

DATA SCIENCE SERIES

STATISTICAL FOUNDATIONS OF DATA SCIENCE



JIANQING FAN
RUNZE LI
CUN-HUI ZHANG
HUI ZOU



CRC Press
Taylor & Francis Group

A CHAPMAN & HALL BOOK

Contents

Preface	xvii
1 Introduction	1
1.1 Rise of Big Data and Dimensionality	1
1.1.1 Biological sciences	2
1.1.2 Health sciences	4
1.1.3 Computer and information sciences	5
1.1.4 Economics and finance	7
1.1.5 Business and program evaluation	9
1.1.6 Earth sciences and astronomy	9
1.2 Impact of Big Data	9
1.3 Impact of Dimensionality	11
1.3.1 Computation	11
1.3.2 Noise accumulation	12
1.3.3 Spurious correlation	14
1.3.4 Statistical theory	17
1.4 Aim of High-dimensional Statistical Learning	18
1.5 What Big Data Can Do	19
1.6 Scope of the Book	19
2 Multiple and Nonparametric Regression	21
2.1 Introduction	21
2.2 Multiple Linear Regression	21
2.2.1 The Gauss-Markov theorem	23
2.2.2 Statistical tests	26
2.3 Weighted Least-Squares	27
2.4 Box-Cox Transformation	29
2.5 Model Building and Basis Expansions	30
2.5.1 Polynomial regression	31
2.5.2 Spline regression	32
2.5.3 Multiple covariates	35
2.6 Ridge Regression	37
2.6.1 Bias-variance tradeoff	37
2.6.2 ℓ_2 penalized least squares	38
2.6.3 Bayesian interpretation	38

2.6.4	Ridge regression solution path	39
2.6.5	Kernel ridge regression	41
2.7	Regression in Reproducing Kernel Hilbert Space	42
2.8	Leave-one-out and Generalized Cross-validation	47
2.9	Exercises	49
3	Introduction to Penalized Least-Squares	55
3.1	Classical Variable Selection Criteria	55
3.1.1	Subset selection	55
3.1.2	Relation with penalized regression	56
3.1.3	Selection of regularization parameters	57
3.2	Folded-concave Penalized Least Squares	59
3.2.1	Orthonormal designs	61
3.2.2	Penalty functions	62
3.2.3	Thresholding by SCAD and MCP	63
3.2.4	Risk properties	64
3.2.5	Characterization of folded-concave PLS	65
3.3	Lasso and L_1 Regularization	66
3.3.1	Nonnegative garrote	66
3.3.2	Lasso	68
3.3.3	Adaptive Lasso	71
3.3.4	Elastic Net	72
3.3.5	Dantzig selector	74
3.3.6	SLOPE and sorted penalties	77
3.3.7	Concentration inequalities and uniform convergence	78
3.3.8	A brief history of model selection	81
3.4	Bayesian Variable Selection	81
3.4.1	Bayesian view of the PLS	81
3.4.2	A Bayesian framework for selection	83
3.5	Numerical Algorithms	84
3.5.1	Quadratic programs	84
3.5.2	Least angle regression*	86
3.5.3	Local quadratic approximations	89
3.5.4	Local linear algorithm	91
3.5.5	Penalized linear unbiased selection*	92
3.5.6	Cyclic coordinate descent algorithms	93
3.5.7	Iterative shrinkage-thresholding algorithms	94
3.5.8	Projected proximal gradient method	96
3.5.9	ADMM	96
3.5.10	Iterative local adaptive majorization and minimization	97
3.5.11	Other methods and timeline	98
3.6	Regularization Parameters for PLS	99
3.6.1	Degrees of freedom	100
3.6.2	Extension of information criteria	102
3.6.3	Application to PLS estimators	102

Statistical Foundations of Data Science gives a thorough introduction to commonly used statistical models and contemporary statistical machine learning techniques and algorithms, along with their mathematical insights and statistical theories. It serves as a graduate-level textbook and a research monograph on high-dimensional statistics, sparsity and covariance learning, machine learning, and statistical inference. It includes ample exercises that involve both theoretical studies as well as empirical applications.

The book begins by introducing the stylized features of big data and their impacts on statistical analysis. It then introduces multiple linear regression and expands the techniques of model building via nonparametric regression and kernel tricks. It provides a comprehensive account of sparsity explorations and model selections for multiple regression, generalized linear models, quantile regression, robust regression, hazards regression, among others. High-dimensional inference is also thoroughly addressed and so is feature screening. The book also provides a comprehensive account of high-dimensional covariance estimation, learning latent factors and hidden structures, as well as their applications to statistical estimation, inference, prediction and machine learning problems. It also introduces thoroughly statistical machine learning theory and methods for classification, clustering, and prediction. These include CART, random forests, boosting, support vector machines, clustering algorithms, sparse PCA, and deep learning.

The authors are international authorities and leaders on the presented topics. All are fellows of the Institute of Mathematical Statistics and the American Statistical Association.

Jianqing Fan is Frederick L. Moore Professor, Princeton University. He is co-editing *Journal of Business and Economics Statistics* and was the co-editor of *The Annals of Statistics*, *Probability Theory and Related Fields*, and *Journal of Econometrics* and has been recognized by the 2000 COPSS Presidents' Award, AAAS Fellow, Guggenheim Fellow, Guy medal in silver, Noether Senior Scholar Award, and Academician of Academia Sinica.

Runze Li is Elberly family chair professor and AAAS fellow, Pennsylvania State University, and was co-editor of *The Annals of Statistics*.

Cun-Hui Zhang is distinguished professor, Rutgers University and was co-editor of *Statistical Science*.

Hui Zou is professor, University of Minnesota and was action editor of *Journal of Machine Learning Research*.



CRC Press

Taylor & Francis Group
an informa business

www.crcpress.com

CRC Press titles are available as eBook editions



ID 20 1001 6752

ISBN 9781466510845

STATISTICS

ISBN: 978-1-4665-1084-5

90000



9 781466 510845