

DOCTORAL THESIS

Kriging model approach to modeling study on relationship between molecular quantitative structures and chemical properties

Yin, Hong

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

Kriging Model Approach to Modeling Study on Relationship between Molecular Quantitative Structures and Chemical Properties

YIN Hong

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

The molecular descriptors include various topological indices, quantum chemical descriptors, physicochemical parameters and so on. They all give structure descriptions of chemical compounds. Chemometrics, especially, quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) attempt to correlate physical, chemical and biological activities or properties with structural descriptors of compounds and find a suitable model, called metamodel, to establish relationships between molecule descriptors and activities or properties. The results are useful in theoretical and computational chemistry, biochemistry, pharmacology and environment research.

Techniques in multivariate analysis and data mining, such as ordinary least squares regression, principal components regression, partial least squares regression, multivariate adaptive regression splines and multivariate additive regression tree, are useful tools for modeling. Metamodels generated by these methods, basically are linear models with independently identical distributed (i.i.d.) random errors. However, the assumption of independent and identical distributed errors in general metamodel is not always true. For instance, many examples show that there can still be unacceptably large residuals compared to measurement errors in many models of QSAR/QSPR research. The reason for this may be diverse. The simplest and the most natural reflection on our mind is that the unaccepted residuals could be dependent. These dependent errors will present more information than independent situation. For instance, we might use a stationary Gaussian process $\{z(\mathbf{x}_i), i = 1, 2, \dots, n\}$ instead of independent random variables ϵ_i 's. In fact, the general Kriging approach just consists of parametric item and a stochastic process. In this thesis, we compared the Kriging models with other metamodels. Experiments showed that the proposed Kriging

approach could improve the regression models used widely in Chemometrics.

It is known that Kriging is an interpolating predictor which would be very beneficial for the fitting of the training data, but is not very so good for the predictions of the testing data when the data were collected with random noise $\epsilon(\mathbf{x})$. So if we add a disturbing input $\epsilon(\mathbf{x})$ in the original Kriging, the new Kriging model called empirical Kriging in some literature will provide more accurate prediction for the noisy data than the Kriging model. Many authors have paid attention to the merit of non-interpolating Kriging model. One of purposes of this thesis is to apply the empirical Kriging model to quantitative structure-activity relationship (QSAR) research. We demonstrate in the case study that the empirical Kriging model can significantly improve the prediction accuracy of other metamodels, including the Kriging models.

Table of Contents

| | |
|--|-----------|
| Declaration | i |
| Abstract | ii |
| Acknowledgements | iv |
| Table of Contents | vi |
| List of Figures | ix |
| List of Tables | xi |
| Chapter 1 Introduction | 1 |
| 1.1 What's QSAR/QSPR? | 1 |
| 1.2 What are Molecular Descriptors? | 3 |
| 1.2.1 Empirical Molecular Descriptors | 3 |
| 1.2.2 Theoretical Molecular Descriptors | 4 |
| 1.2.3 Topological Indices and the Graph Theory | 7 |
| 1.2.4 Basic Definitions | 8 |
| 1.3 Development of Modeling Strategies and Problems in QSAR | 12 |
| 1.4 Two Data Sets and Their Referenced Results | 14 |
| 1.4.1 Data Set I | 14 |
| 1.4.2 Data Set II | 18 |
| 1.5 Model Selection and Bias-Variance Tradeoff | 18 |

| | | |
|--|--|-----------|
| 1.6 | Cross-validation and Model Accuracy | |
| | Assessment | 20 |
| Chapter 2 Literature Review | | 23 |
| 2.1 | Review of Surrogate Modeling Techniques | 23 |
| 2.1.1 | Least Squares Regression | 24 |
| 2.1.2 | Ridge Regression | 26 |
| 2.1.3 | Principal Components Regression | 29 |
| 2.1.4 | Partial Least Squares Regression | 30 |
| 2.1.5 | Sliced Inverse Regression | 33 |
| 2.1.6 | Projection Pursuit Regression | 36 |
| 2.2 | Comparisons of PCR, PLSR, SIR and PPR | 40 |
| 2.2.1 | Analysis of Prediction Results | 40 |
| Chapter 3 Understanding Kriging | | 44 |
| 3.1 | Introduction | 44 |
| 3.2 | History | 45 |
| 3.3 | Kriging Model | 46 |
| 3.3.1 | Derivation of the Prediction Formula | 47 |
| 3.3.2 | Parameter Estimation | 50 |
| 3.3.3 | Applications of Kriging Models to Data I | 51 |
| 3.3.4 | Experiments and Comparisons | 56 |
| 3.4 | Empirical Kriging | 60 |
| 3.4.1 | Models, Estimation, and Prediction | 64 |
| 3.4.2 | Improved Results by Empirical Kriging | 66 |
| 3.4.3 | Analysis of Data Set II by Kriging and Empirical Kriging | 68 |
| 3.5 | The applications of Kriging and empirical Kriging models based on the variables selected by SCAD | 69 |
| 3.5.1 | Computations | 70 |
| 3.5.2 | Results analysis | 70 |
| 3.6 | Penalized Likelihood in Gaussian Kriging Models | 72 |
| 3.6.1 | Penalized Likelihood and Kriging | 74 |

| | | |
|-------------------------|---|-----------|
| 3.6.2 | Theoretic Aspects | 75 |
| 3.6.3 | Penalty Function and Algorithm for Parameter Estimation . . | 76 |
| Chapter 4 | Future Work | 78 |
| 4.1 | Sensitivity Analysis | 78 |
| 4.2 | The combinations of Bagging or Boosting Techniques with Kriging . . | 79 |
| Bibliography | | 80 |
| Curriculum Vitae | | 91 |