Better predictions when models are wrong or underspecified

Ommen, M. van

Better Predictions when Models are Wrong or Underspecified

Thijs van Ommen

Better Predictions when Models are Wrong or Underspecified

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 10 juni 2015
klokke 16.15 uur

door

Matthijs van Ommen

geboren te Rotterdam
in 1984

Promotiecommissie

Promotor:
prof.dr. P.D. Grünwald

Overige leden:
prof.dr. R.D. Gill
prof.dr. N.L. Hjort (University of Oslo)
prof.dr. A.W. van der Vaart

These investigations were performed at the Centrum Wiskunde & Informatica (CWI) and were supported by Vici grant 639.073.04 from the Netherlands Organization for Scientific Research (NWO). Part of the work was done while the author was visiting UC San Diego.

Cover design by Gracia Murris
Printed and bound by Ipskamp Drukkers
ISBN: 978-94-6259-689-4

This dissertation is based on the following publications and manuscripts:

- Chapter 2 is based on

  > T. van Ommen. Combining predictions from linear models when training and test inputs differ. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI 2014)*, pages 653–662, 2014.

- Chapters 3, 4 and 5 are based on the technical report

  > P. D. Grünwald and T. van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *arXiv preprint arXiv:1412.3730*, 2014.

- Chapters 6 and 7 are joint, as yet unpublished, work with T. Feenstra, P. D. Grünwald and W. M. Koolen. Chapter 6 is a significant extension of

  > T. E. Feenstra. Conditional prediction without a coarsening at random condition. Master's thesis, Leiden University, 2012. Thesis adviser: P. D. Grünwald.

- Chapter 8 is based on work that is not currently available elsewhere.

Contents

1 Introduction **1**
- 1.1 Regression . . . 3
  - 1.1.1 Extra-sample prediction . . . 5
  - 1.1.2 Bayesian inconsistency . . . 6
  - 1.1.3 Details on XAIC and SafeBayes . . . 8
- 1.2 Probability updating with underspecified distributions . . . 11
  - 1.2.1 The Monty Hall problem . . . 11
  - 1.2.2 Generalizing the problem . . . 12
  - 1.2.3 Our approach . . . 13
- 1.3 Overview of this dissertation . . . 15

2 Extra-Sample Prediction in Linear Models **17**
- 2.1 Introduction . . . 17
  - 2.1.1 Goals of model selection . . . 18
  - 2.1.2 In-sample and extra-sample error . . . 19
  - 2.1.3 Contents . . . 21
- 2.2 Estimating the extra-sample error . . . 21
  - 2.2.1 Preliminaries . . . 21
  - 2.2.2 Main results . . . 22
  - 2.2.3 The $\kappa_{X'}$ and $o(1)$ terms for linear models . . . 23
- 2.3 Model selection for extra-sample prediction . . . 24
  - 2.3.1 Nonfocused versions of XAIC . . . 24
  - 2.3.2 Focused model selection . . . 25
- 2.4 AIC vs. XAIC (k vs. κ_x) in linear models . . . 25
- 2.5 Experiments . . . 27
  - 2.5.1 Description of experiments . . . 27
  - 2.5.2 Results . . . 28
- 2.6 Discussion . . . 32
  - 2.6.1 Relation to the Bayesian predictive distribution . . . 32
  - 2.6.2 Relation to covariate shift methods . . . 33
- 2.7 Conclusions and future work . . . 33
- 2.A Regularity conditions and proofs . . . 34

3 Bayesian Inconsistency under Misspecification **39**
3.1 Introduction . . . 39
3.1.1 Overview of Chapters 3 to 5 . . . 43
3.2 Preliminaries . . . 44
3.2.1 Setting, logarithmic risk, optimal distribution . . . 44
3.2.2 A special case: The linear model . . . 46
3.2.3 KL-associated prediction tasks for the linear model . . . 46
3.3 The generalized posterior . . . 47
3.3.1 Instantiation to linear model selection and averaging . . . 49
3.4 The SafeBayesian algorithm . . . 50
3.4.1 Introducing SafeBayes via the prequential view . . . 50
3.4.2 Instantiating SafeBayes to the linear model . . . 53
3.4.3 SafeBayes learns to predict as well as the optimal distribution . . . 55
3.5 Main experiment: Varying σ^2 . . . 55
3.5.1 Preparing the main experiments: Model, priors, method, 'truth' . . . 56
3.5.2 The statistics we report . . . 57
3.5.3 Main model selection/averaging experiment . . . 60
3.5.4 Second experiment: Ridge regression, varying σ^2 . . . 61
3.5.5 Executive summary: Joint conclusions from main and additional experiments . . . 69

4 Bayesian Inconsistency: Explanations and Discussion **71**
4.1 Bayes' behaviour explained . . . 71
4.1.1 Explanation I: Variance issues . . . 71
4.1.2 Explanation II: Good vs. bad misspecification . . . 73
4.1.3 Hypercompression . . . 75
4.1.4 Explanation III: The mixability gap & the Bayesian belief in concentration . . . 79
4.2 How SafeBayes works . . . 81
4.3 Discussion, open problems and conclusion . . . 84
4.3.1 Related work I: Learning theory and MDL . . . 88
4.3.2 Related work II: Analysis of Bayesian behaviour under misspecification . . . 89
4.3.3 Future work and open problems . . . 90
4.A More on mix loss . . . 93
4.A.1 Implementing SafeBayes . . . 93
4.A.2 Belief in concentration (proof of Theorem 4.1) . . . 94

5 Bayesian Inconsistency: More Experiments **99**
5.1 Experiments on variations of the prior and the model . . . 99
5.1.1 Experiments with fixed σ^2 . . . 99
5.1.2 Slightly informative prior . . . 102
5.1.3 Prior as advised by Raftery et al. . . . 104
5.1.4 The g-prior . . . 105

5.2 Experiments on variations on the method . . . 106
5.2.1 An idea to be explored further: Discounting initial observations . . . 106
5.2.2 Other methods for model selection: AIC, BIC, (generalized) cross-validation . . . 107
5.2.3 Other methods for learning η: Cross-validation on log-loss and on squared loss . . . 108
5.3 Experiments on variations of the truth . . . 108

6 Worst-Case Optimal Probability Updating 113
6.1 Introduction . . . 113
6.1.1 Caveats on the use of the word 'conditioning' . . . 117
6.1.2 Contents . . . 118
6.2 Definitions and problem formulation . . . 118
6.2.1 Strategies . . . 119
6.2.2 Three standard loss functions . . . 122
6.2.3 Notes on our definition . . . 124
6.3 Worst-case optimal strategies for the quizmaster . . . 124
6.3.1 Application to standard loss functions . . . 127
6.4 Worst-case optimal strategies for the contestant . . . 130
6.4.1 Realizable hyperplanes . . . 130
6.4.2 Existence . . . 132
6.4.3 Characterization and nonuniqueness . . . 133
6.5 Results for well-behaved loss functions . . . 135
6.5.1 Proper continuous loss functions . . . 135
6.5.2 Local loss functions . . . 138
6.6 Conclusion . . . 141
6.A Proofs . . . 142

7 Properties of Message Structures in Probability Updating Games 149
7.1 Introduction . . . 149
7.2 Decomposition of games . . . 152
7.2.1 Decomposition and connected games . . . 152
7.2.2 Substitution decomposition and modules . . . 153
7.3 Outcome symmetry . . . 154
7.3.1 Symmetry of loss functions . . . 154
7.3.2 Symmetry of KT-vectors . . . 156
7.4 The RCAR characterization for general loss functions . . . 157
7.4.1 Graph games . . . 157
7.4.2 Matroid games . . . 158
7.4.3 Loss invariance . . . 160
7.5 Finding RCAR strategies . . . 161
7.5.1 Induced colourings . . . 161
7.5.2 A computational procedure . . . 164
7.5.3 Subclasses of matroid games . . . 168
7.6 Discussion and conclusion . . . 169

7.6.1 Connections to CAR . . . 169
7.6.2 Conclusion . . . 171
7.A Proofs . . . 173

8 Algorithms for Probability Updating Games 181
8.1 Introduction . . . 182
8.2 Path graphs and the taut string algorithm . . . 182
8.2.1 Correspondence . . . 183
8.2.2 Algorithm . . . 184
8.3 Intermezzo: Proportional flows . . . 186
8.3.1 Motivating example: Electrical circuits . . . 186
8.3.2 Definitions: Networks and flows . . . 189
8.3.3 Definitions: Proportional and maximum flows . . . 191
8.3.4 Componentwise rescaling of flows . . . 193
8.3.5 The capacitated Edmonds-Gallai decomposition . . . 194
8.3.6 Characterization in terms of lexicographic maximality . 196
8.3.7 Proportional flows and economic fairness . . . 197
8.3.8 Algorithms . . . 197
8.4 General graph games . . . 199
8.4.1 Bipartite graph games . . . 199
8.4.2 Extension to general graph games . . . 200
8.5 Matroid games . . . 201
8.6 Conclusion . . . 205
8.6.1 Future work . . . 205
8.A Proofs . . . 207

Bibliography 217

Index 229

Samenvatting 233

Acknowledgements 237

Curriculum Vitae 239

List of Figures

1.1 A simple example of a regression problem 3
1.2 Bayesian inconsistency in regression 7

2.1 Squared risk of different model selection methods as a function of x when the true function is $f_1(x) = x + 2$ 29
2.2 Squared risk of different model selection methods as a function of x when the true function is $f_2(x) = |x|$ 29

3.1 The conditional expectation $\mathbf{E}[Y \mid X]$ according to Bayes and SafeBayes in a polynomial regression example 41
3.2 The expected squared error risk for Bayes and SafeBayes as a function of sample size 41
3.3 The square-risk, MAP model order, overconfidence (lack of reliability), and selected $\hat{\eta}$ at each sample size for the wrong-model experiment with $p_{\max} = 50$ 62
3.4 Same graphs as in Figure 3.3 for the correct-model experiment with $p_{\max} = 50$ 63
3.5 Same graphs as in Figure 3.3 for the wrong-model experiment with $p_{\max} = 100$ 64
3.6 Same graphs as in Figure 3.3 for the correct-model experiment with $p_{\max} = 100$ 65
3.7 Bayesian ridge regression: Results for model-wrong experiment conditioned on $p := p_{\max} = 50$ 67
3.8 Bayesian ridge regression: Results for model-correct experiment conditioned on $p := p_{\max} = 50$ 68

4.1 Benign vs. bad misspecification 74
4.2 Cumulative standard, R-, and I-log-loss of standard Bayesian prediction for the model-averaging experiment of Figure 3.3 . . 77
4.3 Instantaneous standard, R- and I-log-loss of standard Bayesian prediction for the run depicted in Figure 4.2 77
4.4 Variance of standard Bayes predictive distribution conditioned on a new input S as a function of S after 50 examples for the polynomial model-wrong experiment 78

4.5 Cumulative losses as a function of η for the experiment of Figure 3.3 83

5.1 Bayesian model averaging, fixed σ^2, for the model-wrong experiment of Figure 3.3 101
5.2 Bayesian ridge regression, fixed σ^2, for the model-wrong experiment of Figure 3.7 102
5.3 Square-risk and self-confidence for two different ridge experiments using the slightly informative prior 103
5.4 Square-risk for model averaging and selection based on the g-prior in the model-wrong experiment of Figure 3.3 105
5.5 Square-risk and selected model order for five different model selection methods 107
5.6 Analogue of Figure 3.3 for determining η by leave-one-out cross-validation with square-loss 109
5.7 Analogue of Figure 3.3 for determining η by leave-one-out cross-validation with log-loss 110

6.1 Logarithmic loss and entropy on a binary prediction 123
6.2 Brier loss and entropy on a binary prediction 123
6.3 Randomized 0-1 loss and entropy on a binary prediction 123
6.4 Characterization of the worst-case optimal strategy for the quizmaster in the Monty Hall game with logarithmic loss 128
6.5 Loss and entropy on a binary prediction for the loss function in Example 6.I 135
6.6 Loss and entropy on Δ_{y_1} for the loss function in Example 6.K . . 137

7.1 Overview of classes of message structures 151
7.2 Underlying graphs of the graph games seen in Chapter 6 158
7.3 Examples of messages structures and their induced colourings . 163
7.4 More messages structures and their induced colourings 168
7.5 A uniform multicover with a multiple message 171

8.1 The taut string problem corresponding to a path game 183
8.2 An electrical circuit containing resistors and diodes 187
8.3 Network and augmented network corresponding to circuit . . . 190
8.4 Two flow networks with their proportional maximum flows . . 193
8.5 Schematic of the capacitated Edmonds-Gallai decomposition . . 195
8.6 Steps in the proportional matroid basis packing algorithm for the matroid game of Example 8.A 204
8.7 Steps in the algorithm for the matroid game of Example 8.B . . 204