Accepted Manuscript

Cross-validation for selecting a model selection procedure

Yongli Zhang, Yuhong Yang

PII: S0304-4076(15)00030-5
DOI: http://dx.doi.org/10.1016/j.jeconom.2015.02.006
Reference: ECONOM 4027

To appear in: *Journal of Econometrics*

Received date: 16 May 2014
Revised date: 13 October 2014
Accepted date: 3 February 2015

Please cite this article as: Zhang, Y., Yang, Y., Cross-validation for selecting a model selection procedure. *Journal of Econometrics* (2015), http://dx.doi.org/10.1016/j.jeconom.2015.02.006

Cross-Validation for Selecting a Model Selection Procedure*

Yongli Zhang
LundQuist College of Business
University of Oregon
Eugene, OR 97403

Yuhong Yang
School of Statistics
University of Minnesota
Minneapolis, MN 55455

Abstract

While there are various model selection methods, an unanswered but important question is how to select one of them for data at hand. The difficulty is due to that the targeted behaviors of the model selection procedures depend heavily on uncheckable or difficult-to-check assumptions on the data generating process. Fortunately, cross-validation (CV) provides a general tool to solve this problem. In this work, results are provided on how to apply CV to consistently choose the best method, yielding new insights and guidance for potentially vast amount of application. In addition, we address several seemingly widely spread misconceptions on CV.

Key words: Cross-validation, cross-validation paradox, data splitting ratio, adaptive procedure selection, information criterion, LASSO, MCP, SCAD

1 Introduction

Model selection is an indispensable step in the process of developing a functional prediction model or a model for understanding the data generating mechanism. While thousands of papers have been published on model selection, an important and largely unanswered question is: How do we select a modeling procedure that typically involves model selection and parameter estimation? In a real application, one usually does not know which procedure fits the data the best. Instead of staunchly following one's favorite procedure, a better idea is to adaptively choose a modeling procedure. In