Model Selection and Model Averaging

Given a data set, you can fit thousands of models at the push of a button, but how do you choose the best? With so many candidate models, overfitting is a real danger. Is the monkey who typed Hamlet actually a good writer?

Choosing a suitable model is central to all statistical work with data. Selecting the variables for use in a regression model is one important example. The past two decades have seen rapid advances both in our ability to fit models and in the theoretical understanding of model selection needed to harness this ability, yet this book is the first to provide a synthesis of research from this active field, and it contains much material previously difficult or impossible to find. In addition, it gives practical advice to the researcher confronted with conflicting results.

Model choice criteria are explained, discussed and compared, including Akaike's information criterion AIC, the Bayesian information criterion BIC and the focused information criterion FIC. Importantly, the uncertainties involved with model selection are addressed, with discussions of frequentist and Bayesian methods. Finally, model averaging schemes, which combine the strengths of several candidate models, are presented.

Worked examples on real data are complemented by derivations that provide deeper insight into the methodology. Exercises, both theoretical and data-based, guide the reader to familiarity with the methods. All data analyses are compatible with open-source `R` software, and data sets and `R` code are available from a companion website.

Gerda Claeskens is Professor in the OR & Business Statistics and Leuven Statistics Research Center at the Katholieke Universiteit Leuven, Belgium.

Nils Lid Hjort is Professor of Mathematical Statistics in the Department of Mathematics at the University of Oslo, Norway.

CAMBRIDGE SERIES IN STATISTICAL AND PROBABILISTIC MATHEMATICS

This series of high-quality upper-division textbooks and expository monographs covers all aspects of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, optimization, and mathematical programming. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books also contain applications and discussions of new techniques made possible by advances in computational practice.

Already published

1. *Bootstrap Methods and Their Application*, by A. C. Davison and D. V. Hinkley
2. *Markov Chains*, by J. Norris
3. *Asymptotic Statistics*, by A. W. van der Vaart
4. *Wavelet Methods for Time Series Analysis*, by Donald B. Percival and Andrew T. Walden
5. *Bayesian Methods*, by Thomas Leonard and John S. J. Hsu
6. *Empirical Processes in M-Estimation*, by Sara van de Geer
7. *Numerical Methods of Statistics*, by John F. Monahan
8. *A User's Guide to Measure Theoretic Probability*, by David Pollard
9. *The Estimation and Tracking of Frequency*, by B. G. Quinn and E. J. Hannan
10. *Data Analysis and Graphics using R*, by John Maindonald and John Braun
11. *Statistical Models*, by A. C. Davison
12. *Semiparametric Regression*, by D. Ruppert, M. P. Wand, R. J. Carroll
13. *Exercises in Probability*, by Loic Chaumont and Marc Yor
14. *Statistical Analysis of Stochastic Processes in Time*, by J. K. Lindsey
15. *Measure Theory and Filtering*, by Lakhdar Aggoun and Robert Elliott
16. *Essentials of Statistical Inference*, by G. A. Young and R. L. Smith
17. *Elements of Distribution Theory*, by Thomas A. Severini
18. *Statistical Mechanics of Disordered Systems*, by Anton Bovier
19. *The Coordinate-Free Approach to Linear Models*, by Michael J. Wichura
20. *Random Graph Dynamics*, by Rick Durrett
21. *Networks*, by Peter Whittle
22. *Saddlepoint Approximations with Applications*, by Ronald W. Butler
23. *Applied Asymptotics*, by A. R. Brazzale, A. C. Davison and N. Reid
24. *Random Networks for Communication*, by Massimo Franceschetti and Ronald Meester
25. *Design of Comparative Experiments*, by R. A. Bailey

Model Selection and Model Averaging

Gerda Claeskens
K.U. Leuven

Nils Lid Hjort
University of Oslo

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521852258

First published 2008

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library.

ISBN 978-0-521-85225-8 hardback

Cambridge University Press
978-0-521-85225-8 - Model Selection and Model Averaging
Gerda Claeskens and Nils Lid Hjort
Frontmatter
More information

To Maarten and Hanne-Sara
– G. C.

To Jens, Audun and Stefan
– N. L. H.

Contents

Preface

Every statistician and data analyst has to make choices. The need arises especially when data have been collected and it is time to think about which model to use to describe and summarise the data. Another choice, often, is whether all measured variables are important enough to be included, for example, to make predictions. Can we make life simpler by only including a few of them, without making the prediction significantly worse?

In this book we present several methods to help make these choices easier. *Model selection* is a broad area and it reaches far beyond deciding on which variables to include in a regression model.

Two generations ago, setting up and analysing a single model was already hard work, and one rarely went to the trouble of analysing the same data via several alternative models. Thus 'model selection' was not much of an issue, apart from perhaps checking the model via goodness-of-fit tests. In the 1970s and later, proper model selection criteria were developed and actively used. With unprecedented versatility and convenience, long lists of candidate models, whether thought through in advance or not, can be fitted to a data set. But this creates problems too. With a multitude of models fitted, it is clear that methods are needed that somehow summarise model fits.

An important aspect that we should realise is that inference following model selection is, by its nature, the second step in a two-step strategy. Uncertainties involved in the first step must be taken into account when assessing distributions, confidence intervals, etc. That such themes have been largely underplayed in theoretical and practical statistics has been called 'the quiet scandal of statistics'. Realising that an analysis might have turned out differently, if preceded by data that with small modifications might have led to a different modelling route, triggers the set-up of *model averaging*. Model averaging can help to develop methods for better assessment and better construction of confidence intervals, p-values, etc. But it comprises more than that.

Each chapter ends with a brief 'Notes on the literature' section. These are not meant to contain full reviews of all existing and related literature. They rather provide some references which might then serve as a start for a fuller search. A preview of the contents of all chapters is provided in Section 1.8.

The methods used in this book are mostly based on likelihoods. To read this book it would be helpful to have at least knowledge of what a likelihood function is, and that the parameters maximising the likelihood are called maximum likelihood estimators. If properties (such as an asymptotic distribution of maximum likelihood estimators) are needed, we state the required results. We further assume that readers have had at least an applied regression course, and have some familiarity with basic matrix computations.

This book is intended for those interested in model selection and model averaging. The level of material should be accessible to master students with a background in regression modelling. Since we not only provide definitions and worked out examples, but also give some of the methodology behind model selection and model averaging, another audience for this book consists of researchers in statistically oriented fields who wish to understand better what they are doing when selecting a model. For some of the statements we provide a derivation or a proof. These can easily be skipped, but might be interesting for those wanting a deeper understanding. Some of the examples and sections are marked with a star. These contain material that might be skipped at a first reading.

This book is suitable for teaching. Exercises are provided at the end of each chapter. For many examples and methods we indicate how they can be applied using available software. For a master's level course, one could leave out most of the derivations and select the examples depending on the background of the students. Sections which can be skipped in such a course would be the large-sample analysis of Section 5.2, the average and Bayesian focussed information criteria of Sections 6.9 and 6.10, and the end of Chapter 7 (Sections 7.8, 7.9). Chapter 9 (certainly to be included) contains worked out practical examples.

All data sets used in this book, along with various computer programs (in `R`) for carrying out estimation and model selection via the methods we develop, are available at the following website: `www.econ.kuleuven.be/gerda.claeskens/public/modelselection`.

Model selection and averaging are unusually broad areas. This is witnessed by an enormous and still expanding literature. The book is not intended as an encyclopaedia on this topic. Not all interesting methods could be covered. More could be said about models with growing numbers of parameters, finite-sample corrections, time series and other models of dependence, connections to machine learning, bagging and boosting, etc., but these topics fell by the wayside as the other chapters grew.

Acknowledgements

The authors deeply appreciate the privileges afforded to them by the following university departments by creating possibilities for meeting and working together in environments conducive to research: School of Mathematical Sciences at the Australian

National University at Canberra; Department of Mathematics at the University of Oslo; Department of Statistics at Texas A&M University; Institute of Statistics at Université Catholique de Louvain; and ORSTAT and the Leuven Statistics Research Center at the Katholieke Universiteit Leuven. N. L. H. is also grateful to the Centre of Advanced Studies at the Norwegian Academy of Science and Letters for inviting him to take part in a one-year programme on biostatistical research problems, with implications also for the present book.

More than a word of thanks is also due to the following individuals, with whom we had fruitful occasions to discuss various aspects of model selection and model averaging: Raymond Carroll, Merlise Clyde, Anthony Davison, Randy Eubank, Arnoldo Frigessi, Alan Gelfand, Axel Gandy, Ingrid Glad, Peter Hall, Jeff Hart, Alex Koning, Ian McKeague, Axel Munk, Frank Samaniego, Willi Sauerbrei, Tore Schweder, Geir Storvik, and Odd Aalen.

We thank Diana Gillooly of Cambridge University Press for her advice and support.

The first author thanks her husband, Maarten Jansen, for continuing support and interest in this work, without which this book would not be here.

Gerda Claeskens and Nils Lid Hjort
Leuven and Oslo

A guide to notation

This is a list of most of the notation used in this book. The page number refers either to the first appearance or to the place where the symbol is defined.

AFIC	average-weighted focussed information criterion	181
AIC	Akaike information criterion	28
AIC_c	corrected AIC	46
$\mathrm{aic}_n(m)$	AIC difference $\mathrm{AIC}(m) - \mathrm{AIC}(0)$	229
a.s.	abbreviation for 'almost surely'; the event considered takes place with probability 1	
BFIC	Bayesian focussed information criterion	186
BIC	Bayesian information criterion	70
BIC^*	alternative approximation in the spirit of BIC	80
$\mathrm{BIC}^{\mathrm{exact}}$	the quantity that the BIC aims at approximating	79
cAIC	conditional AIC	271
$c(S)$, $c(S \mid D)$	weight given to the submodel indexed by the set S when performing model average estimation	193
D	limit version of D_n, with distribution $\mathrm{N}_q(\delta, Q)$	148
D_n	equal to $\sqrt{n}(\widehat{\gamma} - \gamma_0)$	125
dd	deviance difference	91
DIC	deviance information criterion	91
E, E_g	expected value (with respect to the true distribution), sometimes explicitly indicated via a subscript	24
FIC	focussed information criterion	147
FIC^*	bias-modified focussed information criterion	150
$g(y)$	true (but unknown) density function of the data	24
g	the link function in GLM	47
GLM	generalised linear model	46
G_S	matrix of dimension $q \times q$, related to J	146

$h(\cdot)$	hazard rate	85
$H(\cdot)$	cumulative hazard rate	85
I_q	identity matrix of size $q \times q$	
$I(y,\theta)$, $I(y \mid x,\theta)$	second derivative of log-density with respect to θ	26
i.i.d.	abbreviation for 'independent and identically distributed'	
infl	influence function	51
J	expected value of minus $I(Y,\theta_0)$, often partitioned in four blocks	26, 127
J_S	submatrix of J of dimension $(p+\lvert S\rvert)\times(p+\lvert S\rvert)$.	146
J_n, K_n	finite-sample version of J and K	27, 153
$\widehat{J}$, $\widehat{K}$	J_n and K_n but with estimated parameters	27
K	variance of $u(Y,\theta_0)$	26
KL	Kullback–Leibler distance	24
$\mathcal{L}$, $\mathcal{L}_n$	likelihood function	24
ℓ, ℓ_n	log-likelihood function	24
mAIC	marginal AIC	270
MDL	minimum description length	94
mse	mean squared error	12, 149
n	sample size	23
$\mathrm{N}(\xi,\sigma^2)$	normal distribution with mean ξ and standard deviation σ	
$\mathrm{N}_p(\xi,\Sigma)$	p-variate normal distribution with mean vector ξ and variance matrix Σ	
narr	indicating the 'narrow model', the smallest model under consideration	120
$O_P(z_n)$	of stochastic order z_n; that $X_n = O_p(z_n)$ means that X_n/z_n is bounded in probability	
$o_P(z_n)$	that $X_n = o_p(z_n)$ means that X_n/z_n converges to zero in probability	
P	probability	
p	most typically used symbol for the number of parameters common to all models under consideration, i.e. the number of parameters in the narrow model	
p_D	part of the penalty in the DIC	91
q	most typically used symbol for the number of additional parameters, so that p is the number of parameters in the narrow model and $p+q$ the number of parameters in the wide model	

Q	the lower-right block of dimension $q \times q$ in the partitioned matrix J^{-1}	127
REML	restricted maximum likelihood, residual maximum likelihood	271
S	subset of $\{1, \ldots, q\}$, used to indicate a submodel	
se	standard error	
SSE	error sum of squares	35
TIC	Takeuchi's information criterion, model-robust AIC	43
Tr	trace of a matrix, i.e. the sum of its diagonal elements	
$u(y, \theta)$, $u(y \mid x, \theta)$	score function, first derivative of log-density with respect to θ	26
$U(y)$	derivative of $\log f(y, \theta, \gamma_0)$ with respect to θ, evaluated at (θ_0, γ_0)	50, 122
$V(y)$	derivative of $\log f(y, \theta_0, \gamma)$ with respect to γ, evaluated at (θ_0, γ_0)	50, 122
Var	variance, variance matrix (with respect to the true distribution)	
wide	indicating the 'wide' or full model, the largest model under consideration	120
x, x_i	often used for 'protected' covariate, or vector of covariates, with x_i covariate vector for individual no. i	
z, z_i	often used for 'open' additional covariates that may or may not be included in the finally selected model	
δ	vector of length q, indicating a certain distance	121
θ_0	least false (best approximating) value of the parameter	25
Λ	limiting distribution of the weighted estimator	196
Λ_S	limiting distribution of $\sqrt{n}(\widehat{\mu}_S - \mu_{\mathrm{true}})$	148
μ	focus parameter, parameter of interest	120, 146
π_S	$\lvert S\rvert \times q$ projection matrix that maps a vector v of length q to v_S of length $\lvert S\rvert$	
τ_0	standard deviation of the estimator in the smallest model	123
$\phi(u)$	the standard normal density	
$\phi(u, \sigma^2)$	the density of a normal random variable with mean zero and variance σ^2, $\mathrm{N}(0, \sigma^2)$	
$\Phi(u)$	the standard normal cumulative distribution function	

Cambridge University Press
978-0-521-85225-8 - Model Selection and Model Averaging
Gerda Claeskens and Nils Lid Hjort
Frontmatter
More information

$\phi(x, \Sigma)$	the density of a multivariate normal $N_q(0, \Sigma)$ variable	
$\chi^2_q(\lambda)$	non-central χ^2 distribution with q degrees of freedom and non-centrality parameter λ, with mean $q + \lambda$ and variance $2q + 4\lambda$	126
ω	vector of length q appearing in the asymptotic distribution of estimators under local misspecification	123
$\xrightarrow{d}$, $\to_d$	convergence in distribution	
$\xrightarrow{p}$, $\to_p$	convergence in probability	
$\sim$	'distributed according to'; so $Y_i \sim \mathrm{Pois}(\xi_i)$ means that Y_i has a Poisson distribution with parameter ξ_i	
$\doteq_d$	$X_n \doteq_d X'_n$ indicates that their difference tends to zero in probability	