MODEL SELECTION AND PREDICTION: NORMAL REGRESSION

T. P. SPEED[1] AND BIN YU[2]

[1]*Department of Statistics, University of California at Berkeley, CA 94720, U.S.A.*
[2]*Department of Statistics, University of Wisconsin-Madison, WI 53706, U.S.A.*

(Received September 27, 1991; revised April 27, 1992)

Abstract. This paper discusses the topic of model selection for finite-dimensional normal regression models. We compare model selection criteria according to prediction errors based upon prediction with refitting, and prediction without refitting. We provide a new lower bound for prediction without refitting, while a lower bound for prediction with refitting was given by Rissanen. Moreover, we specify a set of sufficient conditions for a model selection criterion to achieve these bounds. Then the achievability of the two bounds by the following selection rules are addressed: Rissanen's accumulated prediction error criterion (APE), his stochastic complexity criterion, AIC, BIC and the FPE criteria. In particular, we provide upper bounds on overfitting and underfitting probabilities needed for the achievability. Finally, we offer a brief discussion on the issue of finite-dimensional vs. infinite-dimensional model assumptions.

Key words and phrases: Model selection, prediction lower bound, accumulated prediction error (APE), AIC, BIC, FPE, stochastic complexity, overfit and underfit probability.