Chapter 3
From "Clothing Standard" to "Chemometrics"

Review of Prof. Kai-Tai Fang's Contributions in Data Mining

Ping He, Xiaoling Peng, and Qingsong Xu

Abstract This paper reviews Prof. Kai-Tai Fang's contributions in data mining. Since the 1970s, Prof. Fang has been committed to applying statistical ideas and methods to deal with large amounts of data in practical projects. By analyzing more than 400,000 pieces of data, he found representative clothing indicators and established the first adult clothing standard in China; through cleaning and modeling steelmaking data from steel mills all over the country, he revised the national standard for alloy structural steel; by studying various data in chemometrics, he introduced many new advanced statistical methods to improve the identification and classification of chemical components, established more effective models for the relationship between quantitative structure and activity, and promoted the application of the traditional Chinese medicine (TCM) fingerprint in TCM quality control. Professor Fang and his team's research achievements in data mining have been highly appreciated by relevant experts. This article is written to celebrate the 80th birthday of Prof. Kaitai Fang.

P. He (✉) · X. Peng
Beijing Normal University and Hong Kong Baptist University United International College, Zhuhai, China
e-mail: heping@uic.edu.hk

X. Peng
e-mail: xlpeng@uic.edu.hk

Q. Xu
Department of Mathematics and Statistics, Central South University, Changsha, China
e-mail: qsxu@csu.edu.cn

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_3

3.1 Introduction

With the rapid development of computer technology, data collection and storage in various industries has greatly improved, and the amount of data has increased dramatically. Various new statistical applications have emerged, showing their vital role in the development of all fields. In this revolutionary data-driven evolution, statistics, as the core foundation and technology of data science, has unprecedentedly developed in recent years. Many new theories and methods of statistics have been proposed to deal with new types of data that have broken through the scales and types of classical data.

In China, as the economy developed, so did the mutual integration of various disciplines and the rapid development of computer technology, which resulted in the government, enterprises and society continuously expanding and promoting the application of statistics. The technology of data mining receives increasing attention and is widely used in various fields. All of these are inseparable from the efforts and contributions of statisticians.

Professor Kai-Tai Fang has always been committed to research on data mining and promoting the application of statistics in practical fields. He, in cooperation with experts in other fields, undertook many practical projects, such as establishing the first Chinese adult clothing standards by analyzing body size measurements, assessing and optimizing the national standard for alloy structural steel through processing the steel-making data from all state-owned steel mills, identifying the composition and structure of the compounds via machine learning put into mass spectral databases, and improving the efficiency of traditional Chinese medicine (TCM) quality control by analyzing TCM fingerprints. As well as these accomplishments, Prof. Fang wrote multiple books. He together with Prof. Yuan Wang published *Number-Theoretic Methods in Statistics* [7]. The theories and methods introduced in the book belong to the intersection of number theory, statistics, and computer science and widely used in computer simulation experiment, agriculture, industry, medicine, and high-tech innovation. Also, the book *Introduction to Multivariate Statistical Analysis* [9] by Profs. Fang and Yaoting Zhang not only has been adopted as a textbook by many universities, but also has been recommended as one of the main reference books for those engaged in work related to statistics and data mining.

3.2 Establishment of the First Chinese Adult Clothing Standard

In the early 1970s, as the Chinese population grew, making clothing needed to be automated, controlled, and standardized for mass production to meet the growing demand. However, at that time, China did not have a national-level clothing standard to provide a reliable basis for the design specifications needed to produce garments. For this reason, Chinas Ministry of Textiles, Ministry of Light Industry and State

Administration of Standards launched a cooperative project to set the first national adult clothing standard.

A clothing standard is a formulation of a series of specifications based on the shape and size of the human body. To create a standard, peoples body types needed to be divided into several categories where the most representative measurements were calculated for each category. Then, clothes would be designed according to these specified parameters to meet the needs of most people. A good clothing standard could not contain too many specifications and also had to accommodate for a wide range of shapes so that people with ordinary bodies could easily buy clothes that fit. Crafting a clothing standard at the national level was an undertaking that required many experts.

Originally, the project members undertaking this monumental task were composed of senior tailors and relevant industry leaders from Beijing, Shanghai, Tianjin and other cities. They conducted surveys on the human body in more than ten provinces across the country. Using stratified sampling, the members selected more than 400,000 men and women for body measurements; men had twelve measurements and women had fourteen. However, they did not know how to set the clothing standard using the vast data, so they just made a preliminary analysis of the figures and then drafted a standard mainly based on their experience. Due to the lack of data-driven evidence and theoretical bases, this clothing standard was not adopted by the State Administration of Standards.

Researchers realized that mathematical and statistical methods should be used to analyze the data and thus provide a reliable basis for the creation of standards. In 1974, Prof. Kaitai Fang was invited to join the standards-setting research team. He found that principal component analysis (PCA), which was widely used in the world to develop the clothing standard, was not suitable for the development of the Chinese standard. The main reason was that the principle components in PCA were too difficult for Chinese clothing workers to understand and master at the time. Therefore, Prof. Fang developed a brand-new method to formulate easy-to-implement Chinese clothing standards [3].

Professor Fang introduced a statistical idea to develop the clothing standard: if a variable can well represent others, then given the variable, the conditional standard deviation of these other variables should be small, and vice versa. Using this idea and optimal design theory in multivariate analysis, He proposed a method to sequentially found the variables with the smallest generalized conditional variance. In the article [3], the data of adult womens clothing in Beijing were adopted as an example. The results showed that height, chest circumference and waist circumference were the most representative measurement sites among all the body measurements and could be used for further body type classification. Then, the most representative values were calculated in each category.

According to the classification of body type and the corresponding representative measurements, manufacturers could mass-produce garments in large batches. The developed clothing standard ensured that most people could buy ready-to-wear clothing and that only a small fraction of people with special body types would need tailored clothing. On behalf of the project team, Prof. Fang reported the Beijing clothing standard formulated by his method to the three state offices (the Ministry of Textiles, the Ministry of Light Industry and the State Administration of Standards) where the standard received unanimous approval.

Later, numerous clothes were produced for most body types according to the project's results. The market feedback was excellent, and thus the method proposed by Prof. Fang to calculate the Chinese adult clothing standard was a success. After many years of development, the team calculated and formulated clothing standards for each of China's sub-regions. Their project "Series of Standards for Chinese Adult Heads" won the 1980 Science and Technology Achievement 3rd Prize Award issued by the Beijing Government; their project "Chinese Adult Clothing Standards" won the 1982 Special Prize of Light Industry of the People's Republic of China. On December 10th, 1988, the national standard GB10000-88 for the Chinese adult body size was officially released.

It is worth mentioning that, in 1982, Prof. Fang came up with a new method to further improve the clothing standard and theoretically defined the concept of representative points for statistical distributions. He also provided numerical algorithm of computing representative points for univariate normal distribution [6]. However, it was a pity that his theory was highly overlapping with that of Prof. Cox [2], a famous British statistician. Professor Fang wasn't discouraged and instead proposed NTLBG algorithm [10] based on number theory and k-means algorithm in 1994, which can obtain the representative points of elliptically contoured distributions. In 2014, he and his students used representative points and random sampling significantly improved the efficiency of the Monte Carlo method [16].

3.3 Revision of the National Standard for Alloy Structural Steel

Alloy structural steel is made metallurgically by adding elements chromium, manganese, nickel, molybdenum silicon and other elements to steel. During the manufacturing process, the contents of these elements must be controlled within a certain scope to ensure that the five mechanical properties, including characteristics such as strength and elasticity, of the produced steel to meet the requirements.

During the 1960s, there already was a national standard for the range that each elements contents should fall into. This standard was used by more than 10 factories all over the country to produce the same kind of alloy structure steel, however, the results were inconsistent. While some steel mills produced a high proportion of qualified alloy steel, other steel mills had low proportions of qualified alloy steel even if the content of all elements, such as carbon, chromium, manganese, and etc. were full compliance with the national standard: Qiqihar Steel Plant, for instance, produced qualified steel only 38% of the time. Thus, many steel mills entertained doubts about the national standard for the scope of contents of elements. This standard was introduced from the Soviet Union. At that time, no one knew its principles, and the reasons for the inconsistent production qualified rate. Since refining a batch of low quality alloy structural steel is expensive, the manufacturers not meeting the national standard would suffer heavy economic losses. Therefore, it was important and urgent to study whether the standard being used was reasonable or whether there was room for improvement.

In 1973, Prof. Kai-Tai Fang and his colleagues took on a project from China's Ministry of Metallurgy to review the national standard for the ranges of each element present when forging alloy structural steel. The volume of data collected in this project was enormous; it included relevant data from all state-owned steel mills. It took the team, composed of professors and engineers, half a year to preprocess the copious amounts of various data. For example, some steel mills used the proportion of steel elements to estimate whether the alloy steel was qualified based on experience, and if the estimate was disqualified, a different treatment would be carried out in the steels quenching process to make it meet the standard. This special kind of treatment make these data need to be eliminated when preprocessing.

After cleaning the data, Prof. Fang and his colleagues built regression models, predicted the five mechanical properties of steel with its elements and did five-fold integral calculation before getting the qualified rate. Finally they found that: the national standard is scientific and reasonable; the combination ratio of elements in steel will affect mechanical properties' rates of meeting the standard, therefore, in the steel making process, one should try to choose the combination which produces the highest qualified rates; and it is correct of some steel mills' experience that whether the steel is qualified or not can be estimate by the element ratio [5].

From this project, Prof. Fang also came up with some new statistics theoretical problems. He refined the metallurgical problem into a extremum problem based on multi-dimensional normal probability density, and proved the necessary condition for the existence of solutions to the extremum problem in 1979 [8]. Also in this paper, he presented an effective algorithm for finding solutions to the extremum problems. While solving these problems, Prof. Fang faced difficulties in calculating the probability of a multidimensional normal distribution. He had to calculate multiple integrals without a computer's aid since the technology at that time could not meet the demands of these problems. Using a suggestion from Prof. Yuan Wang, a famous mathematician, Prof. Fang employed the good lattice method (GLM) [4] and efficiently calculated the multiple integrals. This method also laid the foundation for Prof. Fang's later invention: uniform experimental design.

Although data mining has become popular in the last three decades, Prof. Fang pioneered similar work in the 1970s, when domestic computer technology was still lagging. Through communicating extensively with experts in related field, he carefully preprocessed data, analyzed data, constructed model, tested model and finally obtained convincing conclusions. He not only successfully completed the actual project, but also refined the specific problems into general theory. The general problems were studied and discussed, and solutions that may arise in other practical work were proposed for similar problems.

Speaking of the standard, Prof. Fang also devoted himself in the promotion of standard of statistics. The International Organization for Standardization (ISO) set the standard of ISO5725 to measure the testing precision of an instrument with repeatability and reproducibility. In order to introduce this standard to China, the Standard Administration of China established a specialized committee, in which Prof. Fang was designated as the chair. He explained the statistical theory of the standard to committee members in detail and spent more than two years completing the national standard GB/T6379 [14]. This project was awarded the second prize of Standardization Administration of China. This is a standard of statistics, which mainly includes model of variance analysis of random effect, the elimination of abnormal data and linear regression. Afterwards, Prof. Fang participated in international ISO5725 committee on behalf of China on many occasions.

3.4 Contributions to Chemometrics

Professor Kai-Tai Fang's contribution to chemometrics began by collaborating with Prof. Yizeng Liang, a celebrated analytical chemist from Central South University: Prof. Liang received the Chemometrics Lifetime Achievement Award at the XVI International Conference on Chemometrics in Analytical Chemistry. He has been committed to working on the application of statistical methods in analytical chemistry and has a profound understanding of statistical theories and methods. In 1995, during Prof. Liang's visit to the chemistry department of Hong Kong Baptist University (HKBU), he met Prof. Fang, who was then a professor in the mathematics department of HKBU.

When he learned that Prof. Fang was working on the application of number-theoretic methods in statistics, Prof. Liang was fascinated and studied Prof. Fang's book *Number-Theoretic Methods in Statistics* [7], which had just been published by Profs. Fang and Yuan Wang. In 1996, Profs. Liang and Fang published an article together in the academic journal *Analyst* [24] which proposed a robust multivariate calibration method based on the least median of squares (LMS) and the number-theoretic methods in optimization (SNTO) algorithm. Compared with the least squares method, the proposed method significantly reduces the computational complexity when the analysis system has two or three components. In addition, the method is more robust when there are more abnormal values.

The SNTO algorithm mentioned above is an optimization algorithm for multivariate functions proposed by Profs. Fang and Wang. It uses number theory to evenly distribute points in the search space, and gradually reduces the optimization search space by sequential compression to find the global optimal solution of multivariate functions [7]. In 1997, Profs. Fang and Liang further applied the SNTO algorithm to the constrained background bilinear decomposition model for the quantitative analysis of analytical samples containing unexpected interference [35]. In this paper, the SNTO algorithm was compared with another global optimization algorithm: variable step size generalized simulated annealing (VSGSA). The results showed that when the two methods achieve the same analytical accuracy, the SNTO algorithm is simpler, clearer and easier to implement, making it a practical tool in chemometrics.

In the same year, Prof. Liang advised his colleagues to adopt a sequential uniform design (SUD) procedure for the separation of five dithiocarbamate (DTC) compounds by capillary electrophoresis (CE). The CE technique was unable to separate these five DTC compounds when changing one variable at a time, whereas they were completely separated by using SUD method [23]. In addition, the SUD procedure was introduced as a promising candidate for experimental design in nonlinear multivariate calibration with ANN [36].

Regarding Prof. Fang’s contribution to chemometrics, the application of uniform experiment design in chemical experiments must be mentioned. In 1998, Prof. Liang and his Ph.D. student Qingsong Xu were invited by Prof. Fang to visit the mathematics department of HKBU. They noticed that Atkinson, a well-known statistician in optimal experimental design, recommended using *D* optimal design and *T* optimal design to estimate the kinetic rate in reversible chemical reactions [1]. Thus, Profs. Fang and Liang decided to compare the performance of orthogonal experimental design (OD), *D* optimal design (DOD) and uniform experimental design (UD) in reversible chemical reactions.

Their studies showed that for nonlinear reversible reaction kinetic models, DOD usually performs best if the initial value is not far from the true parameter and the random error is not large. It’s sensitivity to the choice of initial values is a drawback: if the initial value is far from the true parameter, then the parameter estimate is likely to fail. When compared with DOD, OD is less sensitive to the location of initial parameters, but as the random error increases to a certain level, OD faces a similar problem: if the initial value is far from the real parameter, the parameter estimation is also likely to fail. When there is no prior information about the location of real parameters and random noise intensity, UD always performs best among the three designs. The results were summarized in the article [30].

In 2001, Profs. Fang, Liang and Dr. Qingsong Xu published an article “Uniform design and its applications in chemistry and chemical engineering” [25]. The article has had a significant impact in chemistry and chemical engineering: currently, the SCI citation rate has reached 255. In the same year, Prof. Fang was invited by the 50th Gordon Research Conference titled “The Statistics in Chemistry and Chemical Engineering” to deliver a one hour lecture to introduce uniform design. Many chemists and chemical engineers have shown interest in UD and hope to develop an in-depth understanding of UD. Since then, Prof. Fang still receives letters from chemists, requesting to construct uniform experimental tables for their research.

Professors Fang and Liang are largely celebrated today for their work together in chemometrics. In 2016, Prof. Ruqin Yu, fellow of the Chinese Academy of Sciences and the former President of Hunan University, invited Prof. Fang to write an article for the special issue of the *Journal of Chemometrics* in China. In his letter to Dr. Qingsong Xu, Prof. Yu said:

> I am thinking about a problem: throughout the history of chemometrics in China, there are few Chinese original contributions, and most of them are applications of methods proposed by others. You and Prof. Liang worked with Prof. Fang to develop the application of uniform experimental design in chemistry. Uniform experimental design is an entirely original Chinese innovation. It must be stressed that, somehow in this special topic, Chinese scientists still have original innovations in chemometrics. The value of uniform experimental design itself is that it is an original innovation in mathematics.

In 2018, Prof. Fang and Dr. Xu published an article in the *Journal of Chemometrics* reviewing the development of uniform design in chemometrics and its various applications [32].

3.5 Research Group's Further Contributions to Chemometrics

After Profs. Kai-Tai Fang, Yizeng Liang, and Dr. Qingsong Xu published their 2001 article about uniform design in chemistry and chemical engineering, Profs. Fang and Liang further strengthened their cooperation in data mining in chemometrics. In 2002, they organized a series of research seminars where Prof. Fang's five Ph.D. students and Prof. Liang's six Ph.D. students attended all of them together. They also invited each other's Ph.D. students to attend the other school for at least a month to deepen their understanding of the other field, strengthen the discussion and exchange ideas. Together, the two groups have done a series of work involving many aspects of chemometrics.

The research group led by Profs. Fang and Liang worked on an important area in traditional chemometrics: the topological structure representation of organic compounds. In the article [20], various matrix representations, topological indices and atomic properties of topological structures were summarized and the shortcomings of topological indices were discussed. Then in the articles [21, 22], they combined projection pursuit and number theory to mine the hidden structural feature information in the space formed by multiple topological indices, which is associated with some certain chemical properties.

Mass spectrometry is another important aspect of chemometrics and has always been one of the essential methods for the identification and characterization of compounds. With the development of mass spectrometry technology, databases containing mass spectra of a large number of compounds and their other chemical information were established, such as NIST Library and Wiley Library. When the mass spectrum of the compound to be identified already exists in the mass spectrometry database, the computer retrieval method usually performs well. However, existing mass spectrometry libraries contain only a small fraction of the number of compounds: the Chemical Abstracts Service describes more than 200 million natural compounds. Therefore, when the mass spectrum of the compound is not in the existing mass spectrometry database, experts hope that some substructures of the compound can be identified by studying the existing mass spectrometry library.

The research group was also engaged in this research area. Their article [28] proposed a method for detecting the corresponding compound substructure by searching the peak combination of the mass spectrum and then using that peak combination for further compound identification and classification. They also proposed a method which applied the combination of the sliced inverse regression (SIR) and a decision tree to the mass spectrometry data for the identification of the substructure of the compound, which is published in [17].

Professors Fang and Liang’s research group also studied quantitative structural activity relationships (QSAR). QSAR research has always been an important branch of chemometrics, which aims to establish a quantitative relationship between active proprieties of compounds and their structural parameters through appropriate mathematical statistics methods. They adopted SCAD, a variable selection method with the oracle property, in [26] and selected 12 out of 128 topological indexes to establish the explainable connection between molecular boiling point and molecular structure. The articles [15, 34] introduced the Kriging model which is derived from geostatistics and the improved empirical Kriging model into the study of QSAR. Then they combined SCAD with the Kriging model and established the empirical Kriging model with selected important variables, as shown in the article [27]. This scheme has been applied in QSAR research and obtained better results than prior research.

As well as applying traditional statistical methods to where they were needed, the research group introduced the latest and most advanced statistical methods into chemometrics. In the article [31], they employed a two-step multivariate adaptive regression spline (TMARS) to show the relationship between the alkane retention index and the molecular structure. Later, in the subsequent three articles [18, 19, 29], they applied a boosting algorithm to improve the classification performance for the different types of chemical data.

The last contribution to chemometrics mentioned in this paper is Profs. Fang, Liang and the research group’s work with Traditional Chinese Medicine (TCM). They promoted and studied the data mining of TCM. Using statistical analysis and pattern recognition to indicate the authenticity of herb medicine and its main components has been widely used in the field of quality control for Chinese herbal medicines. A TCM fingerprint refers to the chromatogram or spectrogram that can be used to identify the chemical characteristics of TCM which has been properly treated by certain analytical means. In the paper [11], Prof. Fang employed a bootstrap method to estimate the probability distribution of the correlation coefficients of TCM fingerprints between the unknown test samples and the standard fingerprints under the assumption that they belong to the same category, and thus provided the test’s critical value for evaluating whether the fingerprint of an unknown test sample is qualified. In addition, he assessed the phylogenetic relationships of Lycium samples via random amplified polymorphic DNA (RAPD) and entropy theory in the paper [33].

In July 2010, Prof. Fang was invited to deliver a speech titled “A Challenge Research Direction in Biostatistics-Chinese Medicine” at the First Joint Biostatistics Symposium. The speech introduced the current status and challenges of research on the fingerprints of traditional Chinese medicine. In order to better communicate with international counterparts, some papers by the research group were selected for publication in a special issue of the *Journal of Data Science* for data mining in chemometrics. In addition, Prof. Fang and his other collaborator Prof. Yu compiled two volumes titled *Data Mining and Bioinformatics in the field of Chemistry and Traditional Chinese Medicine* [12, 13], which have been highly regarded by international peers.

3.6 Summary

Professor Kai-Tai Fang has been engaged in the field of statistics for decades. He not only has devoted himself to the research of statistical theory, but also has been committed to promoting the development of statistical applications.

With a statistician's astute insight, Prof. Fang saw data mining would be an important research field of statistics and addressed the topic earlier than most statisticians. He has put a lot of energy into actively learning new theories, developing new methods, and courageously putting them into practice. Through in-depth communication with experts in other fields, he has studied and solved many practical problems. Even in the 1970s when computer technology wasn't developed in China, Prof. Fang persisted in overcoming various difficulties and successfully completed national projects.

Professor Fang realized early that in the era of data, statisticians need to adapt themselves to the development of the world, actively embrace data science and carry out research on statistical theories and methods based on actual needs. This attitude is demonstrated as various collaborative research by Prof. Fang's research groups. They made important contributions to the further application and development of statistics in chemometrics.

Acknowledgments This work was partially supported by Guangdong Natural Science Foundation (No. 2018A0303130231) and Guangdong Innovation and Enhancement Project: Education Research Programme (R5201920).

References

1. Atkinson, A.C., Bogacka, B., Bogacki, M.B.: D- and T-optimum designs for the kinetics of a reversible chemical reaction. Chemom. Lab. Syst. **43**, 185–198 (1998)
2. Cox, D.R.: Note on grouping. J. Am. Stat. Assoc. **52**, 543–547 (1957)
3. Fang, K.T.: Using conditional distributions to formulate a national clothing standard. J. Appl. Math. **2**, 63–74 (1976). (Chinese)
4. Fang, K.T.: Uniform design—application of number theory method in experimental design. J. Appl. Math. **3**, 363–372 (1980). (Chinese)
5. Fang, K.T.: Endless and unbending journey to statistics research: the oral autobiography of Kai-tai Fang. Hunan Education Press (2016) (in Chinese)

6. Fang, K.T., He, S.: The problem of selecting a given number of representative points in a normal population and a generalized mill's ratio. Technical report, No. 5, Department of Statistics, Stanford University, USA (1982)
7. Fang, K.T., Wang, Y.: Number-Theoretic Methods in Statistics. Chapman and Hall, London (1994)
8. Fang, K.T., Wu, C.Y.: A probability extremum problem. J. Appl. Math. **2**, 132–148 (1979) (in Chinese)
9. Fang, K.T., Zhang Y.T.: Introduction to Multivariate Statistical Analysis. Science Press (1982) (in Chinese)
10. Fang, K.T., Bentler, P.M., Yuan, K.H.: Applications of number-theoretic methods to quantizers of elliptically contoured distributions. In: Multivariate Analysis and Its Applications, IMS Lecture Notes—Monograph Series, pp. 211–225 (1994)
11. Fang, K.T., Liang, Y.Z., Yin, X.L., Chen, K., Lu, G.H.: Critical value determination on similarity of fingerprints. Chemom. Intell. Lab. Syst. **82**, 236–240 (2006)
12. Fang, K.T., Liang, Y.Z., Yu, R.Q. (eds.): Data Mining and Bioinformatics in Chemistry and Chinese Medicines. Hong Kong Baptist University (2003)
13. Fang, K.T., Liang, Y.Z., Yu, R.Q. (eds.): Data Mining and Bioinformatics in Chemistry and Chinese Medicines, vol. 2, Hong Kong Baptist University (2004)
14. Fang, K.T., Xiang, K.F., Liu, G.Y.: Precision of Test Method. China Standard Press, Beijing (1988)
15. Fang, K.T., Yin, H., Liang, Y.Z.: New approach by Kriging methods to problems in QSAR. J. Chem. Inform. Model. **44**, 2106–2113 (2004)
16. Fang, K.T., Zhou, M., Wang, W.J.: Applications of the representative points in statistical simulations. Sci. China Ser. A **57**, 2609–2620 (2014)
17. He, P., Fang, K.T., Xu, C.J.: The classification tree combined with SIR and its applications to classification of mass spectra. J. Data Sci. **1**, 425–445 (2003)
18. He, P., Fang, K.T., Liang, Y.Z., Li, B.Y.: A Generalized boosting algorithm and its application to two-class chemical classification problem. Analytica Chimica Acta **543**, 181–191 (2005)
19. He, P., Xu, C.J., Liang, Y.Z., Fang, K.T.: Improving the classification accuracy in chemistry via boosting technique. Chem. Intell. Lab. Syst. **70**, 39–46 (2004)
20. Hu, Q.N., Liang, Y.Z., Fang, K.T.: The matrix expression, topological index and atomic attribute of molecular topological structure. J. Data Sci. **1**, 361–389 (2003)
21. Hu, Q.N., Liang, Y.Z., Peng, X.L., Yin, H., Fang, K.T.: Structural interpretation of a topological index. 1. External factor variable connectivity index (EFVCI). J. Chem. Inf. Comput. Sci. **44**, 437–446 (2004)
22. Hu, Q.N., Liang, Y.Z., Yin, H., Peng, X.L., Fang, K.T.: Structural interpretation of a topological index. 2. The molecular connectivity index, the Kappa index, and the atom-type E-state index. J. Chem. Inf. Comput. Sci. **44**, 1193–1201 (2004)
23. Lee, A.W.M., Chan, W.F., Yuen, F.S.Y., Tse, P.K., Liang, Y.Z., Fang, K.T.: An example of a sequential uniform design: application in capillary electrophoresis. Chem. Intell. Lab. Syst. **39**, 11–18 (1997)
24. Liang, Y.Z., Fang, K.T.: Robust multivariate calibration algorithm based on least median squares and sequential number theoretic optimization method. Anal. Chem. **121**, 1025–1029 (1996)
25. Liang, Y.Z., Fang, K.T., Xu, Q.S.: Uniform design and its applications in chemistry and chemical engineering. Chem. Intell. Lab. Syst. **58**, 43–57 (2001)
26. Peng, X.L., Hu, Q.N., Liang, Y.Z.: Variable selection via nonconcave penalty function in structure-boiling points correlations. J. Mol. Struct.: THEOCHEM, **714**, 235–242 (2005)
27. Peng, X.L., Yin, H., Li, R., Fang, K.T.: The application of Kriging and empirical Kriging based on the variables selected by SCAD. Analytica Chimica Acta **578**, 178–185 (2006)
28. Tang, Y., Liang, Y.Z., Fang, K.T.: Data mining in chemometrics: sub-structures learning via peak combinations searching in mass spectra. J. Data Sci. **1**, 481–496 (2003)
29. Varmuza, K., He, P., Fang, K.T.: Boosting applied to classification of mass spectral data. J. Data Sci. **1**, 391–404 (2003)

30. Xu, Q.S., Liang, Y.Z., Fang, K.T.: The effects of different experimental designs on parameter estimation in the kinetics of a reversible chemical reaction. Chem. Intell. Lab. Syst. **52**, 155–166 (2000)
31. Xu, Q.S., Massart, D.L., Liang, Y.Z., Fang, K.T.: Two-step multivariate adaptive regression spline for modeling a quantitative relationship between gas chromatography retention indices and molecular descriptors. J. Chromatogr. A **998**, 155–167 (2003)
32. Xu Q.S., Xu Y.D., Li L., Fang K.T.: Uniform experimental design in Chemometrics. J. Chem. e3020 (2018). https://doi.org/10.1002/cem.3020
33. Yin, X.L., Fang, K.T., Liang, Y.Z., Wong, R.N.S., Ha, A.W.Y.: Assessing phylogenetic relationships of Lycium samples using RAPD and entropy theory. Acta Pharmacologica Sinica **26**, 1217–1224 (2005)
34. Ying, H., Li, R.Z., Fang, K.T., Liang, Y.Z.: Empirical Kriging models and their applications to QSAR. J. Chem. **20**, 1–10 (2007)
35. Zhang, L., Liang, Y.Z., Yu, R.Q., Fang, K.T.: Sequential number-theoretic optimization (SNTO) method applied to chemical quantitative analysis. J. Chem. **11**, 267–281 (1997)
36. Zhang, L., Liang, Y.Z., Jiang, J.H., Yu, R.Q., Fang, K.T.: Uniform design applied to nonlinear multivariate calibration by ANN. Analytica Chimica Acta **370**, 65–77 (1998)