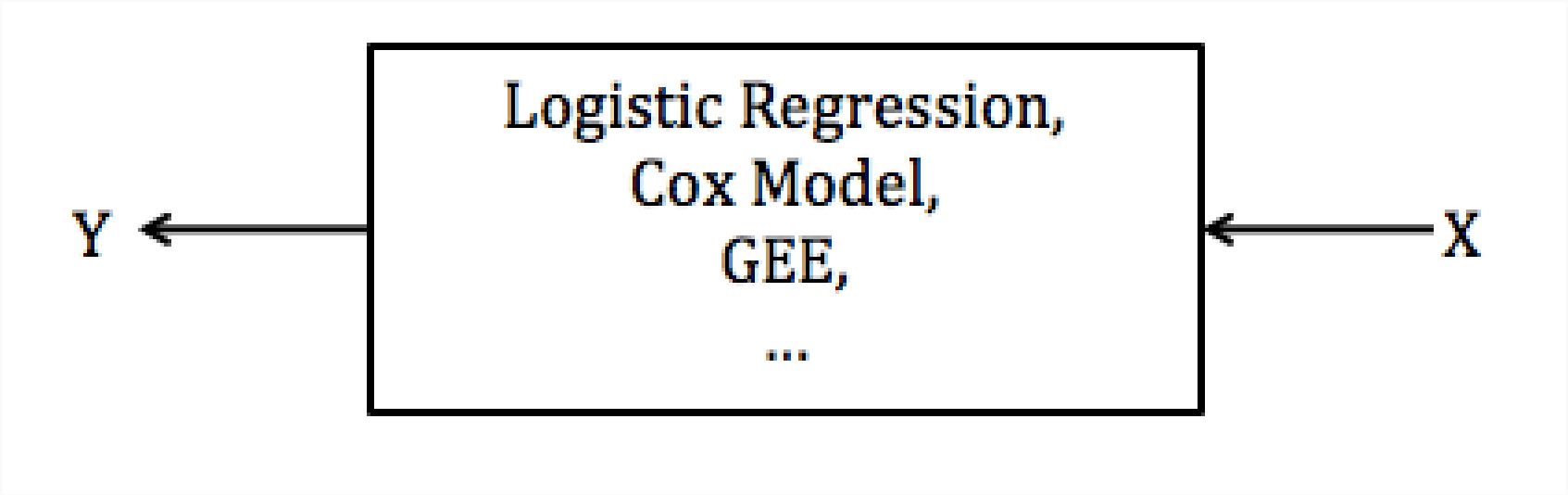
TABLE OF CONTENTS

THE TWO CULTURES

NATURE



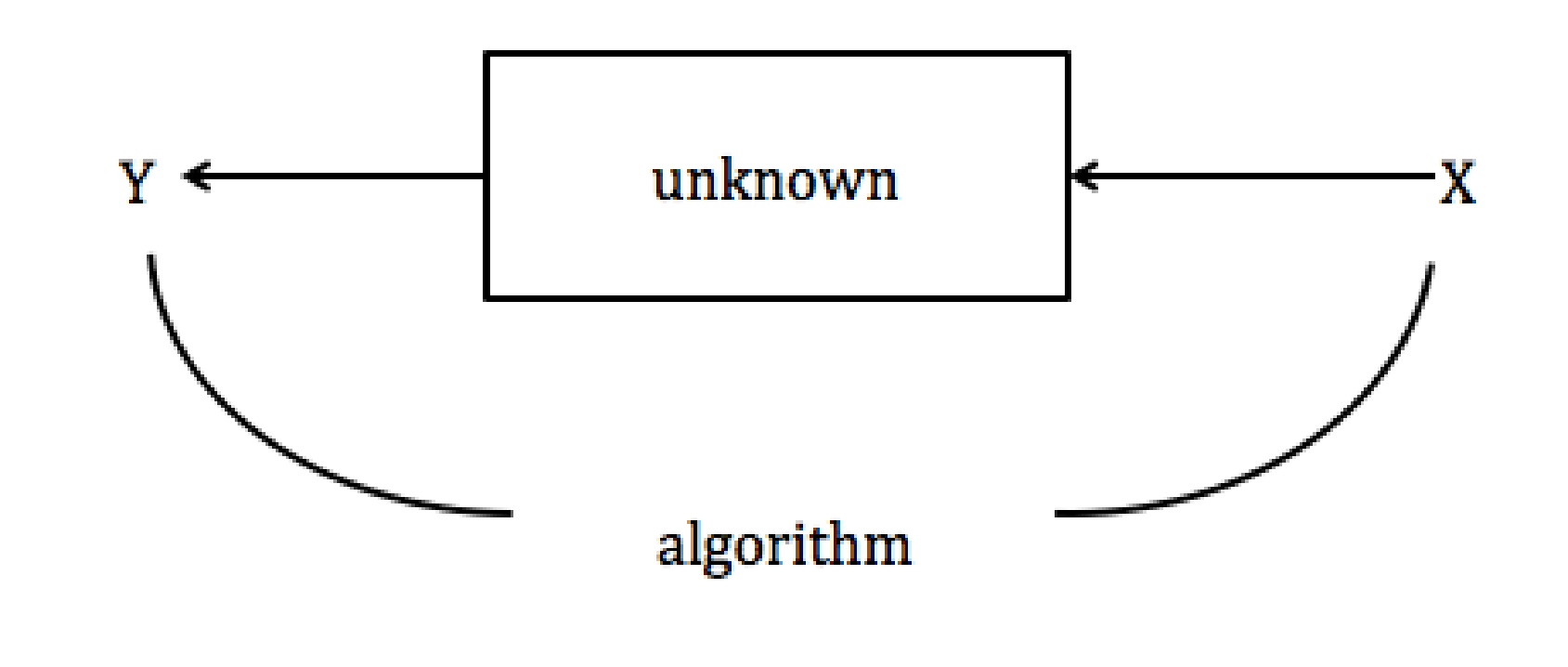
DATA MODEL



Assumptions:

- Stochastic model
- Distribution of residuals
- Further model specific assumptions

ALGORITHMIC MODEL



Goal:

Function $f(x)$ that minimizes loss $L(Y, f(x))$

EXAMPLES FOR ALGORITHMIC MODELS

Methods:

- Support vector machines
- Random forests
- Artificial neural networks
- ...

BREIMAN'S ARGUMENT

THE DATA MODEL—TOO SIMPLE A PICTURE

- Critical model assumptions
- Conclusions about model, not about nature
- Wrong model $\rightarrow$ wrong conclusions about nature
- Algorithmic models only assume iid. variables

THE MODEL'S FIT (1/3)

"A few decades ago (...) the belief in data models was such that even simple precautions such as residual analysis or goodness-of-fit tests were not used" (Breiman 2001, p. 199)

THE MODEL'S FIT (2/3)

- Necessity of checking the model's fit
- Discussion of the fit is superficial
- Most popular: goodness-of-fit tests, residual analysis

The model's fit (3/3)

Goodness-of-Fit Tests

- Not useful if direction of alternative not precisely defined
- Extreme discrepancy to the data is needed

Residual Analysis

- For more than four dimensions: interactions between variables
  $\rightarrow$ manipulation of residual plots

Algorithmic modeling: cross-validation is standard procedure

MULTIPLICITY OF MODELS

- Different models $\rightarrow$ different assumptions
  $\rightarrow$ different conclusions
- Neither model is able to trump
- Further problem: variable selection based on model
- Algorithmic modeling: only iid. assumption

INFERENCE

- Common assumption: $n \to \infty$ never fulfilled
- Testing on 5% level is arbitrary

  (“suspect way to arrive at conclusions”, Breiman 2001, p. 203)
- Algorithmic modeling: no inference

CURSE OF DIMENSIONALITY

- Originally: $n \gg p$
  $\leftrightarrow$ nowadays: $p \gg n$

- Data models become too complex

- Common procedure: reducing dimensionality (e.g. principal component analysis) $\rightarrow$ loss of information

- Algorithmic modeling: the more variables the more information

PREDICTION

- Prediction is more important than interpretation—always
- If prediction is bad, how can interpretation be good?
- Breiman's experience: algorithmic models are best predictors

Breiman's conclusion

- Everyone's choice which model is best

  "The best solution could be an algorithmic model, or maybe a data model, or maybe a combination" (Breiman 2001, p. 206)

- Openness for new methods

DISCUSSION

Bias–variance trade-off

"[The Bias] has to be lurking somewhere inside the theory" (Brad Efron, in Breiman 2001, p. 219)

- In algorithmic modeling, small variance at cost of bias?

- Breiman avoids answer

MULTIPLICITY OF MODELS

- Does not concern prediction
- Just as well in algorithmic models
- Main difference between models: distribution
- Breiman manipulates reader

MODEL ASSUMPTIONS

- Why not use known information (e.g. distribution)?
- Critical iid. assumption in data models and algorithmic models
- Alternatives if iid. assumption is violated?

PREDICTION VERSUS INTERPRETABILITY

- Rivaling abilities of models
- Often interpretation required
- Prediction sometimes indirectly related to data

  "The whole point of science is to open up black boxes, understand their insides, and build better boxes for the purposes of mankind" (Brad Efron, in Breiman 2001, p. 219)

PERSONAL IMPRESSIONS AND CONCLUSION

REFERENCES

Leo Breiman

Statistical Modeling: The Two Cultures.

Statistical Science 16(3), 2001: 199–231.

T. Hastie, R. Tibshirani and J. Friedman

The Elements of Statistical Lernaning. Data Mining, Inference and Prediction.

Heidelberg: Springer, 2009.

QUESTIONS AND DISCUSSION