Frontiers in Probability and the Statistical Sciences

More information about this series at http://www.springer.com/series/11957

Dimitris N. Politis

Model-Free Prediction and Regression

A Transformation-Based Approach to Inference

Springer

Dimitris N. Politis
Department of Mathematics
University of California, San Diego
La Jolla, CA, USA

Frontiers in Probability and the Statistical Sciences
ISBN 978-3-319-21346-0 ISBN 978-3-319-21347-7 (eBook)
DOI 10.1007/978-3-319-21347-7

Library of Congress Control Number: 2015948372

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Für die zwei Violinen und die Viola des Model-Freien Quartetts

Preface

Prediction has been one of the earliest forms of statistical inference. The emphasis on parametric estimation and testing seems to only have occurred about 100 years ago; see Geisser (1993) for a historical overview. Indeed, parametric models served as a cornerstone for the foundation of Statistical Science in the beginning of the twentieth century by R.A. Fisher, K. Pearson, J. Neyman, E.S. Pearson, W.S. Gosset (also known as "Student"), etc.; their seminal developments resulted into a complete theory of statistics that could be practically implemented using the technology of the time, i.e., pen and paper (and slide-rule!).

While some models are inescapable, e.g., modeling a polling dataset as a sequence of independent Bernoulli random variables, others appear contrived, often invoked for the sole reason to make the mathematics work. As a prime example, the ubiquitous—and typically unjustified—assumption of Gaussian data permeates statistics textbooks to the day. Model criticism and diagnostics were developed as a practical way out; see Box (1976) for an account of the model-building process by one of the pioneers of applied statistics.

With the advent of widely accessible powerful computing in the late 1970s, computer-intensive methods such as resampling and cross-validation created a revolution in modern statistics. Using computers, statisticians became able to analyze big datasets for the first time, paving the way towards the "big data" era of the twenty-first century. But perhaps more important was the realization that the way we do the analysis could/should be changed as well, as practitioners were gradually freed from the limitations of parametric models. For instance, the great success of Efron's (1979) bootstrap was in providing a complete theory for statistical inference under a nonparametric setting much like Maximum Likelihood Estimation had done half a century earlier under the restrictive parametric setup.

Nevertheless, there is a further step one may take, i.e., going beyond even nonparametric models, and this is the subject of the monograph at hand. To explain this, let us momentarily focus on regression, i.e., data that are pairs: $(Y_1, X_1), (Y_2, X_2), \ldots,$ (Y_n, X_n), where Y_i is the measured response associated with a regressor value of X_i. There are several ways to model such a dataset; three main ones are listed below.

They all pertain to the standard, homoscedastic additive model:

$$Y_i = \mu(X_i) + \varepsilon_i \tag{1}$$

where the random variables ε_i are assumed to be independent, identically distributed (i.i.d.) from a distribution $F(\cdot)$ with mean zero.

- **Parametric model:** Both $\mu(\cdot)$ and $F(\cdot)$ belong to parametric families of functions, e.g., $\mu(x) = \beta_0 + \beta_1 x$ and $F(\cdot)$ is $N(0, \sigma^2)$.
- **Semiparametric model:** $\mu(\cdot)$ belongs to a parametric family, whereas $F(\cdot)$ does not; instead, it may be assumed that $F(\cdot)$ belongs to a smoothness class, etc.
- **Nonparametric model:** Neither $\mu(\cdot)$ nor $F(\cdot)$ can be assumed to belong to parametric families of functions.

Despite the nonparametric aspect of it, even the last option constitutes a model, and is thus rather restrictive. To see why, note that Eq. (1) with i.i.d. errors is not satisfied in many cases of interest even after allowing for heteroscedasticity of the errors. For example, consider the model $Y_i = G(X_i, \varepsilon_i)$, where the ε_i are i.i.d., and $G(\cdot,\cdot)$ is a nonlinear/non-additive function of two variables. It is for this reason, i.e., to render the data amenable to an additive model such as (1), that a multitude of transformations in regression have been proposed and studied over the years, e.g., Box-Cox, ACE, AVAS, etc.; see Linton et al. (1997) for a review.

Nevertheless, it is possible to shun Eq. (1) altogether and still conduct inference about a quantity of interest such as the conditional expectation function $E(Y|X = x)$. In contrast to nonparametric model (1), the following model-free assumption can be made:

- **Model-free regression:**
  - **Random design.** The pairs $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$ are i.i.d.
  - **Deterministic design.** The variables $X_1, \ldots, X_n$ are deterministic, and the random variables $Y_1, \ldots, Y_n$ are independent with common conditional distribution, i.e., $P\{Y_j \leq y | X_j = x\} = D_x(y)$ not depending on j.

Inference for features, i.e., functionals, of the common conditional distribution $D_x(\cdot)$ is still possible under some regularity conditions, e.g., smoothness. Arguably, the most important such feature is the conditional mean $E(Y|X = x)$ that can be denoted $\mu(x)$. While $\mu(x)$ is crucial in the model (1) as the function explaining Y on the basis of $X = x$, it has a key function in model-free prediction as well: $\mu(x_f)$ is the mean squared error (MSE) optimal predictor of a future response Y_f associated with a regressor value x_f.

As will be shown in the sequel, it is possible to accomplish the goal of point and interval prediction of Y_f under the above model-free setup; this is achieved via the **Model-free Prediction Principle** described in Part I of the book. In so doing, the solution to interesting estimation problems is obtained as a by-product, e.g., inference on features of $D_x(\cdot)$; the prime example again is $\mu(x)$. Hence, a Model-free approach to frequentist statistical inference is possible, including prediction and confidence intervals.

In nonparametric statistics, it is common to try to develop some asymptotic theory for new methods developed. In addition to offering justification for the accuracy of these methods, asymptotics often provide insights on practical implementation, e.g., on the optimal choice of smoothing bandwidth, etc. All of the methods discussed/employed in the proposed Model-free approach to inference will be based on estimators that have favorable large-sample properties—such as consistency—under regularity conditions. Furthermore, asymptotic information on bandwidth rates, MSE decay rates, etc. will be given whenever available in the form of Facts or Claims together with suggestions on their proof and/or references. However, formal theorems and proofs were deemed beyond the scope of this monograph in order to better focus on the methodology, as well as keep the book's length (and time of completion) under control. Perhaps more importantly, note that it is still unclear how to properly judge the quality of prediction intervals in an asymptotic setting; some preliminary ideas on this issue are given in Sects. 3.6.2 and 7.2.3, and the Rejoinder of Politis (2013).

Interestingly, the emphasis on prediction seems to be coming back full-circle in the twenty-first century with the recent boom in machine learning and data mining; see, e.g., the highly influential book on statistical learning by Hastie et al. (2009), and the recent monograph on predictive modeling by Kuhn and Johnson (2013). The Model-free prediction methods presented here are of a very different nature but share some similarities, e.g., in employing cross-validation and sample re-use for fine-tuning and optimization, and may thus complement well the popular model-based approaches to prediction and classification. Furthermore, ideas from statistical learning and model selection could eventually be incorporated in the Model-free framework as well, e.g., selecting a subset of regressors; this is the subject of ongoing work. Notably, the methods presented in this monograph are very computer-intensive; relevant R functions and software are given at: `http://www.math.ucsd.edu/~politis/DPsoftware.html`.

I would like to thank my colleagues in the Departments of Mathematics and Economics of UCSD for their support, and my Ph.D. students for bearing with some of the material. I have benefited immensely from suggestions and discussions with colleagues from all over the world; a very partial list includes: Ian Abramson, Ery Arias-Castro, Brendan Beare, Patrice Bertail, Ricardo Cao, Anirban DasGupta, Richard Davis, Brad Efron, Peter Hall, Xuming He, Nancy Heckman, Göran Kauermann, Claudia Klüppelberg, Piotr Kokoszka, Jens-Peter Kreiss, Michele La Rocca, Jacek Leskow, Tim McMurry, George Michailidis, Stathis Paparoditis, Mohsen Pourahmadi, Jeff Racine, Joe Romano, Dimitrios Thomakos, Florin Vaida, Slava Vasiliev, Philippe Vieu, and Michael Wolf. Further acknowledgements are given at the end of several chapters.

In closing, I would like to thank the Division of Mathematical Sciences of the National Science Foundation for their continuing support with multiple grants, the most recent ones being DMS-10-07513 and DMS 13-08319, and the John Simon Guggenheim Memorial Foundation for a 2011–2012 fellowship that helped me get started on this monograph. I would also like to thank Marc Strauss and Hannah

Bracken of Springer for a wonderful collaboration, and Somnath Datta and the Editorial Board of the Frontiers Series for hosting this project.

The impetus for putting together this monograph was to show how very different statistical problems can be approached afresh in a Model-free setting. Due to time and space limitations, I could only explicitly address a handful of areas of practical implementation, e.g., regression, autoregression, Markov processes, etc. It is my sincere hope that the monograph will incite the interest of readers to take another look at their favorite problem—either theoretical or applied—in this new light; the insights gained may be well worth it.

San Diego, CA, USA
Spring 2015

Dimitris N. Politis

Contents

Acronyms

acf	Autocorrelation function
AIC	Akaike Information Criterion
AR	Auto regressive
ARCH	Auto regressive conditional heteroscedasticity
ARMA	Auto regressive with moving average residuals
cdf	Cumulative distribution function
CVR	Coverage (of interval)
DFT	Discrete Fourier Transform
FSO	Full-Sample Optimal
GARCH	Generalized ARCH
i.i.d.	Independent, identically distributed
i.i.d.–ness	The property of a dataset being i.i.d.
i.i.d. (μ, σ^2)	i.i.d. with mean μ and variance σ^2
LEN	Length (of interval)
LMF	Limit Model Free
LS	Least squares
MA	Moving average
MAD	Mean absolute deviation
MB	Model-based
MF	Model-free
MF^2	Model-free model-fitting
MLE	Maximum likelihood estimation
MSE	Mean squared error
$N(\mu, \sigma^2)$	Normal with mean μ and variance σ^2
NoVaS	Normalizing and variance stabilizing transformation
PMF	Predictive model-free
st.err.	Standard error