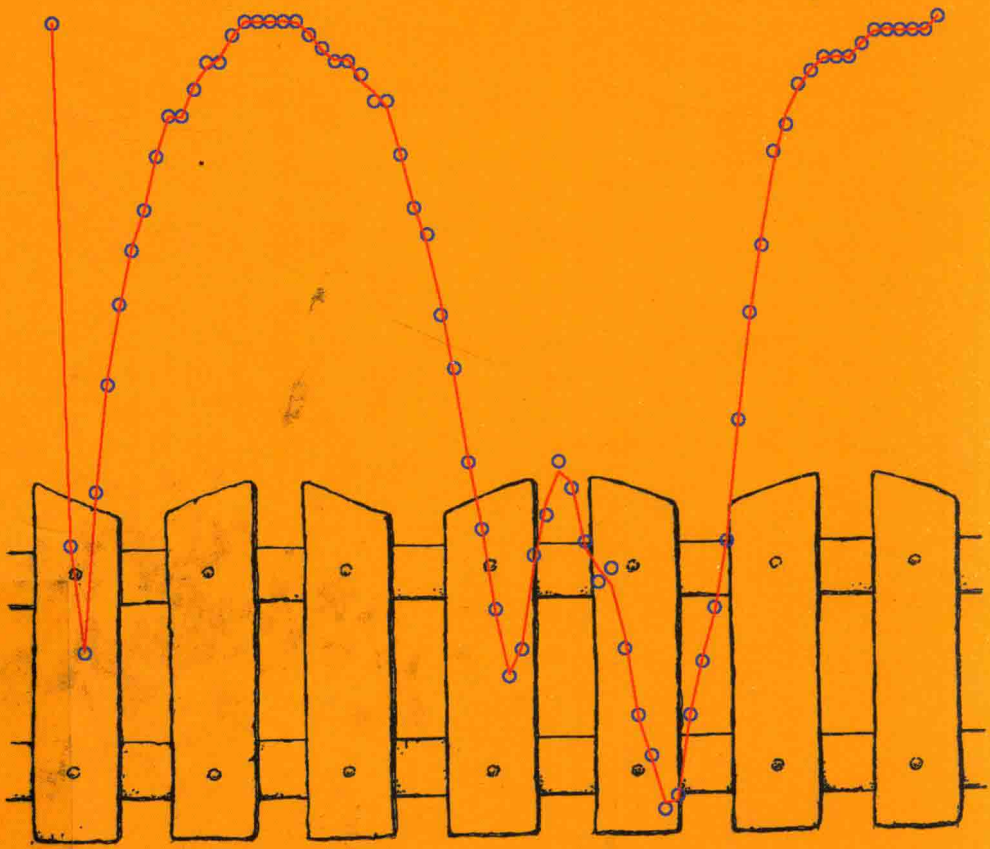
The Fence Methods

Jiming Jiang
Thuan Nguyen



World Scientific

The Fence Methods

Jiming Jiang
University of California, Davis, USA

Thuan Nguyen
Oregon Health & Science University, USA

World Scientific
NEW JERSEY · LONDON · SINGAPORE · BEIJING · SHANGHAI · HONG KONG · TAIPEI · CHENNAI · TOKYO

Published by

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601

UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data

Jiang, Jiming.

The fence methods / by Jiming Jiang (University of California, Davis, USA), Thuan Nguyen (Oregon Health & Science University, USA).

pages cm

Includes bibliographical references and index.

ISBN 978-9814596060 (hardcover : alk. paper)

1. Partially ordered sets. 2. Group theory. 3. Multiple criteria decision making. I. Nguyen, Thuan (Professor of biostatistics) II. Title.

QA171.485.J53 2015

519.5--dc23

2015026036

British Library Cataloguing-in-Publication Data

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

Copyright © 2016 by World Scientific Publishing Co. Pte. Ltd.

All rights reserved. This book, or parts thereof, may not be reproduced in any form or by any means, electronic or mechanical, including photocopying, recording or any information storage and retrieval system now known or to be invented, without written permission from the publisher.

For photocopying of material in this volume, please pay a copying fee through the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, USA. In this case permission to photocopy is not required from the publisher.

Printed by FuIsland Offset Printing (S) Pte Ltd Singapore

The Fence Methods

To beloved children living in poverty

此为试读,需要完整PDF请访问: www.ertongbook.com

Preface

The past two decades have seen an explosion of interest in statistical model selection that is largely driven by practical needs. The traditional methods of model selection, a core of which are the information criteria, encounter difficulties in dealing with high-dimensional and complex problems.

The main difficulty brought by the high-dimensional problems is computational. For example, in regression variable selection when the number of candidate variable, p, is large, it is computationally expensive, or even infeasible, to carry out all-subset selections, as required by the information criteria. Furthermore, when p is larger than n, the sample size, the standard method of fitting the least squares, which is needed in computing the information criterion function, is not possible. There have been major breakthroughs in high-dimensional model selection, thanks to the proposal of shrinkage selection/estimation methods [Tibshirani (1996), Fan and Li (2001), among others].

On the other hand, complex selection problems are often encountered. For example, in many problems of practical interest, the observations are correlated in complex ways. Mixed effects models, such as linear and generalized linear mixed models, have been used in analyzing such complex data, but little was known about model selection in such situations. In particular, the use of information criteria was largely *ad hoc* for mixed model selection. In some other cases, the distribution of data is not fully specified. Here, once again, one encounters difficulty in using the information criteria, because the likelihood function is not available. Another complication came when the data involve missing values, which occurs frequently in practice. The standard methods for model selection, including associated software package, were built for the complete-data situations. As such, these methods and software cannot be directly applied to cases of missing

or incomplete data.

In a breakthrough in complex model selection, Jiang, Rao, Gu and Nguyen [Jiang *et al.* (2008)] proposed a new class of strategies for model selection, which they coined *fence methods*. The idea consists of a procedure to isolate a subgroup of candidate models, known as "correct models", by building a statistical fence, or barrier. Once the fence is built, an optimal model is selected from those within the fence according to a criterion that can be made flexible. In particular, the criterion of optimality can take practical considerations into account. A number of variations of the fence have since been developed. The fence was motivated by the need to overcome the difficulties encountered by the information criteria. Major features of the fence include (i) flexibility in choosing both the measure of lack-of-fit and the criterion of optimality for selection within the fence; (ii) it is data-driven, giving the data plenty of opportunities to "speak out" in making some difficult decisions, namely, the choice of tuning parameters; and (iii) it leaves a room for practical considerations that is specialized to the current problem. A recent review by Jiang (2014) has provided an overview of the fence, including major advances, applications, and open problems. A software package, recently developed by T. Nguyen, J. Zhao, J. S. Rao and J. Jiang and available online at http://fencemethods.com/, has implemented most of the methods associated with the fence.

This monograph is devoted to giving a detailed account of the fence, including its variations, and related topics. It is mainly intended for researchers and graduate students, at M.S. or higher level. The monograph is mostly self-contained. A first course in mathematical statistics, the ability to use computer for data analysis, and familiarity with calculus and linear algebra are prerequisites.

Our research on complex model selection was first initiated by Dr. J. Sunil Rao, who brought up the problem of mixed model selection in the late 1990s. The collaborative research between Dr. Rao and Dr. Jiming Jiang has led to the 2008 paper that, for the first time, introduced the fence methods. Part of the topics presented in the monograph is based on the Ph.D. dissertation by Dr. Thuan Nguyen, who has made important contributions to the development of the fence. Other former students who have contributed, at various points, to the developments include former Ph.D. students Zhonghua Gu, Jiani Mu, Senke Chen, and Erin Melcon, and former M.S. students Mei-Chin Lin, Jianyang Zhao, Xi Ai, Xiaoyun Wang, and Haomiao Meng. In addition, we would like to thank Qui Tran for initiating the LaTex typesetting for the monograph, and Pete Scully

and Michael Lin for helping with the design of the front cover. Our thanks also go to Professors Alan Welsh and Samuel Müller, who invited the authors to visit their institutes in Australia in 2013 that has led to many constructive discussions, especially related to the fence methods. We also thank Professors Welsh and Müller, Professors Partha Lahiri, Danny Pfeffermann, and J. N. K. Rao, and Dr. Long Ngo for their comments, and encouragement, on the fence methods, either in their published articles and books or through personal communications.

Jiming Jiang and Thuan Nguyen
Davis, California and Portland, Oregon
March 2015

Contents

此为试读,需要完整PDF请访问: www.ertongbook.com

Chapter 1

Introduction

On the morning of March 16, 1971, Hirotugu Akaike, as he was taking a seat on a commuter train, came out with the idea of a connection between the relative Kullback-Liebler discrepancy and the empirical log-likelihood function, a procedure that was later named Akaike's information criterion, or AIC [Akaike (1973, 1974); see Bozdogan (1994) for the historical note]. The idea has allowed major advances in model selection and related fields. See, for example, de Leeuw (1992).

To introduce the idea of a new model selection strategy, it is important to understand the "old" strategies, at the center of which are the information criteria. Below we provide a brief review of the criteria. But before we do this, let us keep in mind one of the best-known quotes in Statistics, or perhaps all of Science. George Box, one of the most influential statisticians of the 20th century, once wrote that "essentially, all models are wrong, but some are useful." What it means is that, even though there may not exist a "true" model, in reality, a suitable choice of one may still provide a good (or, perhaps, the best) approximation, from a practical standpoint.

1.1 The information criteria

Suppose that one wishes to approximate an unknown probability density function (pdf), g, by a given pdf, f. The Kullback–Leibler (K-L) discrepancy, or information, defined as

$$I(g;f) = \int g(x) \log g(x)\, dx - \int g(x) \log f(x)\, dx, \tag{1.1}$$

provides a measure of lack of approximation. It can be shown, by Jensen's inequality, that the K-L information is always nonnegative, and it equals

zero if and only if $f = g$ a.e. [i.e., $f(x) = g(x)$ for all x except on a set of Lebesgue measure zero]. However, K-L information is not a distance (Exercise 1.1). Note that the first term on the right side of (1.1) does not depend on f. Therefore, to best approximate g, one needs to find an f that minimizes $-\int g(x) \log f(x)\, dx = -\mathrm{E}_g\{\log f(X)\}$, where E_g means that the expectation is taken with $X \sim g$. Since we do not know g, the expectation is not computable. However, suppose that we have independent observations $X_1, \ldots, X_n$ from g. Then we may replace the expectation by the sample mean, $n^{-1} \sum_{i=1}^{n} \log f(X_i)$, which is an unbiased estimator for the expectation. In particular, under a parametric model, denoted by M, the pdf f depends on a vector θ_M of parameters, denoted by $f = f_M(\cdot|\theta_M)$. For example, in a linear regression model, M may correspond to a subset of predictors, and θ_M the vector of corresponding regression coefficients. Then the AIC is a two-step procedure. The first step is to find the θ_M that minimizes

$$-\frac{1}{n} \sum_{i=1}^{n} \log f_M(X_i|\theta_M) \tag{1.2}$$

for any given M. Note that (1.2) is simply the negative log-likelihood function under M. Therefore, the θ_M that minimizes (1.2) is the maximum likelihood estimator (MLE), denoted by $\hat{\theta}_M$. Then, the second step of AIC is to find the model M that minimizes

$$-\frac{1}{n} \sum_{i=1}^{n} \log f_M(X_i|\hat{\theta}_M). \tag{1.3}$$

However, there is a serious drawback in this approach: Expression (1.3) is no longer an unbiased estimator for $-\mathrm{E}_g\{\log f_M(X|\theta_M)\}$ due to overfitting. The latter is caused by double-use of the same data—for estimating the expected log-likelihood and for estimating the parameter vector θ_M. Akaike (1973) proposed to rectify this problem by correcting the bias, which is

$$\frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_g\{\log f_M(X_i|\hat{\theta}_M)\} - \mathrm{E}_g\{\log f_M(X|\theta_M)\}.$$

He showed that, asymptotically, the bias can be approximated by $|M|/n$, where $|M|$ denotes the dimension of M defined as the number of estimated parameters under M. For example, if M is an ARMA(p, q) model in time series [e.g., Shumway (1988)], then $|M| = p + q + 1$ (the 1 corresponds to the unknown variance). Thus, a term $|M|/n$ is added to (1.3), leading to

$$-\frac{1}{n} \sum_{i=1}^{n} \log f_M(X_i|\hat{\theta}_M) + \frac{|M|}{n}.$$

The expression is then multiplied by the factor $2n$, which does not depend and affect the choice of M, to come up with the AIC:

$$\text{AIC}(M) = -2\sum_{i=1}^{n} \log f_M(X_i|\hat{\theta}_M) + 2|M|. \tag{1.4}$$

In words, the AIC is minus two times the maximized log-likelihood plus two times the number of estimated parameters.

A number of similar criteria have been proposed since the AIC. These include the Bayesian information criterion [BIC; Schwarz (1978)], and a criterion due to Hannan and Quinn (1979). All of these criteria may be expressed, in a general form, as

$$\text{GIC}(M) = \hat{D}_M + \lambda_n|M|, \tag{1.5}$$

where $\hat{D}_M$ is a measure of lack-of-fit by the model M and λ_n is a penalty for complexity of the model, which may depend on the effective sample size, n. The measure of lack-of-fit is such that a model of greater complexity fits better, therefore has a smaller $\hat{D}_M$; on the other hand, such a model receives more penalty for having a larger $|M|$. The effective sample size is equal to the sample size, if the samples are i.i.d.; otherwise, it might not be the same as the sample size (see below). Therefore, criterion (1.5), known as the generalized information criterion, or GIC [Nishii (1984); Shibata (1984)], is a trade-off between model fit and model complexity. In particular, the AIC corresponds to (1.5) with $\hat{D}_M$ being -2 times the maximized log-likelihood and $\lambda_n = 2$; the BIC and HQ have the same $\hat{D}_M$, but $\lambda_n = \log n$ and $c \log\log n$, respectively, where c is a constant greater than 2. We consider another special cases below.

Example 1.1: Hurvich and Tsai (1989) argued that in the case of the ARMA(p, q) model, a better bias correction could be obtained if one replaces $p+q+1$ by an asymptotically equivalent quantity, $n(p+q+1)/(n-p-q-2)$. This leads to a modified criterion known as AICC. The AICC corresponds to (1.5) with $\lambda_n = 2n/(n-p-q-2)$. So, if $n \to \infty$ while the ranges of p and q are bounded, AICC is asymptotically equivalent to AIC.

One concern about AIC is that it does not lead to consistent model selection if the dimension of the optimal model is finite. Here, an optimal model is defined as a correct model with minimum dimension. For example, suppose that the true underlying model is AR(2); then AR(3) is a correct model (by letting the coefficient corresponding to tha additional term equal to zero), but not an optimal model. On the other hand, AR(1) is an incorrect model, or wrong model [that the true underlying model is AR(2)

implies that the true coefficient corresponding to the second-order term is nonzero]. So, if one considers all AR models as candidates, the only optimal model is AR(2). Furthermore, consistency of model selection is defined as that the probability of selecting an optimal model goes to 1 as $n \to \infty$.

On the other hand, the BIC and HG are consistent model selection procedures. One may wonder what causes such a difference. The idea is quite simple, and it has something to do with the choice of λ_n. The AIC is not consistent because it has not given enough penalty for complex models. For example, suppose that the true underlying model is AR(p). Then AIC tends to choose an order higher than p in selecting the order for the AR model. This problem is called overfitting. It can be shown that AIC does not have the other kind of problem, namely, underfitting, meaning that the procedure tends to select an order less than p. In other words, asymptotically, AIC is expected to select, at least, a correct model; but the selected model may not be optimal in that it can be further simplified. For the same reason, AICC is not consistent.

For a procedure to be consistent, one needs to control both overfitting and underfitting. Thus, on the one hand, one needs to increase the penalty λ_n in order to reduce overfitting; on the other hand, one cannot overdo this; otherwise, the underfitting will again make the procedure inconsistent. The question then is: What is the "right" amount of penalty? As far as the consistency is concerned, this is determined by the order of λ_n. It turns out that BIC and HG have the right order for λ_n that guarantees the consistency. See, for example, Jiang (2010) for further explanation.

Furthermore, the lack of consistency does not necessarily imply that AIC is inferior to BIC or HQ, from a practical standpoint. The reason is that the concept of consistency, as defined above, applies only to the "ideal" situation, where the true underlying model is of finite dimension, and is among the candidate models. In practice, however, such an ideal situation almost never occurs (remember "all models are wrong"). What one has instead is a collection of candidate models as approximation to the true underlying model, which is not one of the candidates. For example, in time series analysis, one may use an AR(p) model as an approximation to the true underlying model, which may be expressed as AR(∞). In such a case, it may be argued that the BIC and HQ are inconsistent, while the AIC is consistent in the sense that the selected order (of the AR model) tends to infinity as the sample size increases, which approximates the order of the true model.

1.2 Difficulties with the information criteria

Although the information criteria are broadly used, difficulties are often encountered, especially in some non-conventional situations. We discuss a number of such cases below.

1. The effective sample size. As mentioned, the λ_n that is involved in the information criteria, (1.5), may depend on n, which is supposed to be the effective sample size. For example, if the data are i.i.d., the effective sample size should be the same as the sample size, because every new observation provides, in a way, the same amount of new information. On the other hand, if all the data points are identical, the effective sample size should be 1, regardless of the number of observations, because every new observation provides no additional information. Of course, the latter case is a bit extreme, but there are many practical situations where the observations are correlated, even though they are not identical. One of those situations is mixed effects models. We illustrate with a simple example.

Example 1.2: Consider a linear mixed model defined as $y_{ij} = x'_{ij}\beta + u_i + v_j + e_{ij}$, $i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$, where x_{ij} is a vector of known covariates, β is a vector of unknown regression coefficients (the fixed effects), u_i, v_j are random effects, and e_{ij} is an additional error. It is assumed that u_i's, v_j's and e_{ij}'s are independent, and that, for the moment, $u_i \sim N(0, \sigma_u^2)$, $v_j \sim N(0, \sigma_v^2)$, $e_{ij} \sim N(0, \sigma_e^2)$. It is well-known (e.g., Harville 1977, Miller 1977) that, in this case, the effective sample size for estimating σ_u^2 and σ_v^2 is not the total sample size $m_1 \cdot m_2$, but m_1 and m_2, respectively, for σ_u^2 and σ_v^2. Now suppose that one wishes to select the fixed covariates, which are components of x_{ij}, under the assumed model structure, using the BIC. It is not clear what should be in place of $\lambda_n = \log n$. For example, it does not make sense to let $n = m_1 \cdot m_2$.

2. The dimension of a model. Not only the effective sample size, the dimension of a model, $|M|$, can also cause difficulties. In some cases, such as the ordinary linear regression, this is simply the number of parameters under M, but in other situations where nonlinear, adaptive models are fitted, this can be substantially different. Ye (1998) developed the concept of generalized degrees of freedom (gdf) to track model complexity. A computational algorithm at heart, the method simply repeats the model fitting on perturbed values of the response, y (via resampling), and observes how the fitted values, $\hat{y}$, change. The sum of the sensitivities to change across all of the observations provides an approximation to the model complexity. It can be shown that, in the case of ordinary linear regression, this results

此为试读,需要完整PDF请访问: www.ertongbook.com