Lecture Notes in Mathematics 1896

Pascal Massart

Concentration Inequalities and Model Selection

Ecole d'Eté de Probabilités
de Saint-Flour XXXIII - 2003

Editor: Jean Picard

Springer

Author

Pascal Massart

Département de Mathématique
Université de Paris-Sud
Bât 425
91405 Orsay Cedex
France
e-mail: pascal.massart@math.u-psud.fv

Editor

Jean Picard

Laboratoire de Mathématiques Appliquées
UMR CNRS 6620
Université Blaise Pascal (Clermont-Ferrand)
63177 Aubière Cedex
France
e-mail: jean.picard@math.univ-bpclermont.fr

Cover: Blaise Pascal (1623-1662)

Library of Congress Control Number: 2007921691

Mathematics Subject Classification (2000): 60C05, 60E15, 62F10, 62B10, 62E17, 62G05, 62G07, 62G08, 62J02, 94A17

ISSN print edition: 0075-8434
ISSN electronic edition: 1617-9692
ISSN Ecole d'Eté de Probabilités de St. Flour, print edition: 0721-5363
ISBN-10 3-540-48497-3 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-48497-4 Springer Berlin Heidelberg New York
DOI 10.1007/978-3-540-48503-2

Typesetting by the author and SPi using a Springer LaTeX macro package

Cover design: *design & production* GmbH, Heidelberg

Printed on acid-free paper SPIN: 11917472 VA41/3100/SPi 5 4 3 2 1 0

Foreword

Three series of lectures were given at the 33rd Probability Summer School in Saint-Flour (July 6–23, 2003) by Professors Dembo, Funaki, and Massart. This volume contains the course of Professor Massart. The courses of Professors Dembo and Funaki have already appeared in volume 1869 (see below). We are grateful to the author for his important contribution.

Sixty-four participants have attended this school. Thirty-one have given a short lecture. The lists of participants and short lectures are enclosed at the end of the volume.

The Saint-Flour Probability Summer School was founded in 1971. Here are the references of Springer volumes where lectures of previous years were published. All numbers refer to the *Lecture Notes in Mathematics* series, expect S-50 which refers to volume 50 of the *Lecture Notes in Statistics* series.

1971: vol 307	1980: vol 929	1990: vol 1527	1998: vol 1738
1973: vol 390	1981: vol 976	1991: vol 1541	1999: vol 1781
1974: vol 480	1982: vol 1097	1992: vol 1581	2000: vol 1816
1975: vol 539	1983: vol 1117	1993: vol 1608	2001: vol 1837 & 1851
1976: vol 598	1984: vol 1180	1994: vol 1648	2002: vol 1840 & 1875
1977: vol 678	1985/86/87: vol 1362 & S-50	1995: vol 1690	2003: vol 1869 & 1896
1978: vol 774	1988: vol 1427	1996: vol 1665	2004: vol 1878 & 1879
1979: vol 876	1989: vol 1464	1997: vol 1717	2005: vol 1897

Further details can be found on the summer school web site
`http://math.univ-bpclermont.fr/stflour/`

Jean Picard
Clermont-Ferrand, April 2006

Preface

These notes would have never existed without the efforts of a number of people whom I would like to warmly thank. First of all, I would like to thank Lucien Birgé. We have spent hours working on model selection, trying to understand what was going on in depth. In these notes, I have attempted to provide a significant account of the nonasymptotic theory that we have tried to build together, year after year. Through our works we have promoted a nonasymptotic approach in statistics which consists in taking the number of observations as it is and try to evaluate the effect of all the influential parameters. At this very starting point, it seems to me that it is important to provide a first answer to the following question: why should we be interested by a nonasymptotic view for model selection at all? In my opinion, the motivation should neither be a strange interest for "small" sets of data nor a special taste for constants and inequalities rather than for limit theorems (although since mathematics is also a matter of taste, it is a possible way for getting involved in it!). On the contrary, the nonasymptotic point of view may turn out to be especially relevant when the number of observations is large. It is indeed to fit large complex sets of data that one needs to deal with possibly huge collections of models at different scales. The nonasymptotic approach for model selection precisely allows the collection of models together with their dimensions to vary freely, letting the dimensions be possibly of the same order of magnitude as the number of observations.

More than ten years ago we have been lucky enough to discover that concentration inequalities were indeed the probabilistic tools that we needed to develop a nonasymptotic theory for model selection. This offered me the opportunity to study this fascinating topic, trying first to understand the impressive works of Michel Talagrand and then taking benefits of Michel Ledoux's efforts to simplify some of Talagrand's arguments to bring my own contribution. It has been a great pleasure for me to work with Gábor Lugosi and Stéphane Boucheron on concentration inequalities. Most of the material which is presented here on this topic comes from our joint works.

Sharing my enthusiasm for these topics with young researchers and students has always been a strong motivation for me to work hard. I would like all of them to know how important they are to me, because not only seeing light in their eyes brought me happiness but also their theoretical works or their experiments have increased my level of understanding of my favorite topics. So many thanks to Sylvain Arlot, Yannick Baraud, Gilles Blanchard, Olivier Bousquet, Gwenaelle Castellan, Magalie Fromont, Jonas Kahn, Béatrice Laurent, Marc Lavarde, Emilie Lebarbier, Vincent Lepez, Frédérique Letué, Marie-Laure Martin, Bertrand Michel, Elodie Nédélec, Patricia Reynaud, Emmanuel Rio, Marie Sauvé, Christine Tuleau, Nicolas Verzelen, and Laurent Zwald.

In 2003, I had this wonderful opportunity to teach a course on concentration inequalities and model selection at the St Flour Summer school but before that I have taught a similar course in Orsay during several years. I am grateful to all the students who followed this course and whose questions have contributed to improve on the contents of my lectures.

Last but not least, I would like to thank Jean Picard for his kindness and patience and all the people who accepted to read my first draft. Of course the remaining mistakes or clumsy turns of phrase are entirely under my responsibility but (at least according to me) their comments and corrections have much improved the level of readability of these notes. You have been often too kind, sometimes pitiless and always careful and patient readers, so many thanks to all of you: Sylvain Arlot, Yannick Baraud, Lucien Birgé, Gilles Blanchard, Stéphane Boucheron, Laurent Cavalier, Gilles Celeux, Jonas Kahn, Frédérique Letué, Jean-Michel Loubes, Vincent Rivoirard, and Marie Sauvé.

Abstract

Model selection is a classical topic in statistics. The idea of selecting a model via penalizing a log-likelihood type criterion goes back to the early 1970s with the pioneering works of Mallows and Akaike. One can find many consistency results in the literature for such criteria. These results are asymptotic in the sense that one deals with a given number of models, and the number of observations tends to infinity. We shall give an overview of a nonasymptotic theory for model selection which has emerged during these last ten years. In various contexts of function estimation it is possible to design penalized log-likelihood type criteria with penalty terms depending not only on the number of parameters defining each model (as for the classical criteria) but also on the "complexity" of the whole collection of models to be considered. The performance of such a criterion is analyzed via nonasymptotic risk bounds for the corresponding penalized estimator which expresses that it performs almost as well as if the "best model" (i.e., with minimal risk) were known. For practical relevance of these methods, it is desirable to get a precise expression of the penalty terms involved in the penalized criteria on which they are based. That is why this approach heavily relies on concentration inequalities, the prototype being Talagrand's inequality for empirical processes. Our purpose is to give an account of the theory and discuss some selected applications such as variable selection or change points detection.

Contents