Mixture models based on power means and generalised Q-fractions

Author:

Maria Helena Ackermann

Supervisor:

Prof. Walter Focke

University of Pretoria: Institute of Applied materials

Co-supervisor:

Dr. Roelof Coetzer

Sasol: Research and Development

A thesis submitted in partial fulfilment of the requirements for the degree of

MSc in Applied Science: Chemical Technology

Faculty of Engineering, Built Environment and Information Technology

University of Pretoria, Pretoria

September 12, 2011

University of Pretoria

Contents

Abstract

Mixture experiments are widely applied. The Scheffé quadratic polynomial is the most popular mixture model in industry due to its simplicity, but it fails to accurately describe the behaviour of response variables that deviate greatly from linear blending. Higher-order Scheffé polynomials do possess the ability to predict such behaviour but become increasingly more complex to use and the number of estimable parameters grow exponentially [15]. A parameter-parsimonious mixture model, developed from the linear blending rule with weighted power means and Wohl's Q-fractions, is introduced. Bootstrap is employed to analyse the model statistically. The model is proved to be flexible enough to model non-linear deviations from linear blending without losing the simplicity of the linear blending rule.

Keywords:
Mixture experiments; Mixture models; Experimental design; Scheffé quadratic polynomial; Bootstrap; Wohl's Q-fractions; Weighted power mean

Acknowledgements

Supervisors

Prof. Walter Focke

"If you wait to do everything until you're sure it is right, you'll probably never do much of anything." -Win Borden-

Dr. Roelof Coetzer

"Years teach us more than books."-Berthold Auerbach-

Family

My husband, Martin Ackermann

"You are my box of Smarties. What-a-lot-I-got" -Mia Ackermann-

My brother-in-law, Etienne Ackermann

"Those that know, do. Those that understand, teach."-Aristotle-

My parents, Johan & Aletta Knoetze and my brother, Johan-Simeon Knoetze

"Family...a group experience of love and support."-Marianne Williamson-

Friends

The "Groendakkie Chemie-Klub": Nina van Jaarsveld
Belinda van der Westhuizen
Elize Smit
Shandré Buitendacht

"I get by with a little help from my friends."-The Beatles-

"I'll lean on you and you lean on me and we'll be okay."-Dave Matthews Band-

List of Figures

List of Tables

List of Abbreviations

AAD	Average absolute deviation
ANOVA	Analysis of variance
BPE	Bootstrap prediction error
MAD	Maximum absolute deviation
MON	Motor octane number
MSE	Mean square error
RON	Research octane number
RSM	Response surface methodology
SSE	Sum of squared error
SSR	Sum of squares of regression
SST	Total sum of squares

List of Symbols

a_i	Suitably chosen parameter characteristics of component i.
B	Number of bootstrap samples.
cov	Covariance.
E	Expected value operator.
$\hat{f}_i$	Frequency with which component w_i from sample W appears in bootstrap replicate sample W^*.
F	Chapter 1: F-ratio.
F	Chapter 4:Unknown population distribution in bootstrap methodology.
l	Number of lattice points.
lb	Lower bound.
L	The sum of the squared errors.
m	Chapter 1:Integer associated with spacing of lattice points. The degree of a Scheffé canonical polynomial.
m	Chapter 4: Number of times a data set is resampled in the bootstrap methodology.
n	Degree of the polynomial.
p	Number of different blends. Number of terms in the model.

PE	Prediction error.
q	Number of mixture components.
Q_i	Generalised composition variable known as a Q-fraction.
r_i	Chapter 1: Number of experimental runs on blend or at lattice point.
r, s	Chapter 3: Adjustable parameters associated with the Q-fraction model.
se	Standard error.
R^2	Correlation coefficient.
R^2_A	Adjusted correlation coefficient.
$t_{f,\alpha}$	Tabulated value of the t-distribution with f degrees of freedom at a level of significance of α.
ub	Upper Bound.
v_i	Volume fraction.
var	Variance.
w_i	Chapter 3: Mass fraction.
w_i	Chapter 4: Component i of sample W.
w_i^*	Chapter 4: Component i of sample W^*.
W	Sample from which bootstrap replications is drawn.
W^*	Bootstrap sample generated by resampling from W.
x_i	Composition proportion of component i in the mixture.
$\mathbf{X'X}$	Information matrix.
y	Observed response variable.
z_i	Mole fraction.
$\hat{}$	Indicates an estimated value
$\bar{}$	Indicates an average value
*	Chapter 2: Redefines a parameter.
*	Chapter 4: Indicates parameters and variables associated with the bootstrap sample.

Greek Symbols

α	Level of significance.
β_i	Parameter associated with a pure component.
$\beta_{ij\ldots k}$	Parameter associated with a blend of components $, i, j, \ldots k$
ε	Experimental error. Difference between the estimated response and the actual response.
η	Mean of the observed experimental response variables
σ^2	Chapter 1:Variance of the distribution of the errors ε.
σ^2	Chapter 4:Variance descriptor.
θ	Set of estimable parameters in bootstrap methodology.
θ^*	Set of estimable parameters generated from bootstrap samples in bootstrap methodology.

Subscripts

AE	Apparent error.
BE	Bootstrap error.
FE	Final error.

Introduction

Mixture models are everywhere. Chemical engineers employ mixture models to predict excess thermodynamic properties such as Gibbs free energy and the enthalpy of polar and non-polar mixtures [30, 44]. Materials science uses them to predict the physical properties of metal alloys and the behaviour of porous media. The physical properties of cement mixtures, such as compressive strength, are modelled using mixture experiments [2, 21, 52, 61]. The food industry is greatly dependent on mixture experiments as all its formulations depend on the proportions of the ingredients [10, 19]. Mixture models have been applied in product quality improvement by identifying the optimal blend [7, 56]. A corporation in France has even employed mixture designs to blend a table wine from various other wines [64]. Other examples for which mixture experiments are used; tensile strength, physical properties of metal alloys, even cake formulations and tobacco blends [15, 35, 57].

A mixture experiment is defined as an experiment in which the response variable is not dependent on the total amount of the mixture but on the relative amounts of the mixture components. The proportions of the components can be expressed by volume, weight, mole, etc., as long as the proportion values are greater than zero and sum to one. Under these constraints the factor space for the measured responses is a $(q-1)$-dimensional simplex, where q is the number of components in the mixture. Mixture experiments are employed either to predict a response variable (such as the fluffiness of a cake), to

screen the components of a mixture to simplify the problem, to measure the effects of the components on the response variable, or to optimise the response variable over the experimental region [3, 15]. An interesting application of mixture experiments was a plant diversity experiment where the response variable was known and used to predict the elementary components [59]. If the focus of a mixture experiment is to optimise a particular response variable, Response Surface Methodology (RSM) is employed. A response surface is generated by plotting the measured response perpendicularly above the corresponding coordinate of component proportions. The shape of the response surface provides insight into the properties of the response variable. An aid in visualising the shape of response surfaces is contour plots. Each contour corresponds to a particular height of the response surface [15]. These concepts will be illustrated in Chapter 1. To analyse a response surface successfully, three key areas need to be addressed: (i) a proper model needs to be developed to approximate the response surface over the experimental region;(ii) a suitable experimental design needs to be implemented during data collection; and (iii) model adequacy has to be tested [43]. These three areas form the foundation of a successful mixture experiment.

Mixture Models: The fundamental purpose of a mixture experiment is to develop a mathematical model that can accurately describe the dependence of the response variable on the mixture components. Various mixture models for different experimental setups have been developed over the years and a selection of these will be addressed in Chapter 2. In 1958, Henry Scheffé published his pioneering article *Experiments with mixtures* and became the first to develop a model for a mixture experiment [57]. Despite an abundance of very sophisticated models in the literature today, it is still Scheffé canonical polynomials that are most commonly used [51]. Their popularity can be ascribed to their simplicity. They are often of low degree, which implies that fewer observations are required to estimate the parameters and there are fewer terms to interpret, making them easier to understand and handle [15]. It is most expedient to keep these properties in mind when developing a mixture model as an alternative for Scheffé canonical polynomials. The proposed parameter-parsimonious mixture models will be introduced in Chapter 3.

Experimental Design: The results and conclusions that can be drawn from an experiment are directly influenced by the manner in which the data are collected. The principle

is simple: an informative statistical analysis of an experiment is dependent on informative data. Informative data are ensured when statistical principles are applied during data collection. Poorly planned experiments often leave important questions unanswered due to a lack of informative data. A properly planned mixture experiment allows more than one, if not all, of the above experimental ideologies to be addressed [64]. Various experimental designs have been developed for mixture experiments. Chapter 1 addresses some of the most popular experimental designs such as the simplex lattice design and the simplex centroid design.

Statistical Analysis: Statistical analysis of experimental data serves various purposes in mixture experiments. It provides a means for model simplification by identifying the most dominant components in the mixture, but it is applied foremost to parameter estimation and model validation. Various mixture models may be appropriate for a given experiment and it is crucial to pick the best model for the system in order to gain proper understanding of the mixture experiment. It is common to use computer-based methods for model selection and the literature cites the use of stepwise regression, backward elimination regression and the "RSQUARE-procedure" from SAS and subset selection methods [3]. This study employs bootstrap for parameter estimation and model validation and this is explained in Chapter 4.

One of the greatest challenges in mixture experiments is to develop a model that can predict the octane number of a fuel blend. Octane number is one of the most important properties of fuel since it provides a measure of the anti-knock property of the fuel blend. It is measured as either the Research Octane Number (RON), which represents low-speed city driving, or as the Motor Octane Number (MON), which represents high-speed freeway driving [4]. It has always been a goal of refiners to accurately predict the octane number and the literature is rich in proposed empirical models [4, 34, 45, 67]. The standard deviation of the prediction error of these models is generally less than 0.8 RON/MON values. Most of these models were developed from a narrow range of gasoline components which hampers their extrapolation capabilities. Another setback is that for many of the proposed models the required input data are cumbersome and expensive to obtain. The models are often difficult to apply and interpret, and therefore less than ideal for practical application [4, 45]. The octane number is dependent on composition and it is well known

that the octane number does not blend linearly [4]. Gasoline consists of hundreds, if not thousands, of compounds, rendering it impossible to account for each compound's contribution to the octane number.

The holistic aim of this project is to develop parameter-parsimonious mixture models that can predict the octane number of fuel blends by incorporating the physical properties of each individual fuel stream that can be measured quickly and cost-effectively into the model. Every single component of gasoline is then indirectly taken into account.

The project is split into two phases. Phase 1 is the subject of this dissertation with the primary objective of constructing a parameter-parsimonious model that is flexible enough to model deviation from linear blending. It focuses on the theory of experiments with mixtures; introduces a new mixture model and applies the statistical method, bootstrap, to verify the model. Phase 2 is intended to include a literature review of the history of octane number as a measure of expressing the anti-knock property of fuel and a review of previously developed octane-prediction models. The purpose of phase 2 is to subject the models to further vigorous testing under more stringent conditions, with the ultimate aim of predicting the octane number of fuel blends.

Chapter 1

Experiments with Mixtures

The first mixture experiment was described by Claringbold in 1955 [11]. He introduced the simplex design as an easier experimental design for studying the joint action of related hormones. *Joint action* is a term used in toxicological studies to refer to the simultaneous action of substances that were administered separately in an organism [11]. In 1958, Henry Scheffé published what is today considered as the pioneering article in mixture experiment research [57]. He expanded and generalised Claringbold's simplex design to all mixtures where the response variable is only dependent on component proportions and not on the total amount of the mixture [57]. This definition for mixture experiments introduces the following constraints:

$$\sum_{i=1}^{q} x_i = 1 \qquad x_i \geq 0 \qquad i = 1, 2, ...q \tag{1.1}$$

where q is the number of components in the mixture.

The most important research conducted in the first 50 years of mixture experiment research has been concisely summarised by Gregory Piepel [47]. It provides a glimpse into the vast research field that mixture experiments have developed into. There are three main areas in mixture experiments that arise from the literature: the *mixture model*,

the *experimental design* and *statistical analysis*. Each of these areas will be separately addressed and explained on the basis of an example. The data in Table 1.1 were taken from a mixture experiment in the pharmaceutical sciences[1] [33].

Piepel and Cornell (1994) summarise the typical steps of a mixture experiment as follows [50]:

1. *Define the objectives of the experiment.*
2. *Select the mixture components.*
3. *Identify any constraints on the mixture components.*
4. *Specify the responses to be measured.*
5. *Propose an appropriate model from the response data as functions of the mixture components.*
6. *Select an experimental design that is sufficient for fitting the model and testing its adequacy.*

In the pharmaceutical example, the objective of the experiment is to measure the influence that non-ionic surfactants have on the size of the nanospheres in a pseudolatex. A three-component mixture consisting of poloxamer 188 NF, polyoxyethylene 40 monostearate NF and polyoxyethylene sorbitan fatty acid ester NF is selected for the study. For simplicity, the components will be referred to as A, B and C respectively in the remaining text. The components are not subjected to any constraints and can assume any proportion between 0 and 1. The size of the nanospheres of the pseudolatex is measured in nanometres (Table 1.1). With the objective of the experiment decided and the response to be measured specified, a suitable model needs to be suggested that describes the relationship between the response variable and the components [50].

1.1 The Mixture Model

Composition variables A, B and C adhere to the restrictions defined in (1.1). Therefore all the possible proportions assumed by the components can be graphically illustrated by a

[1] The data from the publication were adapted and altered for illustrative purposes.

TABLE 1.1: Observed nanosphere size in nm for 11 different blends

Blend	Composition Variables			Observed Responses		
#	A (X_1)	B (X_2)	C (X_3)	Run 1	Run 2	Run 3
1	1.000	0.000	0.000	250.1	250.4	250.2
2	0.000	1.000	0.000	274.2	274.2	274.3
3	0.000	0.000	1.000	533.5	533.2	533.3
4	0.500	0.500	0.000	255.2	255.9	253.8
5	0.500	0.000	0.500	267.3	267.5	267.4
6	0.000	0.500	0.500	294.3	294.5	294.5

triangle with pure-components on the vertices, as illustrated in Figure 1.1. The complete set of all possible proportions is referred to as the *factor space*.

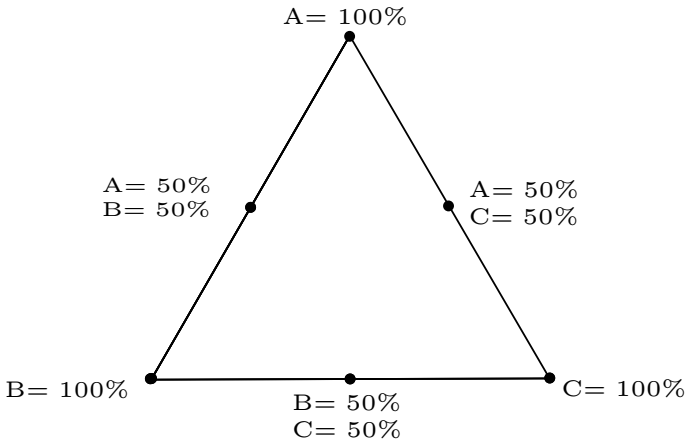


FIGURE 1.1: Triangular factor space for components A, B and C

The *response variable* (nanosphere size) measured at various mixture blends, generates a *response surface* perpendicular above the factor space (Figure 1.2). The height of the response above the factor space indicates the value of the response. The proposed mixture model should describe the behaviour of the response variable over the factor space; in other words, it should describe the shape of the response surface.

In theory, polynomials can represent continuous response surfaces with great accuracy if enough terms are included [35]. The general form of a regression polynomial of degree n with q variables is:

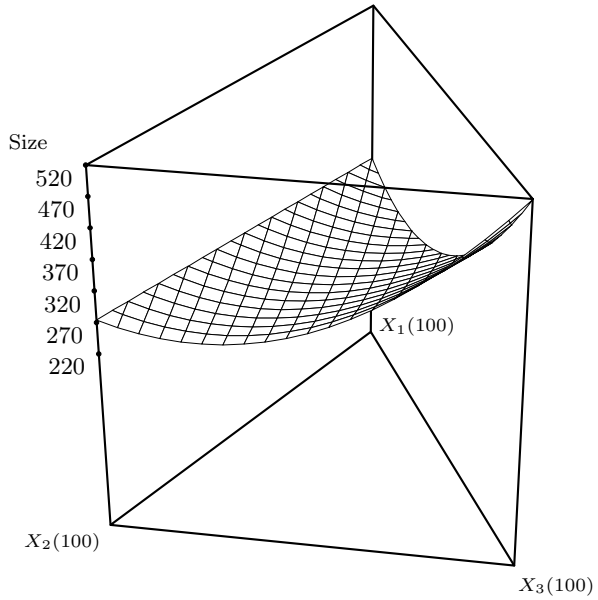


FIGURE 1.2: Response surface perpendicular above the factor space

$$y = \beta_0 + \sum_{1\leq i\leq q} \beta_i x_i + \sum\sum_{1\leq i\leq j\leq q} \beta_{ij} x_i x_j + \sum\sum\sum_{1\leq i\leq j\leq k\leq q} \beta_{ijk} x_i x_j x_k + \cdots + \sum\sum\cdots\sum_{1\leq i_1\leq i_2\leq\cdots\leq i_n\leq q} \beta_{i_1 i_2 \ldots i_n} x_{i_1} x_{i_2} \ldots x_{i_n} \tag{1.2}$$

The interpretations of the $\binom{q+n}{n}$ number of parameters in the above polynomial are subject to the restriction that the sum of the components must equal one (1.1). Under these restrictions there exists a high collinearity among the x_i's which implies that the parameters, $\beta_i, \beta_{ij}, \ldots, \beta_{i_1 i_2 \ldots i_n}$ are not unique [15, 57]. It is possible to remove the dependency of the x_i variable by substituting

$$x_q = 1 - \sum_{i=1}^{q-1} x_i \tag{1.3}$$

into (1.2). This reduces the number of parameters to $\binom{q+n-1}{n}$ but does not affect the degree of the polynomial. This is not an ideal approach in mixture experiments as the influence that component q has on the response variable is now removed from the model. Scheffé proposed that the polynomials be rewritten in what he called their *canonical form* [57]. The canonical polynomials still have only $\binom{q+n-1}{n}$ parameters with the additional benefit that the coefficients, β, have simple interpretations [57]. The deduction of only a second-

order canonical polynomial will be considered in this dissertation. For the deduction of a higher order canonical polynomial, refer to Appendix A.

The second-order canonical polynomial is obtained by multiplying the constant term in (1.2) by the identity $x_1 + x_2 + x_3 + \cdots + x_q = 1$ and by replacing all the square terms with the identity

$$x_i^2 = x_i(1 - \sum_{\substack{j=1 \\ j \neq i}}^{q} x_j)$$

The constant term, β_0, in (1.2) is now absorbed into the β_i coefficients and the quadratic parameters β_{ii} are absorbed into the β_i and β_{ij} parameters [20].

These principles can be applied to a regression polynomial of any degree to deduce the mathematically equivalent canonical form. The resulting first-, second- and third-order canonical polynomials are [57]:

Linear blending:
$$y = \sum_{i=1}^{q} \beta_i x_i \tag{1.4}$$

Quadratic model:
$$y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1<i<j<q} \beta_{ij} x_i x_j \tag{1.5}$$

Cubic model:
$$y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1<i<j<q} \beta_{ij} x_i x_j + \sum\sum_{1<i<j} \delta_{ij} x_i (x_i - x_j) + \sum\sum\sum_{1<i<j<k} \beta_{ijk} x_i x_j x_k \tag{1.6}$$

Special cubic model:
$$y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1<i<j<q} \beta_{ij} x_i x_j + \sum\sum\sum_{1<i<j<k} \beta_{ijk} x_i x_j x_k \tag{1.7}$$

The special cubic model is an extension of the quadratic model and is obtained by adding the term $\beta_{ijk} x_i x_j x_k$ [35]. The models are known as the Scheffé canonical polynomials and are widely used in mixture experiment applications due to their simplicity of use and ease of interpretation. They are typically of low degree since increasing the degree of the function increases the number of parameters to be estimated and interpreted exponentially. The accuracy of these polynomials depends on the number of terms included in the polynomial, as mentioned earlier. This shortcoming of Scheffé polynomials creates the need for models that are high in accuracy but with few estimable parameters. Various other models have been introduced since Scheffé first introduced the canonical

polynomials either as improvements or for other experimental requirements. Some of the most noteworthy models will be addressed in Chapter 2.

Quadratic polynomials are in most applications sufficient to model the response surface. They are also preferred for the small number of parameters that need to be estimated and interpreted. For these reasons they are still one of the most popular mixture models in industry. Therefore a quadratic polynomial will be used to model the pharmaceutical data. The response (dependent) variable, y, is the nanosphere size in nanometres, with the composition variables x_i being the respective proportions of components A, B and C.

Mixture experiments are employed either to predict a response variable, to screen components of a mixture to simplify the problem, to measure the effects of the components on the response variable, or to optimise the response variable over the experimental region [3, 15]. If the experiment is properly planned, more than one, if not all, of these goals can be met. The experimental design specifies the minimum number of data points needed to describe the relationship between the response variable and the components, to estimate the parameters and to assess model accuracy [50].

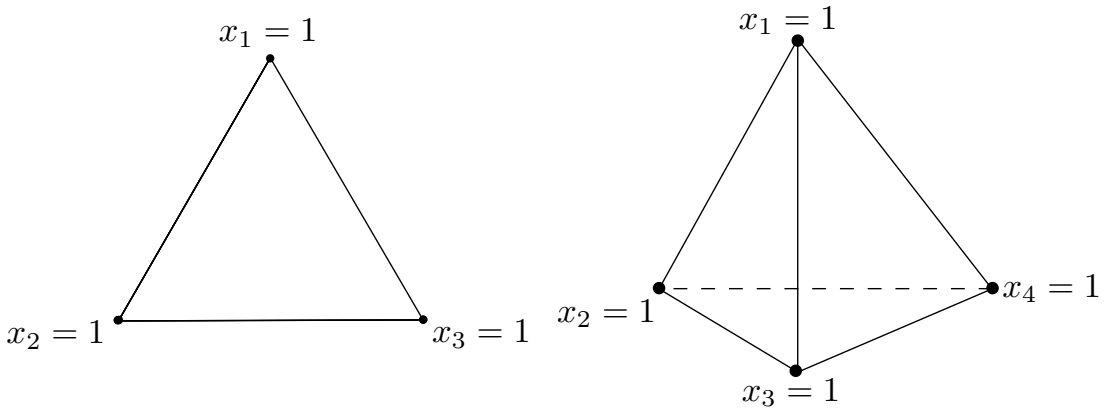
1.2 Experimental Design

The experimental region or factor space of any mixture experiment is defined by the constraints in (1.1) and can be represented by a $(q-1)$-dimensional simplex. For our three-component example, the factor space is a triangle. For a four-component mixture, it will be a tetrahedron, as illustrated in Figure 1.3. The pure components are positioned at the vertices.

In most mixture experiments the behaviour of the response variable over the whole factor space is of interest. Uniformly distributed points covering the whole factor space need to be selected. Scheffé labelled this uniform distribution of points as a $\{q, m\}$-simplex lattice, where q is the number of components and corresponds to the corners of the simplex and m is an integer that depicts the spacing of the lattice points.

The number of points, or experimental blends, required for any lattice is:

$$\text{Number of lattice points } (l) = \binom{q+m-1}{m} = \frac{(q+m-1)!}{m!(q-1)!} \tag{1.8}$$



(a) Triangular design for three components (b) Tetrahedral design for four components

FIGURE 1.3: Simplex lattice design for three and four components respectively

The lattice consists of mixture composition points and therefore m is associated with the proportions each component can assume. The $m + 1$ different proportion values for each component are $0, \frac{1}{m}, \frac{2}{m}, \ldots, 1$.

The convenience of the simplex lattice design is that if the integer m depicts the degree of the model, the number of lattice points in the design corresponds to the number of parameters that must be estimated.

Consider our example. We have a three-component mixture ($q = 3$) and we want to employ a second-order Scheffé polynomial ($m = 2$) to model the response variable. From the previous section we know that the number of parameters to be estimated is $\binom{3+2-1}{2} = 6$. Our $2D$-simplex is a triangle and since $m = 2$, the number of lattice points required is six (1.8). The proportions each of the components must assume are $0, \frac{1}{2}, 1$. The six lattice points are depicted in Figure 1.4.

The simplex lattice design is one of the most common experimental designs used in mixture experiments. It provides an equally spaced distribution of points which covers the whole factor space and has just enough data points to fit a canonical polynomial of degree m uniquely. The simplex lattice design is applied for the prediction of response variables of mixtures with $q, q-1, q-2 \ldots$ components. A drawback of this design is that the prediction is based on mixtures of, at most, m components. It is always desirable to fit a polynomial with the lowest possible degree. This implies that a second-order polynomial fitted to a $\{q, 2\}$-simplex lattice entails only binary and pure-component mixture data,

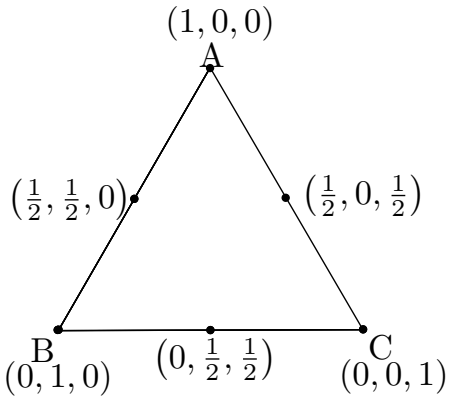


FIGURE 1.4: The six lattice points for the pharmaceutical experiment

regardless of the number of components [58]. To address this drawback of the simplex lattice design, Scheffé introduced the simplex centroid design in 1963 [58].

The simplex centroid design consists of $2^q - 1$ data points. These observations are taken at each of the following lattice points: q pure components, $\binom{q}{2}$ binary mixtures, $\binom{q}{3}$ ternary mixtures, up to the mixture $(\frac{1}{q}, \frac{1}{q}, \ldots, \frac{1}{q})$. This design contains all possible subsets of the q-components where the mixtures have equal proportions. In essence, the simplex centroid lattice consists of the centroid of the design and the centroid of every sub-dimensional simplex it contains [58]. Consider a $\{3, 2\}$-simplex centroid design. The lattice will consist of seven observations taken at the following blends (Figure 1.5):

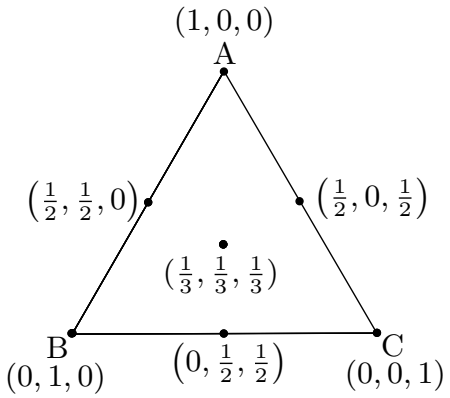


FIGURE 1.5: The simplex centroid design

A polynomial that has the same number of estimable parameters as lattice points in the simplex centroid design is [15, 58]:

$$y = \sum_{i=1}^{q} \beta_i x_i + \sum \sum_{i<j}^{q} \beta_{ij} x_i x_j + \sum \sum \sum_{i<j<k}^{q} \beta_{ijk} x_i x_j x_k + \cdots + \beta_{12\ldots q} x_1 x_2 \ldots x_q \qquad (1.9)$$

Both the simplex lattice design and the simplex centroid design are boundary designs. This means that, with the exception of the centroid, all the design points lie on the edges, faces and vertices of the simplex [15]. These designs are not ideal for mixture experiments where it is known beforehand that the blends must contain all the components present, for example, when the aim is to measure the effects that the components have on the response variable or when the behaviour of components relative to other components are investigated. Cornell introduced the axial design in 1975 to address these types of experiments [13]. In the axial design, the experimental points lie on the component axes. Component axes are imaginary lines that extend from the vertex, $x_i = 1$ to the base point $x_i = 0, x_j = \frac{1}{(q-1)}$ for all $j \neq i$ as shown in Figure 1.6.

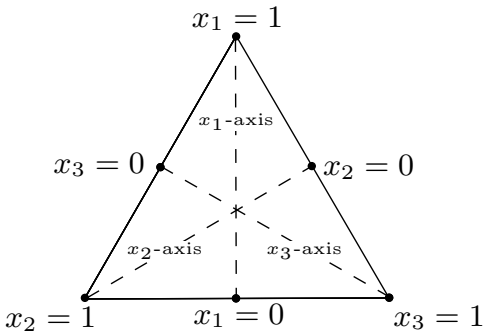


FIGURE 1.6: The axial design

The length of an axis is considered to be one unit in the simplex coordinate system. The simplest example of an axial design is one where the points are equidistant from the centroid of the simplex. The distance from the centroid is denoted Δ and has a maximum value of $\frac{(q-1)}{q}$ (Figure 1.7).

All the above-mentioned designs are applicable to experiments where the whole factor space is under consideration. Conversely, there are experiments where the component proportions are subject to constraints, which results in a smaller region within the factor space being applicable. Mclean and Anderson [41] proposed the extreme vertices design for constrained components:

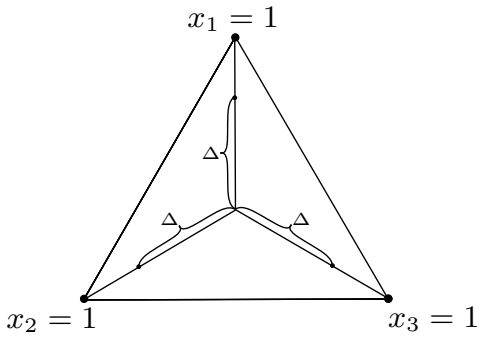


FIGURE 1.7: Design points equidistant from the axial design centroid

$$\sum_{i=1}^{q} x_i = 1 \qquad 0 \leq lb_i \leq x_i \leq ub_i \leq 1 \tag{1.10}$$

where lb_i is the lower bound and ub_i is the upper bound of component x_i. Constraints on the components imply that a smaller region of the original factor space defined by (1.1) is considered. The chosen experimental design should only allow for data points from the appropriate region. This smaller region contained within the factor space is referred to as an *irregular hyper-polyhedron* [41]. This approach requires that all of the vertices of the hyper-polyhedron needs to be calculated. Once the vertices are known, the centroids of each face, as well as the centroid of the hyper-polyhedron, are determined. Experimental mixtures are then blended at these composition points and an appropriate model is fitted. This design is discussed in more detail by way of an example in Mclean and Anderson [41].

1.3 Statistical Analysis

In this example it was decided to consider the Scheffé quadratic polynomial to describe the relationship between the response variable and the components. This choice was based on the simplicity of the model. The model is of low degree, which implies that there are few parameters to be estimated. The parameters are simple to interpret and it is one of the most common models applied in the literature. Although we can justify the decision, it is not guaranteed that the model will accurately describe the responses.

Even though parameters can be estimated by the click of a button today, there are a few statistical principles and assumptions that one needs to be aware of.

Consider a mixture experiment that was repeated N times. The observed response, y_u, of the $u'th$ $(u = 1, 2, \ldots N)$ experimental run varies around a mean, η_u, with a constant variance of σ^2 for all $u = 1, 2, \ldots N$. The deviation from the mean, η_u, can be expressed as the experimental error, ε_u. Therefore

$$y_u = \eta_u + \varepsilon_u \qquad 1 \leq u \leq N \tag{1.11}$$

It is assumed that the experimental errors are uncorrelated and identically distributed around zero with a variance of σ^2. This assumption defines the following properties of the experimental error:

$$E(\varepsilon_u) = 0 \tag{1.12}$$

$$E(\varepsilon_u^2) = \sigma^2 \tag{1.13}$$

$$E(\varepsilon_u \varepsilon_{u'}) = 0 \tag{1.14}$$

$$u \neq u', \qquad u, u' = 1, 2, \ldots, N$$

Based on these properties, it can be shown that the expected value of the observed response, y_u, is equal to the mean around which it varies.

$$E(y_u) = \eta_u \qquad u = 1, 2, \ldots, N \tag{1.15}$$

This can be graphically represented as Figure 1.8:

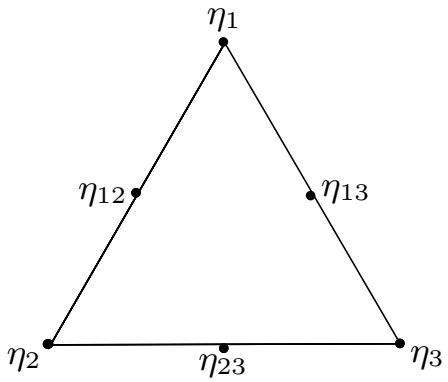


FIGURE 1.8: The response variables at the corresponding lattice points

The one-to-one relationship between the number of parameters and the number of

design points, allows the parameters to be expressed as functions of the expected response variables. To be consistent with our example, consider a $\{3, 2\}$-canonical polynomial:

$$\eta = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 \tag{1.16}$$

The experimental design consists of six lattice points ($l = 6$), (1.8), as shown in Figure 1.8. Substituting the design points into the polynomial yields the desired expressions. For the pure-component blends $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$, the β_i coefficient simply takes the value of the corresponding pure-component expected response:

$$\beta_1 = \eta_1 \qquad \beta_2 = \eta_2 \qquad \beta_3 = \eta_3 \tag{1.17}$$

The coefficients for the binary interactions at lattice points $\left(\frac{1}{2}, \frac{1}{2}, 0\right)$; $\left(\frac{1}{2}, 0, \frac{1}{2}\right)$ and $\left(0, \frac{1}{2}, \frac{1}{2}\right)$ are determined as follows:

$$\eta_{ij} = \beta_i x_i + \beta_j x_j + \beta_{ij} x_i x_j \tag{1.18}$$

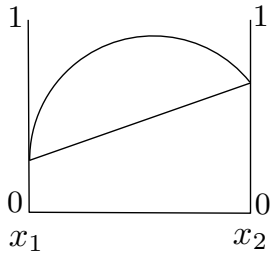
$$= \frac{1}{2}(\beta_i + \beta_j) + \frac{1}{4}\beta_{ij} \tag{1.19}$$

$$\beta_{ij} = 4\eta_{ij} - 2\eta_i - 2\eta_j \tag{1.20}$$

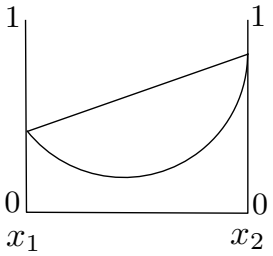
Note the changes in notation in the above equations. The expected response variables measured for pure components are denoted η_i, where $i = 1, 2, 3$, and the expected responses for binary blends are denoted η_{ij} where $i, j = 1, 2, 3$ for $i \neq j$.

The interpretation of the coefficients is straightforward. Mixtures exhibit three different types of behaviour in the response. Components either blend linearly or they exhibit antagonism or synergism when blending.

Consider a two-component mixture. Linear blending occurs when the response variable of all possible blends of the two components can be plotted on a straight line between pure-component A and pure-component B. Synergism occurs when blending the two components has an additive effect on the response variable and a positive deviation from linear blending is observed. Antagonism occurs when mixing the two components has a subtractive effect on the response variable and a negative deviation from linear blending is observed. Synergism and antagonism are illustrated in Figure 1.9.



(a) Positive deviation from the linear blending rule



(b) Negative deviation from the linear blending rule

FIGURE 1.9: Synergism and antagonism

β_i is the contribution made by pure-component i to the response surface. β_i is the height of the response surface above the vertex $x_i = 1$ and has the largest contribution at this lattice point. The term $\beta_{ij}x_ix_j$ contributes to the response surface everywhere where x_i and x_j are not zero, but it makes its maximum contribution with the binary blend when $x_i = \frac{1}{2}$ and $x_j = \frac{1}{2}$. β_{ij} accounts for the synergism or antagonism that is observed in the response surface. If β_{ij} is positive, it depicts synergistic behaviour and if β_{ij} is negative, it accounts for antagonistic behaviour in the response surface. If the response surface is most accurately represented by the cubic model, the binary mixture has an additional term to describe the response behaviour $\delta_{ij}x_ix_j(x_i - x_j)$. This term allows synergism and antagonism to be observed along the $i - j$ edge [15, 57].

The parameters of the Scheffé polynomial have now been expressed as simple functions of the expected response variables. The parameter estimates, $\hat{\beta}_i$ and $\hat{\beta}_{ij}$, can be written in terms of the observed response variables, y_i and y_{ij}.

$$\hat{\beta}_1 = y_1 \qquad \hat{\beta}_2 = y_2 \qquad \hat{\beta}_3 = y_3 \tag{1.21}$$

$$\hat{\beta}_{ij} = 4y_{ij} - 2y_i - 2y_j \qquad i, j = 1, 2, \ldots q, \qquad i \leq j \tag{1.22}$$

The number of responses measured at each pure-component is denoted $r_i = 1, 2, \ldots n$, $i = 1, 2, \ldots q$. The number of observed responses at the binary blends is $r_{ij} = 1, 2, \ldots n$, with $i, j = 1, 2, \ldots q$ for $i \neq j$. It is advisable to take more than one observation at every lattice point ($r_l > 1$). This improves the accuracy of the parameter estimates, as shown later.

$$\hat{\beta}_1 = \bar{y}_1 \qquad \hat{\beta}_2 = \bar{y}_2 \qquad \hat{\beta}_3 = \bar{y}_3 \tag{1.23}$$

$$\hat{\beta}_{ij} = 4\bar{y}_{ij} - 2\bar{y}_i - 2\bar{y}_j \text{ where} \tag{1.24}$$

$$\bar{y}_i = \frac{\sum_{l=1}^{r_i} y_{i_l}}{r_i} \tag{1.25}$$

$$\bar{y}_{ij} = \frac{\sum_{l=1}^{r_{ij}} y_{ij_l}}{r_{ij}} \tag{1.26}$$

These equations, (1.23) to (1.26), are the least-squares calculations for the parameter estimates. Note that the scalar quantities, 4, 2 and 2, in the formula for $\hat{\beta}_{ij}$ come from the proportions $x_i = \frac{1}{2}$ and $x_j = \frac{1}{2}$ that were used in the design and are not dependent on the r_i,r_j and r_{ij} [15].

Revisit equation (1.11). In the relationship $y_u = \eta_u + \varepsilon_u$, η_u is replaced by the proposed model. In this case, the Scheffé quadratic model becomes

$$y_u = (\sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1<i<j<q} \beta_{ij} x_i x_j)_u + \varepsilon_u \qquad u = 1, 2 \ldots N \tag{1.27}$$

The estimated response model is:

$$\hat{y}_u = (\sum_{i=1}^{q} \hat{\beta}_i x_i + \sum\sum_{1<i<j<q} \hat{\beta}_{ij} x_i x_j)_u + \varepsilon_u \qquad u = 1, 2 \ldots N \tag{1.28}$$

where q indicates the number of components in the mixture and $\hat{y}_u$ is the estimated response variable at the $u'th$ experimental trial. The parameters of the model, β_i and β_{ij}, are estimated by the method of least-squaress.

The least-squares estimators of the parameters are the estimated values that minimise the sum of the squared of errors [42]:

$$L = \sum_{u=1}^{N} \varepsilon_u^2 \tag{1.29}$$

$$= \sum_{u=1}^{N} (y_u - \hat{y}_u)^2 \tag{1.30}$$

Differentiating L with respect to the parameters in question, equating it to zero and simplifying the equations yields the least-squares normal equations as expressed in (1.23)

[15, 42]. Refer to Cornell (2002) for an explanation of normal equations in matrix notation [15].

The properties of the parameter estimates are determined by the properties of the random errors, ε_u. If the errors are assumed to be normally distributed with a mean of zero and a variance of σ^2, then from (1.11) it is clear that y_u also has a normal distribution with a variance of σ^2 but with a mean of η_u. The parameters are linear functions of the observations and are therefore also normally distributed [42]. If the observations used to estimate the parameters were collected only at the lattice points, the means and variances of the parameter distributions are [15]:

$$E(\hat{\beta}_i) = E(\bar{y}_i) = \beta_i \tag{1.31}$$

$$var(\hat{\beta}_i) = var(\bar{y}_i) = \frac{\sigma^2}{r_i} \tag{1.32}$$

$$E(\hat{\beta}_{ij}) = E[4\bar{y}_{ij} - 2(\bar{y}_i + \bar{y}_j)] = \beta_{ij} \tag{1.33}$$

$$var(\hat{\beta}_{ij}) = var[4\bar{y}_{ij} - 2(\bar{y}_i + \bar{y}_j)] = \frac{16\sigma^2}{r_{ij}} + \frac{4\sigma^2}{r_i} + \frac{4\sigma^2}{r_j} \tag{1.34}$$

$$cov(\hat{\beta}_i, \hat{\beta}_j) = 0 \qquad i \neq j \tag{1.35}$$

$$cov(\hat{\beta}_i, \hat{\beta}_{ij}) = -\frac{2\sigma^2}{r_i} \tag{1.36}$$

$$cov(\hat{\beta}_{ij}, \hat{\beta}_{ik}) = \frac{4\sigma^2}{r_i}, j \neq k \tag{1.37}$$

Parameters $\hat{\beta}_i$ are therefore distributed as $N \sim (\beta_i, \frac{\sigma^2}{r_i})$ and parameters $\hat{\beta}_{ij}$ have $N \sim (\beta_{ij}, \frac{16\sigma^2}{r_{ij}} + \frac{4\sigma^2}{r_i} + \frac{4\sigma^2}{r_j})$ distributions. Here N indicates the normal distribution [15]. These properties provide insight in determining whether the parameter values are significantly different from zero during hypothesis testing. The positive square root of the estimated variances of the parameters is known as the standard error $se(\hat{\beta}_i) = \sqrt{var(\hat{\beta}_i)}$ and $se(\hat{\beta}_{ij}) = \sqrt{var(\hat{\beta}_{ij})}$. The standard error is used to calculate $100(1-\alpha)\%$ *confidence intervals* for the estimated parameter, where typically $\alpha = 5\%$. If the estimated value of the parameter falls withing the boundaries of the interval, then the estimated parameter is acceptable. The narrower the confidence intervals, the more accurate the estimation:

$$\hat{\beta}_i - t_{n-p,\frac{\alpha}{2}} se(\hat{\beta}_i) \leq \beta_i \leq \hat{\beta}_i + t_{n-p,\frac{\alpha}{2}} se(\hat{\beta}_i) \tag{1.38}$$

$$\hat{\beta}_{ij} - t_{n-p,\frac{\alpha}{2}} se(\hat{\beta}_{ij}) \leq \beta_{ij} \leq \hat{\beta}_{ij} + t_{n-p,\frac{\alpha}{2}} se(\hat{\beta}_{ij}) \tag{1.39}$$

In these equations n denotes the total amount of observations in the experiment and p is the number of estimable parameters. The quantity $n-p$ is the degrees of freedom. The value $t_{\frac{\alpha}{2},n-p}$ is the tabulated value for the t-distribution, with $n-p$ degrees of freedom at a level of significance of $\frac{\alpha}{2}$.

In typical regression it might be possible to reduce the models by omitting the non-significant terms, but it is seldom justifiable to remove parameters from a mixture model [19]. Snee and Marquardt (1976) introduced screening techniques for mixture experiments to identify the most important components [62]. This is mostly applicable to mixtures with more than six components where the aim is not to model a response surface for predictive purposes, but rather to identify compounds with the greatest effects in order to reduce the overall experiment [14, 62]. The variances and covariances of the parameters are all dependent on the precision of the experimental observations via σ^2. The value for σ^2 is often unknown but it can be replaced by its estimate $\hat{\sigma}^2$, which is calculated as:

$$\hat{\sigma}^2 = \sum_{k=1}^{l} \sum_{u=1}^{r_l} \frac{(y_{ku} - \bar{y}_k)^2}{\sum_{k=1}^{l}(r_k - 1)} \tag{1.40}$$

From (1.40) it is clear that the accuracy of the parameter estimates can be improved by improving the experimental precision and/or increasing the number of observations at every lattice point.

It is desirable to estimate the responses within a certain level of accuracy. This can be done by calculating the $(1-\alpha)100\%$ confidence interval for η , the expected value of the response (1.15).

$$\hat{y}(x) - \Delta < \eta < \hat{y}(x) + \Delta \tag{1.41}$$

$$\Delta = [t_{f,\frac{\alpha}{2}}] \sqrt{\widehat{var}[\hat{y}(x)]} \tag{1.42}$$

where f is the degree of freedom associated with $\hat{\sigma}^2$ and $t_{f,\frac{\alpha}{2}}$ is the tabled t-value with f degrees of freedom at the $\frac{\alpha}{2}$ level of significance. A typical value for α is 5%. As long as $\hat{y}(x)$ falls within the limits of the confidence interval, the required level of accuracy is met.

To calculate the confidence intervals, the estimated variance of the estimated response, $\widehat{var}[\hat{y}(x)]$, needs to be determined. The variance of the estimates response, $var[\hat{y}(x)]$, is calculated by substituting the parameter estimates into the model:

$$\hat{y}(x) = \sum_{i=1}^{q} \hat{\beta}_i x_i + \sum\sum_{i<j}^{q} \hat{\beta}_{ij} x_i x_j \tag{1.43}$$

$$= \sum_{i=1}^{q} \bar{y}_i x_i + \sum\sum_{i<j}^{q} (4\bar{y}_{ij} - 2\bar{y}_i - 2\bar{y}_j) x_i x_j \tag{1.44}$$

This can be simplified to:

$$\hat{y}(x) = \sum_{i=1}^{q} a_i \bar{y}_i + \sum\sum_{i<j}^{q} a_{ij} \bar{y}_{ij} \tag{1.45}$$

where $a_i = x_i(2x_i - 1)$ and $a_{ij} = 4x_i x_j \; i, j = 1, 2, \ldots, q, i < j$. The values of a_i and a_{ij} are specified by the composition point x and are therefore fixed without error [15]. The variance of the estimated response can be written as:

$$var[\hat{y}(x)] = \sigma^2 \left\{ \sum_{i=1}^{q} \frac{a_i^2}{r_i} + \sum\sum_{i<j}^{q} \frac{a_{ij}^2}{r_{ij}} \right\} \tag{1.46}$$

An estimate for $var[\hat{y}(x)]$ is calculated when σ^2 is replaced by its estimate $\hat{\sigma}^2$. The variance of the predicted response $\hat{y}(x)$ can be written in terms of the variances and covariances of $\hat{\beta}_i$ and $\hat{\beta}_{ij}$, implying that the variance of the predicted response at a given composition is directly influenced by the properties of the parameter estimates. A smaller $var[\hat{y}(x)]$ results in a narrower confidence interval and a more accurate predicted response.

It is necessary to understand the variation between the measured response variables in mixture experiments. The overall variation among measured responses can be divided into two sources: variation among the average responses between blends and variation between measured responses at any given blend.

Scheffé {q,m}-canonical polynomials have the same number of terms as experimental design points. The variation explained by the fitted model is the variation between the observed responses at the various design points. This variation is referred to as the sum of squares of regression (SSR) and is calculated as:

$$SSR = \sum_{u=1}^{N} (\hat{y}_u - \bar{y})^2 \tag{1.47}$$

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_N}{N} \tag{1.48}$$

where $\hat{y}_u$ is the estimated value of the $u'th$ response and $\bar{y}$ is the overall average of the observations. The SSR has $p-1$ degrees of freedom, where p is the total number of different blends or the number of terms in the model.

The variation among the replicate observations at the respective design points is accounted for separately from the variation among the blends and is known as the sum of squared errors (SSE):

$$SSE = \sum_{u=1}^{N} (y_u - \hat{y}_u)^2 \tag{1.49}$$

The SSE has $N-p$ degrees of freedom.

The total sum of squares (SST) is the total amount of variation for the complete data set of N blends and is the sum of the SSE and SSR, which implies $N-1$ degrees of freedom.

$$SST = \sum_{u=1}^{N} (y_u - \bar{y})^2 \tag{1.50}$$

$$SST = SSE + SSR \tag{1.51}$$

This breakdown of the variance is known as the analysis of variance (ANOVA) and can be summarised as in table Table 1.2:

TABLE 1.2: Analysis of variance table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Regression (fitted model)	$p-1$	$SSR = \sum_{u=1}^{N} (\hat{y}_u - \bar{y})^2$	$SSR/(p-1)$
Residual	$N-p$	$SSE = \sum_{u=1}^{N} (y_u - \hat{y}_u)^2$	$SSE/(N-p)$
Total	$N-1$	$SST = \sum_{u=1}^{N} (y_u - \bar{y})^2$	$SST/(N-1)$

The model fit needs to be assessed to determine whether it appropriately describes the data. Typical model fit criteria are the F-test, the correlation coefficient, R^2, and

the adjusted correlation coefficient, R_A^2. These criteria are calculated from the above-mentioned ANOVA statistics.

The F-ratio is defined as:

$$F = \frac{\text{Mean Square of Regression}}{\text{Mean Square of Errors}} = \frac{SSR/(p-1)}{SSE/(N-p)} \tag{1.52}$$

The correlation coefficient, R^2, is defined as

$$R^2 = \frac{SSR}{SST} \tag{1.53}$$

and can be interpreted as the proportion of the variability in the data that is accounted for by the ANOVA model [42]. The adjusted correlation coefficient, R_A^2, measures the reduction in the estimate of the error variance due to fitting the model relative to the estimate of the error variance when the simple model $y = \beta_0 + \varepsilon$ is fitted [15]:

$$R_A^2 = 1 - \frac{SSE/(N-p)}{SST/(N-1)} \tag{1.54}$$

These test statistics can be calculated by appropriate statistical software. It is important to ensure that the software used to assess model fit supports mixture experiments. The constraints (1.1) that define a mixture experiment result in high colinearity among the parameters, which renders t-tests unstable. The lack of an intercept variable in mixture models also tends to inflated values for R^2 and R_A^2, which does not reflect true model fit [15, 19, 40]. If a normal regression through an intercept is applied to fit a mixture model, there are two possible ways of addressing the statistics. The test statistics can be calculated separately using the correct formula. Alternatively, the mixture model, as well as its mathematical equivalent intercept model, can be fitted. The mixture model provides the correct parameter estimates and the intercept equivalent provides the correct statistics [19, 40].

1.4 An Example

Now that the statistical groundwork has been laid, let's return to the pharmaceutical example. The effects that various blends of a three-component mixture have on the size

of pseudolatex nanospheres are measured. The Scheffé quadratic polynomial was chosen to model the data and the simplex lattice design was used to collect the data. The data are summarised in Table 1.1.

The model to fit is:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 \tag{1.55}$$

From (1.23) the estimated parameters are:

$$\beta_1 = 250.23 \qquad \beta_2 = 274.23 \qquad \beta_3 = 533.33$$
$$\beta_{12} = 4(254.90) - 2(250.23) - 2(274.23) = -29.07$$
$$\beta_{13} = 4(267.40) - 2(250.23) - 2(533.33) = -497.53$$
$$\beta_{23} = 4(294.43) - 2(274.23) - 2(533.32) = -437.40$$

From the parameter values we can draw the following conclusions:

$$\hat{\beta}_1 < \hat{\beta}_2 < \hat{\beta}_3 \tag{1.56}$$
$$|\hat{\beta_{12}}| < |\hat{\beta_{23}}| < |\hat{\beta_{13}}| \tag{1.57}$$

implying that pure-component C has the greatest influence on the nanosphere size and component A the smallest influence. The binary coefficients indicate overall antagonistic deviation from linear blending, as depicted in Figure 1.2. Binary blends that include component C have the greatest influence on the response. Therefore, if larger nanosphere sizes are required, the proportion of component C should be increased. If small nanospheres are desired, component C should be as small as possible.

Substitute the parameter values into (1.55) to obtain the model that will to be used to estimate the response variable:

$$\hat{y} = 250.23x_1 + 274.23x_2 + 533.33x_3 - 29.07x_1x_2 - 497.53x_1x_3 - 437.40x_2x_3 \tag{1.58}$$

The error variance σ^2 was estimated from equation (1.40) to be $\hat{\sigma}^2 = 0.2028$. Estimates for properties of the parameters (1.31) are calculated by simply replacing σ^2 with its estimate $\hat{\sigma}^2$:

$$\widehat{var}(\hat{\beta}_i) = \frac{\hat{\sigma}^2}{r_i} = \frac{0.2028}{3} = 0.06759 \qquad i = 1,2,3 \text{ since } r_i = 3 \text{ for all i} \tag{1.59}$$

$$\widehat{var}(\hat{\beta}_{ij}) = \frac{24\hat{\sigma}^2}{r} = 1.622 \text{ since } r_{ij} = r_i = r_j = 3 \tag{1.60}$$

$$\widehat{cov}(\hat{\beta}_i, \hat{\beta}_{ij}) = -\frac{2\hat{\sigma}^2}{r_i} = \frac{2(0.2028)}{3} = -0.1352 \text{ for all } i = 1,2,3 \text{ since } r_i = r_j = 3 \tag{1.61}$$

$$\widehat{cov}(\hat{\beta}_{ij}, \hat{\beta}_{ik}) = \frac{4\hat{\sigma}^2}{r_i} = -\frac{4(0.20278)}{3} = -0.2704 \qquad j \neq k \tag{1.62}$$

TABLE 1.3: The 95% confidence intervals for the estimated parameters

Parameter	df	Lower Limit	$\hat{\beta}$	Upper Limit	$se(\hat{\beta})$
β_1	1	249.67	250.23	250.80	0.26
β_2	1	273.67	274.23	274.80	0.26
β_3	1	532.77	533.33	533.90	0.26
β_{12}	1	-31.84	-29.07	-26.29	1.27
β_{13}	1	-500.31	-497.53	494.76	1.27
β_{23}	1	-440.18	-437.40	-434.62	1.27

All the parameters fall within the limits of their calculated confidence intervals as calculated from (1.38). We can infer that all parameters are significant. As expected, the covariance between parameters $\hat{\beta}_i$ and $\hat{\beta}_{ij}$ is negative. This will always be the case for Scheffé quadratic polynomials and it follows from the expression of β_i and β_{ij} in (1.21):

$$\widehat{cov}(\hat{\beta}_i, \hat{\beta}_{ij}) = E[(\hat{\beta}_i - \beta_i)(\hat{\beta}_{ij} - \beta_{ij})] \tag{1.63}$$

$$= E[(\bar{y}_i - \beta_i)(4\bar{y}_{ij} - 2\bar{y}_i - 2\bar{y}_j - \beta_{ij})] \tag{1.64}$$

$$= 4E(\bar{y}_i\bar{y}_{ij}) - 2E(\bar{y}_i^2) - 2E(\bar{y}_i\bar{y}_j) \tag{1.65}$$

$$= \frac{-2\hat{\sigma}^2}{r_i} \tag{1.66}$$

The model's adequacy can be tested at various composition points within the design space. For six different composition points, an estimated response is calculated and the confidence interval determined (1.41):

TABLE 1.4: Calculated values for $\hat{y}(x)$ and 95% confidence limits for the true nanosphere size for six arbitrary blends

	x			Lower		Upper	
Blend	x_1	x_2	x_3	Limit	$\hat{y}(x)$	Limit	$(\widehat{var}[\hat{y}(x)])^{\frac{1}{2}}$
1	0.31	0.67	0.012	257.66	258.17	258.67	0.5058
2	0.25	0.24	0.51	282.48	282.92	283.35	0.4924
3	0.18	0.30	0.53	289.39	289.83	290.27	0.4934
4	0.30	0.51	0.19	240.56	241.01	241.45	0.4938
5	0.36	0.012	0.63	312.05	312.57	313.09	0.5104
6	0.13	0.061	0.81	406.21	406.60	407.00	0.4849

The model seems adequate since all the estimated responses fall within the limits of their respective confidence intervals at a 95% level of certainty (Table 1.4).

The next step is an ANOVA analysis to assess the model fit. The composition vectors of the six design points are substituted into (1.58) to calculate the estimated responses for the measured responses on which the model was developed. The SSE, SSR and SST can now be calculated with equations (1.47), (1.49),(1.50). Refer to Table 1.5 for a summary of the calculations.

The model was analysed with Design Expert Version 8.0.4 and Table 1.6 provides a summary of the ANOVA statistics. The *p-value* < 0.0001 indicates that the model is significant at a 5% level of significance. The R^2 and R^2_A are both calculated as 1.000 and indicate that the model is a good fit to the data. To illustrate the fit graphically, the predicted response variable is plotted against the measured response variable (Figure 1.10). The highly linear relationship confirms the good fit.

Another effective method for gaining insight into the fitted model is contour plots. The response surface generated by the model can be projected onto a contour plot. Contour plots are a common and effective method for visually studying the behaviour of the estimated response surface. They are particularly useful for systems with two or three components [21, 39, 61]. For mixtures where $q > 3$, dissections of the response surfaces can be studied [15]. Examining contour plots can indicate equal coefficients to help with model reduction [62]. This indicates the region of the response surface where the optimal

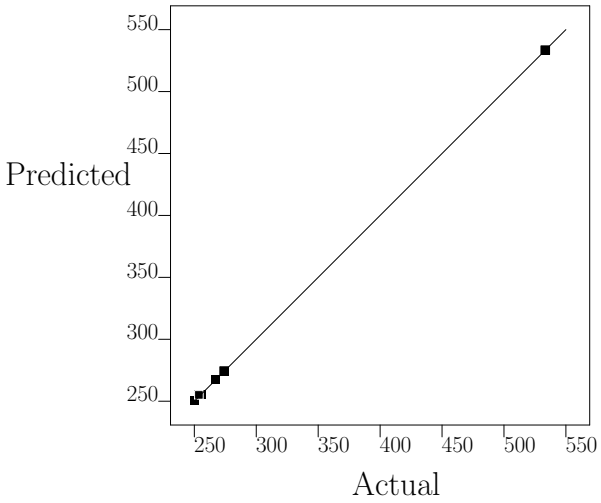


FIGURE 1.10: Predicted vs. actual response variables

values for response can be located. It can be used to compare different mixture models where more than one model was fitted to the data [19, 35]

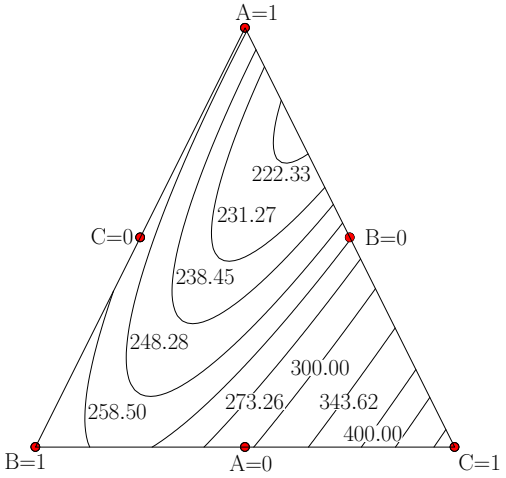


FIGURE 1.11: Contour plot of the estimated responses from our fitted model

Figure 1.11 depicts the contour plot for our example. It confirms the conclusions drawn from the parameters: the greater the proportion of component C, the greater the nanosphere size. The smaller nanosphere sizes are obtained in the vicinity of pure-component A, where C and B have small proportions.

This chapter has focused on the Scheffé quadratic polynomial and the simplex lattice design as an introduction to mixture experiments. Various other mixture models have been developed over the years to address other experimental conditions which the Scheffé polynomials cannot model adequately. Some of these models are addressed in Chapter 2.

TABLE 1.5: Sum of squared deviations for nanosphere data

Observed Values (y_u)	Predicted Values ($\hat{y}_u$)	Residuals ($y_u - \hat{y}_u$)	Deviations ($y_u - \bar{y}$)	Regression Deviations ($\hat{y}_u - \bar{y}$)
250.1	250.23	-0.13	-62.33	-62.20
274.2	274.23	-0.03	-38.23	-38.20
533.5	533.33	0.17	221.07	220.90
255.2	254.97	0.23	-57.23	-57.47
267.3	267.40	-0.10	-45.13	-45.03
294.3	294.43	-0.13	-18.13	-18.00
250.4	250.23	0.17	-62.03	-62.20
274.2	274.23	-0.03	-38.23	-38.20
533.2	533.33	-0.13	220.77	220.90
255.9	254.97	0.93	-56.53	-57.47
267.5	267.40	0.10	-44.93	-45.03
294.5	294.43	0.07	-17.93	-18.00
250.2	250.23	-0.03	-62.23	-62.20
274.3	274.23	0.07	-38.13	-38.20
533.3	533.33	-0.03	220.87	220.90
253.8	254.97	-1.17	-58.63	-57.47
267.4	267.40	0.00	-45.03	-45.03
294.5	294.43	0.07	-17.93	-18.00
$\bar{y}$=312.43		SSE=2.43	SST=179340.36	SSR=179337.93

TABLE 1.6: Analysis of variance table for the nanosphere data

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F-value $\lvert Prop\rvert > F$	p-value
Regression (fitted model)	5	179337.93	35867.59	1.769E+05	< 0.0001
Residual	12	2.43	0.2		
Total	17	179340.36	10549.43		

Chapter 2

Mixture Models

Different experimental conditions call for different types of mixture models. Scheffé's polynomials are only applicable to experimental situations where the whole factor space is under consideration and where the boundaries of the design are included. These experimental conditions are less than ideal for experiments that, for example, are designed to screen components or measure the interaction effects of the components [15, 12].

Regardless of why the mixture model was developed, there are mathematical axioms that these models must obey to be consistent in their application. Some of these consistency rules follow naturally from the definition of a mixture experiment [36]. Some of the axioms are preferred properties of the models. The consistency rules that we deemed important for the model we developed are [28]:

- The mixing property, y, reduces to the pure component value when any fraction approaches unity.
- The relation for a q-component mixture reduces to the corresponding $(q-1)$-component form in the limit of infinite dilution of one of the component proportions.
- Symmetry: The predicted property values are independent of the way in which

component indices are assigned. All of the models considered in this dissertation comply with this requirement.

- REFLEXIVITY: If all the parameter values are equal, the function reduces to a pure component mixture: $f(\beta, \beta, \ldots, \beta; x_1, x_2, \ldots, x_q) = \beta$

- DECOMPOSABILITY: The model should be invariant if one component is divided into two or more identical subcomponents. Or, in other terms, if two components are identical, the q-component model should reduce to a $(q-1)$- component model.

- HOMOGENEITY: A mixture model is homogenous of degree one if
  $f(\lambda\beta_1, \lambda\beta_2, \ldots, \lambda\beta_n; x_1, x_2, \ldots, x_q) = \lambda f(\beta_1, \beta_2, \ldots, \beta_q; x_1, x_2, \ldots, x_q)$
  This ensures dimensional homogeneity.

In addition to the consistency rules, another property our mixture model should adhere to is that parameter values remain constant when new components are introduced to the mixture [28]. In reality, adding a component to a mixture introduces new interactions and affects the current interactions among components. The aim is to incorporate these effects into the new parameters introduced to the model and keep the previous parameter estimates constant. This simplifies model application in industry where it is often expensive and/or time-consuming to re-estimate model parameters every time a new component is introduced to a mixture.

Many of the models introduced as improvements or expansions of the Scheffé polynomials do not adhere to these consistency rules. It is our belief that a model that does adhere to these consistency rules is more flexible in application and can be applied to a broader range of problems. The models considered in this text are Scheffé (1958) [57], Becker (1968) [6], Cox (1971) [17], Piepel (2007) [49], Draper and John (1977) [22], Aitchison and Bacon-Shone (1984) [1], Darroch and Waller (1984) [18] and Draper and Pukelsheim (1998) [23].

Scheffé polynomials adhere to all the above-mentioned consistency rules. To prove this, consider the $\{3, 2\}$-canonical polynomial that was fitted to the pharmaceutical example

in Chapter 1:

$$y = \sum_{i=1}^{3} \beta_i x_i + \sum \sum_{i<j}^{3} \beta_{ij} x_i x_j \tag{2.1}$$

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 \tag{2.2}$$

It is easier to prove the consistency if the polynomial is written in its quadratic form. Using the relation $\sum_{i=1}^{q} x_i = 1$ we get:

$$y = (\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)(x_1 + x_2 + x_3) + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3$$

$$y = \beta_1 x_1^2 + \beta_2 x_2^2 + \beta_3 x_3^2 + \beta_{12}^* x_1 x_2 + \beta_{13}^* x_1 x_3 + \beta_{23}^* x_2 x_3$$

$$\text{where} \qquad \beta_{ij}^* = \beta_i + \beta_{ij} + \beta_j \tag{2.3}$$

To prove that Scheffé quadratic polynomials reduce to pure component properties if any component approaches unity, let $x_1 = 1, x_2 = x_3 = 0$

$$y = \beta_1 x_1^2 + \beta_2 x_2^2 + \beta_3 x_3^2 + \beta_{12}^* x_1 x_2 + \beta_{13}^* x_1 x_3 + \beta_{23}^* x_2 x_3 \tag{2.4}$$

$$y = \beta_1 \tag{2.5}$$

These polynomials also reduce to $(q-1)$-component form if one of the components tends to zero. Let $x_3 = 0$:

$$y = \beta_1 x_1^2 + \beta_2 x_2^2 + \beta_3 x_3^2 + \beta_{12}^* x_1 x_2 + \beta_{13}^* x_1 x_3 + \beta_{23}^* x_2 x_3$$

$$y = \beta_1 x_1^2 + \beta_2 x_2^2 + \beta_{12}^* x_1 x_2 \tag{2.6}$$

REFLEXIVITY: Let $\beta_i = \frac{\beta_{ij}^*}{2} = \beta \qquad i, j = 1, 2, 3$

$$y = \beta x_1^2 + \beta x_2^2 + \beta x_3^2 + 2\beta x_1 x_2 + 2\beta x_1 x_3 + 2\beta x_2 x_3$$

$$y = \beta(x_1^2 + x_2^2 + x_3^2 + 2x_1 x_2 + 2x_1 x_3 + 2x_2 x_3)$$

$$= \beta(x_1 + x_2 + x_3)^2 \qquad \text{but } \sum_{i=1}^{3} x_i = 1$$

$$y = \beta$$

DECOMPOSABILITY: Let $x_2 + x_3 = x_5$ since $q_2 = q_3 = q_5$ therefore $\beta_2 = \beta_3 = \beta_5$

$$\begin{aligned} y &= \beta_1 x_1^2 + \beta_2 x_2^2 + \beta_3 x_3^2 + \beta_{12}^* x_1 x_2 + \beta_{13}^* x_1 x_3 + \beta_{23}^* x_2 x_3 \\ &= (\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)^2 \\ &= (\beta_1 x_1 + \beta_5 x_5)^2 \\ &= \beta_1 x_1 + \beta_5 x_5 + 2\beta_{15} x_1 x_5 \\ &= \beta_1 x_1 + \beta_5 x_5 + \beta_{15}^* x_1 x_5 \end{aligned}$$

HOMOGENEITY:

$$\begin{aligned} y &= \sum_{i=1}^{3} (\lambda \beta_i) x_i + \sum \sum_{i<j}^{3} (\lambda \beta_{ij}) x_i x_j \\ &= \lambda (\sum_{i=1}^{3} \beta_i x_i + \sum \sum_{i<j}^{3} \beta_{ij} x_i x_j) \end{aligned}$$

After Scheffé published his models in 1958, he and Quenouille debated the interpretation of his coefficients. Quenouille pointed out that Scheffé's interpretation of coefficients did not allow for models where one component is inert or has an additive effect. Scheffé did not accept Quenouille's criticism. He claimed that his coefficients should be interpreted in terms of antagonism and synergism, and not as interactions in the strictest sense [6, 54]. However, the presence of an inert component or one with an additive effect renders Scheffé polynomials inadequate regarding fit and interpretation of coefficients [6]. In 1968 Becker introduced three models homogeneous of degree one that can account for the inert or additive effects of components in mixtures [6].

H1
$$y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1 \le i < j \le q} \beta_{ij} min(x_i x_j) \tag{2.7}$$

$$y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1 \le i < j \le q} \beta_{ij} min(x_i x_j) + \cdots + \sum\sum\cdots\sum_{1 \le i_1 < i_2 < \ldots i_n \le q} \beta_{i_1, i_2, \ldots i_n} min(x_{i_1} x_{i_2} \ldots x_{i_n})$$

H2
$$y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1 \le i < j \le q} \frac{\beta_{ij} x_i x_j}{x_i + x_j} \tag{2.8}$$

$$y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1 \le i < j \le q} \frac{\beta_{ij} x_i x_j}{x_i + x_j} + \cdots + \sum\sum\cdots\sum_{1 \le i_1 < i_2 < \ldots i_n \le q} \beta_{i_1, i_2, \ldots i_n} \frac{x_{i_1} x_{i_2} \ldots x_{i_n}}{(x_{i_1} + x_{i_2} + \cdots + x_{i_n})^{n-1}}$$

H3
$$y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1 \le i < j \le q} \beta_{ij} (x_i x_j)^{\frac{1}{2}} \tag{2.9}$$

$$y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1 \le i < j \le q} \beta_{ij} (x_i x_j)^{\frac{1}{2}} + \cdots + \sum\sum\cdots\sum_{1 \le i_1 < i_2 < \ldots i_n \le q} \beta_{i_1, i_2, \ldots i_n} (x_{i_1} x_{i_2} \ldots x_{i_n})^{\frac{1}{n}}$$

The first model, H1, is completely additive and should be applied in deciding the significance of joint effects. A drawback of H1 is that it requires a maximum or minimum at the centroid of the factor space under consideration. H2 and H3 should be applied for smoothing the response. H2 can be regarded as a special version of Scheffé polynomials since setting the denominators equal to one reduces H2 to a Scheffé polynomial. This allows the coefficients in the H2 model to be estimated by using similar simple linear forms as coefficients in Scheffé polynomials. From inspection it is clear that none of the three models complies with the requirement of decomposability.

In 1971 Cox introduced mixture models with the aim of addressing three drawbacks of the Scheffé polynomials [17]:

1. Scheffé polynomials give different parameters, β_i, for replicate experiments on exactly the same system, where the responses differ by a constant between the replicates.

2. The absence of squared terms in Scheffé polynomials of order two and higher does not provide any information on the direction or magnitude of the curvature of the response.

3. The parameter interpretations are in terms of very simple mixtures.

Cox's approach to mixture models requires a starting formulation $\mathbf{s} = (s_1, s_2, \ldots s_q)$. Depending on the application, $\mathbf{s}$ might be either a convenient starting reference mixture, which could be quite arbitrary, or it might be the centroid or in the vicinity of the centroid of the region of interest. The latter is preferred. Piepel (1983) proposed different methods of calculating centroids for experiments where the region of interest is not the entire simplex [46]. Cox introduced first- and second-order polynomials of the form [13, 17]:

$$\text{First-degree model:} \qquad y = \beta_0 + \sum_{i=1}^{q} \beta_i x_i \tag{2.10}$$

$$\text{Second-degree model:} \qquad y = \beta_0 + \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q}\sum_{j=1}^{q} \beta_{ij} x_i x_j \tag{2.11}$$

$$\text{with } \beta_{ij} = \beta_{ji}$$

Cox's approach entails an increase in one of the components of $\mathbf{s}$ from x_i to $x_i + \Delta_i$. All other components are adjusted accordingly, which implies that $x_j, j \neq i$, changes to $\frac{x_j - \Delta_i s_i}{(1-s_i)}$. The expected responses for the respective Cox polynomials are:

$$\Delta y = \frac{\beta_i \Delta_i}{1 - s_i} - \frac{\Delta_i}{1 - s_i} \sum_{\substack{j=1 \\ j \neq i}}^{q} \beta_j s_j \tag{2.12}$$

$$\Delta y = \beta_i \frac{\Delta_i}{1 - s_i} + \beta_{ii} \left(\frac{\Delta_i}{1 - s_i} \right)^2 \tag{2.13}$$

Imposing the constraints

$$\sum_{j=1}^{q} \beta_j s_j = 0 \tag{2.14}$$

$$\sum_{k=1}^{q} \beta_{jk} s_k = 0 \qquad j = 1, 2, \ldots q \tag{2.15}$$

on the two models implies that

$$y = \beta_0. \tag{2.16}$$

Equation (2.14) gives the change in the expected response at any composition point. It is interpreted as the slope of the standard mixture at a given composition. Constant replicates among experiments are absorbed by β_0. β_i is the slope of the response surface for

changes in the $i'th$ component if the other component proportions are adjusted accordingly and β_{ij} is the linear-linear interaction of components i and j [17]. These parameter interpretations are closer to those of ordinary polynomial parameters. The factor $1/(1-s_i)$ can be absorbed into the variables by redefining them as

$$z_j = \frac{x_j}{(1-s_j)} \tag{2.17}$$

This implies that the expected response is

$$\Delta y = \sum_{i=1}^{q} \beta_i \Delta z_i + \sum\sum_{1 \leq i,j \leq q} \beta_{ij} \Delta z_i \Delta z_j \tag{2.18}$$

Cornell introduced the axial design, discussed in the previous chapter, specifically for Cox's polynomials. These polynomials measure the effect that changes in components have on the expected response, which makes them ideal for experiments where the aim is to measure component effects [13]. A possible disadvantage to this application of Cox's polynomial is the interpretation of the coefficients. The pure components coefficient, β_i, depends on $1 - s_i$ which can vary greatly between different components. This affects the comparison of the effects of different components. Piepel (2007) addressed this setback by proposing component slope linear models as an alternative to first-order Cox polynomials [49]:

Version 1: $$y = \gamma_0 + \sum_{i=1}^{q} \gamma_i (1-s_i) x_i = \gamma_0 + \sum_{i=1}^{q} \gamma_i x_i^{'} \tag{2.19}$$

where $x_i^{'} = (1-s_i)x_i$ and the coefficients γ_i are subject to the constraint

$$\sum_{j=1}^{q} \gamma_j (1-s_j) s_j = 0 \tag{2.20}$$

Version 2: $$y = \gamma_0 + \sum_{i=1}^{q} \gamma_i 100(1-s_i) x_i = \gamma_0 + \sum_{i=1}^{q} \gamma_i x_i^{*} \tag{2.21}$$

where $x_i^{*} = 100(1-s_i)x_i$ and the coefficients γ_i are subject to the constraint

$$\sum_{j=1}^{q} \gamma_j (1-s_j) s_j = 0 \tag{2.22}$$

In these models the coefficients are the rate of change in the response rather than the amount of change. The component slope models do not provide higher prediction accuracy, nor do they fit data better. They merely provide parameter interpretations

that are more applicable to practice. The component effects can be assessed directly from the coefficients [49]. For a comparison between the Scheffé linear model, Cox first-order model and component slope models, refer to Piepel (2006) and Piepel (2007) [48, 49].

The similarity between these models and Scheffé's models implies that these models also adhere to the consistency rules but they are limited in their application. They are typically not applied to the whole factor space but only to a region surrounding the starting formulation. Cox stated that these polynomials are not the best choice for experimental situations where the response variable depicts extreme behaviour as components' proportions tend to zero [17]. This kind of behaviour was addressed by Draper and John (1977) [22]. They introduced augmented Scheffé polynomials specifically for these situations. The experimental region they considered does not include the boundaries where $x_i = 0$. Experimental runs are allowed close to the boundary but not actually on it. They merely extended Scheffé polynomials with an additional term, $\sum_{i=1}^{q} \beta_i x_i^{-1}$, as shown in (2.23) [22]:

Linear Blending:
$$y = \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \beta_{-i} x_i^{-1} \tag{2.23}$$

Quadratic model:
$$y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1<i<j<q} \beta_{ij} x_i x_j + \sum_{i=1}^{q} \beta_{-i} x_i^{-1} \tag{2.24}$$

Cubic model:
$$y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1<i<j<q} \beta_{ij} x_i x_j + \sum\sum_{1<i<j} \delta_{ij} x_i (x_i - x_j) + \sum\sum\sum_{1<i<j<k} \beta_{ijk} x_i x_j x_k + \sum_{i=1}^{q} \beta_{-i} x_i^{-1} \tag{2.25}$$

Special Cubic model:
$$y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1<i<j<q} \beta_{ij} x_i x_j + \sum\sum\sum_{1<i<j<k} \beta_{ijk} x_i x_j x_k + \sum_{i=1}^{q} \beta_{-i} x_i^{-1} \tag{2.26}$$

The x_i^{-1} allows for the extreme changes in the response behaviour as any component proportion approaches zero. The coefficients of these terms are considered *edge effects*. These models are mostly applied to predict the response variable and it is advised that no specific meaning should be placed on any of the coefficients [22].

Aitchison and Bacon-Shone (1984) also introduced a model for this kind of experimental situation. In their log-contrast model, the composition $\mathbf{x}$ of the mixture is transformed

into a log-ratio composition **z** by

$$z_i = log(\frac{x_i}{x_q}) \qquad i = 1, 2, \ldots, q-1 \tag{2.27}$$

The proposed models are:

$$y = \beta_0 + \sum_{i=1}^{q-1} \beta_i log\left(\frac{x_i}{x_q}\right) \tag{2.28}$$

$$y = \beta_0 + \sum_{i=1}^{q-1} \beta_i log\left(\frac{x_i}{x_q}\right) + \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \beta_{ij} log\left(\frac{x_i}{x_q}\right) log\left(\frac{x_j}{x_q}\right) \tag{2.29}$$

Since it is not possible to take the logarithm of zero, this also models extreme behaviour where components tend to zero and serves as an alternative to the models of Draper and John [1].

Since neither Draper and John (1977) nor Aitchison and Bacon-Shone (1984) include the boundaries of the experimental region, their models cannot reduce to the pure component if the composition of $(q-1)$ components tends to zero. These models therefore do not obey the desired consistency rules.

In 1985 Darroch and Waller introduced another form of additive model in which the response variable is the sum of separate functions of every proportion [18]. For a three-component mixture, the model is defined as:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + f(x_1) + g(x_2) + h(x_3) \tag{2.30}$$

under the condition

$$f(0) = f(1) = g(0) = g(1) = h(0) = h(1) = 0 \tag{2.31}$$

The condition (2.31) allows the constants, $\beta_1, \beta_2, \beta_3$, and the functions, $f(x_1), g(x_2), h(x_3)$, to be uniquely defined. The functions of the pure components are interpreted as the *"non-blandness"* of the component. A component is bland if it *lacks distinctive character* in the mixture. Deviation from linear blending for synergistic behaviour is given by $f(x_1) + g(x_2) + h(x_3)$, the sum of the non-blandness [18]. For this model, interaction is defined as non-additivity, i.e. if the additive model cannot describe the response, there is non-additivity in the response and therefore interaction exists among the components.

Darroch and Waller's additive model is closely related to Scheffé's canonical polynomials. Scheffé replaced the squared terms in a quadratic regression model to favour product terms as shown in Chapter 1, whereas Darroch and Waller's additive model favoured the squared terms. When the relation $2x_ix_j = (1 - x_k)^2 - x_i^2 - x_j^2$ is applied to a three-component mixture, x_i, x_j, x_k, the additive form of the quadratic regression model is [18]:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \theta x_1(1 - x_1) + \phi x_2(1 - x_2) + \psi x_3(1 - x_3) \tag{2.32}$$

From the close relation between the Darroch and Waller and Scheffé polynomials it follows that Darroch and Waller's model adheres to all the consistency requirements. Unfortunately, it also shares the same setback as Scheffé's models, i.e. that as the number of components in the mixture increases, the number of parameters that must be estimated increases significantly as well.

Draper and Pukelsheim (1998) introduced replacements for Scheffé canonical polynomials that are based on Kronecker algebra and vector and matrices [23]. These models are referred to as *K-models* or *K-polynomials*. The models introduced to replace first-, second- and third-order Scheffé polynomials respectively are:

$$y = x'\theta = \sum_{i=1}^{q} \theta_i x_i \tag{2.33}$$

$$y = (x \otimes x)'\theta = \sum_{i=1}^{q}\sum_{j=1}^{q} \theta_{ij} x_i x_j = \sum_{i=1}^{q} \theta_i x_i^2 + 2\sum_{1\le i<j\le q}\sum \theta_{ij} x_i x_j \tag{2.34}$$

$$y = (x \otimes x \otimes x)'\theta = \sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{k=1}^{q} \theta_{ijk} x_i x_j x_k \tag{2.35}$$

Consider the second-order K-polynomial. It is fully homogeneous in second-order terms. It is assumed that $\theta_{ij} = \theta_{ji}$, which means that in comparison with the Scheffé polynomial, the multiplicity of mixed terms x_ix_j has been doubled. This may seem disadvantageous, but the gain in symmetry more than compensates for these duplications. The K-polynomial is simply obtained from second-order Scheffé models by replacing all the x_i terms with x_i^2. Scheffé polynomials and K-models have the same number of parameters, i.e. $\binom{q+1}{3}$ for the second-order functions. Scheffé models and K-models are closely related. Consider the second-order Scheffé model. In order to relate it to the K-model, it must

first be converted to being homogenous of the second degree.

$$y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1 \leq i < j \leq q} \beta_{ij} x_i x_j \tag{2.36}$$

$$\sum_{i=1}^{q} \beta_i x_i \sum_{i=1}^{q} x_i = \sum_{i=1}^{q} \beta_i x_i^2 + \sum\sum_{1 \leq i < j \leq q} (\beta_i + \beta_j) x_i x_j \text{ where } \sum_{i=1}^{q} x_i = 1 \tag{2.37}$$

The difference between the Scheffé polynomials and the K-models is:

$$\begin{aligned} &\sum_{1 \leq i,j \leq q} \theta_{ij} x_i x_j - \sum_{i=1}^{q} \beta_i x_i - \sum\sum_{1 \leq i < j \leq q} \beta_{ij} x_i x_j \\ &= \sum_{1 \leq i \leq q} (\theta_{ii} - \beta_i) x_i^2 + \sum\sum_{1 \leq i < j \leq q} (2\theta_{ij} - \beta_i - \beta_j - \beta_{ij}) x_i x_j \end{aligned} \tag{2.38}$$

This difference disappears for all x if and only if

$$\beta_i = \theta_{ii} \qquad \text{and} \qquad \beta_{ij} = 2\theta_{ij} - \theta ii - \theta_{jj} \tag{2.39}$$

A similar approach can be followed to show the relationship between higher-order Scheffé polynomials and K-models. Since K-models can be seen as a re-parametrisation of Scheffé's models, they adhere to all the consistency requirements. K-models provide an advantage in experiments where the model is dependent on the total amount of the mixture. Unlike Scheffé polynomials, they remain homogeneous as the total amount of component proportions exceeds one [23].

K-polynomials are especially advantageous for experimental designs where additional constraints are imposed on the components [53]:

$$0 < lb_i < x_i < ub_i < 1 \tag{2.40}$$

where lb_i and ub_i are the lower and upper bounds respectively imposed on component x_i. To explain the advantages that K-polynomials present under these additional constraints, consider the matrix notation of a mixture model. A general mixture model is expressed in matrix terms as:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \text{ or} \tag{2.41}$$

$$E[\mathbf{Y}] = \mathbf{X}\beta \tag{2.42}$$

The normal equations for the least-squares parameter estimates, $\hat{\beta}$, is expressed as:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2.43}$$

The $\mathbf{X}'\mathbf{X}$ matrix is known as the *information matrix* for estimating $\hat{\beta}$. The vector of fitted values is given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} \tag{2.44}$$

If columns of the $\mathbf{X}$ matrix are linearly dependent on each other, the $\mathbf{X}'\mathbf{X}$ matrix is singular, implying that $(\mathbf{X}'\mathbf{X})^{-1}$ does not exist and the parameter estimates as expressed in (2.43) cannot be calculated [24]. This is often the case in mixture experiments where the variables are subjected to additional constraints. Their linear dependence increases, which results in linearly dependent columns of the $\mathbf{X}$ matrix [53]. However, regression programs have become so accurate that if the columns of $\mathbf{X}$ are not exactly linearly dependent, the $\mathbf{X}'\mathbf{X}$ matrix can still be calculated. The $\mathbf{X}'\mathbf{X}$ is then referred to as being *ill-conditioned*. Even though this is, in general terms, a computational improvement, it does raise concerns regarding the stability of the parameter estimates of the fitted model [53]. If $\mathbf{X}'\mathbf{X}$ is a severely ill-conditioned matrix, a change as small as 0.001 in a single design point can result in a completely different model, to the extent that parameters initially estimated to be large and positive alter to become negative [53].

An ill-conditioned $\mathbf{X}'\mathbf{X}$ matrix therefore results in least-squares estimators with large standard errors that are highly correlated and highly dependent on the precise position of the design points. This creates difficulties in models where the parameters have individual interpretations [53]. In Chapter 1 it was mentioned that there is more than one parametrisation for a model but that all parametrisations lead to the same predictions and the same prediction intervals. Different parametrisations, however, lead to different degrees of ill-conditioning of the $\mathbf{X}'\mathbf{X}$ matrix. It is therefore important to find the best-conditioned model. Generally, improving the conditioning of the information matrix reduces the variances of the estimated coefficients and their correlations, which in turn makes the model less dependent on the position of the design points [53].

The conditioning of the $\mathbf{X}'\mathbf{X}$ matrix can be assessed by a *condition number*:

$$cond(\mathbf{X}'\mathbf{X}) = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \tag{2.45}$$

where λ_{max} and λ_{min} are the largest and smallest eigenvalues calculated from the determinant equation

$$|\mathbf{X}'\mathbf{X} - \lambda\mathbf{I}| = 0 \tag{2.46}$$

If $\mathbf{X}'\mathbf{X}$ is close to singular, λ_{min} will be close to zero and the condition number will be very large. Smaller condition numbers imply greater stability in the least-squares estimates [53].

The quadratic K-polynomials always provide the smallest maximum eigenvalue, λ_{max}, of the information matrix and therefore these models are always the best conditioned [53].

Many of the models developed in the literature were developed for specific experimental situations [1, 22]. Although these models may have met with great success within the boundaries of their application, they cannot be applied to different experimental situations. Scheffé quadratic polynomials are very popular for experiments with two to four components. However, complex mixtures that give rise to complex response surfaces are seldom accurately estimated by Scheffé polynomials of such low degree. Even though success may be achieved by simply increasing the degree of the polynomial, this significantly increases the number of parameters that need to be estimated and interpreted. This in turn requires more design points, which implies a more time-consuming and costly experiment. The ultimate goal is to develop a mixture model that has few estimable parameters but is still flexible enough to be applied to a variety of experimental situations.

Chapter 3

A New Model

There are various mixture models available in the literature, as shown in Chapter 2, but few that meet the requirements that we desire in a model. These models are often effective for the different experimental designs to which they are applicable, but not flexible enough to model beyond the limitations and restrictions they were developed on [1, 22]. Many models are mere re-parametrisations of Scheffé models with the aim of extending the application of the Scheffé models, but they also share the setbacks [18]. Scheffé polynomials are convenient for mixtures with few components where the responses can be accurately described by low-order models. Most popular is the Scheffé quadratic model.

The composition dependence of many mixture properties shows non-linear deviations from predictions made by the quadratic Scheffé polynomial. The conventional approach is to use higher-order Scheffé polynomials [5, 9]. However, this significantly increases the experimental effort required to fit the additional model parameters.

Global models that are able to correlate mixture behaviour over the entire factor space are desirable. Empirical models are the norm since predictive mechanistic theories are seldom available. The mathematical form of the empirical model should be flexible enough to correlate the underlying information with no unnecessary restrictions; it should be

consistent with available physical theory; the parameters should be easy to interpret and estimable with common multivariate estimation techniques, and it should be parameter parsimonious. Ideally, the coefficients should be obtainable from pure component properties or, at most, from binary mixture data. The model should then be predictive for the general multivariate case.

A new parameter-sparse mixture model is introduced in this chapter. The model is an expansion of the simple linear blending rule. Wohl's Q-fractions and power means are employed as a means to this end.

3.1 Composition Descriptors

Consider a mixture composed of q different components. The mixture composition is quantified by the vector $\mathbf{x} \in \mathbf{R}_+^n$ where $\mathbf{x}$ is some measure of concentration of the species in the mixture that adheres to the restrictions

$$0 \leq x_i \leq 1 \qquad \sum_{i=1}^{q} x_i = 1 \tag{3.1}$$

Various composition descriptors can be used. In thermodynamics, mole fractions, z_i, are commonplace since they have theoretical significance. Mass fractions have been used as composition variables to correlate experimental data for liquid thermal conductivity [63]. Mass fractions are defined by:

$$w_i = \frac{M_i x_i}{\sum_{j=1}^{q} M_j x_j} \tag{3.2}$$

where M_i is the molar mass of component i. Gasoline octane numbers are usually correlated using volume fractions which are given by:

$$v_i = \frac{V_i x_i}{\sum_{j=1}^{q} V_j x_j} \tag{3.3}$$

where the molar volume is calculated from the ratio of molar mass to density by $V_i = \frac{M_i}{\rho_i}$ [34].

In 1946 Wohl introduced generalised composition variables by defining unique Q-fractions as follows [68]:

$$Q_i = \frac{a_i}{\sum_{j=1}^{q} a_j x_j} \tag{3.4}$$

where a_i is a suitably chosen parameter characteristic of component i. Note that the a_i are not all independent. Owing to the scaling property of the Q-fractions, one value must be predetermined (it may be set to unity) in order to fix unique values for the others. Alternatively, the a_i values may be normalised by forcing their sum to equal unity.

3.2 Modifications of the Linear Blending Rule

The simplest model that can describe mixture properties is the linear blending rule [18]

$$y = \sum_{i=1}^{q} \beta_i x_i \tag{3.5}$$

where β_i represents the physical property value for pure component i and x_i is a concentration descriptor in fractional units. The advantages of this model are that it has no adjustable parameters and that knowledge of only pure components suffices to predict multicomponent behaviour. This model provides a reasonable estimate for the variation of molar volume of a liquid whose composition is expressed in terms of mole fractions. From this it follows that the density of the liquid also obeys the linear blending rule with composition expressed in volume fractions:

$$\rho = \sum_{i=1}^{q} \rho_i v_i \tag{3.6}$$

The linear blending rule with composition defined by volume fractions has been applied in predicting fuel octane numbers [38]. As for many other physical properties of mixtures, the octane number deviates from the linear blending rule in practice. Adjusting to the linear blending rule to account for the deviations might increase its predictive ability, while the advantages of such a simple model remain.

The linear blending rule represents a weighted arithmetic mean over the pure component property values. Other, more general means can be used, such as the weighted power mean. The weighted power mean is especially significant since many other means are simply special cases, as shown in Table 3.1. The precise definition of the weighted power mean is:

$$y_r(\beta, \mathbf{x}) = \lim_{t \to r^+} \left(\sum_{i=1}^{q} x_i \beta_i^t \right)^{\frac{1}{t}} \qquad (r \in \mathbf{R}) \tag{3.7}$$

TABLE 3.1: Special forms of weighted power means

r	Defintion	Weighted Mean
-1	$y = [\sum_{i=1}^{q} \frac{x_i}{a_i}]^{-1}$	Harmonic mean
0	$y = \prod_{i=1}^{q} a_i^{x_i}$	Geometric mean
1	$y = \sum_{i=1}^{q} a_i x_i$	Arithmetic mean; Linear blending rule
2	$y = \sqrt{\sum_{i=1}^{q} x_i a_i^2}$	Quadratic weighted root mean square

This allows for the special case where $r = 0$.

Suppose $r, s \in \mathbf{R}$ with $r > s$. Then the following fundamental inequality holds for the weighted power mean for any given vector of positive numbers $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$ and fixed normalised weights $\mathbf{x} = (x_1, x_2, \ldots, x_n)$:

$$y_r(\mathbf{x}, \beta) \geq y_s(\mathbf{x}, \beta) \tag{3.8}$$

In layman's terms, this means that the bias increases towards the higher pure component property values as r increases. Values for $r > 1$ describe synergistic deviations from the linear blending rule where values of $r < 1$ describe antagonistic deviations. This is illustrated in Figure 3.1. A disadvantage of the weighted power mean model is that all binaries in a given mixture must exhibit either synergism or antagonism. Mixture data where the binary data exhibit both synergism and antagonism cannot be modelled by power mean models.

Another approach to constructing flexible mixture models is rational approximations [31]. The simplest is the ratio of two linear forms:

$$y = \frac{\sum_{i=1}^{q} a_i \beta_i x_i}{\sum_{j=1}^{q} a_j x_j} \tag{3.9}$$

where a_i are pure component adjustable parameters. In essence, this proposal replaces the mole, volume and mass fractions, as defined by (3.2) and (3.3), with Wohl's Q-fractions (3.4). Equation (3.9) still represents the linear blending rule but with the composition variables substituted by Wohl's Q-fractions:

$$y = \sum_{i=1}^{q} \beta_i Q_i \tag{3.10}$$

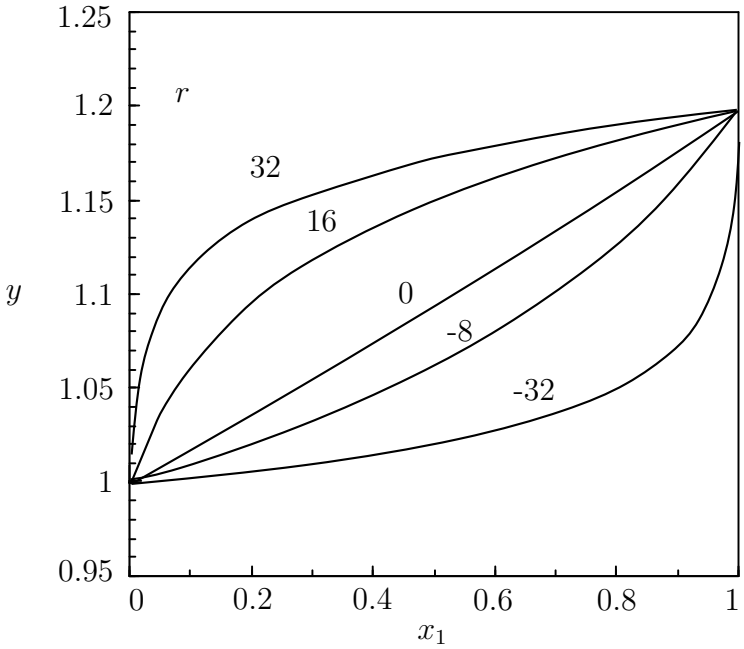


FIGURE 3.1: The effect of r on the composition trends for an arbitrary blend modelled by the power mean mixture model in (3.7)

where

$$Q_i = \frac{a_i x_i}{\sum_{j=1}^{q} a_j x_j} \tag{3.11}$$

This model has $q-1$ adjustable parameters. Figure 3.2 illustrates the range of behaviours that can be predicted by this simple Q-fraction model.

To avoid discontinuities it is necessary that $a_i > 0$ for all i. This model will reduce to the linear blending rule if all the $a_i's$ assume the same value. If all the β_i values are identical, it will be unable to predict deviation from the linear blending rule. This implies that (3.9) relies on amplifying differences in pure component property values. The model also cannot predict values that are either greater than the greatest or smaller than the smallest pure component property value. The observed trends shown in Figure 3.2 resemble those of the power mean mixture model. The Q-fraction model, on the other hand, can handle binary data that exhibit both synergistic and antagonistic behaviour.

Two approaches to adjusting the linear blending rule were considered. The first was to employ weighted power means and the second was to replace the composition descriptor with Wohl's Q-fractions. A natural expansion of these concepts is to combine them. This yields a model with q adjustable parameters:

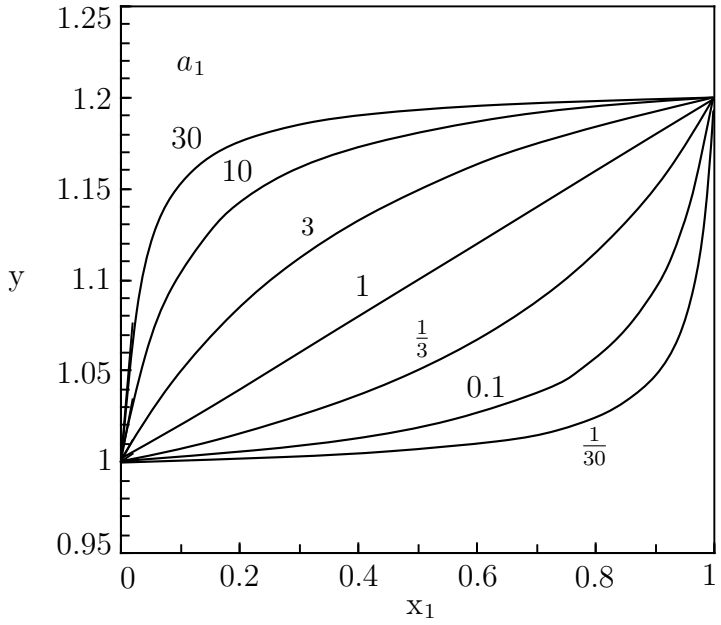


FIGURE 3.2: Trends predicted by the Q-fraction model defined by (3.9): The effect of a_1 on the response variable y in a binary blend with $\beta_1 = 1.2$; $\beta_2 = 1.0$ and $a_2 = 1$

$$y = \left(\sum_{i=1}^{q} \frac{a_i x_i \beta_i^r}{\sum_{j=1}^{q} a_j x_j} \right)^{\frac{1}{r}} \tag{3.12}$$

Equation (3.12) still describes a linear mixture model. If the equation is rewritten in the form:

$$y^r = \sum_{i=1}^{q} \frac{a_i x_i \beta_i^r}{\sum_{j=1}^{q} a_j x_j} = \sum_{i=1}^{q} Q_i \beta_i^r \tag{3.13}$$

then the response variable, y^r, follows the linear blending rule with composition quantified by Q-fractions. Equation (3.13) has the ability to model binary data that exhibit both synergism and antagonism but it is still limited to modelling only values that fall within the boundaries of the smallest and largest pure component values. To address this, consider the denominator of (3.9). This denominator corresponds to an arithmetic mean over the a_i values. A revised model is obtained by changing the denominator to a power mean of order s. The following model is obtained:

$$y = \frac{\sum_{i=1}^{q} a_i x_i \beta_i^r}{\left(\sum_{j=1}^{q} a_j^s x_j \right)^{\frac{1}{s}}} \tag{3.14}$$

and it proves a significant step forward, as shown in Figure 3.3.

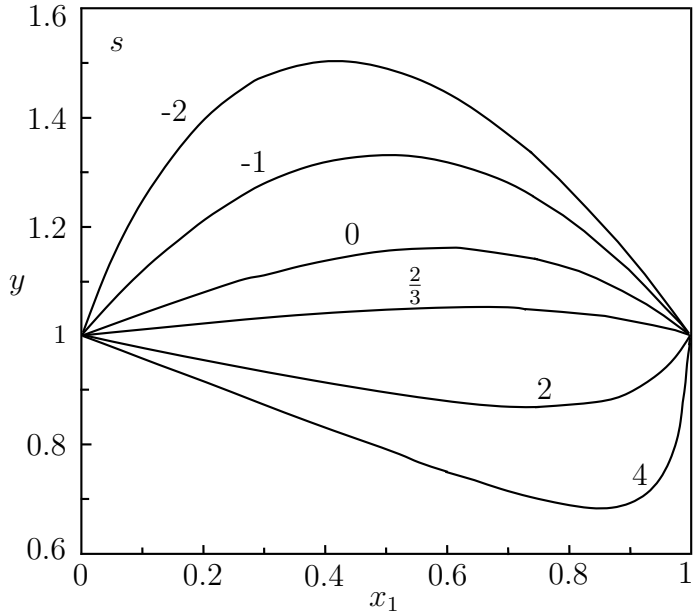


FIGURE 3.3: Trends predicted by the Q-fraction model defined by (3.14): The effect of exponent s on the response variable, y, in a binary blend with $\beta_1 = 1.2$; $\beta_2 = 1$; $a_1 = 1$ and $a_2 = 3$

Equation (3.14) is able to model the physical properties of mixtures where their values are larger than the largest and smaller than the smallest pure component values, even when two or more pure component values (β_i) are identical. We label the non-linear groupings, $a_i x_i / \left(\sum_{j=i}^{q} a_j^s x_j\right)^{\frac{1}{s}}$, as *generalised Q-fractions* but note that they do not represent normalised composition descriptors anymore.

A generalised Q-fraction model is obtained by modifying equation (3.12) in a similar way. This yields a rational expression compromising two weighted power means:

$$y = \left(\frac{\sum_{i=1}^{q} a_i x_i \beta_i^r}{\left(\sum_{j=1}^{q} a_j^s x_j\right)^{\frac{1}{s}}} \right)^{\frac{1}{r}} \tag{3.15}$$

This equation includes all the models discussed above. For example, setting $r = s = 1$ yields equation (3.9) and setting $a_i = 1$ for all the values of i reduces the model to equation (3.7) with mole fractions as composition descriptors.

With an appropriate mixture model having been identified, it was necessary to decide on a fitting experimental design. However, due to the availability of large, accurate data

sets, we did not develop an experimental design. Instead we used the physical property data reported by Ridgeway and Butler (1967) for the three-component mixture of benzene, cyclohexane and hexane [55].

Chapter 4

Bootstrap Analysis

Consider the illustration in Figure 4.1 comparing parameter estimation in the real world with parameter estimation in the bootstrap world [26]. The real-world estimation procedure was explained in Chapter 1. Observations, $\mathbf{x}$, are drawn from a population with an unknown distribution, F, and estimates, $\hat{\theta}$, for the parameter of interest, θ are calculated. In Chapter 1, we assumed F to be a normal distribution. This assumption directly influenced the properties of the estimates. If the underlying assumption of normality is wrong, we cannot be sure of the accuracy of the estimates or the validity of the model.

Bootstrap is a statistical method employed to estimate variances, confidence intervals and various other properties of statistics without any knowledge of the distribution of the data set [25]. It only assumes that the observations are from an independent identically distributed population. The estimates and confidence intervals are obtained from an approximating distribution, $\hat{F}$, that is constructed by resampling the sample [25].

The bootstrap world is concerned with determining a point estimate, $\hat{F}$, for the unknown distribution F. The estimate $\hat{F}$ yields bootstrap data vectors $\mathbf{x}^*$. For each bootstrap data vector, a bootstrap replication of the parameter estimate, $\hat{\theta}^* = s(\mathbf{x}^*)$, is calculated. Since $\hat{F}$ is known, as many as needed replications of $\hat{\theta}^*$ can be calculated. The observed variability in $\hat{\theta}^*$ can then be used to assess the accuracy of $\hat{\theta}$ [26].

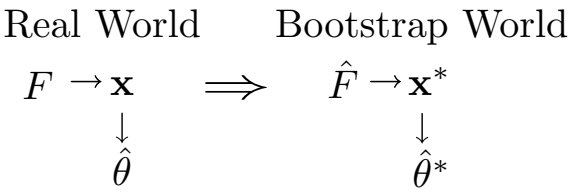


FIGURE 4.1: Typical bootstrap application

The double arrow in Figure 4.1 represents the estimation of F from $\mathbf{x}$. The jump from the real word to the bootstrap world is done with the *plug-in-principal*. This is the only inference step in bootstrap [26] and is discussed in detail in the following section.

4.1 Non-parametric Bootstrap Methodology

Consider a data set of m independent data points denoted by the matrix $W = (w_1, w_2, \ldots, w_m)^T$ where w_i is the pair; $w_i = (x_i, y_i)$. Here x_i denotes the composition descriptor and y_i is the response variable or observed experimental result, e.g. the viscosity or surface tension of the mixture. The distribution of the data set, $F(W)$, is unknown. The aim is to estimate a set of parameters $\theta = t(F)$ for a mixture model, e.g. β_i and β_{ij} for the quadratic Scheffé model. The parameter set $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ is a function of the unknown distribution F. The results for θ therefore depend on the data drawn from F. The estimate of θ, denoted as $\hat{\theta} = s(W)$, is a function of the data set W. It is determined by minimising the sum of square errors:

$$SSE = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{4.1}$$

where $\hat{y}_i$ denotes the predicted value of the response variable y_i. The precision of the estimate $\hat{\theta}$ is determined by estimating its standard error, $se_F(\hat{\theta})$, as explained in Chapter 1. This corresponds to the standard deviation of the estimated values of θ.

The bootstrap approach is a computer-based method for estimating standard errors [27]. The advantage of non-parametric bootstrap is that no information about F is required. Bootstrap substitutes the unknown distribution F with an *empirical distribution* $\hat{F}$ which yields a so-called *plug-in estimate* of the standard error. The *ideal bootstrap estimate* of the standard error $se_F(\hat{\theta})$ is denoted as $se_{\hat{F}}(\hat{\theta}^*)$. No elegant formula exists to

calculate the exact numerical value of the ideal estimator. However, bootstrap utilises a computational procedure to obtain a good approximation for $se_{\hat{F}}(\hat{\theta}^*)$. In short, the bootstrap approach is to draw B independent *bootstrap samples*. For each of these B *bootstrap replicates*, an estimate of $\hat{\theta}$, denoted by $\hat{\theta}_b^*, b = 1, 2, \ldots, B$, is determined by minimising the SSE (4.1). The standard deviation of the bootstrap replications $\hat{\theta}_b^*$ provides an estimate for the standard error of $\hat{\theta}$.

A *bootstrap sample* is generated by resampling m times with replacement from the sample $W = (w_1, w_2, \ldots, w_m)^T$ and is denoted by $W^* = (w_1^*, w_2^*, \ldots, w_m^*)^T$. Each w_i in the sample has the same probability, $1/m$, to be sampled and it can therefore be sampled more than once. The empirical distribution $\hat{F}$ is then defined as the vector of observed frequencies of the sampled w_i:

$$\hat{F} = \left(\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_m\right) \qquad \text{with} \qquad \hat{f}_i = \frac{\#(w_i = j)}{m}, i = 1, 2, \ldots, m \tag{4.2}$$

$\hat{F}$ is only one possible realisation of the true probability distribution F. The number of bootstrap samples, B, must be sufficiently large for $\hat{F}$ to empirically represent the true underlying distribution F [25]. The magnitude of B generally depends on the type of study but Efron and Tibshirani state two general rules [27]: a small number of bootstrap replications such as $B = 25$, is usually informative and even as few as 50 replications can be enough to give a good estimate of the standard error. A bootstrap sample size greater than $B = 200$ is seldom necessary to determine the standard error estimate. In the present study, B =1 000 was used in order to calculate accurate bootstrap confidence intervals.

Bootstrap confidence intervals are calculated to quantify parameter accuracy further. The bootstrap confidence intervals reported in this study are the BC_a-confidence intervals. BC_a-confidence intervals are *second-order accurate* and *transformation respecting* [27]. *Second-order accuracy* implies that the error in the calculated interval endpoint approaches zero at a rate inversely proportional to m, the sample size. This is an order of magnitude faster than the rate of first-order accurate methods where the rate is proportional to $1/\sqrt{m}$ [25]. This property is of particular value when m is small [25]. The approximations of the interval endpoints of BC_a-confidence intervals are therefore much more accurate than for other reported confidence intervals, such as the percentile

bootstrap confidence intervals [27]. *Transformation respecting* implies that the interval endpoints transform correctly when a parameter θ is replaced by some function of θ. A drawback of BC_a confidence intervals is the large number of bootstrap replications required for accurate predictions [27].

Bootstrap does not only provide measures of parameter accuracy, but can also be employed to identify the best-fitted model. The prediction error of a model is defined as the expected squared difference between the response variable, y_i, and its estimated value, $\hat{y}_i$, but thus it tends to be too optimistic. Bootstrap provides a measure that adjusts the model prediction error for its downward bias to yield a more accurate prediction error to indicate the model best fitted to the data. Efron and Tibshirani (1993) distinguish between two different bootstrap approaches [27]: the bootstrap method can be applied either to the data set or to the residuals after the SSE has been optimised. A *random sample* consists of elements from a larger population that had an equal probability of being selected. This implies that every time the experiment is repeated, a different result is obtained. A *deterministic sample* is the opposite in that it provides a single outcome given a specific input. Residual bootstrapping is more appropriate when the explanatory variables are deterministic in nature. The observed physical property, surface tension or viscosity depends on the mixture composition. Thus the explanatory variables are the mole fractions which must sum to unity. Similar values should be observed for repeated measurements on a given sample with due allowance for measurement error. Residual bootstrapping is implemented under the very strong assumption that the error between y_i and its mean u_i is not dependent on x_i, i.e. it has the same distribution F regardless of the value of x_i. We are therefore assuming that no heteroscedascity exists among the errors. In step 2 of the residual bootstrap algorithm it is shown that in the bootstrap data set, W^*, the composition descriptor, x_i, is exactly the same as for the actual data set. This is because they are treated as fixed and not as random variables, even though they were generated randomly. The standard error obtained by treating them as fixed constants reflects the precision associated with the sample of x_i actually observed [27].

It was therefore decided to apply residual bootstrapping in this study as the data set W is more deterministic than random in nature.

4.2 Residual bootstrapping Algorithm

The procedure for residual bootstrapping is as follows:

1. Estimate the parameter set, $\hat{\theta} = s(W)$ by minimising the sum of least squares, SSE (4.1). The residuals are the difference between the estimated response variable $\hat{y}_i = f(\hat{\theta}, w_i)$ and the actual response variable, y_i:

$$\varepsilon_i = y_i - \hat{y}_i, i = 1, 2, \ldots, m \tag{4.3}$$

2. Generate a bootstrap sample of the residuals, ε^*, by sampling m times with replacement from the residuals calculated in step 1. Each residual has an equal probability $(1/m)$ of being selected and may be selected more than once. The empirical distribution, $\hat{F}$, is defined as the vector of observed frequencies of the residuals. Next, the bootstrap sample of residuals is added to the estimated response variable:

$$y_i^* = \hat{y}_i + \varepsilon_i^*, i = 1, 2, \ldots, m \tag{4.4}$$

   This generates the bootstrap response variable, y_i^*, which, when combined with the corresponding mole fraction vector x_i, results in a bootstrap data set of the form $W^* = ((x_1, y_1^*), (x_2, y_2^*), \ldots, (x_m, y_m^*))^T = (w_1^*, w_2^*, \ldots, w_m^*)^T$. Now repeat the procedure B times to generate B replications of W^*.

3. For each bootstrap data set, $W_b^*, b = 1, 2, \ldots, B$, estimate the parameter set $\hat{\theta}_b^* = (\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_p^*)$ by minimising the SSE (4.1):

$$\hat{\theta}_b^* = s(W_b^*), b = 1, 2, \ldots, B \tag{4.5}$$

   Note that the bootstrap is independent of the underlying probability distribution and the nature of $s(W)$.

4. Obtain the bootstrap estimates for the set of parameters by calculating $\hat{\theta}_b^* = (\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_p^*)$ which is the mean of the B bootstrap replications of the parameters that were generated in step 3:

$$\bar{\theta}_j^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_{b,j}^* \qquad j = 1, 2, \ldots, p \tag{4.6}$$

This provides a set of bootstrap estimates for the parameters:

$$\bar{\theta}^* = (\bar{\theta}_1^*, \bar{\theta}_2^*, \ldots, \bar{\theta}_p^*) \tag{4.7}$$

The median can also be used as a more robust estimator of location.

5. For each estimation in step 4, calculate:

$$\hat{\sigma}_j^* = \left(\frac{1}{B-1} \sum_{b=1}^{B} [\hat{\theta}_{b,j}^* - \bar{\theta}_j^*]^2 \right)^{\frac{1}{2}} \qquad j = 1, 2, \ldots, p \tag{4.8}$$

This provides p bootstrap estimates of the standard error of the parameters:

$$\hat{\sigma^*} = (\sigma_1^*, \sigma_2^*, \ldots, \sigma_p^*) \tag{4.9}$$

6. The joint confidence interval between $\hat{\theta}_i$ and $\hat{\theta}_j$ can be mapped by plotting $\hat{\theta}_{b,i}^*$ against $\hat{\theta}_{b,j}^*$ for $b = 1, 2, \ldots, B$.

7. Calculate the bootstrap *Bias Corrected and accelerated (BC_a)* intervals. BC_a-confidence intervals are given by:

$$BC_a : (\hat{\theta}_{Lo}, \hat{\theta}_{Up}) = (\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)}) \tag{4.10}$$

where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}\left(\hat{z}_0 + z^{(1-\alpha)}\right)} \right) \tag{4.11}$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}\left(\hat{z}_0 + z^{(1-\alpha)}\right)} \right) \tag{4.12}$$

with

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}_b^* < \hat{\theta}\}}{B} \right) \tag{4.13}$$

$$\hat{a} = \frac{\sum_{i=1}^{m} (\bar{\theta} - \hat{\theta}_i)^3}{6\left(\sum_{i=1}^{m} (\bar{\theta} - \hat{\theta}_i)^2\right)^{\frac{3}{2}}} \tag{4.14}$$

Here Φ is the standard normal cumulative distribution function and $z^{(\alpha)}$ is the $100\alpha^{th}$ percentile point of a standard normal distribution. The quantity $\hat{z}_0$ adjusts

for bias in the estimate by measuring the discrepancy of $\hat{\theta}^*$ and $\hat{\theta}$ in normal units. If exactly half of the $\hat{\theta}_b^*$ values are less than $\hat{\theta}$, then $\hat{z}_0 = 0$. The quantity $\hat{a}$ is known as *acceleration* because it estimates the rate of change of the standard error (standard deviation) of $\hat{\theta}$ with respect to the true parameter value θ. It corrects for the often unrealistic assumption that $\hat{\theta} \sim N(\theta, se^2)$, i.e. has a normal distribution yielding symmetrical confidence intervals. If $\hat{z}_0$ and $\hat{a}$ were both zero, the equations would simplify to $1 - 2\alpha$ percentile intervals defined by the α and the $1 - \alpha$ percentiles of $\hat{G}$, the cumulative distribution function of $\hat{\theta}^*$ [27]. In layman's terms, the BC_a-confidence intervals improve the performance of percentile confidence intervals by accounting for transformations, bias corrections and acceleration improvements [25].

8. The apparent prediction error for data set $W = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))^T$ is [60]:

$$\hat{\sigma}_{AE}^2 = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \tag{4.15}$$

As mentioned previously, the $\hat{\sigma}_{AE}^2$ measure of prediction error is not sufficient on its own as it tends to underestimate the true prediction error. This is due to the fact that the same data set, on which the parameters were developed, is used to assess the fit of the model.

9. Calculate an estimate of the prediction error using the estimated bootstrap parameters:

$$PE_b^2 = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_{b,i}^*)^2 \tag{4.16}$$

Here y_i is the response variable of the data set W and $\hat{y}_{b,i}^*$ is the predicted value of y_i using the model and the b^{th} set of bootstrap-fitted parameters. The B bootstrap repetitions provide B error estimates and the bootstrap estimate of prediction error is the overall average of the error estimates [60]:

$$\hat{\sigma}_{BE}^2 = \frac{1}{B} \sum_{b=1}^{B} PE_b^2 \tag{4.17}$$

10. Estimate the downward bias, or *optimism*, in (4.15) with bootstrap as follows:

$$\varphi_{BE} = \hat{\sigma}^2_{BE} - \frac{1}{Bm}\sum_{b=1}^{B}\sum_{i=1}^{m}(\hat{y}_i - \hat{y}^*_{b,i})^2 \tag{4.18}$$

The final estimate of prediction error is the apparent error (4.15) plus the optimism (4.18)

$$\hat{\sigma}^2_{FE} = \hat{\sigma}^2_{AE} + \varphi_{BE} \tag{4.19}$$

The data points, w_i, that are omitted in each bootstrap sample are, in effect, used as a validation data set to determine the predictive capability of the model [60]. The parameter $\hat{\theta}$ as determined on the full data set will always produce a smaller prediction error than the parameters estimated on the bootstrap data $\hat{\theta}^*_b$:

$$\hat{\sigma}^2_{AE} < PE^2_b \qquad b = 1, 2, \ldots, B \tag{4.20}$$

If the bootstrap parameters, $\hat{\theta}^*_b$, are consistently worse than the best fit, then the model's predictive power is poor and the *optimism* will be large [60].

Chapter 5

Model Analysis

In Chapter 3 a new model was introduced. This model was developed systematically from the linear blending rule by introducing weighted power means and Q-fractions. Weighted power means allowed for either synergistic or antagonistic deviation from the linear blending rule without compromising the linearity of the model. Unfortunately, binary blends that exhibit both synergism and antagonism cannot be modelled by weighted power means. The next option explored was replacing the composition variable with Wohl's Q-fractions. This model can handle binaries with opposite behaviour but can only model physical properties within the limits of the physical property values set by the largest and smallest pure component values. The two concepts were combined in the following model:

$$y = \left(\frac{\sum_{i=1}^{q} a_i x_i \beta_i^r}{\left(\sum_{j=1}^{q} a_j^s x_j \right)^{\frac{1}{s}}} \right)^{\frac{1}{r}} \tag{5.1}$$

This model can describe synergistic and antagonistic behaviour simultaneously; it can predict physical properties larger than the largest and smaller than the smallest pure component physical properties and it retains its linear simplicity:

$$
\begin{aligned}
y &= \left(\frac{\sum_{i=1}^{q} a_i x_i \beta_i^r}{\left(\sum_{j=1}^{q} a_j^s x_j \right)^{\frac{1}{s}}} \right)^{\frac{1}{r}} \\
y^r &= \frac{\sum_{i=1}^{q} a_i x_i \beta_i^r}{\left(\sum_{j=1}^{q} a_j^s x_j \right)^{\frac{1}{s}}} = \sum_{i=1}^{q} Q_i^* \beta_i^r \text{ where} \\
Q_i^* &= \frac{a_i x_i}{(\sum_{j=1}^{q} a_j^s x_j)^{1/s}}
\end{aligned} \tag{5.2}
$$

Q_i^* is known as the *generalised Q-fraction*; β_i represents the physical property of every pure component and a_i are adjustable parameters associated with every pure component.

5.1 Consistency Requirements

In Chapter 2 a set of consistency requirements was introduced that we believe a truly flexible mixture should adhere to. It will now be shown that the proposed model complies with all these requirements. Consider a three-component mixture:

$$
\begin{aligned}
y^r &= \frac{\sum_{i=1}^{3} a_i x_i \beta_i^r}{\left(\sum_{j=1}^{3} a_j^s x_j \right)^{\frac{1}{s}}} \\
&= \frac{a_1 x_1 \beta_1^r + a_2 x_2 \beta_2^r + a_3 x_3 \beta_3^r}{(a_1^s x_1 + a_2^s x_2 + a_3^s x_3)^{\frac{1}{s}}}
\end{aligned}
$$

Let $x_1 = 1, x_2 = x_3 = 0$, then:

$$
\begin{aligned}
y^r &= \frac{a_1 \beta_1^r}{(a_1^s)^{\frac{1}{s}}} = \beta_1^r \\
y &= \beta_1
\end{aligned}
$$

The model reduces to $(q-1)$ components if one the components tends to zero. Let $x_3 = 0$.

$$
\begin{aligned}
y^r &= \frac{a_1 x_1 \beta_1^r + a_2 x_2 \beta_2^r + a_3 x_3 \beta_3^r}{(a_1^s x_1 + a_2^s x_2 + a_3^s x_3)^{\frac{1}{s}}} \\
y^r &= \frac{a_1 x_1 \beta_1^r + a_2 x_2 \beta_2^r}{(a_1^s x_1 + a_2^s x_2)^{\frac{1}{s}}} \\
&= \frac{\sum_{i=1}^{2} a_i x_i \beta_i^r}{\left(\sum_{j=1}^{2} a_j^s x_j\right)^{\frac{1}{s}}}
\end{aligned}
$$

REFLEXIVITY: If all the pure component property parameters are equal, $\beta_1 = \beta_2 = \beta_3 = \beta$, it is safe to assume that the adjustable property parameters, a_i, are also equal since they are associated with a specific pure component. Therefore, let $\beta_1 = \beta_2 = \beta_3 = \beta$ and $a_1 = a_2 = a_3 = a$:

$$
\begin{aligned}
y^r &= \frac{a_1 x_1 \beta_1^r + a_2 x_2 \beta_2^r + a_3 x_3 \beta_3^r}{(a_1^s x_1 + a_2^s x_2 + a_3^s x_3)^{\frac{1}{s}}} \\
&= \frac{a x_1 \beta^r + a x_2 \beta^r + a x_3 \beta^r}{(a^s x_1 + a^s x_2 + a^s x_3)^{\frac{1}{s}}} \\
&= \frac{a \beta^r (x_1 + x_2 + x_3)}{a (x_1 + x_2 + x_3)^{\frac{1}{s}}} \text{ since } \sum_{i=1}^{3} x_i = 1 \\
y &= \beta
\end{aligned}
$$

DECOMPOSABILITY: Let $x_2 + x_3 = x_5$ then $\beta_2 = \beta_3 = \beta_5$ and $a_2 = a_3 = a_5$:

$$
\begin{aligned}
y^r &= \frac{a_1 x_1 \beta_1^r + a_2 x_2 \beta_2^r + a_3 x_3 \beta_3^r}{(a_1^s x_1 + a_2^s x_2 + a_3^s x_3)^{\frac{1}{s}}} \\
&= \frac{a_1 x_1 \beta_1^r + a_5 \beta_5^r (x_2 + x_3)}{(a_1^s x_1 + a_5^s (x_2 + x_3))^{\frac{1}{s}}} \\
&= \frac{a_1 x_1 \beta_1^r + a_5 \beta_5^r x_5}{(a_1^s x_1 + a_5^s x_5)^{\frac{1}{s}}}
\end{aligned}
$$

HOMOGENEITY:

$$
\begin{aligned}
y^r &= \frac{a_1x_1(\lambda\beta_1)^r + a_2x_2(\lambda\beta_2)^r + a_3x_3(\lambda\beta_3)^r}{(a_1^sx_1 + a_2^sx_2 + a_3^sx_3)^{\frac{1}{s}}} \\
&= \frac{\lambda^r(a_1x_1\beta_1^r + a_2x_2\beta_2^r + a_3x_3\beta_3^r)}{(a_1^sx_1 + a_2^sx_2 + a_3^sx_3)^{\frac{1}{s}}} \\
y &= \lambda\left(\frac{\sum_{i=1}^3 a_ix_i\beta_i^r}{\left(\sum_{j=1}^3 a_j^sx_j\right)^{\frac{1}{s}}}\right)^{\frac{1}{r}}
\end{aligned}
$$

It is not necessary to confirm homogeneity in the adjustable parameters, a_i, since the constant multiple will simply cancel out in a similar fashion to that in *reflexivity*.

5.2 Experimental Data and Model Testing

The physical property data for viscosity and surface tension of a three-component system, benzene+cyclohexane+n-hexane, as reported by Ridgeway and Butler (1967), were utilised for parameter estimation and model testing [55]. These are extensive data sets consisting of pure component, binary and ternary interaction data that negate the need for an experimental design.

Every step of the model development was statistically analysed using the bootstrap method as set out in Chapter 4, and implemented in the statistical program *R* [66]. Table 5.1 is a summary of the models and their parameters analysed in this study. These models represent the different steps of model development followed. The analysis of every step provided an indication of whether the adjustments to the linear blending rule were serving the purpose they were intended for. Figure 5.1 and Figure 5.2 show that with every adjustment, an improvement was made to the linear blending rule.

The mean square error (MSE) and the bootstrap prediction error (BPE) are reported for every model tested. The best-fit model was chosen on the basis of the BPE. Figure 5.1 and Figure 5.2 compare the MSE and BPE obtained for the various models for surface tension and viscosity respectively. Our new model, labelled Q-fraction (r, s) (5.1), unmistakably outperformed the other models.

TABLE 5.1: The Q-fraction and power mean mixture models

Model [Parameters]	Definition	
Linear blending rule [none]	$y = \sum_{i=1}^{q} \beta_i x_i$	(5.3)
Power mean [r]	$y = (\sum_{i=1}^{q} x_i \beta_i^r)^{\frac{1}{r}}$ $(r \neq 0)$	(5.4)
Q-fractions $[a_i]$	$y = \sum_{i=1}^{q} a_i \beta_i x_i / \sum_{j=1}^{q} a_j x_j$	(5.5)
Q-fractions(s) $[a_i, s]$	$y = \sum_{i=1}^{q} a_i \beta_i x_i / \left(\sum_{j=1}^{q} a_i^s x_j\right)^{\frac{1}{s}}$ $(s \neq 0)$	(5.6)
Q-fractions(r, s) $[a_i, r, s]$	$y = \left(\sum_{i=1}^{q} a_i \beta_i^r x_i / \left(\sum_{j=1}^{q} a_i^s x_j\right)^{\frac{1}{s}}\right)^{\frac{1}{r}}$ $(r, s \neq 0)$	(5.7)
Scheffé quadratic polynomial $[\beta_{ij}]$	$y = \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1 \leq i < j \leq q} \beta_{ij} x_i x_j$	(5.8)

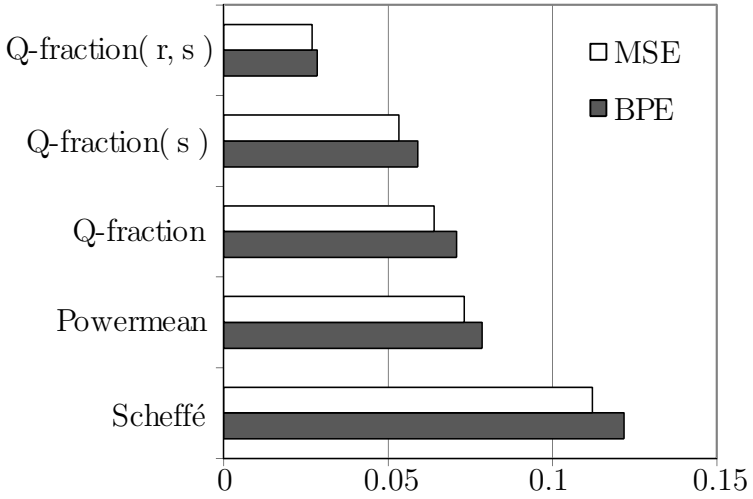


FIGURE 5.1: MSE and BPE of the mixture models for surface tension

The Scheffé quadratic model was effective for viscosity but performed surprisingly poorly for surface tension. Even the power mean, with only one adjustable parameter, fared better. Further statistical analysis was focused only on the Q-fraction (r, s) model due to its outstanding data-correlating performance for both viscosity and surface tension. A summary of the results is given in Table 5.2 and Table 5.3. The optimised least-squares and bootstrap-determined model parameters, together with the bootstrap bias-corrected and accelerated (BC_a) 95% confidence intervals, are reported.

The ample width of the 95% confidence intervals listed in Table 5.2 and Table 5.3 reveals that the r and s parameters are not precisely estimated and neither are the a_i parameters for surface tension. However, the a_i parameters for viscosity have quite narrow

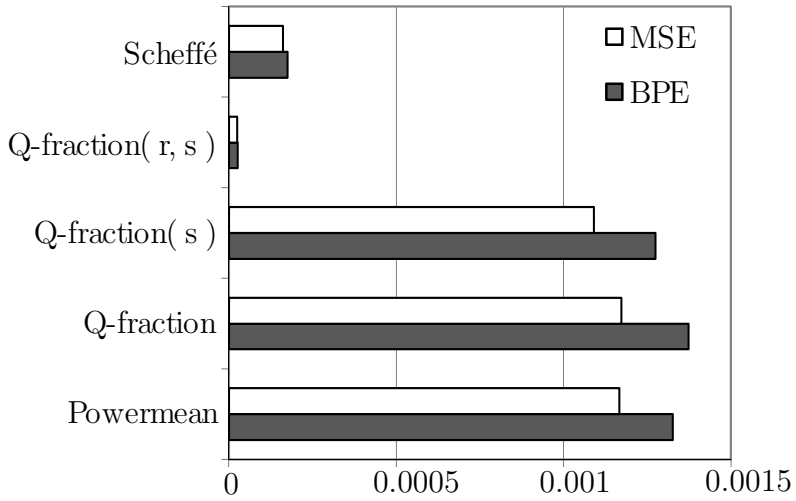


FIGURE 5.2: MSE and BPE of the mixture models for viscosity

confidence intervals and are therefore estimated with greater precision. Joint confidence intervals, visualised by plotting $\hat{\theta}_j^*$ versus $\hat{\theta}_i^*$ for all $i, j = 1, 2, \ldots, p, i \neq j$, expose the extent of the correlations between the model parameters. This information is readily available from the bootstrap. Figure 5.3 shows the results for the parameters of (5.1) obtained for the viscosity data set.

Strong linear parameter correlations are revealed for the combinations r & s, and a_1 & a_2. This is confirmed by the fact that in both cases the correlation coefficients exceed $R^2 > 0.8$. The other parameters' combinations are uncorrelated, as indicated by the very low absolute values of the correlation coefficients. Figure 5.4 shows the results for the surface tension data.

In this instance all the fitted parameters are highly correlated, albeit non-linearly. From Table 5.3, the wide 95% confidence intervals indicate that all the parameters of the Q-fractions (r, s) model are weakly estimated for surface tension data. Large uncertainties in the parameter values can be caused by deficiencies in the model or by considerable scatter in the data, especially when the deviations from the linear blending rule are minor. On the model side there may be a true lack of fit or of significant parameter correlation as a consequence of a model structure that is too flexible [27]. The wide confidence intervals reported in Table 5.2 and Table 5.3, illustrated in Figure 5.4 and Figure 5.3, are attributed to the latter explanation. The physical property data for the present binary mixtures all

TABLE 5.2: Parameter values and confidence intervals for surface tension determined for the Q-fraction (r, s) model from the full data set and the binary data only. The MSE and BPE are for the full data set only

	Full Data Set (Binaries only)		BC_a Confidence Intervals	
	Direct			
Parameters	Optimised	Bootstrap	Full Set	Binaries Only
r	-7.585 (-6.590)	-8.606 (-7.264)	(-17.26,-4.432)	(13.72,-4.583)
s	-0.878 (-1.392)	-1.104 (-1.636)	(-3.612,-0.269)	(-5.043,-0.407)
a_1	0.366 (0.353)	0.389 (0.369)	(0.304,0.586)	(0.310,0.510)
a_2	0.517 (0.486)	0.491 (0.478)	(0.460,0.556)	(0.422,0.529)
a_3	0.117 (0.161)	0.119 (0.153)	(0.052,0.275)	(0.017,0.273)
MSE	0.027	0.028		
BPE	0.027	0.057		
AAD,%	0.54 (0.56)	0.88 (0.82)		
MAD,%	1.86 (2.12)	3.34 (3.08)		

show similar antagonistic trends, with the experimental values being lower than expected from the linear blending rule. This presents a problem as both the power indices r and s, as well as the a_i values, can generate this type of concave composition dependence, as explained in Chapter 3. Thus the strong parameter correlation observed here, is attributed to reciprocal parameter compensation. Such severe parameter correlation is less likely to occur for data sets with binaries showing mixed behaviour, i.e. some showing antagonism and others synergism.

The flexibility of the model affects the accuracy of the estimates of the surface tension data set more than that of the viscosity data set. This can be attributed to the fact that viscosity presents greater deviation from linear blending than surface tension, as depicted in Figure 5.5 and Figure 5.6. The viscosity data therefore require a more flexible model.

One way of addressing the high parameter correlation and the wide confidence intervals is to reduce the dimension of the model. The most advantageous approach would be to replace all the a_i parameters with other measurable physical properties (or property

TABLE 5.3: Parameter values and confidence intervals for viscosity determined for the Q-fraction (r, s) model from the full data set and the binary data only. The MSE and BPE are for the full data set only

	Full Data Set (Binaries only)		BC_a Confidence Intervals	
Parameters	Direct Optimisation	Bootstrap	Full Set	Binaries Only
r	-1.701 (-1.448)	-1.701 (-1.458)	(-1.870,-1.518)	(-1.565,-1.314)
s	-1.287 (-2.013)	-1.287 (-2.018)	(-1.637,-1.010)	(-2.416,-1.623)
a_1	0.564 (0.520)	0.564 (0.522)	(0.536,0.588)	(0.499,0.540)
a_2	0.277 (0.281)	0.277 (0.281)	(0.274,0.282)	(0.278,0.285)
a_3	0.159 (0.199)	0.159 (0.198)	(0.135,0.187)	(0.178,0.217)
MSE	$2.4E10^{-5}$	$2.6E10^{-5}$		
BPE	$3.5E10^{-5}$	$3.7E10^{-5}$		
AAD,%	0.82 (0.97)	0.82 (0.98)		
MAD,%	2.13 (3.11)	2.13 (3.25)		

combinations) of pure component i in the mixture. In the case of surface tension, inspection revealed that the refractive index provides a reasonable replacement. The refractive index does not deviate greatly from the linear blending rule, implying that a model such as the one proposed in (5.1) is excessive. The refractive index is more than sufficiently modelled by the power mean model with $r = -10.89$ (5.6), as illustrated in Figure 5.7.

The denominator of equation (5.1) now corresponds to the power mean model for the refractive index, implying that the index s in equation (5.1) must correspond to r in the power mean model. Therefore $s = -10.89$, which results in only one adjustable parameter, r, in (5.1) to be estimated. Figure 5.8 shows that a reasonably good fit is obtained if $r = -4.42$. The average absolute deviation (AAD) between predicted and measured values is just 0.85% with a maximum absolute deviation (MAD) of 3.2%. The means that the surface tension of a mixture can be predicted fairly well, within 1 mJ/m^2, when refractive index data are available for the pure components of the mixture. More importantly, the 95% confidence interval for r has narrowed considerably to $(-4.77, -4.40)$.

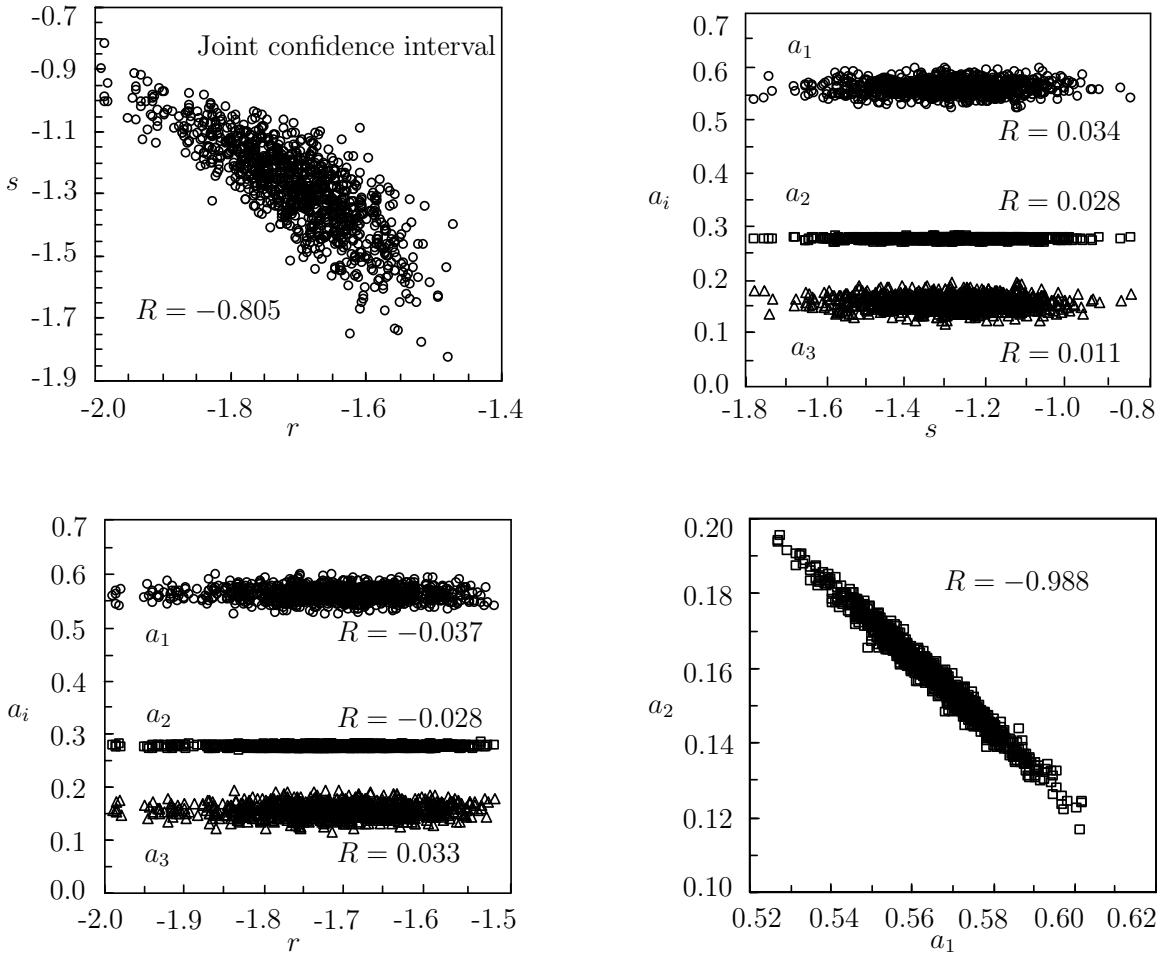


FIGURE 5.3: Joint confidence intervals and correlation coefficients for the parameters of the Q-fraction (r, s) model for the viscosity of the ternary system benzene+cyclohexane+hexane

An approach similar to that above was employed for the binary data. In this case the bootstrap samples were drawn from the binary data alone. The Q-fractions (r, s) model also provided the best fit for the binary data. The predictive qualities of the model, with the parameters estimated from regression of the binary data only, were checked using the full data set, including the ternary data. The calculated prediction errors are shown in Table 5.2 and Table 5.3. The parameters determined from the binary data alone do differ numerically from those determined from the full data set. However, Table 5.2 and Table 5.3 show that acceptable property predictions are achieved for the completed data set. Predictions for both viscosity and surface tension are, on average, accurate to within 1% (AAD), with a maximum deviation of 3.3% (MAD). This supports the contention

that multi-component behaviour can be predicted using the Q-fraction model parameters determined from binary mixture data.

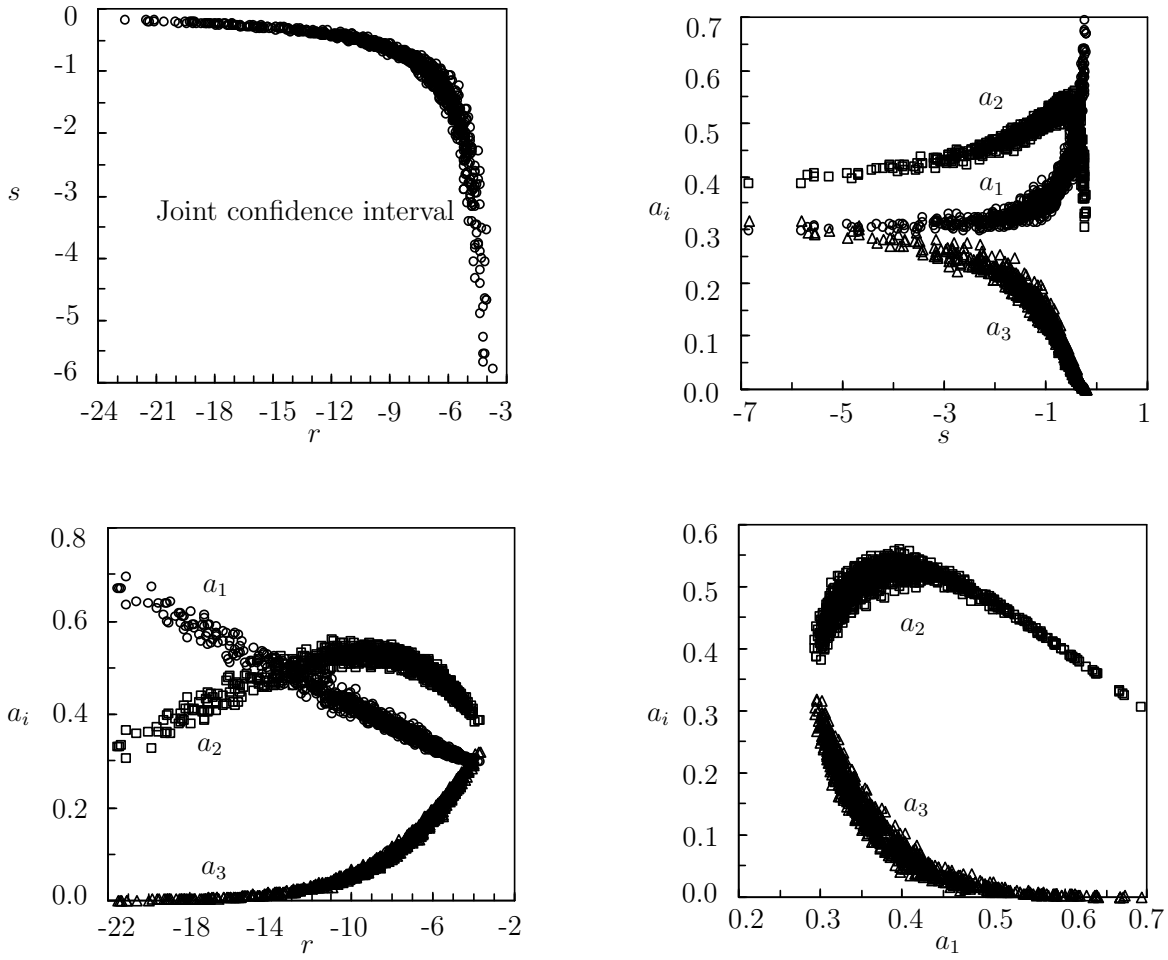


FIGURE 5.4: Joint confidence intervals and correlation coefficients for the parameters of the Q-fraction (r, s) model for the surface tension of the ternary system benzene+cyclohexane+hexane

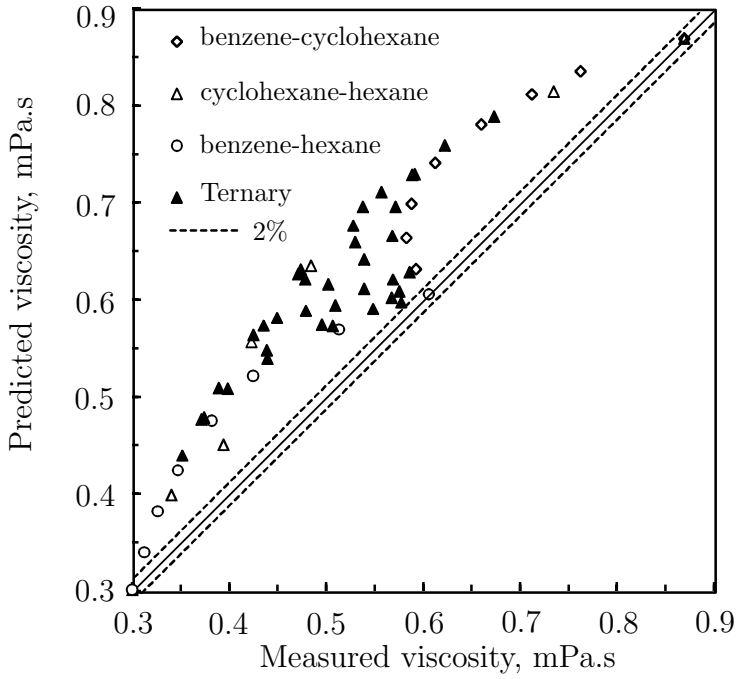


FIGURE 5.5: Deviation from linear blending for the viscosity data set

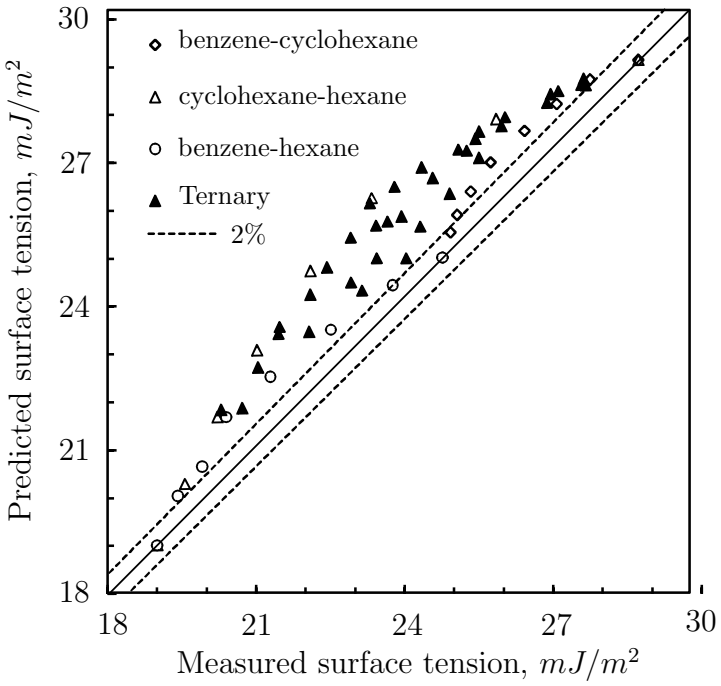


FIGURE 5.6: Deviation from linear blending for the surface tension data set

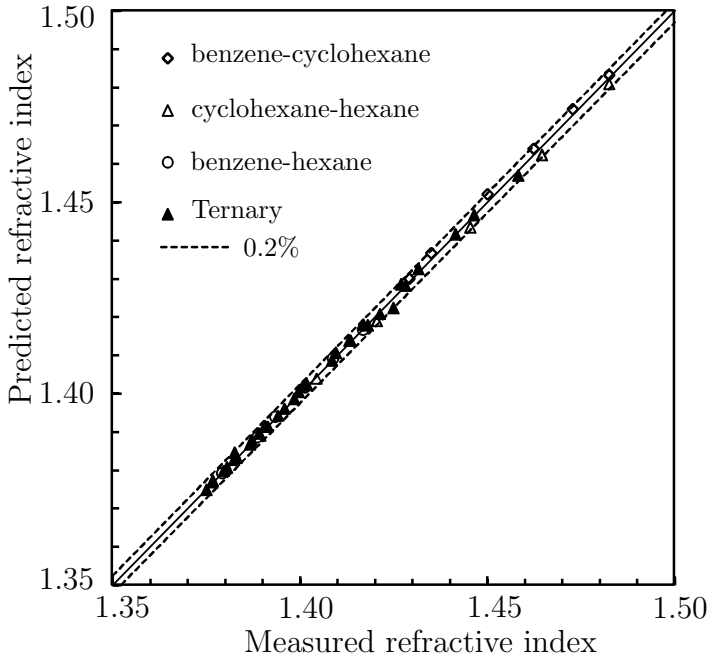


FIGURE 5.7: Refractive index modelled by the power mean model (5.6) with $r = -10.89$

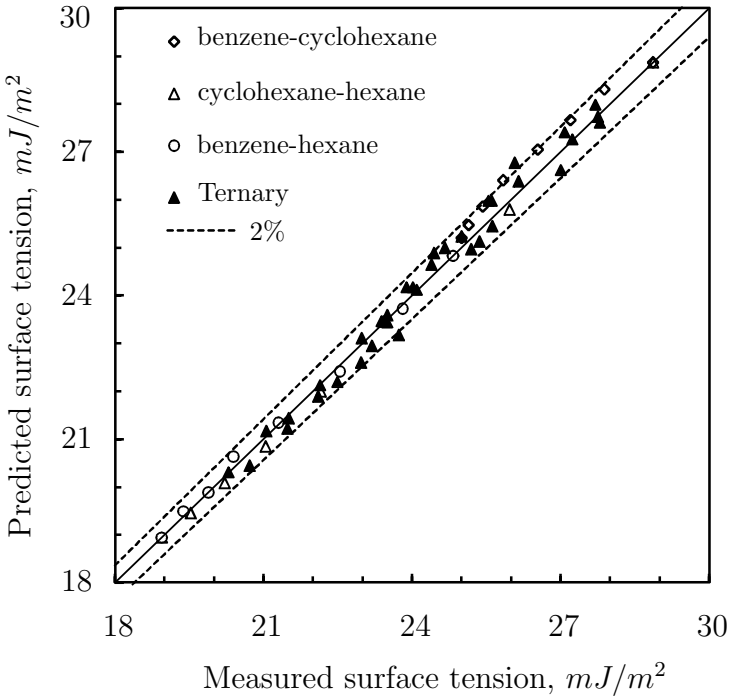


FIGURE 5.8: Surface tension modelled with (5.1) where $s = -10.89$ and $r = -4.42$, and the a_i parameters correspond to refractive index values for pure components

5.3 Parameter Estimation Procedure

The objective was to determine the set of parameter estimates with the highest probability of being correct. These parameter estimates are known as the maximum likelihood estimates. We employed the method of least squares to calculate the parameter estimates [37].

For linear models the maximum likelihood estimates are the least square parameter estimates if certain assumptions are met [8, 37]:

1. Assume that all the experimental uncertainty can be attributed to the dependent variable.

2. The measurement errors can be described by a Gaussian distribution.

3. Assume that no systematic error exists in the data.

4. The functional form is correct.

5. There are enough data points to represent the experimental uncertainties accurately.

6. All the observations are independent.

The least square estimates of a linear model can be calculated in a single step with the normal equations. Even though the proposed model can be expressed as a linear model (5.2), it is not linear in the parameters. This implies that the parameter estimation cannot occur in one step, as shown in Chapter 1, but that an iterative process is necessary [8].

The iterative procedure of the nonlinear least square analysis employs an algorithm that uses an initial estimation of the parameter set to generate a better approximation. This improved approximation is then used to generate an even better approximation. The process repeats until a stable set of parameter estimates is obtained. A better approximation is one where the weighted sum of squares of the differences between the fitted model and the experimental data decreases. The iterative process repeats until the weighted least squares function reaches a minimum. The weighting factors represent the relative precision of each data point [37].

An optimal set of least square parameter estimates for a nonlinear model yields the minimum variance of fit but that does not necessarily imply maximum likelihood parameter estimates were obtained [37]. Nonlinear models introduce the possibility that more than one set of parameters, each yielding a minimum in the sum of squares of residuals, exists. Once the optimisation procedure reaches a minimum, there is no guaranteed method of ensuring that this is the global minimum [37].

All the parameter estimates given in this chapter, are estimates that yielded the lowest bootstrap prediction error (BPE). The initial estimates for the bootstrap optimisation procedure was obtained from a direct optimisation of the model. Various initial parameter values over a wide range were explored for the direct optimisation. The initial values that yielded optimised values with the smallest mean square error was then fed into the bootstrap algorithm. The final results obtained in this experiment were independent from the initial values which is a good indication that the set of least square parameter estimates determined by the bootstrap is the global optimal estimates.

Chapter 6

Conclusion and Future work

Physical property data sometimes show severe deviations from linear blending. Although complex models exist to model such data, they typically involve estimation of a large number of parameters estimates [29, 32]. The primary objective of this study was to construct a relatively simple mixture model that would be parameter parsimonious but still flexible enough to model synergistic and antagonistic non-linear deviation simultaneously. A Q-fraction (r, s) model was developed that embodies several mixture rules as special cases. The simplest of these is the linear blending rule. It assumes that a mixture property is the composition-weighted arithmetic mean over pure component properties. The linear blending rule is desirable since it requires only pure component data for mixture property prediction.

Excellent data fits were obtained with the Q-fractions (r, s) model for surface tension and viscosity. Despite the excellent fit, surface tension presented with wide confidence intervals and strong parameter correlations. There are three possible reasons for this:

- Data for these properties do not deviate enough from the linear blending rule.
- The binary mixtures show rather similar property trends with respect to their variation with composition.

- The model structure is too flexible, allowing good correlations to be obtained with every different parameter value.

The high correlation between the parameters indicates that the model is over-parameterised. To address these problems the pure component descriptors were set to a measurable physical property of the mixture. For the surface tension data set, the descriptors was set to the refractive index property. This reduced the estimable parameters to only one, r, and removed the parameter dependence. The confidence interval for r narrowed significantly.

Experimenters should be cautious when selecting a mixture model for an experimental design. The Scheffé quadratic model is considered the standard model for mixture experiments but showed poor correlating power for surface tension. The single-parameter weighted power mean model proved superior. The Q-fraction (r, s) model proved flexible enough to accurately model data with strong non-linear tendencies, e.g. viscosity, but proved too flexible for data that still retain linear behaviour, e.g. surface tension. The model may not be suited to deal with most highly non-linear data.

Nevertheless, the model's flexibility and potential to model multicomponent data that feature non-linear composition dependence was amply demonstrated. The refractive index, a physical property of the mixture, was successfully incorporated into the model to predict surface tension more accurately. These models also have the advantage of being able to predict multicomponent behaviour from knowledge of binary behaviour.

Future Work

The research was commenced with the aim of developing a parameter-parsimonious mixture model that could ultimately model the octane number of fuel blends. Current octane-prediction models are often difficult to apply and interpret, and the required input data are cumbersome and expensive to obtain [4, 34]. The Q-fraction (r, s) model is a parameter-sparse model that proved itself able to model non-linear deviation accurately from composition dependence. The number of estimable parameters in the model can be further reduced by setting the pure component descriptors to a measurable physical property of the mixture.

The Q-fraction (r, s) should now be tested on modelling the octane number, either

RON or MON, for a fuel blend. The pure component descriptors should be set to physical properties of the components, preferably properties that can be measured on line during blending, in an attempt to improve prediction accuracy.

Fuel consists of hundreds of thousands of compounds and it is impossible to account for every single compound in a mixture model. It is common practise to lump components with similar traits and/or physical properties together in categories, e.g. aromatics and oxygenates, for such large mixtures, and then develop the mixture model not on the pure components, but on the categories. The relative proportion of each category should then sum to one. Each category is represented in the mixture by one or more of its member components [16].

The ability of the Q-fraction (r, s) model to handle categorised components should be analysed extensively. Fuels are often blended from different fuel streams. The aim is to predict the octane number of the final blend from the physical property data of each stream without any knowledge of the composition of the streams. The different streams will serve as the respective categories.

Fuel is also subject to stringent constraints on various categories [65]. The Q-fraction (r, s) model should therefore also be tested on mixtures where constraints are imposed on the components. These constraints imply that a smaller region of the factor space is under consideration and therefore an appropriate experimental design should be selected to model this region.

The Q-fraction (r, s) model proved to be promising. Its behaviour with regard to the viscosity and surface tension data sets justifies further research into its ability to model more complex mixtures. The decision of initial values for the estimation procedure requires a closer look as well. The procedure followed as set out in Chapter 5 was sufficient for model validation but for model application the decision of initial values is crucial in ensuring that the maximum likelihood estimators for the parameters are calculated.

Bibliography

[1] Aitchison, J., and Bacon-Shone, J. Log contrast models for experiments with mixtures. *Biometrika 71*, 2 (1984), 323–330.

[2] Akalin, O., Akay, K. U., Sennaroglu, B., and Tez, M. Optimization of chemical admixture for concrete on mortar performance tests using mixture experiments. *Chemometrics and Intelligent Laboratory Systems* (2010). doi: 10.1016/j.chemolab.2010.08.013.

[3] Akay, K. U. A note on model selection in mixture experiments. *Journal of Mathematics and Statistics 3*, 3 (2007), 93–99.

[4] Albahari, T. A. Structural group contribution method for predicting the octane number of pure hydrocarbon liquids. *Industrial and Engineering Chemistry Research 42* (2003), 657–662.

[5] Atikler, U., Demir, H., Tokatli, F., Tihminlioglu, F., Balkose, D., and Ulku, S. Optimisation of the effect of colemanite as a new synergistic agent in an intumescent system. *Polymer Degradation and Stability 91*, 7 (2006), 1563–1570.

[6] Becker, N. G. Models for the response of a mixture. *Journal of the Royal Statistical Society Series B (Methodological) 30*, 2 (1968), 349–358.

[7] Bello, L. H. A. D., and Vieira, A. F. C. Optimization of a product performance using mixture experiments. *Journal of Applied Statistics 37*, 1 (2010), 105–117.

[8] Britt, H. I., and Luecke, R. H. The estimation of parameters in nonlinear, implicit models. *Technometrics 15*, 2 (1973), 233–247.

[9] Chardon, J., Nony, J., Sergent, M., Mathieu, D., and Luu, R. P. T. Experimental research methodology applied to the development of a formulation for use with textiles. *Chemometrics and Intelligent Laboratory Systems 6* (1989), 313–321.

[10] Chu, C. A., and Resurreccion, A. V. A. Optimization of a chocolate peanut spread using response surface methodology (RSM). *Journal of Sensory Studies 19*, 3 (2004), 237–260.

[11] Claringbold, P. J. Use of the simplex design in the study of joint action of related hormones. *Biometrics 11*, 2 (June 1955), 174–185.

[12] Cornell, J. A. Experiments with mixtures: A review. *Technometrics 15*, 3 (August 1973), 437–455.

[13] Cornell, J. A. Some comments on designs for Cox's mixture polynomial. *Technometrics 17*, 1 (February 1975), 25–35.

[14] Cornell, J. A. Experiments with mixtures: An update and bibliography. *Technometrics 21*, 1 (1979), 95–106.

[15] Cornell, J. A. *Experiments with Mixtures. Designs, Models, and the Analysis of Mixture Data*, 3rd ed. Wiley: New York, 2002.

[16] Cornell, J. A., and Good, I. J. The mixture problem for categorized components. *Journal of the American Statistical Association 65*, 329 (1970), 339–355.

[17] Cox, D. R. A note on polynomial response functions for mixtures. *Biometrika 58*, 1 (1971), 155–159.

[18] Darroch, J. N., and Waller, J. Additivity and interaction in three component experiments with mixtures. *Biometrika 72*, 1 (1985), 153–163.

[19] Dingstad, G. I., Egelandsdal, B., and Naes, T. Modeling methods for crossed mixture experiments. A case study from sausage production. *Chemometrics and Intelligent Laboratory Systems 66* (2003), 175–190.

[20] Dingstad, G. I., Westad, F., and Naes, T. Three case studies illustrating the properties of ordinary and partial least squares. *Chemometrics and Intelligent Laboratory Systems 71* (2004), 33–45.

[21] Douglas, E., and Pouskouleli, G. Prediction of compressive strength of mortars made with portland cement-blast-furnace slag-fly ash blends. *Cement and Concrete Research 21*, 2 (1991), 523–534.

[22] Draper, N. R., and John, R. C. S. A mixture model with inverse terms. *Technometrics 19*, 1 (1977), 37–46.

[23] Draper, N. R., and Pukelsheim, F. Mixture models based on homogeneous polynomials. *Journal of Statistical Planning and Inference 71*, 1–2 (1998), 303–311.

[24] Draper, N. R., and Smith, H. *Applied Regression Analysis*, 3rd ed. Wiley Series in Probability and Statistics. Wiley: New York, 1998.

[25] Efron, B. Better bootstrap confidence intervals. *Journal of the American Statistical Association 82*, 397 (1987), 171–185.

[26] Efron, B. Second thoughts on the bootstrap. *Statistical Science 18*, 2 (2003), 135–140.

[27] Efron, B., and Tibshirani, R. J. *An Introduction to the Bootstrap*. Chapman and Hall/CRC: New York, 1993.

[28] Focke, W. Mixture models based on neural network averaging. *Neural Computation 18* (2006), 1–9.

[29] Focke, W. W. Mixture model for correlation excess enthalpy data. *Journal of Chemical Engineering of Japan 40*, 4 (2007), 295–303.

[30] Focke, W. W. Weighted-power-mean mixture model for the gibbs energy of fluid mixtures. *Industrial and Engineering Chemical Research 48*, 11 (2009), 5537–5541.

[31] Focke, W. W., and du Plessis, B. Correlating multicomponent mixture properties with homogeneous rational functions. *Industrial and Engineering Chemistry Research 43*, 26 (2004), 8369–8377.

[32] Focke, W. W., Sandrock, C., and Kok, S. Weighted-power-mean mixture model: Empirical mixing laws for liquid viscosity. *Industrial and Engineering Chemistry Research 46*, 13 (2007), 4600–4666.

[33] Frisbee, S. E., and McGinity, W. Influence of nonionic surfactants on the physical and chemical properties of a biodegradable pseudolatex. *European Journal of Pharmaceutics and Biopharmaceutics 40*, 6 (1994), 355–363.

[34] Ghosh, P., Hickey, K. J., and Jaffe, S. B. Development of a detailed gasoline composition-based octane model. *Industrial and Engineering Chemistry Research 45* (2006), 337–345.

[35] Gorman, J. W., and Hinman, J. E. Simples lattice designs for multicomponent systems. *Technometrics 4*, 4 (November 1962), 463–487.

[36] Hamad, E. Z. Exact limits of mixture properties and excess thermodynamic functions. *Fluid Phase Equilibria 142*, 1–2 (1998), 163–184.

[37] Johnson, M. L., and Faunt, L. M. *Parameter Estimation by Least-Squares Methods*, vol. 210 of *Methods in Enzymology*. Academic press, San Diego, 1992, ch. 1.

[38] Jr, M. E. M., Stollsteimer, J., and Wims, A. M. Determination of gasoline octane numbers from chemical composition. *Analytical Chemistry 47*, 13 (1975), 2301–2301.

[39] Jumaa, M., Kleinebudde, P., and Müller, B. Mixture experiments with the oil phase of parenteral emulsions. *European Journal of Pharmaceutics and Biopharmaceutics 46*, 2 (1998), 161–167.

[40] Marquardt, D. W., and Snee, R. D. Test statistics for mixture models. *Technometrics 16*, 4 (1974), 533–537.

[41] Mclean, R. A., and Anderson, V. L. Extreme vertice design of mixture experiments. *Technometrics 8*, 3 (August 1966), 447–454.

[42] Montgomery, D. C. *Design and Analysis of Experiments*, 7th ed. Wiley: New York, 1976.

[43] Murty, J. S., and Das, M. N. Design and analysis of experiments with mixtures. *The Annals of Mathematical Statistics 39*, 5 (1968), 1517–1539.

[44] Pando, C., Renuncio, J. A. R., Calzon, J. A. G., Christensen, J. J., and Izatt, R. M. Correlation and prediction of ternary excess enthalpy data. *Journal of Solution Chemistry 16*, 7 (1987), 503–527.

[45] Pasadakis, N., Gaganis, V., and Foteinopoulos, C. Octane number prediction for gasoline blends. *Fuel Processing Technology 87* (2006), 505–509.

[46] Piepel, G. F. Calculating centroids in constrained mixture experiments. *Technometrics 25*, 3 (1983), 279–283.

[47] Piepel, G. F. *50 Years of Mixture Experiment Research: 1955-2004*. World Scientific Publishing, Singapore, 2006, ch. 12 In: Response Surface Methodology and Related Topics.

[48] Piepel, G. F. A note comparing component-slope Scheffé and Cox parameterizations of the linear mixture experiment model. *Journal of Applied Statistics 33*, 4 (2006), 397–403.

[49] Piepel, G. F. A component slope linear model for mixture experiments. *Quality Technology and Quantitative Management 4*, 3 (2007), 331–343.

[50] Piepel, G. F., and Cornell, J. A. Mixture experiment approaches: Examples, discussion, and recommendations. *Journal of Quality Technology 26*, 3 (1994), 177–196.

[51] Piepel, G. F., and Landmesser, S. M. Mixture experiment alternatives to the slack variable approach. *Quality Engineering 21* (2009), 262–276.

[52] Piepel, G. F., and Redgate, T. A mixture experiment analysis of hald cement data. *The American Statistician 52*, 1 (1998), 23–30.

[53] Prescott, P., Dean, N. R., Dean, A. M., and Lewis, S. M. Mixture experiments: ILL-conditioning and quadratic model specification. *Technometrics 44*, 3 (2002), 260–268.

[54] Quenouille, M. H. Experiments with mixtures. *Journal of the Royal Statistical Society Series B (Methodological) 21*, 1 (1953), 201–202.

[55] Ridgeway, K., and Butler, P. A. Some physical properties of the ternary system benzene-cyclohexane-n-hexane. *Journal of Chemical and Engineering Data 12* (1967), 509.

[56] Saikaew, C., and Sripaya, P. Quality improvement of recycled plastic products using mixture experiments. *Proceedings of the Second International Conference on Environmental and Computer Science* (2009), 312–316. doi: 10.1109/ICECS.2009.13.

[57] Scheffé, H. Experiments with mixtures. *Journal of the Royal Statistical Society Series B (Methodological) 20*, 2 (1958), 344–360.

[58] Scheffé, H. The simplex centroid design for experiments with mixtures. *Journal of the Royal Statistical Society Series B (Methodological) 25*, 2 (1963), 235–263.

[59] Semmar, N., Jay, M., Farman, M., and Roux, M. A new approach to plant diversity assessment combining HPLC data, simplex mixture design and discriminant analysis. *Environmental Modeling and Assessment 13*, 1 (2007), 17–33.

[60] Sharrock, C. J., and Coetzer, R. J. L. Selecting robust kinetic models on noisy data using the bootstrap. *International Journal of Chemical Reactor Engineering 5* (2007), A103.

[61] Silva, A., Pinto, D., Segadães, A., and Devezas, T. C. Designing particle sizing and packing for flowability and sintered mechanical strength. *Journal of the European Ceramic Society 30*, 14 (2010), 2955–2962.

[62] Snee, R. D., and Marquardt, D. W. Screening concepts and designs for experiments with mixtures. *Technometrics 18*, 1 (1976), 19–29.

[63] Spindler, K., Hoffmann, N., Sohns, J., and Hahne, E. Thermal conductivity of binary and ternary refrigament mixtures: Experimental results and correlations. *High Temperatures-High Pressures 29*, 6 (1997), 659–664.

[64] Steinberg, D. M., and Hunter, W. G. Experimental design: Review and comment. *Technometrics 26*, 2 (May 1984), 71–97.

[65] Stikkers, D. E. Octane and the environment. *The Science of the Total Environment 299* (2002), 37–56.

[66] Team, R. D. C. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. 3-900051-07-0.

[67] Twu, C. H., and Coon, J. E. A generalized interaction method for the prediction of octane numbers for gasoline blends. *Simulation Sciences Inc: Brea, California* (1998), 1–2.

[68] Wohl, K. Thermodynamic evaluation of binary and ternary liquid systems. *Transactions of the American Institute of Chemical Engineers 42* (1946), 215–249.

Appendix A

Deduction of a third-order Scheffé Polynomial

$$
\begin{aligned}
y &= \beta_0 + \sum_{i=1}^{q} \beta_i x_i + \sum\sum_{1\le i\le j\le q} \beta_{ij} x_i x_j + \sum\sum\sum_{1\le i\le j\le k\le q} \beta_{ijk} x_i x_j x_k \\
&= \beta_0 \sum_{i=1}^{q} x_i + \sum_{i=1}^{q} \beta_i x_i + \sum_{i=1}^{q} \beta_{ii} x_i^2 + \sum\sum_{1\le i<j\le q} \beta_{ij} x_i x_j + \sum_{i=1}^{q} \beta_{iii} x_i^3 \\
&+ \sum\sum_{1\le i<j<\le q} \beta_{iij} x_i^2 x_j + \sum\sum_{1\le i<j\le q} \beta_{ijj} x_i x_j^2 + \sum\sum\sum_{1\le i<j<k\le q} \beta_{ijk} x_i x_j x_k \qquad (6.1)
\end{aligned}
$$

The following relations hold:

$$
\begin{aligned}
x_i^2 &= x_i - \sum_{\substack{j=1 \\ i\neq j}}^{q} x_i x_j \\
x_i^3 = x_i &= \frac{1}{2} \sum_{\substack{j=1 \\ i\neq j}}^{q} [3x_i x_j + x_i x_j (x_i - x_j) - \sum_{\substack{k=1 \\ k\neq j,j}}^{q} x_i x_j x_k] \\
x_i^2 x_j + x_i x_j^2 &= \frac{1}{2} [x_i x_j + x_i x_j (x_i - x_j) - \sum_{\substack{k=1 \\ k\neq i,j}}^{q} x_i x_j x_k]
\end{aligned}
$$

Substituting these relations into (6.1) and simplifying yields:

$$y = \sum_{i=1}^{q} \beta_i^* x_i + \sum\sum_{1 \leq i < j \leq q} \beta_{ij}^* x_i x_j + \sum\sum_{1 \leq i < j \leq q} \delta_{ij} x_i x_j (x_i - x_j) + \sum\sum\sum_{1 \leq i < j < k \leq 1} \beta_{ijk}^* x_i x_j x_k \tag{6.2}$$

Where

$$\begin{aligned} \beta_i^* &= \beta_0 + \beta_i + \beta_{ii} + \beta_{iii} \\ \beta_{ij}^* &= \beta_{ij} - \beta_{ii} - \frac{3}{2}\beta_{iii} + \frac{1}{2}\beta_{iij} \\ \delta_{ij} &= \frac{1}{2}\left(\beta_{iij} - \beta_{iii}\right) \\ \beta_{iii}^* &= \beta_{iii} - \beta_{iij} + \beta_{ijk} \end{aligned}$$

If the terms $\delta_{ij} x_i x_j (x_i - x_j)$ are not considered, (6.2) reduces to the special cubic polynomial [15].