ELSEVIER
爱思唯尔

系统参数辨识的信息准则及算法

SYSTEM PARAMETER IDENTIFICATION: INFORMATION CRITERIA AND ALGORITHMS

陈霸东，朱煜，胡金春，[美] 乔斯 · C. 普伦斯派 著
Badong Chen, Yu Zhu, Jinchun Hu, Jose C. Principe

清华大学出版社

系统参数辨识的
信息准则及算法

SYSTEM PARAMETER IDENTIFICATION:
INFORMATION CRITERIA AND ALGORITHMS

陈霸东，朱煜，胡金春，[美] 乔斯·C. 普伦斯派 著
Badong Chen, Yu Zhu, Jinchun Hu, Jose C. Principe

清华大学出版社
北京

内 容 简 介

本书系统地介绍系统参数辨识信息准则及算法的最新研究成果，主要内容包括信息论基本概念、基于信息论的参数估计、基于最小误差熵准则的系统辨识、基于最小信息距离的系统辨识、基于互信息准则的系统辨识。

本书在2011年出版的同名中文版图书基础上进行了修订和增补，适合高等院校或研究机构从事系统辨识、信号处理、机器学习等研究工作的师生以及其他科技工作者阅读参考。

This edition of **System parameter identification: information criteria and algorithms** by Badong Chen, Yu Zhu, Jinchun Hu, Jose C. Principe is published by arrangement with **ELSEVIER INC**,. a Delaware corporation having its principal place of business at 360 Park Avenue South, New York, NY 10010, USA

图书在版编目(CIP)数据

系统参数辨识的信息准则及算法＝System parameter identification: information criteria and algorithms: 英文/陈霸东等著. --北京：清华大学出版社，2014

ISBN 978-7-302-35941-8

Ⅰ.①系…　Ⅱ.①陈…　Ⅲ. ①参数分析–英文　Ⅳ. ①O212.1

中国版本图书馆CIP数据核字(2014)第067314号

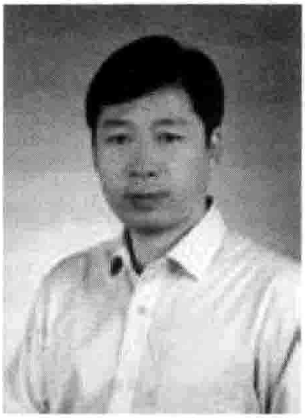
责任编辑： 黎　强
封面设计： 常雪影
责任印制： 刘海龙

出版发行： 清华大学出版社
　　　网　　址： http://www.tup.com.cn, http://www.wqbook.com
　　　地　　址： 北京清华大学学研大厦A座　　　**邮　　编：** 100084
　　　社 总 机： 010-62770175　　　**邮　　购：** 010-62786544
　　　投稿与读者服务： 010-62776969, c-service@tup.tsinghua.edu.cn
　　　质量反馈： 010-62772015, zhiliang@tup.tsinghua.edu.cn
印 装 者： 三河市春园印刷有限公司
经　　销： 全国新华书店
开　　本： 153mm×235mm　　**印　　张：** 16.5
版　　次： 2014年5月第1版　　**印　　次：** 2014年5月第1次印刷
定　　价： 120.00元

产品编号：056918-01

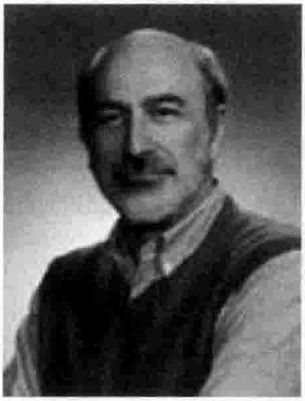
About the Authors

Badong Chen received the B.S. and M.S. degrees in control theory and engineering from Chongqing University, in 1997 and 2003, respectively, and the Ph.D. degree in computer science and technology from Tsinghua University in 2008. He was a post-doctoral researcher with Tsinghua University from 2008 to 2010 and a post-doctoral associate at the University of Florida Computational NeuroEngineering Laboratory during the period October 2010 to September 2012. He is currently a professor at the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. His research interests are in system identification and control, information theory, machine learning, and their applications in cognition and neuroscience.



Yu Zhu received the B.S. degree in radio electronics in 1983 from Beijing Normal University, and the M.S. degree in computer applications in 1993 and the Ph.D. degree in mechanical design and theory in 2001, both from China University of Mining and Technology. He is currently a professor with the Department of Mechanical Engineering, Tsinghua University. His research field mainly covers IC manufacturing equipment development strategy, ultra-precision air/maglev stage machinery design theory and technology, ultra-precision measurement theory and technology, and precision motion control theory and technology. He has more than 140 research papers and 100 (48 awarded) invention patents.

Jinchun Hu, associate professor, born in 1972, graduated from Nanjing University of Science and Technology. He received the B.E. and Ph.D. degrees in control science and engineering in 1994 and 1998, respectively. Currently, he works at the Department of Mechanical Engineering, Tsinghua University. His current research interests include modern control theory and control systems, ultra-precision measurement principles and methods, micro/nano motion control system analysis and realization, special driver technology and device for precision motion systems, and super-precision measurement and control.



Jose C. Principe is a distinguished professor of electrical and computer engineering and biomedical engineering at the University of Florida where he teaches advanced signal processing, machine learning, and artificial neural networks modeling. He is BellSouth Professor and the founding director of the University of Florida Computational NeuroEngineering Laboratory. His primary research interests are in advanced signal processing with information theoretic criteria (entropy and mutual information) and adaptive models in reproducing kernel Hilbert spaces, and the application of these advanced algorithms to brain machine interfaces. He is a Fellow of the IEEE, ABME, and AIBME. He is the past editor in chief of the IEEE Transactions on Biomedical Engineering, past chair of the Technical Committee on Neural Networks of the IEEE Signal Processing Society, and past President of the International Neural Network Society. He received the IEEE EMBS Career Award and the IEEE Neural Network Pioneer Award. He has more than 600 publications and 30 patents (awarded or filed).

Preface

System identification is a common method for building the mathematical model of a physical plant, which is widely utilized in practical engineering situations. In general, the system identification consists of three key elements, i.e., the data, the model, and the criterion. The goal of identification is then to choose one from a set of candidate *models* to fit the *data* best according to a certain *criterion*. The criterion function is a key factor in system identification, which evaluates the consistency of the model to the actual plant and is, in general, an objective function for developing the identification algorithms. The identification performances, such as the convergence speed, steady-state accuracy, robustness, and the computational complexity, are directly related to the criterion function.

Well-known identification criteria mainly include the least squares (LS) criterion, minimum mean square error (MMSE) criterion, and the maximum likelihood (ML) criterion. These criteria provide successful engineering solutions to most practical problems, and are still prevalent today in system identification. However, they have some shortcomings that limit their general use. For example, the LS and MMSE only consider the second-order moment of the error, and the identification performance would become worse when data are non-Gaussian distributed (e.g., with multimodal, heavy-tail, or finite range). The ML criterion requires the knowledge of the conditional probability density function of the observed samples, which is not available in many practical situations. In addition, the computational complexity of the ML estimation is usually high. Thus, selecting a new criterion beyond second-order statistics and likelihood function is attractive in problems of system identification.

In recent years, criteria based on information theoretic descriptors of entropy and dissimilarity (divergence, mutual information) have attracted lots of attentions and become an emerging area of study in signal processing and machine learning domains. Information theoretic criteria (or briefly, information criteria) can capture higher order statistics and information content of signals rather than simply their energy. Many studies suggest that information criteria do not suffer from the limitation of Gaussian assumption and can improve performance in many realistic scenarios. Combined with nonparametric estimators of entropy and divergence, many adaptive identification algorithms have been developed, including the practical gradient-based batch or recursive algorithms, fixed-point algorithms (no step-size), or other advanced search algorithms. Although many elegant results and techniques have been developed over the past few years, till now there is no book devoted to a systematic study of system identification under information theoretic criteria. The primary focus of this book is to provide an overview of these developments, with emphasis on the nonparametric estimators of information criteria and gradient-based identification algorithms. Most of the contents of this book originally appeared in the recent papers of the authors.

The book is divided into six chapters: the first chapter is the introduction to the information theoretic criteria and the state-of-the-art techniques; the second chapter presents the definitions and properties of several important information measures; the third chapter gives an overview of information theoretic approaches to parameter estimation; the fourth chapter discusses system identification under minimum error entropy criterion; the fifth chapter focuses on the minimum information divergence criteria; and the sixth chapter changes the focus to the mutual information-based criteria.

It is worth noting that the information criteria can be used not only for system parameter identification but also for system structure identification (e.g., model selection). The Akaike's information criterion (AIC) and the minimum description length (MDL) are two famous information criteria for model selection. There have been several books on AIC and MDL, and in this book we don't discuss them in detail. Although most of the methods in this book are developed particularly for system parameter identification, the basic principles behind them are universal. Some of the methods with little modification can be applied to blind source separation, independent component analysis, time series prediction, classification and pattern recognition.

This book will be of interest to graduates, professionals, and researchers who are interested in improving the performance of traditional identification algorithms and in exploring new approaches to system identification, and also to those who are interested in adaptive filtering, neural networks, kernel methods, and online machine learning.

The authors are grateful to the National Natural Science Foundation of China and the National Basic Research Program of China (973 Program), which have funded this book. We are also grateful to the Elsevier for their patience with us over the past year we worked on this book. We also acknowledge the support and encouragement from our colleagues and friends.

Xi'an
P.R. China
March 2013

Symbols and Abbreviations

The main symbols and abbreviations used throughout the text are listed as follows.

$\lvert . \rvert$	absolute value of a real number
$\lVert \cdot \rVert$	Euclidean norm of a vector
$\langle .,. \rangle$	inner product
$\mathbb{I}(.)$	indicator function
$\boldsymbol{E}[.]$	expectation value of a random variable
$\boldsymbol{f'(x)}$	first-order derivative of the function $f(x)$
$\boldsymbol{f''(x)}$	second-order derivative of the function $f(x)$
$\nabla_{\boldsymbol{x}} \boldsymbol{f(x)}$	gradient of the function $f(x)$ with respect to x
$\mathbf{sign}(.)$	sign function
$\boldsymbol{\Gamma}(.)$	Gamma function
$(.)^{\boldsymbol{T}}$	vector or matrix transposition
$\boldsymbol{I}$	identity matrix
$\boldsymbol{A}^{-1}$	inverse of matrix A
$\mathbf{det}\,\boldsymbol{A}$	determinant of matrix A
$\mathbf{Tr}\boldsymbol{A}$	trace of matrix A
$\mathbf{rank}\boldsymbol{A}$	rank of matrix A
$\mathbf{log}(.)$	natural logarithm function
z^{-1}	unit delay operator
$\mathbb{R}$	real number space
$\mathbb{R}^{n}$	n-dimensional real Euclidean space
$\boldsymbol{\rho(X, Y)}$	correlation coefficient between random variables X and Y
$\mathbf{Var}[X]$	variance of random variable X
$\mathbf{Pr}[A]$	probability of event A
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution with mean vector μ and covariance matrix Σ
$\mathbf{U}[\boldsymbol{a}, \boldsymbol{b}]$	uniform distribution over interval $[a, b]$
$\boldsymbol{\chi^2(k)}$	chi-squared distribution with k degree of freedom
$\boldsymbol{H}(X)$	Shannon entropy of random variable X
$\boldsymbol{H}_{\phi}(X)$	ϕ-entropy of random variable X
$\boldsymbol{H}_{\alpha}(X)$	α-order Renyi entropy of random variable X
$\boldsymbol{V}_{\alpha}(X)$	α-order information potential of random variable X
$\boldsymbol{S}_{\alpha}(X)$	survival information potential of random variable X
$\boldsymbol{H}_{\Delta}(X)$	Δ-entropy of discrete random variable X
$\boldsymbol{I}(X; Y)$	mutual information between random variables X and Y
$\boldsymbol{D}_{\mathbf{KL}}(X \Vert Y)$	KL-divergence between random variables X and Y
$\boldsymbol{D}_{\phi}(X \Vert Y)$	ϕ-divergence between random variables X and Y
$\boldsymbol{J}_{\mathbf{F}}$	Fisher information matrix
$\bar{\boldsymbol{J}}_{\mathbf{F}}$	Fisher information rate matrix

$\boldsymbol{p}(.)$	probability density function
$\kappa(.,.)$	Mercer kernel function
$\boldsymbol{K}(.)$	kernel function for density estimation
$\boldsymbol{K}_h(.)$	kernel function with width h
$\boldsymbol{G}_h(.)$	Gaussian kernel function with width h
$\mathcal{H}_k$	reproducing kernel Hilbert space induced by Mercer kernel κ
$\mathbb{F}_\kappa$	feature space induced by Mercer kernel κ
$\boldsymbol{W}$	weight vector
$\boldsymbol{\Omega}$	weight vector in feature space
$\tilde{\boldsymbol{W}}$	weight error vector
η	step size
$\boldsymbol{L}$	sliding data length
MSE	mean square error
LMS	least mean square
NLMS	normalized least mean square
LS	least squares
RLS	recursive least squares
MLE	maximum likelihood estimation
EM	expectation-maximization
FLOM	fractional lower order moment
LMP	least mean p-power
LAD	least absolute deviation
LMF	least mean fourth
FIR	finite impulse response
IIR	infinite impulse response
AR	auto regressive
ADALINE	adaptive linear neuron
MLP	multilayer perceptron
RKHS	reproducing kernel Hilbert space
KAF	kernel adaptive filtering
KLMS	kernel least mean square
KAPA	kernel affine projection algorithm
KMEE	kernel minimum error entropy
KMC	kernel maximum correntropy
PDF	probability density function
KDE	kernel density estimation
GGD	generalized Gaussian density
$S\alpha S$	symmetric α-stable
MEP	maximum entropy principle
DPI	data processing inequality
EPI	entropy power inequality
MEE	minimum error entropy
MCC	maximum correntropy criterion
IP	information potential
QIP	quadratic information potential
CRE	cumulative residual entropy
SIP	survival information potential
QSIP	survival quadratic information potential

KLID	Kullback–Leibler information divergence
EDC	Euclidean distance criterion
MinMI	minimum mutual information
MaxMI	maximum mutual information
AIC	Akaike's information criterion
BIC	Bayesian information criterion
MDL	minimum description length
FIM	Fisher information matrix
FIRM	Fisher information rate matrix
MIH	minimum identifiable horizon
ITL	information theoretic learning
BIG	batch information gradient
FRIG	forgetting recursive information gradient
SIG	stochastic information gradient
SIDG	stochastic information divergence gradient
SMIG	stochastic mutual information gradient
FP	fixed point
FP-MEE	fixed-point minimum error entropy
RFP-MEE	recursive fixed-point minimum error entropy
EDA	estimation of distribution algorithm
SNR	signal to noise ratio
WEP	weight error power
EMSE	excess mean square error
IEP	intrinsic error power
ICA	independent component analysis
BSS	blind source separation
CRLB	Cramer–Rao lower bound
AEC	acoustic echo canceller

Contents

1 Introduction

1.1 Elements of System Identification

Mathematical models of systems (either natural or man-made) play an essential role in modern science and technology. Roughly speaking, a mathematical model can be imagined as a mathematical law that links the system inputs (causes) with the outputs (effects). The applications of mathematical models range from simulation and prediction to control and diagnosis in heterogeneous fields. System identification is a widely used approach to build a mathematical model. It estimates the model based on the observed data (usually with uncertainty and noise) from the unknown system.

Many researchers try to provide an explicit definition for system identification. In 1962, Zadeh gave a definition as follows [1]: "System identification is the determination, on the basis of observations of input and output, of a system within a specified class of systems to which the system under test is equivalent." It is almost impossible to find out a model completely matching the physical plant. Actually, the system input and output always include certain noises; the identification model is therefore only an approximation of the practical plant. Eykhoff [2] pointed out that the system identification tries to use a model to describe the essential characteristic of an objective system (or a system under construction), and the model should be expressed in a useful form. Clearly, Eykhoff did not expect to obtain an exact mathematical description, but just to create a model suitable for applications. In 1978, Ljung [3] proposed another definition: "The identification procedure is based on three entities: the data, the set of models, and the criterion. Identification, then, is to select the model in the model set that describes the data best, according to the criterion."

According to the definitions by Zadeh and Ljung, system identification consists of three elements (see Figure 1.1): data, model, and equivalence criterion (equivalence is often defined in terms of a criterion or a loss function). The three elements directly govern the identification performance, including the identification accuracy, convergence rate, robustness, and computational complexity of the identification algorithm [4]. How to optimally design or choose these elements is very important in system identification.

The model selection is a crucial step in system identification. Over the past decades, a number of model structures have been suggested, ranging from the simple

System Parameter Identification. DOI: http://dx.doi.org/10.1016/B978-0-12-404574-3.00001-4

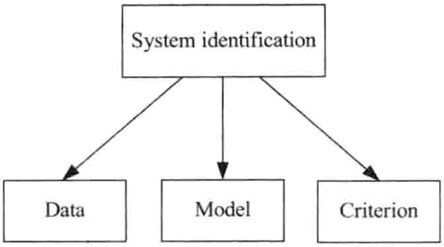


Figure 1.1 Three elements of system identification.

linear structures [FIR (finite impulse response), AR (autoregressive), ARMA (autoregressive and moving average), etc.] to more general nonlinear structures [NAR (nonlinear autoregressive), MLP (multilayer perceptron), RBF (radial basis function), etc.]. In general, model selection is a trade-off between the quality and the complexity of the model. In most practical situations, some prior knowledge may be available regarding the appropriate model structure or the designer may wish to limit to a particular model structure that is tractable and meanwhile can make a good approximation to the true system. Various model selection criteria have also been introduced, such as the cross-validation (CV) criterion [5], Akaike's information criterion (AIC) [6,7], Bayesian information criterion (BIC) [8], and minimum description length (MDL) criterion [9,10].

The data selection (the choice of the measured variables) and the optimal input design (experiment design) are important issues. The goal of experiment design is to adjust the experimental conditions so that maximal information is gained from the experiment (such that the measured data contain the maximal information about the unknown system). The optimality criterion for experiment design is usually based on the information matrices [11]. For many nonlinear models (e.g., the kernel-based model), the input selection can significantly help to reduce the network size [12].

The choice of the equivalence criterion (or approximation criterion) is another key issue in system identification. The approximation criterion measures the difference (or similarity) between the model and the actual system, and allows determination of how good the estimate of the system is. Different choices of the approximation criterion will lead to different estimates. The task of parametric system identification is to adjust the model parameters such that a predefined approximation criterion is minimized (or maximized). As a measure of accuracy, the approximation criterion determines the performance surface, and has significant influence on the optimal solutions and convergence behaviors. The development of new identification approximation criteria is an important emerging research topic and this will be the focus of this book.

It is worth noting that many machine learning methods also involve three elements: model, data, and optimization criterion. Actually, system identification can be viewed, to some extent, as a special case of supervised machine learning. The main terms in system identification and machine learning are reported in Table 1.1. In this book, these terminologies are used interchangeably.

Table 1.1 Main Terminologies in System Identification and Machine Learning

System Identification	Machine Learning
Model, filter	Learning machine, network
Parameters, coefficients	Weights
Identify, estimate	Learn, train
Observations, measurements	Examples, training data
Overparametrization	Overtraining, overfitting

1.2 Traditional Identification Criteria

Traditional identification (or estimation) criteria mainly include the least squares (LS) criterion [13], minimum mean square error (MMSE) criterion [14], and the maximum likelihood (ML) criterion [15,16]. The LS criterion, defined by minimizing the sum of squared errors (an error being the difference between an observed value and the fitted value provided by a model), could at least dates back to Carl Friedrich Gauss (1795). It corresponds to the ML criterion if the experimental errors have a Gaussian distribution. Due to its simplicity and efficiency, the LS criterion has been widely used in problems, such as estimation, regression, and system identification. The LS criterion is mathematically tractable, and the linear LS problem has a closed form solution. In some contexts, a regularized version of the LS solution may be preferable [17]. There are many identification algorithms developed with LS criterion. Typical examples are the recursive least squares (RLS) and its variants [4]. In statistics and signal processing, the MMSE criterion is a common measure of estimation quality. An MMSE estimator minimizes the mean square error (MSE) of the fitted values of a dependent variable. In system identification, the MMSE criterion is often used as a criterion for stochastic approximation methods, which are a family of iterative stochastic optimization algorithms that attempt to find the extrema of functions which cannot be computed directly, but only estimated via noisy observations. The well-known least mean square (LMS) algorithm [18–20], invented in 1960 by Bernard Widrow and Ted Hoff, is a stochastic gradient descent algorithm under MMSE criterion. The ML criterion is recommended, analyzed, and popularized by R.A. Fisher [15]. Given a set of data and underlying statistical model, the method of ML selects the model parameters that maximize the likelihood function (which measures the degree of "agreement" of the selected model with the observed data). The ML estimation provides a unified approach to estimation, which corresponds to many well-known estimation methods in statistics. The ML parameter estimation possesses a number of attractive limiting properties, such as consistency, asymptotic normality, and efficiency.

The above identification criteria (LS, MMSE, ML) perform well in most practical situations, and so far are still the workhorses of system identification. However, they have some limitations. For example, the LS and MMSE capture only the second-order statistics in the data, and may be a poor approximation criterion, especially in nonlinear and non-Gaussian (e.g., heavy tail or finite range distributions) situations. The ML criterion requires the knowledge of the conditional distribution (likelihood function) of the data given parameters, which is unavailable in many practical problems. In some complicated problems, the ML estimators are unsuitable or do not exist. Thus, selecting a new criterion beyond second-order statistics and likelihood function is attractive in problems of system identification.

In order to take into account higher order (or lower order) statistics and to select an optimal criterion for system identification, many researchers studied the non-MSE (nonquadratic) criteria. In an early work [21], Sherman first proposed the non-MSE criteria, and showed that in the case of Gaussian processes, a large family of non-MSE criteria yields the same predictor as the linear MMSE predictor of Wiener. Later, Sherman's results and several extensions were revisited by Brown [22], Zakai [23], Hall and Wise [24], and others. In [25], Ljung and Soderstrom discussed the possibility of a general error criterion for recursive parameter identification, and found an optimal criterion by minimizing the asymptotic covariance matrix of the parameter estimates. In [26,27], Walach and Widrow proposed a method to select an optimal identification criterion from the least mean fourth (LMF) family criteria. In their approach, the optimal choice is determined by minimizing a cost function which depends on the moments of the interfering noise. In [28], Douglas and Meng utilized the calculus of variations method to solve the optimal criterion among a large family of general error criteria. In [29], Al-Naffouri and Sayed optimized the error nonlinearity (derivative of the general error criterion) by optimizing the steady state performance. In [30], Pei and Tseng investigated the least mean *p*-power (LMP) criterion. The fractional lower order moments (FLOMs) of the error have also been used in adaptive identification in the presence of impulse alpha-stable noises [31,32]. Other non-MSE criteria include the M-estimation criterion [33], mixed norm criterion [34–36], risk-sensitive criterion [37,38], high-order cumulant (HOC) criterion [39–42], and so on.

1.3 Information Theoretic Criteria

Information theory is a branch of statistics and applied mathematics, which is exactly created to help studying the theoretical issues of optimally encoding messages according to their statistical structure, selecting transmission rates according to the noise levels in the channel, and evaluating the minimal distortion in messages [43]. Information theory was first developed by Claude E. Shannon to find fundamental limits on signal processing operations like compressing data and on reliably storing and communicating data [44]. After the pioneering work of Shannon, information theory found applications in many scientific areas, including physics, statistics, cryptography, biology, quantum computing, and so on. Moreover, information theoretic measures (entropy, divergence, mutual information, etc.) and principles (e.g., the principle of maximum entropy) were widely used in engineering areas, such as signal processing, machine learning, and other