

## DOCTORAL THESIS

### Methods of variable selection and their applications in quantitative structure-property relationship (QSPR)

Peng, Xiaoling

*Date of Award:*  
2005

[Link to publication](#)

#### General rights

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

**Methods of Variable Selection and Their  
Applications in Quantitative Structure - Property  
Relationship (QSPR)**

**PENG Xiaoling**

A thesis submitted in partial fulfillment of the requirements  
for the degree of  
Doctor of Philosophy

**Principal Supervisor: Prof. FANG Kai-Tai**

**Hong Kong Baptist University**

**April 2005**

# Abstract

Quantitative structure-activity/property relationships (QSAR/QSPR) has become an important branch of modern chemistry in past decades. A fundamental goal of QSAR/QSPR studies is to predict complex physical, chemical, biological, and technological properties of chemicals from simpler descriptors, preferably those calculated solely from molecular structure. Topological indices (TIs) are numerical descriptors derived from the molecular graphs. They provide a convenient and inexpensive means of quantifying molecular structure, measuring molecular characters such as branching, shape and size. However, with the development of the topological indices, a large number of such descriptors have been proposed and their definition become more and more complex. They bring new problems to QSAR/QSPR studies. Based on the problems we met, this thesis includes two parts, one is the generalization and structural interpretation of topological indices, and the other is the applications of variable selection methods in QSPR.

In the first part of this thesis, we investigate a large amount of famous topological indices and decompose them into sets of topological character bases, different sets of character bases indicate different information of molecular structures, such as bond, atom, etc. Using the topological character bases of connectivity index  $\chi$ , we illustrate the great success of the connectivity index on many QSAR or QSPR researches in a new point of view-the impersonality of  $\chi$ 's bond weighting formula. Then, it is suggested to recompose some topological indices by adjusting the weights upon character bases according to different properties/activities. Using the method of orthogonal block variables, the character base sets are blocked to extract the most useful

information from different information subspaces (constructed by different character bases). The regression of only a few new orthogonal block variables shows large improvements both in fitting and prediction ability of the model.

The second part of my thesis is about the variable selection methods and their applications in QSPR. A new variable selection approach based on smoothly clipped absolute deviation (SCAD) penalized least squares is employed for interpretation and prediction of boiling points (BPs) of 530 alkanes. All the saturated hydrocarbons with carbon numbers from 2 to 10 and 128 common topological indices are taken into account. As a result, only 12 topological indices are selected from 95 pretreated ones but they still present a satisfying fitting and prediction effects. On the other hand, the proposed variable selection method is based on linear models. However, most of the existing relationships in QSPR cannot be described well by simple linear models. In such cases, some non-linear models should be taken into account. In the following part of the thesis, the Kriging method is considered to construct a non-linear model on variables selected by SCAD. The sequential combination of the two methods shows large improvement on the prediction ability when compared with the simple linear regression on the selected variables and the Kriging model on randomly chosen variables.

**Keywords:** Variable selection, Topological character bases, Block variables, Penalized least squares, Kriging models.

# Table of Contents

<b>Declaration</b>	<b>i</b>
<b>Abstract</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>Table of Contents</b>	<b>vi</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>ix</b>
<b>Chapter 1 Introduction and Background</b>	<b>1</b>
1.1 QSAR/QSPR and Topological Indices . . . . .	2
1.1.1 What is QSAR/QSPR? . . . . .	2
1.1.2 Topological Indices and the Graph Theory . . . . .	4
1.2 Variable Selection/Model Selection in Linear Regression . . . . .	14
1.2.1 Criteria for Model Selection and Stepwise Procedures in Linear Regression . . . . .	17
1.2.2 Selecting Variables by Linear Combination and Orthogonalization	28
1.2.3 Penalized Least-squares and Variable Selection . . . . .	31
1.3 Outline of the Thesis . . . . .	33
<b>Chapter 2 Block Variables of Topological Information Space in QSPR</b>	<b>35</b>
2.1 Impersonality of the Connectivity Index $\chi$ . . . . .	35

2.2	Recombination of Topological Indices According to Different Properties	41
2.3	Topological Character Bases and the Topological Information Space .	49
2.4	Orthogonal Block Variables of Topological Information Space in QSPR	53
2.4.1	The Method of Orthogonal Block Variables . . . . .	54
2.4.2	QSPR Models in Topological Information Space . . . . .	55
<b>Chapter 3 Variable Selection via SCAD Penalized Least-squares in BP-structure Relationship</b>		<b>59</b>
3.1	Introduction . . . . .	59
3.2	Variable Selection via SCAD Penalized Least-squares . . . . .	62
3.3	Case Study . . . . .	65
3.4	Results and Discussion . . . . .	68
<b>Chapter 4 The Combination of SCAD and Kriging Model in QSPR</b>		<b>75</b>
4.1	Introduction . . . . .	75
4.2	Construction of Kriging Models on Selected Variables in QSPR . . .	78
<b>Bibliography</b>		<b>85</b>
<b>Curriculum Vitae</b>		<b>96</b>