Sélection de modèle PLS par rééchantillonnage bootstrap

Abdelaziz Faraj, Hicham Noçairi, Michel Constant

Institut Français du Pétrole
1&4 Av. de Bois-Préau
92500 Rueil Malmaison

{ abdelaziz.faraj, hicham.nocairi, michel.constant }@ifp.fr

Résumé. Le problème de la sélection de modèle en régression PLS est primordial pour la modélisation de phénomènes physiques même si le nombre des variables, pouvant être supérieur à celui des individus, paraît au premier abord peu important pour la mise en œuvre de la méthode. Les techniques de sélection consistent à retenir, parmi les modèles ayant un bon pouvoir de prédiction, ceux qui font intervenir le minimum de variables explicatives. La méthode que nous présentons dans ce papier est basée sur l'utilisation du bootstrap. Elle permet de calculer la distribution empirique des coefficients du modèle et de n'en conserver que les plus significatifs grâce à des tests statistiques. Elle mesure, par ailleurs, le pouvoir prédictif des modèles de régression construits aussi bien pour chaque individu que globalement. Nous illustrons cette approche en l'appliquant à un jeu de données.

Mots-Clés. Régression PLS, bootstrap, validation croisée, sélection de variables, sélection de modèles.

1 Introduction

Dans l'industrie pétrolière, la plupart des processus se présentent sous la forme d'un système à entrées-sorties (données de forage, simulations de gisement pétrolier, chimiométrie, données de procédés, etc.). Il est souvent nécessaire de faire recours à des modèles pour expliciter les relations pouvant exister entre les variables d'entrée et les réponses qui leur sont associées. De tels modèles doivent être explicatifs, c'est-à-dire éclairer les mécanismes des phénomènes physiques qu'ils décrivent. Ils doivent être prédictifs, c'est-à-dire donner, pour des valeurs des variables explicatives fixées, des sorties aussi proches que possible de celles obtenues par le processus expérimental. Et enfin – point primordial pour le praticien – ils doivent être opérationnels, c'est-à-dire pouvoir participer à l'amélioration du processus physique auquel ils sont rattachés (orienter le forage, diminuer le risque d'éboulement, augmenter la production, prédire les propriétés chimiques d'un composé, améliorer une méthode, etc.). Par ailleurs, pour des raisons de coût, le nombre des expériences qui servent à la construction de ces modèles est souvent faible voire inférieur à celui des variables en entrée