SFB
823

Numerical algebraic fan of a design for statistical model building

Discussion Paper

Nikolaus Rudak, Sonja Kuhnt,
Eva Riccomagno

Nr. 4/2013

SFB
823

Contents

1 Introduction

In this article we develop methods for the analysis of non-standard experimental designs by using techniques from algebraic statistics. Our work is motivated by a thermal spraying process used to produce a particle coating on a surface, e.g. for wear protection or durable medical instruments. In this application non-standard designs occur as intermediate results from initial standard designs in a two-stage production process. We investigate algebraic methods to derive better identifiable models with particular emphasis on the second stage of two-stage processes.

Ideas from algebraic statistics are explored where the design as finite set of distinct experimental settings is expressed as solution of a system of polynomials. Thereby the design is identified by a polynomial ideal and features and properties of the ideal are explored and provide inside into the structures of models identifiable by the design [Pistone et al., 2001, Riccomagno, 2009]. Holliday et al. [1999] apply these ideas to a problem from the automotive industry with an incomplete standard factorial design, Bates et al. [2003] to the question of finding good polynomial metamodels for computer experiments.

In our thermal spraying application, designs for the controllable process parameters are run and properties of particles in flight measured as intermediate responses. The final output describes the coating properties, which are very time-consuming and expensive to measure as the specimen has to be destroyed. It is desirable to predict coating properties either on the basis of process parameters and/or from particle properties. Rudak et al. [2012] provides a first comparison of different modeling approaches. There are still open questions: which models are identifiable with the different choices of input (process parameters, particle properties, or both)? Is it better to base the second model between particle and coating properties on estimated expected values or the observations themselves? The present article is a contribution in this direction. Especially in the second stage particle properties as input variables are observed values from the originally chosen design for the controllable factors. The resulting design on the particle property level can be tackled with algebraic statistics to determine identifiable models. However, it turns out that resulting models contain elements which are only identifiable due to small deviations of the design from more regular points, hence leading to unwanted unstable model results.

We tackle this problem with tools from algebraic statistics. Because of the fact that data in the second stage are very noisy, we extend existing theory by switching from symbolic, exact computations to numerical computations in the calculation of the design ideal and of its fan. Specifically, instead of polynomials whose solution are the design points, we identify a design with a set of polynomials which "almost vanish" at the design points using results and algorithms from Fassino [2010].

The paper is organized as follows. In Section 2 three different approaches towards the modeling of a final output in a two-stage process are introduced and compared. The algebraic treatment and reasoning is the same whatever the approach. Section 3 contains the theoretical background of algebraic statistics for experimental design, always exemplified for the special application. Section 4 is the case study itself.

2 Direct, indirect and composite model

Aiming at a prediction model of the final response Z in a two stage model, we consider three different approaches where the prediction model is either based on the initial input X, the intermediate outcomes Y or a prediction $\hat{Y}$ of them. After introducing the different approaches in general we discuss them in more detail for main effect models from Y to Z.

2.1 Three model strategies

To fix notation assume X has q components, Y has p components, and Z has m components. Model building is based on an initial design D_x and we have observed values D_y and D_z.

A first model building strategy, which we name **direct model**, assumes $Z = h(X) + \delta$ with $E(\delta|X) = E(\delta) = 0$ and given $Var(\delta|X) = Var(\delta)$ and hence

$$E(Z|X = x) = h(x).$$

Our **composite model** is based on the assumptions that $Z = g(f(X)) + \eta$ and $Y = f(X) + \epsilon$, thus

$$E(Z|X = x) = g(E(Y|X = x))$$

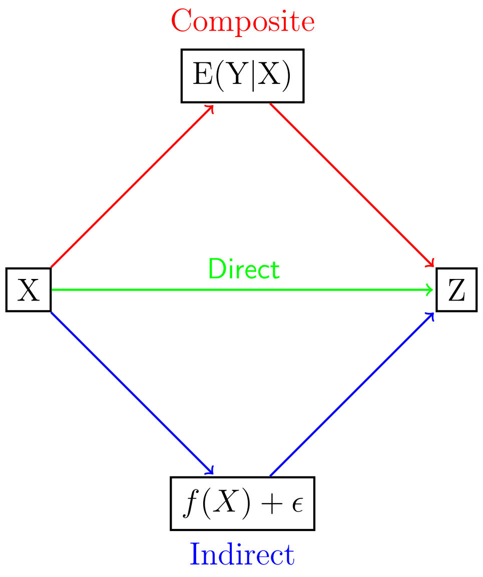

Figure 1: Modeling strategies

and the **indirect model** takes $Z = g(f(X) + \epsilon) + \tilde{\eta}$ and $Y = f(X) + \epsilon$, hence

$$E(Z|X = x) = E(g(f(X) + \epsilon)|X = x). \tag{1}$$

We assume throughout

$$E(\epsilon|X) = E(\epsilon) = 0 \text{ and } Var(\epsilon|X) = Var(\epsilon) \text{ given}$$
$$E(\eta|Y) = E(\eta|X) = 0 \text{ and } Var(\eta|Y) = Var(\eta|X) = Var(\eta) \text{ given}$$
$$E(\tilde{\eta}|Y) = E(\tilde{\eta}|X) = 0 \text{ and } Var(\tilde{\eta}|Y) = Var(\tilde{\eta}|X) = Var(\tilde{\eta}) \text{ given.}$$

Figure 1 illustrates these three model approaches.

If g is a linear function then Equation (1) becomes $E(Z|X) = g(f(x))$ by linearity of expectation and the indirect and composite model coincide.

2.2 Main effect linear models from Y to Z

We next compare the different approaches on the model level for the special case of linear models and main effects in going from Y to Z. Note that we still allow models beyond main effects in the direct strategy as well as from X to Y for the other two strategies. Without loss of generality we set $m = 1$, hence $Z \in \mathbb{R}$.

Under these restrictions the **direct model** becomes

$$Z = h(X) + \delta \underbrace{=}_{\text{linear model}} X_z^T \gamma^* + \delta$$

with γ^* an unknown parameter vector and X_z a vector of monomials of the original X-variables, to model intercept, main effects, interactions, quadratic terms and so on, as required. Thereby it follows from the assumptions in Section 2.1 that

$$E(Z|X = x) = X_z^T \gamma^*. \tag{2}$$

We next introduce a notation to represent polynomial models that will be expedient in this section and later on. The symbol x^α stands for the monomial $x_1^{\alpha_1} \dots x_q^{\alpha_q}$ where for $i = 1, \dots, q$, α_i is a non negative integer number and $\alpha = (\alpha_1, \dots, \alpha_q) \in \mathbb{Z}^q_{\geq 0}$. For example, the intercept is given by the zero vector $(0, \dots, 0) = 0_q$, and a main effect model by $\sum_{\alpha \in L} \theta_\alpha x^\alpha$ with $L = \{0_q, (1, 0, \dots, 0), (01, 0, \dots, 0), \dots, (0, \dots, 0, 1)\}$ and θ_α real numbers, while a generic linear model is of the form

$$\sum_{\alpha \in L} \theta_\alpha x^\alpha$$

with $\theta_\alpha \in \mathbb{R}^q$ and L a finite subset of $\mathbb{Z}^q_{\geq 0}$. In this notation model (2) above becomes

$$E(Z|X = x) = x_z^T \gamma^* = \sum_{\alpha \in L} \gamma^*_\alpha x^\alpha$$

Note that the support of the model $x_z = [x^\alpha]_{\alpha \in L}$ is identified with the exponents of the monomials in the model and the parameter vector is $\gamma^* = [\gamma^\alpha]_{\alpha \in L}$.

For the **composite model** when we assume a main effect linear model between Y and Z, equation (1) simplifies to

$$E(Z|X = x) = E(\gamma_0 + f(X)^T \gamma + \eta | X = x) = \gamma_0 + f(x)^T \gamma \tag{3}$$

with $\gamma_0 \in \mathbb{R}$ and $\gamma \in \mathbb{R}^p$ unknown parameters, for some suitable $p \in \mathbb{Z}_{\geq 1}$. Similarly when g gives a main effect linear model the **indirect model** gives

$$E(Z|X = x) = E(\tilde{\gamma_0} + (f(X) + \epsilon)^T \tilde{\gamma} + \tilde{\eta}|X) = \tilde{\gamma_0} + f(x)^T \tilde{\gamma} \tag{4}$$

From (3) and (4) we can conclude that the indirect and composite strategies are structurally the same if and only if $\gamma = \tilde{\gamma}$ and $\gamma_0 = \tilde{\gamma_0}$.

Next, we replace each component of $f(x)$ in (3) and (4) by a multivariate linear model. So for $i = 1, \dots, p$ let the i-th component of f be written as

$$f(x)_i = \sum_{\alpha \in L_i} x^\alpha \beta^i_\alpha = x_{y,i}^T \beta^i$$

where L_i identifies the support vector $x_{y,i} = [x_\alpha]_{\alpha \in L_i}$ for the X to Y_i regression model and $\beta^i = [\beta^i_\alpha]_{\alpha \in L_i}$ gives the unknown parameter vector. Hence equation (3) becomes

$$\begin{aligned} E(Z|X=x) &= \gamma_0 + f(x)^T\gamma = \gamma_0 + \left[\sum_{\alpha \in L_i} x^\alpha \beta^i_\alpha\right]^T_{i=1,\dots,p} \gamma \\ &= \gamma_0 + \sum_{i=1}^{p} \sum_{\alpha \in L_i} x^\alpha \beta^i_\alpha \gamma_i. \end{aligned} \tag{5}$$

By assuming equality of $E(Z|X=x)$ in all modeling approaches, from (2) and (5) we obtain an equality of polynomials

$$\sum_{\alpha \in L} \gamma^*_\alpha x^\alpha = \gamma_0 + \sum_{i=1}^{p} \sum_{\alpha \in L_i} x^\alpha \beta^i_\alpha \gamma_i$$

This holds true if and only if coefficients of the same monomial on the left hand side and right hand side are equal. To expand on this we further assume that all X-to-Y models admit intercept, so that $0_q \in L_i$ for all i from 1 to p, and define $L^*_i = L_i \setminus \{0_q\}$. The above become

$$\gamma^*_{0_q} + \sum_{\alpha \in L^*} \gamma^*_\alpha x^\alpha = \gamma_0 + \sum_{i=1}^{p} \beta^i_{0_q}\gamma_i + \sum_{i=1}^{p} \sum_{\alpha \in L_i} x^\alpha \beta^i_\alpha \gamma_i$$

Equating coefficients of the intercept gives

$$\gamma^*_{0_q} = \gamma_0 + \sum_{i=1}^{p} \beta^i_{0_q}\gamma_i$$

Similarly for each $\alpha \in L^*$ we have

$$\gamma^*_\alpha = \sum_{i=1}^{p} \beta^i_\alpha \gamma_i$$

where β^i_α is zero if α is not in L_i. Finally for $\alpha \notin L^*$ we have

$$0 = \sum_{i=1}^{p} \beta^i_\alpha \gamma_i$$

The above can be intended as theoretical aliasing relationships among the parameters for the indirect/composite case and the direct case. For a generalization to a multivariate linear model with higher order terms for the Y-variables, further assumptions on the structure of the error terms are necessary.

Fitting above models to real data sets, e.g. by common estimating and model selection procedures, indicates that the obtained models have different monomials in the x variables. This prompts us to adopt new ways to compare the three approaches by algebraic statistics. Besides, in any approach it is of interest to know which models may be identified from the given design on the X, Y or $\hat{Y}$'s. One aim is to find out if information is lost or models are missed by considering any of the three possible input types. For the model selection procedure the knowledge of possible maximal models is extremely useful as an all-subset selection is usually unfeasible.

3 Computational polynomial algebra and designed experiments

A design or a set of observations can be seen as the zeros of polynomial equations. This simple observation is the entry key for algebraic geometry to the design and analysis of experiments.

In the case study we analyse in Section 4 we consider a full factorial design with central point in four factors, D_x. In total we have 17 points at which four different outputs $Y = (Y_1, \dots, Y_4)$ are measured. The observed or the estimated values of Y at D_x are the input points for the next stage of the modeling process from Y to Z (see Figure 1). To start with we ignore the output, concentrate on the input and consider the two dimensional analogue of D_x.

Example 1. The design $D_x = \{(\pm 1, \pm 1), (0, 0)\}$ is the solution set of

$$\begin{cases} p_1 = x_1^3 - x_1 = 0 \\ p_2 = x_2^3 - x_2 = 0 \\ p_3 = (x_1 - x_2)(x_1 + x_2) = 0. \end{cases}$$

From classical theory we know that only two saturated models, with the hierarchical (or order ideal) property, are identifiable by this design.[1] The order ideal property states that any lower order term of an interaction term in the model is in

[1] Here a model with support $[f_1(x), \dots, f_r(x)]$ is identified by the design with distinct points $d_1, \dots, d_s$ if the design/model matrix $[f_j(d_i)]_{i=1,\dots,s; j=1,\dots,r}$ is full rank. It is saturated if the rank is $r = s$.

the model as well. Peixoto [1990] among many authors advocates the desirability of the hierarchical property for a statistical model. In practice the final model will have less terms than there are design points and, very often, its terms are chosen from a larger set satisfying the order ideal property.

The two models for Example 1 are $\{1, x_1, x_2, x_1, x_1, x_2, x_1^2\}$ and $\{1, x_1, x_2, x_1, x_1, x_2, x_2^2\}$. The corresponding design/model matrices coincide and are

$$X = \left(\begin{array}{ccccc} 1 & x_1 & x_2 & x_1x_2 & x_1^1/x_2^2 \\ \hline 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{array}\right) \tag{6}$$

Clearly X is invertible and the two models are identifiable. In this example it is evident that there are no other saturated identifiable hierarchical models. These two models give the so-called statistical fan of D_x. Note that x_1^2 and x_2^2 cannot be part of the same models because they are aliased. Algebraically this follows from the fact that $p_3 = (x_1 - x_2)(x_1 + x_2) = 0$ is equivalent to $x_1^2 = x_2^2$. Statistically this also means that both effects are not distinguishable by data observed from this design. The notion of a statistical fan goes back to [Pistone et al., 2001, Def. 35].

Definition 1. *The statistical fan of a design is the set of hierarchical (support vectors for polynomial) models identified by the design with as many terms as distinct design points.*

Main properties of the statistical fan are:

- it is finite;
- each of its elements, called leaves, is formed by as many monomials as there are points in the design;
- each leaf is an order ideal and hence it contains 1, the constant term;
- the design/model matrix for each leaf is invertible.

In designs with a less regular structure than D_x, the statistical fan might not be as easy to determine as in Example 1. Many authors advocate the importance of hierarchical models (see e.g. Cox and Reid [2000], p.104). (Subsets of) fans provide lists of saturated hierarchical models each of which can be be input to a selection procedure for determination of a well-fitting parsimonious submodel. Furthermore, if we have different hierarchical models in the fan which differ only by a few terms this gives an indication of confounding within these terms.

However, the space of hierarchical models is often too large for an exhaustive search of saturated and identifiable models by the design, namely of the statistical fan. Still it is useful to have a large selection of saturated hierarchical models from which to select a submodel. A systematic method to investigate at least an "interesting" part of that space is provided by algebraic methods. The obtained subset of the statistical fan is called the algebraic fan of a design, or of the design ideal. For the analysis of the relation between these two fans see Maruri-Aguilar [2007] and Section 3.1 below. The technical tool at the basis of the computation is a term-ordering on the set of monomials. The technique from computational commutative algebra that allows this, also provides a theory that develops the observation about aliasing written up before Definition 1 for general designs. The key notion is the design ideal discussed below in Subsection 3.1.

3.1 Term ordering, matrices and fans

A good reference for this section is Cox et al. [1996]. A mathematical reference for the connection between matrices and term ordering is Robbiano [1985] and for the algebraic fan of a polynomial ideal see Mora and Robbiano [1988].

The set of polynomials in the variables $x_1, \ldots, x_n$ and with real coefficients is indicated with $\mathbb{R}[x_1, \ldots, x_n]$ and the set of monomials with T^n. More generally instead of real coefficients we might consider coefficients in an algebraic field and in the applications we often have rational coefficients. Polynomials are a linear combination of monomials and in turn, monomials are special polynomials formed by just one power product with coefficient equal to one. Note that a monomial $x^\alpha = x_1^{\alpha_1} \ldots x_n^{\alpha_n}$ is represented by its exponent vector $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{Z}^n_{\geq 0}$ with α_i non-negative integers for all $i = 1, \ldots, n$. Hence ordering monomials correspond to order non-negative integer vectors, more precisely a term order τ on T^n is a well-order relation on $\mathbb{Z}^n_{\geq 0}$. This can be extended to $\mathbb{Z}^n$ but we do not need to consider this generalisation here.

Definition 2. *A term order τ is a total order on T^n such that*

1. *1 is the smallest term (i.e. $1 <_\tau x^\alpha$ for all x^α in $T^n \setminus \{1\}$) and*
2. *τ is compatible with simplification of monomials (i.e. $x^\alpha <_\tau x^\beta$ then $x^{\alpha+\gamma} <_\tau x^{\beta+\gamma}$ for all $\gamma \in \mathbb{Z}^n_{\geq 0}$).*

Terms in a polynomial p can be ordered according to a τ and in particular the largest term in p is called the leading term of p.

Let A be a $n \times n$ matrix with integer entries whose rows are linearly independent and such that in every column the top non-zero entry is positive. Then A induces a term order on T^n by setting $x^\alpha < x^\beta$ if and only if $A\alpha < A\beta$ if and only if the first non-zero component of $A(\alpha - \beta)$ is positive. Furthermore, every term order can vice versa be described by an appropriate associated matrix A.

After this brief introduction to term orders we outline the link between polynomials and designs using our running example.

Example 2 (Example 1 contd.)**.** The design D_x is the zero set of the three polynomials $p_1 = p_2 = p_3 = 0$. However its points satisfy also the following equation

$$s_1(x_1^3 - x_1) + s_2(x_1^2 x_2 - x_2) + s_3(x_1 - x_2)(x_1 + x_2) = 0$$

for any polynomials s_1, s_2, s_3. These polynomials are elements of the polynomial ideal generated by p_1, p_2 and p_3 defined as

$$I(D_x) = \left\{ s_1(x_1^3 - x_1) + s_2(x_2^3 - x_2) + s_3(x_1 - x_2)(x_1 + x_2) : s_1, s_2, s_3 \in \mathbb{R}[x_1, x_2] \right\}$$

$I(D_x)$ is called the ideal generated by p_1, p_2 and p_3. It is also referred to as the design ideal of D_x or the vanishing ideal at D_x.

More generally the ideal generated by $p_1, \ldots, p_t \in \mathbb{R}[x_1, \ldots, x_n]$ is indicated with $I = \langle p_1, \ldots, p_t \rangle$ and the set of common zeros of the elements in I, equivalently the zero set of $p_1 = \ldots = p_t = 0$, is referred to as the variety of I. We work with a special case of polynomial ideals and varieties, namely varieties formed by a finite number of distinct points, also referred to as zero dimensional varieties. For a generalization to designs with replicated points see Notari and Riccomagno [2010]. There is no need to consider it here because all of our designs in Section 4 turn out to be without replications.

Observe that $x_1^2x_2 - x_2$ in Example 2 above has been obtained by substituting in p_2 the condition $x_1^2 = x_2^2$ obtained from p_3. This is an example of aliasing: x_1^2 and x_2^2 take the same values over D_x, leading to a rewriting rule within $I(D_x)$: $x_2^3 = x_1^2x_2$. Term orders allow us to determine and perform these rewritings systematically ensuring that the process ends univocally and returns a set of polynomials which generate the same ideal and are formed by monomials of lowest term with respect to the chosen term order. This is formalised by the notion of Gröbner bases, which are special types of generators of polynomial ideals whose introduction by Buchberger [1970] was at the core of the development of computational commutative algebra. They provide a general method by which many problems requiring solutions of polynomial system of equations in mathematics and engineering, and more recently statistics, can be solved by structurally simple algorithms.

Definition 3. *A set of polynomials G is a Gröbner basis (or G-basis) with respect to the term order τ if*

$$\langle Lt_\tau(f) : f \in \langle G \rangle \rangle = \langle Lt_\tau(p) : p \in G \rangle$$

where $Lt_\tau(f)$ is the highest term in f with respect to τ. A τ-Gröbner basis G is reduced if for all $g \in G$

1. *the coefficients of $LT(g)$ is equal to 1 and*

2. *no term of g lies in $\langle LT(G \setminus \{g\} \rangle$.*

The definition of G-bases states the equality between two monomial ideals: the ideal generated by the leading terms of the elements in G and the ideal generated by the leading terms of the polynomials in the ideal generated by G. Roughly spoken, a reduced G-basis is written as economically as possible. A well-known theorem states that a τ reduced G-basis is unique. Given τ, (reduced) Gröbner bases are computed via the Buchberger algorithm whose efficiency is largely improved when the underlying variety is a finite set of points as in our case (see e.g Faugere et al. [1993], Möller and Buchberger [1982]).

Example 3. For any term ordering τ for which x_2 is smaller than x_1, the three polynomials $g_1 = \underline{x_1^3} - x_1$, $g_2 = \underline{x_1^2 x_2} - x_2$, $g_3 = (x_1 - x_2)(x_1 + x_2) = \underline{x_1^2} - x_2^2$ form a Gröbner basis. The leading terms are underlined.

In this example there is only one other possible Gröbner basis of the ideal. It is obtained for term orders in which x_1 is smaller than x_2. By symmetry argument it is seen to be $(x_2^3 - x_2)$, $(x_2^2 x_1 - x_1)$, $(x_1 - x_2)(x_1 + x_2)$.

Definition 4. *The set of all reduced Gröbner bases of an ideal as the term order varies is called the algebraic fan of the ideal.*

A saturated hierarchical model identifiable by the design is determined from a τ-Gröbner basis G as those monomials in T^n which are not divisible by any of the leading terms of the elements of G. This is sometimes referred to as the Fundamental Theorem of Algebra. The obtained set is called a quotient basis, and in some literature it is known as an Est set (Pistone et al. [2001]). We indicate it as $\mathcal{O}_\tau(D)$. It belongs to the statistical fan of D. Hence its main properties are: it has as many terms as there are points in D, it is a (real) vector space basis of the space of interpolating (real valued) polynomial functions at D and the design/model matrix

$$X = [d^\alpha]_{d \in D, \alpha \in \mathcal{O}_\tau(D)}$$

for D and $\mathcal{O}_\tau(D)$ is invertible. Hence $\mathcal{O}_\tau(D)$ is one of the elements of the algebraic fan. As there is a one-to-one relationship between reduced G-bases and order ideals, the set we are most interested in $F_D = \{\mathcal{O}_\tau(D) : \tau\}$ is also called the algebraic fan of D.

Example 4. In Example 3 the leading terms are underlined. The set $\mathcal{O}_\tau(D_x)$ is $1, x_1, x_2, x_1^2, x_1 x_2$ corresponding to the design/model matrix in Example 6. The full algebraic fan of D_x is $\{\{1, x_1, x_2, x_1^2, x_1 x_2\}, \{1, x_1, x_2, x_2^2, x_1 x_2\}\}$.

Note that for D_x the algebraic and the statistical fans coincides. This is not usually the case and in general the algebraic fan is much smaller than the statistical fan (see Maruri-Aguilar [2007]). It follows that the algebraic fan is finite. The computations of Gröbner bases, order ideals, and algebraic fan can be performed using specialised software such as `gfan` (Jensen) and it can be computationally very demanding.

3.2 Algebraic analysis of the direct case

The case study in Section 4 involves a full factorial design with central point in the four factors k, l, d, f. By generalising Example 3 to four dimensions we deduce that its algebraic fan has four leaves obtained by permutation of the four factors. Each leaf has seventeen elements as there are seventeen distinct points in the design.

For any term order τ on T^4 for which f is lowest, the reduced Gröbner basis is

$$\left\{f^3 - f, d \cdot f^2 - d, d^2 - f^2, k \cdot f^2 - k, k^2 - f^2, l \cdot f^2 - l, l^2 - f^2\right\} \tag{7}$$

and the corresponding saturated model is

$$\mathcal{O}_\tau(D) = \left\{\begin{array}{r} 1, f, f^2, d, l, k, \\ df, lf, ld, kf, kd, kl, \\ ldf, kdf, klf, kld, \\ kldf \end{array}\right\}$$

It includes f^2 and all square free terms of total degree at most four. As for its two dimensional analogue, this is a special case where the algebraic fan equals the statistical fan, providing all four hierarchical models with 17 monomials and for which the design matrix is invertible. This statement follows by observing that

1. a monomial in $\mathcal{O}_\tau(D)$ cannot have degree three or more in any variable because the four factors have three levels each (the leading terms in (7) are of total degree less than four) and that

2. d^2, k^2 and l^2 are aliased with f^2, indeed the two evaluation vectors $[f^2(d)] = [l^2(d)]$ are equal. Hence as f^2 is in the model, d^2, k^2 and l^2 cannot be.

The intersection of the four models in the fans gives a hierarchical model with all 16 interactions up to order four.

3.3 Motivations for a numerical fan of a design

Usually the necessary computations to obtain the algebraic fan cannot be done by hand even for designs which exhibit regular geometric structure. A study for the class of Latin hypercube designs carried out in Bernstein et al. [2010] shows different situations that can occur when computing the algebraic fan. Moreover by Theorem 30 in Pistone et al. [2001] for a design whose points are chosen at random (with respect to any Lebesgue absolute continuous measure) the algebraic fan equals the statistical fan with probability one. These are examples where the algebraic fan is very large, albeit it can be much smaller than the statistical fan, for few points in many dimensions (Maruri-Aguilar [2007]). Furthermore for practical purposes it might not be desirable to compute the full algebraic or statistical fans. We argue this here with special reference to the real case driving our work.

In our two stage problem the design in the first stage has a nice regular structure and in Section 3.2 we computed easily its fan. However, the four "designs" D^*_y in the second stage treated in Section 4, look, although are not, random and have a fairly complex geometrical structure. Standard statistical techniques go only so far (see e.g. Rudak et al. [2012]) in their analysis and do not provide information on the aliasing structure imposed by the D^*_y on the space of polynomial models. This is where, we believe, the algebraic method adopted in this paper becomes worthwhile. The aliasing structure, described by a reduced G-basis, is clearly term-ordering dependent and clever application of Euclidean division of polynomials allows us to substitute terms in a model in order to include physically meaningful interactions or to exclude the simultaneous presence of some terms in the model.

In our application we encounter yet another problem. The complexity of D^*_y carries over to its ideal and to its fans which could have many leaves. Example 5 shows another reason why it might be desirable to consider only a subset of the fans by excluding numerically unstable leaves. Example 5 shows some of the issues we encounter and overcome by the approximated version of the design ideal and of its fan described in Subsection 3.5. A measure of stability of a system of linear equations $Ax = b$ with $A \in \mathbb{R}^{n\times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^n$ is the condition number $||A|| \cdot ||A^{-1}||$ of a matrix A [Allaire, 2009]. The condition number is at least 1 and if it is 1 or approximately 1, then the matrix A is said to be well conditioned. The matrix A is said to be ill conditioned if the condition number is large ($>> 1$). In case of an ill conditioned matrix A, the solution of the $Ax = b$ will be sensitive to errors in the matrix A or the right hand side b.

Example 5. Let D be the 2^2 full factorial design with levels ±1. The algebraic and statistical fans have only one leaf $\{1, x_1, x_2, x_1x_2\}$ and a generating set of the design ideal (the only reduced G-basis) is given by the two polynomials $x_1^2 - 1$ and $x_2^2 - 1$. Now suppose to substitute the point $(1, -1)$ with $(1, -1.001)$. A reduced G-basis is

$$\underline{x_1^2} - 1,\ \underline{x_2^3} - x_2 + 1.001x_2^2 - 1.001,\ \underline{x_1x_2} + x_2 - x_1 - 2001 + 2000\underline{x_2^2}$$

where underlined are the possible leading terms. The algebraic and statistical fans are formed by two leaves

$$\{1, x_1, x_2, x_1x_2\} \text{ and } \{1, x_1, x_2, x_2^2\}$$

The corresponding design/model matrices are

$$X_1 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1.001 & -1.001 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \text{ and } X_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1.001 & 1.002001 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix},$$

respectively. The condition number of X_1 is almost 1.000707180 and of X_2 is 3.737445584. The X_1 matrix is well conditioned and so are problems relying on it, e.g. stability of most commonly used algorithms in statistical analysis is ensured. But no statistician will be comfortable with the results of an analysis based on X_2.

A switch is required from symbolic, exact computations to numerical computations. Few key points summarise this section and lead us into Subsection 3.4:

1. the generating set of a design ideal embeds the design itself;

2. a (τ-reduced) G-basis of the design ideal also embeds a full identifiable model: the set of terms not divisible by its leading terms. The tail of each polynomial in a reduced G-basis is a linear combination of these terms;

3. starting for a generating set the FGLM algorithm computes the algebraic fan (Faugere et al. [1993]).

More poignantly we can state that a (reduced) G-basis gives a simultaneous and implicit description of a design and of its fan.

3.4 Numerical BM-algorithm

This section deals with the designs in the second stage of the analysis. Two cases can occur. We have a 17 point design D_y^{obs} of measured values or we have estimated designs D_y^{est} obtained by prediction from the first stage analysis. Theoretically both types of designs could include replicated points, but this does not occur in our application. The strategy we develop next applies to both types of designs which are characterised by the fact that the coordinates of their points are known up to a certain precision. We might think that there are measurement errors for D_y^{obs} or prediction errors in D_y^{est}.

We seek a set of polynomials which "almost vanish" at the design points, namely evaluated at the design points are close enough to zero. To do that, we use the numerical Buchberger-Möller (NBM) algorithm in Fassino [2010] and its implementation in CoCoA4 (CoCoATeam). The NBM algorithm is from the field of approximate computational algebraic geometry and is based on a least square approximation. It is a variation of a purely symbolic algorithm: the Buchberger-Möller algorithm (Möller and Buchberger [1982]) and its spirit is numerical.

The inputs to the NBM algorithm are a finite set of distinct points in n dimensions, say $\mathcal{D} \in \mathbb{R}^n$, a term-ordering τ and a precision vector $(\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$. The outputs are a set of polynomials $\mathcal{G}$ and a hierarchical set of monomials $\mathcal{O}$. The output includes also a flag stating whether the X matrix build from $\mathcal{O}$ and $\mathcal{D}$ is numerically stable in the sense of Example 5.

We recall from Fassino [2010] the basic definitions, see also references and discussion therein.

Definition 5. *Let* $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n$ *and* $\epsilon_M = \max\{\epsilon_i : i = 1, \ldots, n\}$.

1. *A point* $\bar{d} = (\bar{d}_1, \ldots, \bar{d}_n) \in \mathbb{R}^n$ *is an* ϵ*-(admissible) perturbation of* $d = (d_1, \ldots, d_n) \in \mathbb{R}^n$ *if* $|d_i - \bar{d}_i| < \epsilon_i$ *for* $i = 1, \ldots, n$. *Let* $\mathcal{D}^\epsilon$ *be the set of all* ϵ*-perturbed points of* $\mathcal{D}$.

2. *Without loss of generality assume* $\mathcal{D} \subset [-1, 1]^n$. *A polynomial* g, *with coefficient vector* c, *is almost vanishing at* $\mathcal{D}$ *if*

$$\frac{||X||_2}{||c||_2} < O(\epsilon_M)$$

where $X = [g(d)]_{d \in \mathcal{D}}$ is the evaluation vector of g at $\mathcal{D}$ and $||a||_2$ is the Euclidean norm of the vector $a \in \mathbb{R}^d$.

3. *The set of polynomials almost vanishing at $\mathcal{D}$ is called the approximate ideal of tolerance ϵ.*

We can assume $\mathcal{D} \subset [-1, 1]^n$ because the support vector of an identifiable polynomial model is invariant by scaling and translation of design points. This holds true also for the numerical fans of Section 3.5. The main properties of $\mathcal{G}$ and $\mathcal{O}$ are listed in Theorems 4.1 and 5.1 in Fassino [2010], respectively. Here we just state them briefly: $\mathcal{G} = \{g\}$ is finite; $\mathcal{G}$ is the approximate ideal of tolerance ϵ of $\mathcal{D}$ and of $\mathcal{D}^\epsilon$; $\mathcal{G}$ can be viewed as an approximation of a Gröbner basis of the polynomial ideal of a "more regular" set of points close to $\mathcal{D}$; $\mathcal{G}$ likely is not a proper (different from $\mathbb{R}[x_1, \ldots, x_n]$ polynomial ideal. If the flag value is true, then $\mathcal{O}$ is stable neglecting errors of order $O(\epsilon_M)$; the tail of the first polynomial in $\mathcal{G}$ is formed by the smallest monomials with respect to τ for which the design matrix is full-rank for $\mathcal{O}$ and for every perturbed design in $\mathcal{D}^\epsilon$. This is particularly interesting for us because it can be interpreted as a high-dimensional surface of a shape which is as simple as possible in τ and which approximates our original designs $\mathcal{D}$ in a least square sense. Indeed the NBM algorithm returns an implicit representation of $\mathcal{D}$ depending on the term ordering τ in input.

3.5 Numerical fan

An algorithm which computes the numerical fan, that is repeats the NBM for every input term-ordering, has not been implemented yet. From the fact that the underlying variety is zero-dimensional, it follows that the fan is finite. Indeed the key technical step in the NBM algorithm and the FGLM algorithm, is to start building the (almost) vanishing polynomial in $\mathcal{G}$ by adding the lowest possible monomials in τ:

1. start with $M := \{1\}$,

2. consider the smallest available monomials in τ, say x^α,

3. solve the least square problem for $\mathcal{D}$, M and x^α,

4. check if the obtained polynomial is zero for all $d \in \mathcal{D}$ (in the exact case) or small enough in some norm, e.g. Euclidean in the NBM algorithm,

5. if yes, add the obtained polynomial to $\mathcal{G}$,

6. if not, add x^α to M and repeat from 2.

For a numerical version of the fan of a design, in Step 2 one needs to consider each possible x^α that preserves the order ideal structure. Clearly this procedure will return the statistical fan. In high dimension this is no trivial task. For a special class of polynomial interpolators this has been attempted in Bates et al. [2003]. In Section 4 we choose to approximate the numerical fan in two ways.

1. Compute a subset by running the NBM algorithm for some significant term orderings. For a similar procedure see Holliday et al. [1999].

2. Compute the exact algebraic fan of the first polynomials in an approximated vanishing ideal, $\mathcal{G}$, and the $\mathcal{O}$ set of each leaf in this fan. The intersection of these $\mathcal{O}$ sets satisfies the hierarchical property and forms the support of a polynomial models identifiable by $\mathcal{D}$ for every term ordering. It is a robust, core, set of terms to include in the input of standard methods for building regression models.

We point out that the polynomials returned by the NBM algorithm do not generate usually a proper polynomial ideal because they might have non common zeros, unless the tolerance parameters are set to zero. However their role for our application, both when giving interpretation in terms of aliasing and when discussing identifiable models, is the same as that of generating set of the exact design ideal. Furthermore, as already mentioned, $\mathcal{G}$ can be viewed as an approximation of a G-basis of the polynomial ideal of a set of points $\hat{\mathcal{D}}$ close to $\mathcal{D}$ and with a less complex geometric structure.

4 Application: Thermal Spraying Process

In this section we apply the theory of Section 3 to the application that motivated it: the analysis of an experiment from a thermal spraying process where a full factorial design with central point with four controllable parameters ($\mathcal{D}_x$) is run. During the experiments four particle properties, $\mathcal{D}_y^{obs}$ are measured online. Afterwards coating

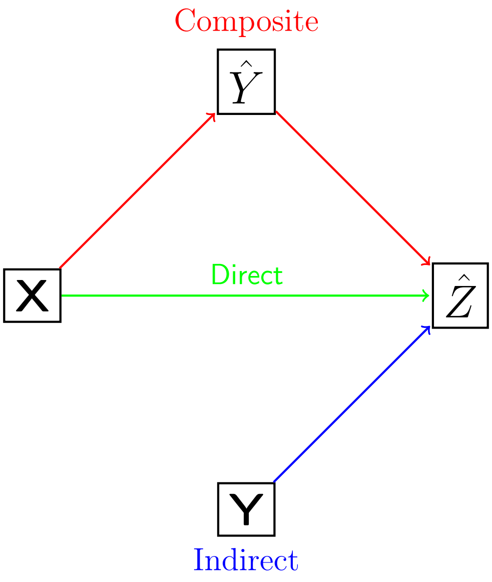

Figure 2: Prediction strategies

properties, $\mathcal{D}_z$ are also measured. The aim of the whole project is to determine X-to-Y models from measured data in order to define good Y-to-Z models for the prediction of the difficult-to-measure Z and hard-to-control Y. The different prediction strategies are summarized in Figure 2.

In Section 4.1 we start the statistical analysis of the thermal spraying data with direct models X-to-Z based on the hierarchical model identified in Section 3.2. In order to build Y-to-Z models, three X-to-Y models are discussed in Section 4.2, which evaluated at the points in $\mathcal{D}_x$ define three "estimated designs" in Y. We compare them with $\mathcal{D}_y^{obs}$ and among themselves by comparing "almost vanishing polynomials" of increasing complexity built with the same criterion, here given by the choice of the same term order. This is done in Section 4.3 and then we compare (part of) their fans in Section 4.4. The computations were done in R 2.15.1 (see R Core Team [2012]), CoCoA 4.7.5 (see CoCoATeam) and `gfan 0.5` (see Jensen).

4.1 Statistical analysis of the direct case

Next we build linear models by forward backward search for each coating property (hardness, thickness, porosity and deposition rate) where the maximal model follows from Section 3.2. We start with a constant linear model and then perform a forward backward selection based on the AIC criterion with the hierarchical model as maximal model. The calculated models can be found in Appendix 6.1. Table 1 contains the R^2 with corresponding adjusted R^2 for each coating property. The best fit is derived for porosity with $R^2 = 0.87$, whereas deposition rate is worst predicted with $R^2 = 0.28$. This can be also observed in Figure 3 which contains the fitted versus predicted values. Here, the red line indicates a perfect fit and the green lines stay for $\pm 10\%$. For hardness all points lie within the $\pm 10\%$ area (or are very close) and for the remaining coating properties some points are outside of the $\pm 10\%$ region.

	R^2	Adjusted R^2
Porosity	0.87	0.82
Hardness	0.62	0.49
Thickness	0.77	0.66
Deposition rate	0.28	0.20

Table 1: R^2 squared and adjusted R^2 for each coating property

4.2 Possible designs for the second stage

Fitting models from Y to Z requires input data on the Y-stage. This can either be the observed Y-values or predicted values from a X-to-Y model, where we compare three different models constructed as follows.

1. *FB*: We select a model by means of forward backward selection based on the AIC criterion (see Akaike [1973]) where the minimal model contains only the intercept and the maximal model is the usual saturated model for $\mathcal{D}_x$ computed also in Section 3.2. The predicted values at $\mathcal{D}_x$ are collected in the estimated design $\mathcal{D}_y^{FB}$.
2. *best*: We perform an all subset selection where the maximal model contains all main effects and interactions up to order 4 to generate $\mathcal{D}_y^{best}$.
3. *simple*: This strategy builds a model that consists only of main effects and the predicted values at $\mathcal{D}_x$ are denoted by $\mathcal{D}_y^{simple}$.

The fitted regression models are stated in Appendix 6.2. Table 2 shows the adjusted and the non-adjusted R^2 values. Not surprisingly the *best* model performs best in

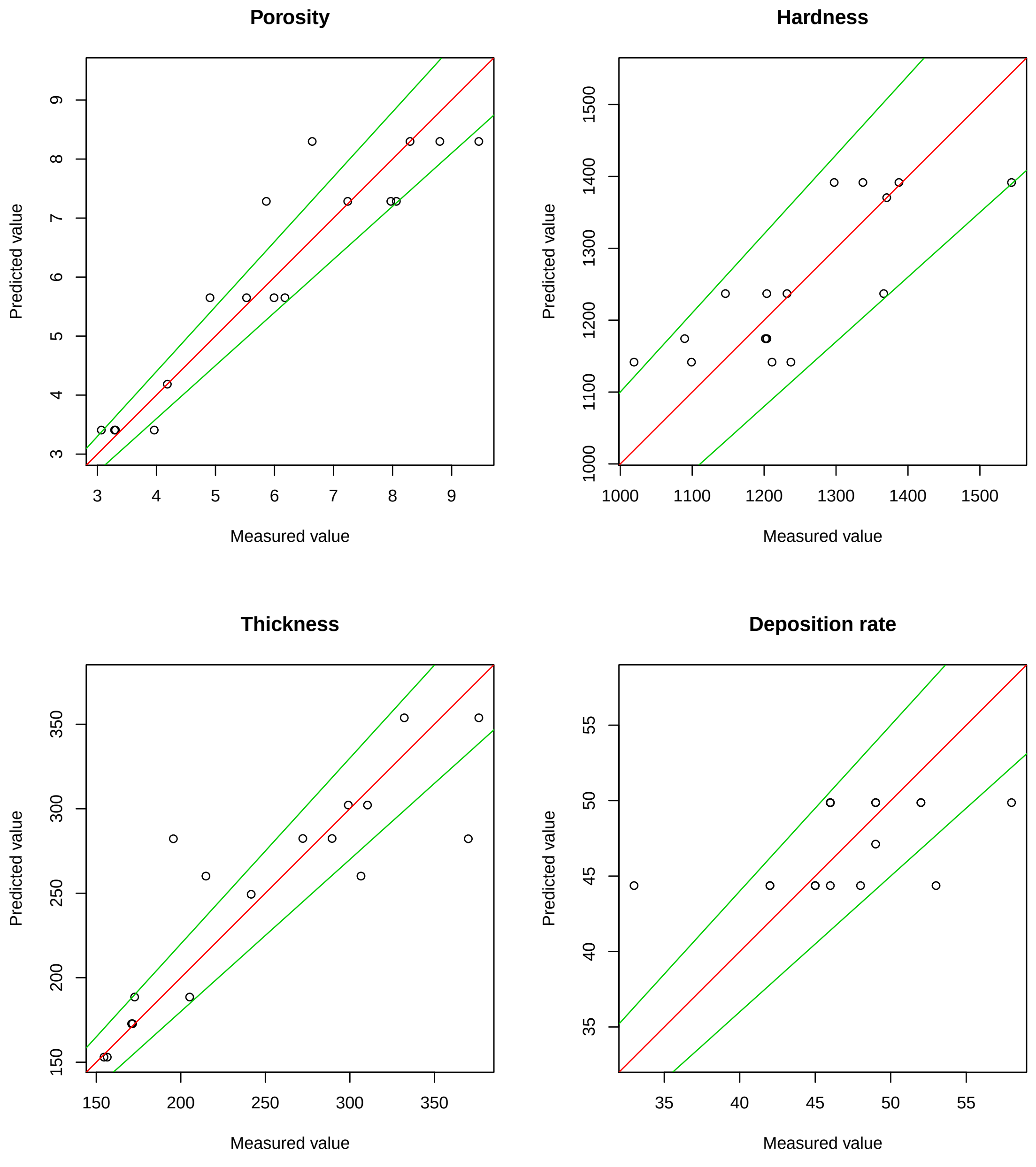

Figure 3: Fitted vs. measured values for direct case

terms of the highest R^2-values and can therefore be considered as good approximation of the observed $\mathcal{D}_y^{obs}$. However, with (adjusted) R^2 values above 0.67 even the *simple* model seems to work quite reasonable, with exception of flame width (adjusted $R^2 = 0.56$).

The generating process of the $\mathcal{D}_y^*$ designs destroys the symmetries of the $\mathcal{D}_x$ design. In particular, both $\mathcal{D}_y^{obs}$ and $\mathcal{D}_y^*$, where $* \in \{best, FB, simple\}$, have a fairly intricate geometry, although $\mathcal{D}_x$ is a very regular design.See the scatter plots in

		FB	*best*	*simple*
Temperature	R^2	0.88	0.98	0.79
	Adjusted R^2	0.81	0.92	0.72
Velocity	R^2	0.94	0.99	0.92
	Adjusted R^2	0.91	0.97	0.89
Flame Width	R^2	0.76	0.97	0.67
	Adjusted R^2	0.66	0.86	0.56
Flame Intensity	R^2	0.85	0.99	0.75
	Adjusted R^2	0.78	0.96	0.67

Table 2: R^2 and adjusted R^2 for the three different modeling strategies

Figure 4. Their irregularity comes from different sources, all traceable back to the measurement errors of the observed Y values, an inherent complexity of the generating process, and modeling approximation. The R^2 values in Table 2 are a measure of this, but we would like to investigate and compare further the geometry of the four $\mathcal{D}_y^*$. Observe furthermore that the designs $\mathcal{D}_y^*$ and $\mathcal{D}_y^{obs}$ all have 17 distinct points. In general this is not necessarily the case.

4.3 Approximated vanishing ideals for the Y-designs

A rough measure of the difference between the $\mathcal{D}_y^*$ designs and $\mathcal{D}_y^{obs}$ is given by the cumulated distances. Recall that each $d^* \in \mathcal{D}_y^*$ is the predicted value of a $d \in \mathcal{D}_x$ with respect to a certain model and that d^{obs} is the observed Y-value at a specific $d \in \mathcal{D}_x$ input. Hence we can define

$$CSS = \sum_{d \in \mathcal{D}_x} ||d^{obs} - d^*||_2^2$$

unambiguously, where $||.||_2$ is the Euclidean distance for vectors. These are given in Table 3. As $\mathcal{D}_y^{best}$ results from the best subset selection, its cumulated distances are lowest as expected. The polynomials in each exact design ideals $I(\mathcal{D}_y^*)$ vanish at the points of the corresponding $\mathcal{D}_y^*$, by definition. Even when the designs are close in a Euclidean distance, e.g. are an ϵ-perturbation of $\mathcal{D}_y^{obs}$, their ideals could be very different. This is an implicit analogue of the well-known problem of overfitting. The

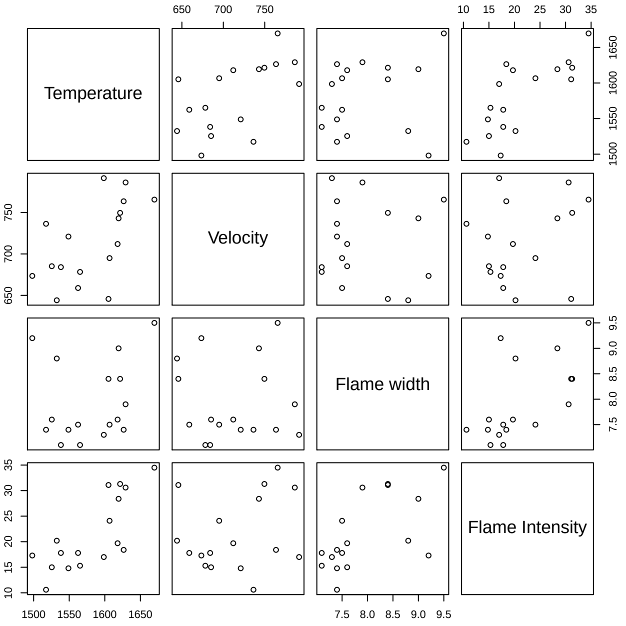

Figure 4: Illustration of design $\mathcal{D}_y$

four design ideals have generating sets with many polynomials and their fans are rather big.

More informative for us is to consider approximated versions of the design ideals and compare them. We fix the term order $degrevlex(t, v, w, i)$ with i smallest, then w, v, t in order. This choice implies that the first polynomials in the output of the NBM algorithm involves as far as possible main effect terms and lower order interaction terms with a preference of i over, say, t.

In order to compute the approximated ideals of the Y-designs $\mathcal{D}_y^*$, the NBM algorithm requires to specify a tolerance vector $\epsilon = (\epsilon_1, \dots, \epsilon_4)$ where ϵ_1 refers to temperature, ϵ_2 to velocity, ϵ_3 to flame width and ϵ_4 to flame intensity. The manufacturer of the measurement system recommends an uncertainty of 2% for temperature and velocity measurements whereas the uncertainty for intensity and width are not

	Distance between $\mathcal{D}_y^{obs}$ and $\mathcal{D}_y^{best}$	Distance between $\mathcal{D}_y^{obs}$ and $\mathcal{D}_y^{FB}$	Distance between $\mathcal{D}_y^{obs}$ and $\mathcal{D}_y^{simple}$
Cumulated distance	1204.20	6951.60	11189.45

Table 3: Cumulated distances $\sum_{d \in \mathcal{D}_x} ||d^{obs} - d^*||_2^2$

known. Therefore, we choose

$$\epsilon = (25, 14, 0.5, 0.3).$$

It turns out that the first polynomial returned by the NBM algorithm contains only main effect terms for all four designs. These are

$$\begin{aligned}
obs : t &- \frac{164}{459}v + 21w - \frac{272}{45}i - 1364,\\
FB : w &- \frac{1168}{16839}i - \frac{685}{106},\\
best : t &- \frac{597}{1567}v + \frac{728}{45}w - \frac{208}{37}i - 1318,\\
simple : w &- \frac{251}{2983}i - \frac{295}{48}.
\end{aligned}$$

The resulting order ideals $\mathcal{O}$ given in Table 4 are not stable in the sense of Fassino [2010] and of Section 3.3.

Case	Order ideal
obs	1, i, w, v, i^2, vi, v^2, i^3, vi^2, i^4, i^5
best	1, i, w, v, i^2, vi, v^2, i^3, i^4, i^5, i^6, i^7
FB	1, i, v, i^2, vi, v^2, i^3, vi^2, v^{2}i, v^3, i^4, i^5, i^6
simple	1, i, v, i^2, vi, v^2, i^3, vi^2, i^4, i^5, i^6, i^7, i^8, i^9

Table 4: Order ideals for $\epsilon = (25, 14, 0.5, 0.3)$

Experiments show that the uncertainty of 2% is strongly overestimated and therefore we decide together with engineers on a different choice of ϵ which corresponds to a realistic uncertainty on the one hand (for t, v) and ensures a stable order ideal on the other hand (for w, i). We finally choose

$$\epsilon = (5, 2, 0.01, 0.01).$$

The number of polynomials in each approximated vanishing set is given in Table 5

	obs	*FB*	*best*	*simple*
Number of Polynomials	17	15	15	12

Table 5: Number of Polynomials in each Ideal

and Table 6 gives the number of different monomials between the supports of the first four polynomials of each approximated vanishing set. In particular, none of the four designs is not an ϵ-perturbation of any among the other three designs.

	obs vs *FB*	*obs* vs *best*	*obs* vs *simple*
Polynomial 1	0	0	3
Polynomial 2	0	1	3
Polynomial 3	1	4	6
Polynomial 4	3	3	4

Table 6: Number of different monomials in the support of the first four polynomials in each approximated vanishing ideal

Consider the unitary version of the approximating sets obtained by multiplying each polynomial with the inverse of the Euclidean norm of its coefficient vector (see e.g. (Heldt et al. [2009])). Due to the fact that the four approximating sets are computed with respect to the same term order, it is reasonable to compare the first polynomials in each approximating set separately from the second polynomials and so on. The Euclidean norm of the difference between the coefficient vectors of the first four polynomials in the approximated vanishing set of $\mathcal{D}_y^{obs}$ and $\mathcal{D}_y^{FB}$ are given in Table 7.

	obs-*FB*	*obs*- *best*	*obs*-*simple*
Polynomial 1	0.16	0.12	0.33
Polynomial 2	0.69	1.97	1.99
Polynomial 3	0.64	1.85	0.77
Polynomial 4	0.05	1.99	0.13

Table 7: Norm of the difference of standardized coefficient vectors of the polynomials in each approximated vanishing set

The polynomials of the almost vanishing sets for the three $\mathcal{D}_y^*$ have to almost vanish, in the sense of Fassino [2010] and Section 3.4, at the observed values $\mathcal{D}_y^{obs}$ if $\mathcal{D}_y^*$ is a good approximation of $\mathcal{D}_y^{obs}$. We have already observed that the four designs are not an ϵ-perturbation of each other for $\epsilon = (5, 2, 0.01, 0.01)$ and that the R^2 values in Table 2 are a measure of this. In order to check this further and also in order to check whether the first polynomials are sufficiently informative to compute the fans, we substitute $\mathcal{D}_y^{obs}$ in the first and second polynomials of each almost vanishing set. The resulting values, which we call **implicit residuals**, should be almost zero. Figures 5 and 6 show the implicit residuals for polynomial 1 and 2 of each ideal. The worst approximations are for the exact vanishing ideal of $\mathcal{D}_y^{simple}$. Figures 7 and 8 show the implicit residuals for polynomial 1 and 2 where the corresponding D_y^* is plugged in. Indeed, the residuals are very small as they have to almost vanish. We can observe that the absolute maximal value of the implicit residuals is lower than $1e^{-3}$, although the residuals for the case *obs* are largest. Furthermore, the residuals lie either over or under the x-axis. The implicit residuals for polynomials 2 in Figure 8 scatter around zero, as expected involving more non-linear terms than the polynomials 1. Note that they are largest for the situations *best* and *simple*.

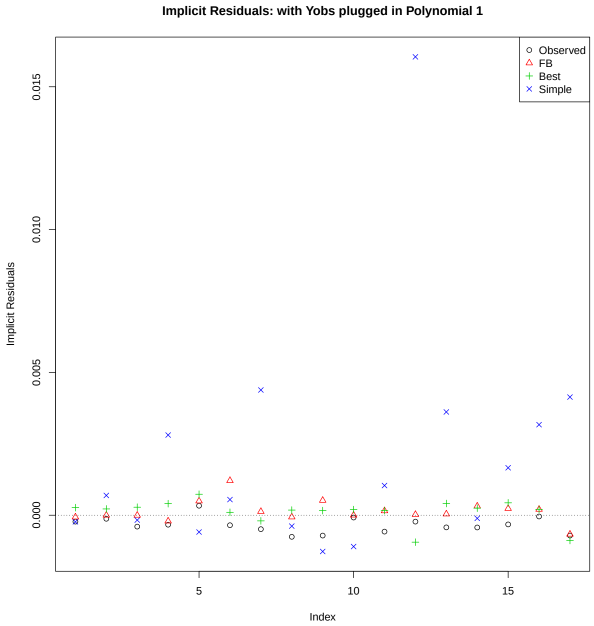

Figure 5: Implicit Residuals for Polynomial 1

4.4 Computation of the Algebraic Fan

In Section 4.3 we have considered one possible $\mathcal{O}$ set for each $\mathcal{D}_y^*$ and for $\mathcal{D}_y^{obs}$; that is, one set of monomial terms from which to start a forward backward model search

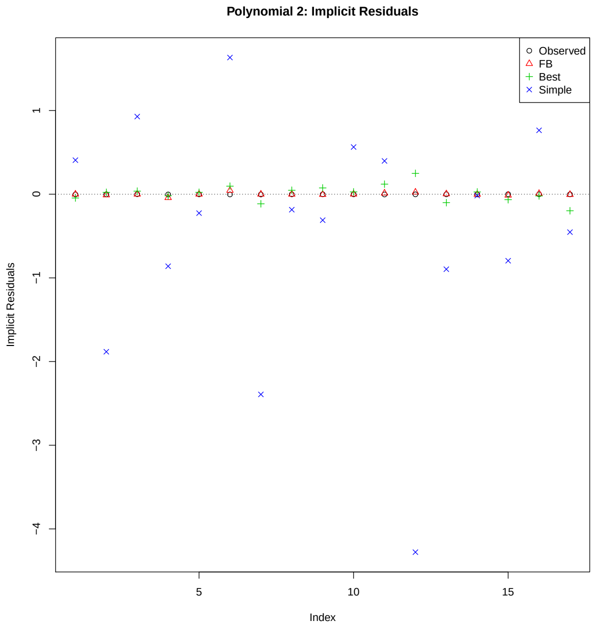

Figure 6: Implicit Residuals for Polynomial 2

depending on a given term ordering. Here, we consider a larger set of $\mathcal{O}$ sets by varying the term ordering. We adopt two intrinsically different strategies.

Strategy A

In strategy A we use `gfan 0.5` in order to derive the algebraic fan of the approximated ideals of $\mathcal{D}_y^*$ and $\mathcal{D}_y^{obs}$. If the variety of a polynomial ideal is empty then the algebraic fan contains only the constant term 1. Therefore, exact computations of the algebraic fan of the approximated ideals of $\mathcal{D}_y^*$ and of $\mathcal{D}_y^{obs}$ fail due to the fact that the polynomials in the approximated ideals have no common zero. So, for each Y-design we consider only a subset of the approximated ideal and compute its algebraic fan. We proceed as follows:

1. Take a subset S of the approximated ideal.
2. Use `gfan 0.5` to compute the algebraic fan corresponding to S.
3. Derive the leading terms of each polynomial in each leaf of the algebraic fan.
4. Consider, L, the union of these leading terms.
5. Compute $\mathcal{O}$ the set of monomials not divisible by any element in L.

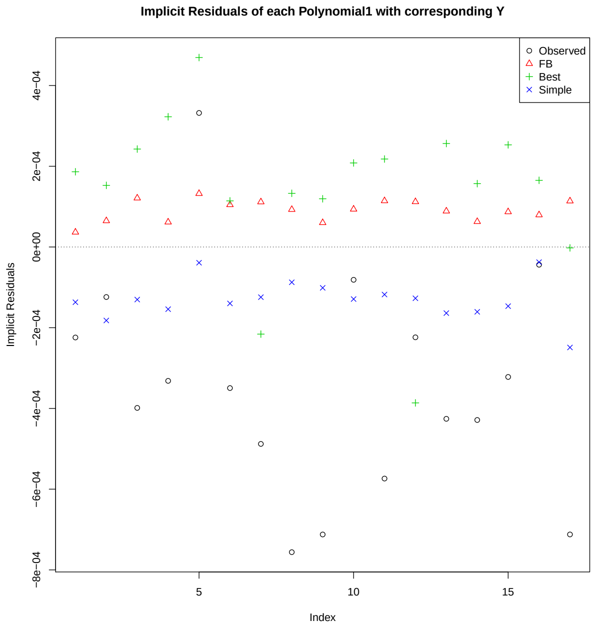

Figure 7: Implicit Residuals for Polynomial 1 where corresponding $\mathcal{D}_y^*$ is plugged in

Equivalently, for each leaf we could consider the set of monomials not divisible by the leading terms in the leaf and take the intersection over the leaves. The final set $\mathcal{O}$ contains one (the intercept). Based on the analysis in Section 4.3 we take S to be the first polynomial for each approximated ideals, namely they are:

best:

$$f^1_{best}(t,v,i,w) = t^2 - 207/199tv + 44/153v^2 + 458/27tw - 979/43vw - 1221w^2 - 97/9ti + 253/45vi + 1217/4wi + 589/67i^2 - 2345t + 2593/2v + 2546w + 10322i + 1284817$$

FB:

$$f^1_{FB}(t,v,i,w) = t^2 - 686/1135tv + 383/2280v^2 + 219/4tw - 383/26vw + 2263/2w^2 - 1017/67ti + 724/157vi - 1211/4wi + 58i^2 - 2835t + 731v - 86942w + 20501i + 2100764$$

obs:

$$f^1_{obs}(t,v,i,w) = t^2 - 865/707tv + 1191/3223v^2 - 733/17tw - 433/37vw - 2009w^2 - 403/175ti + 232/47vi + 803wi - 701/15i^2 - 1934t + 1404v + 90095w - 7869/2i + 738463$$

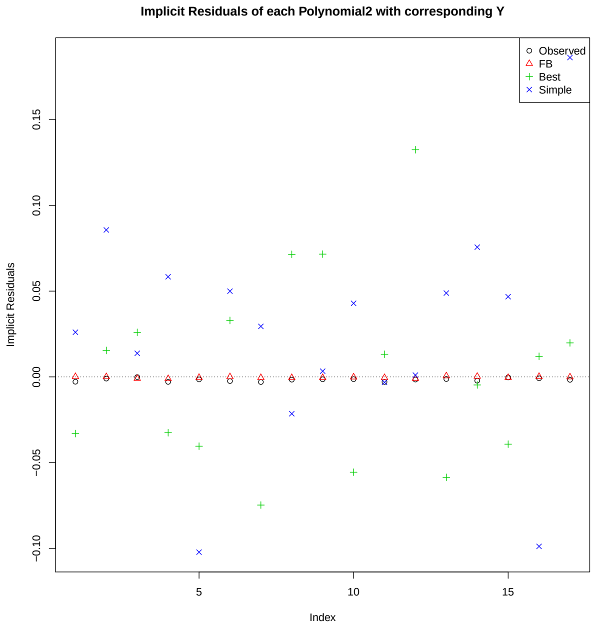

Figure 8: Implicit Residuals for Polynomial 2 where corresponding $\mathcal{D}_y^*$ is plugged in

simple:

$f^1_{simple}(t, v, i, w) = tw + 1277/2364vw + 621/4w^2 - 779/13225ti - 61/972vi - 419/12wi + 299/183i^2 - 341/51t - 292/99v - 7375/2w + 691/2i + 17291$

The L sets in point 4. of the algorithm above are listed in Table 8. Clearly, the set of monomials that are not divisible by the leading terms consists of all main effects and square free interactions up to order four for the situations *FB*, *best* and *obs*. This is different for the situation *simple*. Here, the set of monomials consists terms not involving the interactions tw,vw, ti and vi. Because t and v appear only in form of an interaction (tw,vw,ti,vi), it is possible to take every possible power of t and v which is not a desirable result the situation *simple*.

Thus the $\mathcal{O}$ sets for *obs*, *best* and *FB* are equal and given by intercept, main terms and all 11 square free interactions, while the $\mathcal{O}$ set for *simple* is given by intercept, main terms and the two way interactions tv, tw. This suggests a different role in statistical analysis and interpretation for interactions common to all situations from the other interactions. In particular, the smaller model is identifiable for all considered designs and term orderings. In this sense it is a robust, core, set of terms to consider when searching for a good model for prediction as well as fitting.

FB	t^2,v^2,w^2,i^2
best	t^2,v^2,w^2,i^2
obs	t^2,v^2,w^2,i^2
simple	tw,vw,w^2,i^2,ti,vi

Table 8: Union of leading terms based on strategy A

Strategy B

In strategy B the StableBBasisNBM5 function of CoCoa4 (CoCoATeam), which implements the numerical Buchberger-Möller algorithm, is used to compute the approximate ideals and the corresponding $\mathcal{O}$ set. In order to get the algebraic fan we have to repeat these calculations for every possible term ordering. This is not implemented in CoCoa4 or elsewhere, yet. Therefore, we compute the approximated vanishing ideals together with the corresponding $\mathcal{O}$ sets for three standard term orderings, namely lexicographical, degree lexicographical and reverse degree lexicographical ordering. These are quite extreme term orderings with respect to the monomials to be included in the leaves ($\mathcal{O}$ sets): lexicographic orderings tend to include first all powers of the smallest variable, while degree compatible term orderings favors the inclusion of the first suitable monomials with lowest total degree (sum of exponents).

We also permute the order of the main factors (see Holliday et al. [1999]). In this way, to each design $\mathcal{D}_y^{obs}$ and $\mathcal{D}_y^*$ with $* \in \{best, FB, simple\}$ there is associated a (subset of its) fan, $\mathcal{F}^{obs}$ or $\mathcal{F}^*$. Each subfan has 72 leaves each of which is labelled by the term ordering with respect to which it has been computed. Figure 9 gives a comparison of the leaves within each subfan by displaying the number of the 20 most frequent monomials in $\mathcal{F}^*$, $* \in \{obs, best, FB, simple\}$. The 20 most frequent terms for $\mathcal{D}_Y^{best}$ and $\mathcal{D}_Y^{obs}$ coincide. There are four differences between the situation *simple* and *obs*, namely t^4, w^4, ti and tv^2. Two differences can be observed between the situation *FB* and *obs* which are v^4 and tv^2. Furthermore, we compare the i-th leaf from situation *obs* with the i-th leaf for the remaining three situations, for $n = 1, \ldots, 72$. The i-th leaf was derived by the same term ordering for every situation. Thus for a good approximation of $\mathcal{D}_y^{obs}$ we should get nearly the same leaf

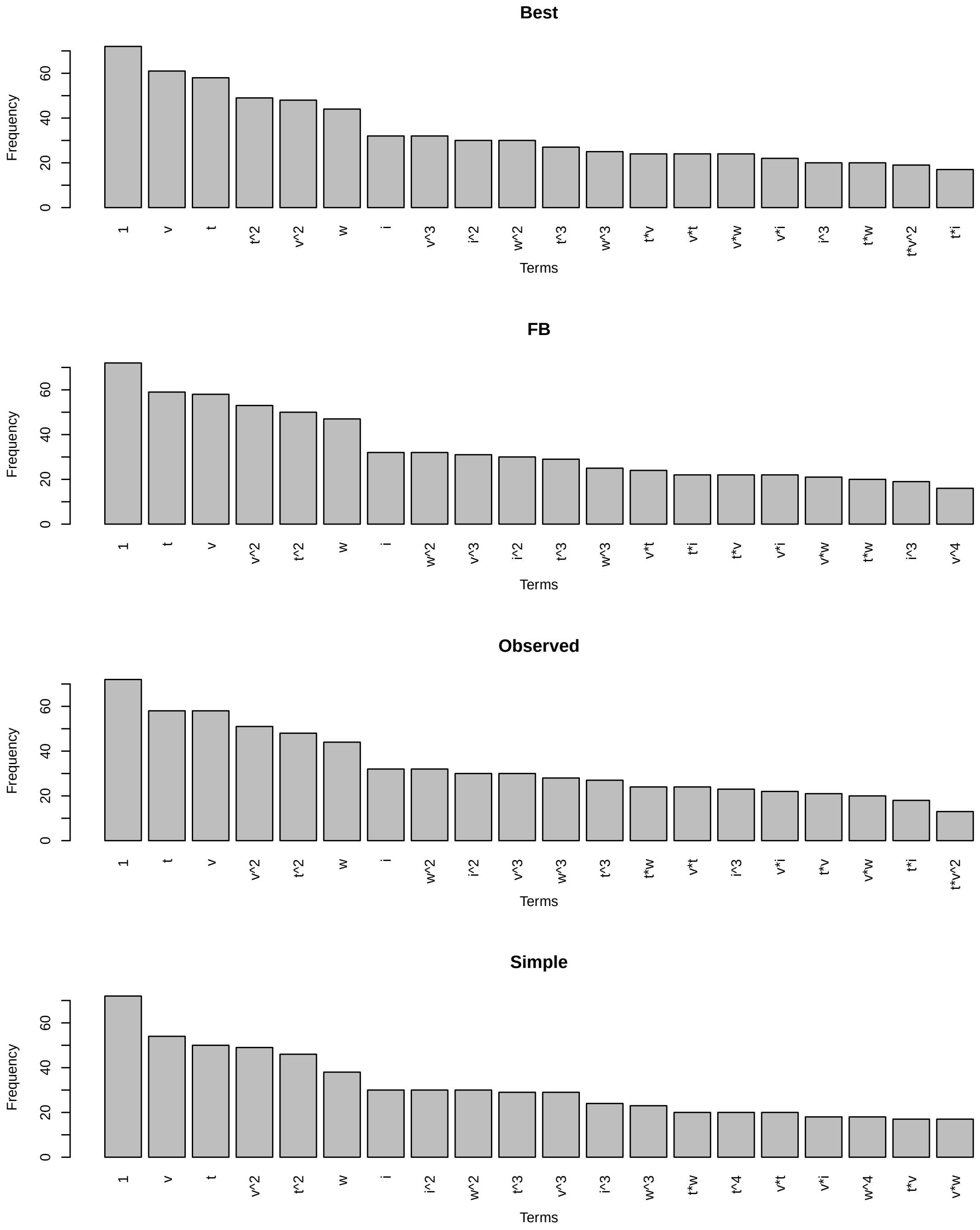

Figure 9: Number of most frequent terms for each situation

in all cases. (Notice that different designs can have the same fan.) Next we count the number of terms for the i-th leaf of situation *best*, *FB* and *simple* that coincide with the terms of the n-th leaf of situation *obs*. By this way, we get information about the similarity of the leafs, $n = 1, \ldots, 72$. Figure 10 presents a boxplot of the number of differences. As expected occur the most agreements for the situation *best*. The median (15) is the same for situation *FB* and *simple*, whereas the box is higher for the situation *FB*.

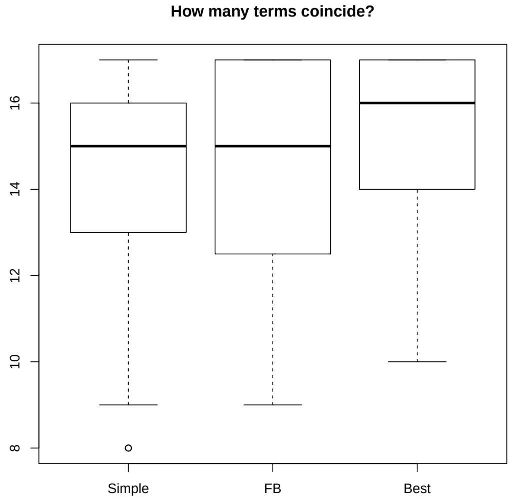

Figure 10: Leaf to leaf comparison (*best*, *FB* , *simple* vs. *obs*). How many terms coincide?

4.5 Composite vs Direct Approach

Finally we compare the goodness of fit for the models resulting from the composite strategy and direct strategy as these are the two models resulting in a prediction of Z based on X. We take each leaf (saturated model) as scope for a forward and backward selection. The model selection is based on the AIC criterion. We obtain a selected model for each leaf for each Y-design and take the model with minimal AIC value. Thus, our procedure is as follows:

1. For each $\mathcal{D}_y^*$, $* \in \{simple, best, fb, obs\}$, we perform a model selection as follows:

- We conduct a forward backward selection based on AIC with
  - ⇒ Minimal model: only intercept
  - ⇒ Maximal model: the saturated model from each leaf
- By this way, we end up with a selected model for each leaf

2. We choose the model with minimal AIC among the selected models for each leaf from Step 1.

		Porosity	Hardness	Thickness	Deposition Rate
Adjusted R^2	*best*	0.67	0.44	0.66	0.24
	FB	0.65	0.31	0.80	0.08
	obs	0.62	0.49	0.65	0.14
	simple	0.72	0.29	0.75	0.19
	direct	0.82	0.49	0.66	0.20

Table 9: Adjusted R^2 values for composite and direct models

Table 9 displays the adjusted R^2 values for the selected models for each Y-design and for the direct case from Section 4.1. In Figure 11 the observed values for every coating property are plotted against the predicted values for each model considered. The red lines indicate a perfect fit and the green lines stand for an uncertainty band $^{+}_{-}10\%$, as in Figure 3. All models lead to low values of the adjusted R^2 for the deposition rate. Here, we might have a problem with the quality of measuring deposition rate such that we discard these results from our comparison of models. The highest adjusted R^2 values for porosity and hardness are achieved by the direct case and the composite strategy leads to highest R^2 values for thickness (situation *FB*). Thereby, the direct approach is not always superior to two-way strategies and vice versa. However, all approaches lead to comparable R^2 values. Depending on the aim at hand we might go along with the respective strategy: prediction of coating properties based on particles in flight especially if day-effect are suspected, prediction of coating properties based on process parameters if no in flight properties are available. As a next step hybrid models might be considered where both process parameters and in flight properties are taken into account in order to get further inside into the process.

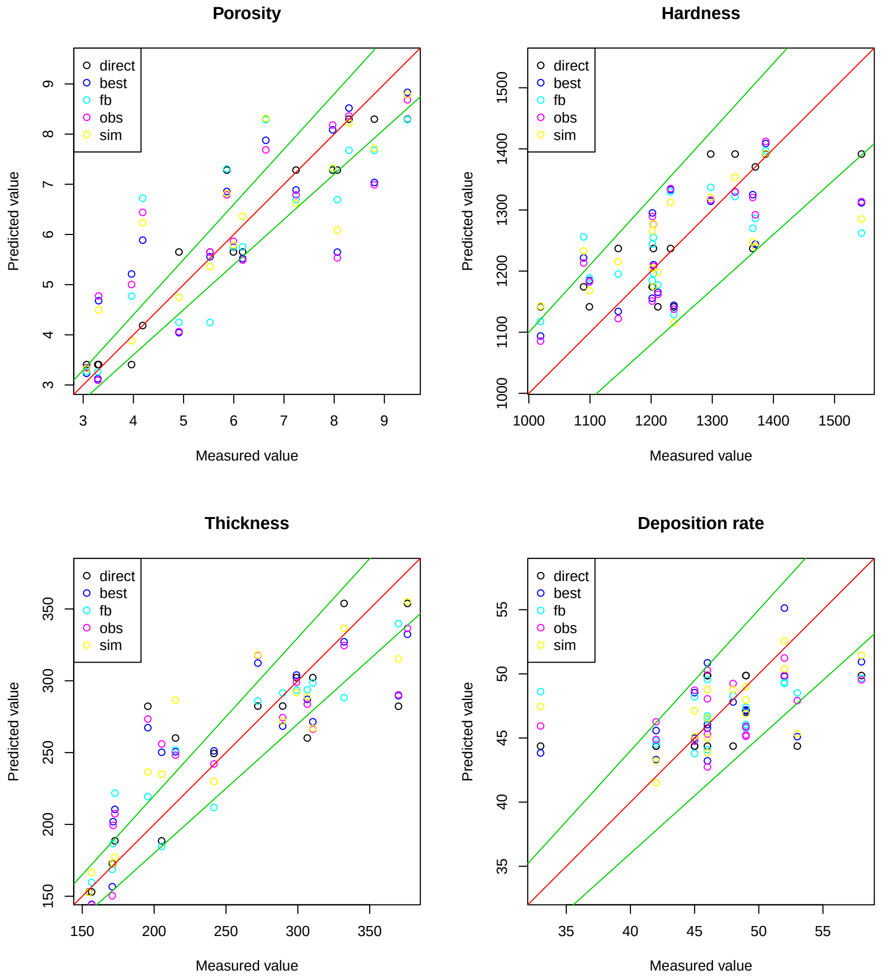

Figure 11: Fitted vs. measured values for direct and all indirect cases

5 Conclusion

In this paper, we treat the question of identifiable models in a two-stage process. The main focus lies on models for the relationship between the intermediate variables and the final output. These models have to be based on data or predictions which result from observations on standards design in the initial input space. We adapt tools from algebraic statistics to this situation. The novelty is the use of approximating ideals in order to deal with the instability in the observed or predicted designs. We employ an algorithm from Fassino [2010], whose use in statistics is completely new. Our work is motivated throughout by a thermal spraying process for which different modeling strategies are compared. The models treated are from the class of linear models. It is known that more elaborate models like generalized linear models, non-linear models or measurement error models might be more appropriate. However, the algebraic treatment would be very much the same, hence we stay with the easier to handle linear models. Overall, we achieve a much improved model selection due to an enhanced knowledge of the space of identifiable models achieved through methods developed within the general framework of algebraic statistics.

Acknowledgments

The financial support of the DFG (SFB 823: Project B1) and the DAAD are gratefully acknowledged.

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1973.

G. Allaire. *Numerical Analysis and Optimization.* Oxford University Press, New York, 2009. Reprinted.

R.A. Bates, B. Giglio, and H.P. Wynn. A global selection procedure for polynomial interpolators. *Technometrics*, 45(3):246–255, 2003.

Y. Bernstein, H. Maruri-Aguilar, S. Onn, E. Riccomagno, and H.P. Wynn. Minimal average degree aberration and the state polytope for experimental designs. *Ann. Inst. Stat. Math.*, 62:673–698, 2010.

B. Buchberger. Ein algorithmisches Kriterium für die Lösbarkeit eines algebraischen Gleichungssystems. *Aequationes Math.*, 4:374–383, 1970.

CoCoATeam. CoCoA: a system for doing Computations in Commutative Algebra. Available at `http://cocoa.dima.unige.it`.

D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties, and Algorithms.* Springer-Verlag, New York, 1996. Second Edition.

D.R. Cox and N. Reid. *The Theory of the Design of Experiments.* Chapman & Hall, Boca Raton, 2000. Second Edition.

C. Fassino. Almost vanishing polynomials for sets of limited precision points. *J. Symbolic Comput.*, 45(1):19–37, 2010.

J. Faugere, P. Gianni, P. Lazard, and T. Mora. Efficient computation of zero-dimensional grobner bases by change of ordering. *J. Symbolic Comput*, 16:329–344, 1993.

D. Heldt, M. Kreuzer, S. Pokutta, and H. Poulisse. Approximate computation of zero-dimensional polynomial ideals. *J. Symbolic Comput.*, 44(11):1566–1591, 2009.

T. Holliday, G. Pistone, E. Riccomagno, and H.P. Wynn. The application of computational algebraic geometry to the analysis of designed experiments: a case study. *Comput. Statist.*, 14(2):213–231, 1999.

A.N. Jensen. Gfan, a software system for Gröbner fans and tropical varieties. Available at `http://home.imf.au.dk/jensen/software/gfan/gfan.html`.

H. Maruri-Aguilar. Methods from computational commutative algebra in design and analysis of experiments. Ph.D. thesis Statistics, Warwick, 2007.

H. M. Möller and B. Buchberger. The construction of multivariate polynomials with preassigned zeros. In *Computer algebra (Marseille, 1982)*, volume 144 of *Lecture Notes in Comput. Sci.*, pages 24–31. Springer, Berlin, 1982.

T. Mora and L. Robbiano. The Gröbner fan of an ideal. *Symb. Comput*, 6(2-3): 183–208, 1988.

R. Notari and E. Riccomagno. Replicated measurements and algebraic statistics. In *Algebraic and geometric methods in statistics*, pages 187–202. Cambridge Univ. Press, Cambridge, 2010.

J.Lh. Peixoto. A property of well-formulated polynomial regression models. *The American Statistician*, 44(1):26–30, 1990.

G. Pistone, E. Riccomagno, and H.P. Wynn. *Algebraic statistics*, volume 89 of *Monographs on Statistics and Applied Probability.* Chapman & Hall/CRC, Boca Raton, FL, 2001. Computational commutative algebra in statistics.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2012. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

E. Riccomagno. A short history of algebraic statistics. *Metrika*, 69:397–418, 2009.

L. Robbiano. Term orderings on the polynomial ring. In *EUROCAL'85*, pages 513–517. Springer, Berlin, 1985.

N. Rudak, S. Kuhnt, B. Hussong, and W. Tillmann. On different strategies for the prediction of coating properties in a hvof process. Sfb 823 discussion paper 29/12, TU Dortmund University, 2012.

6 Appendix

6.1 Direct Models

Porosity: $Po = 4.1850 - 1.6311 \cdot k - 0.8145 \cdot l + 1.9748 \cdot l^2 - 0.3071 \cdot k \cdot l$

Hardness: $Ha = 1370.45 + 78.15 \cdot k - 46.82 \cdot l - 134.35 \cdot l^2 - 30.47 \cdot k \cdot l$

Thickness: $Th = 249.40 + 50.25 \cdot f + 28.37 \cdot d - 21.80 \cdot k - 18.47 \cdot d \cdot k + 14.46 \cdot f \cdot k$

Deposition Rate: $Dr = 47.12 + 2.75 \cdot d$

6.2 First Stage Models

Method FB

Temperature: $$t = 1606.700 + 32.763 \cdot k - 20.088 \cdot l - 17.925 \cdot d + 9.463 \cdot f - 26.925 \cdot l^2 + 12.950 \cdot k \cdot f$$

Velocity: $$v = 695.000 + 41.075 \cdot k - 14.588 \cdot d + 13.525 \cdot l + 19.925 \cdot l^2 - 6.162 \cdot k \cdot d$$

Flame Width: $$w = 7.9471 + 0.5625 \cdot f + 0.2125 \cdot d - 0.1875 \cdot l - 0.2000 \cdot f \cdot l + 0.1500 \cdot d \cdot l$$

Flame Intensity: $$i = 21.406 + 5.163 \cdot f + 2.575 \cdot k - 1.938 \cdot l - 1.200 \cdot d + 2.225 \cdot f \cdot k$$

Method Best

Temperature:

$$\begin{aligned} t =\,& 1581.359 - 20.088 \cdot l + 32.763 \cdot k - 17.925 \cdot d 9.463 \cdot f 2.625 \cdot l \cdot f \\ & +12.950 \cdot k \cdot f - 3.437 \cdot d \cdot f - 2.625 \cdot l \cdot k \cdot d \\ & +8.637 \cdot l \cdot k \cdot f + 10.525 \cdot l \cdot d \cdot f + 7.400 \cdot k \cdot d \cdot f \\ & +3.887 \cdot l \cdot k \cdot d \cdot \end{aligned}$$

Velocity:

$$\begin{aligned} v =\,& 713.753 + 13.525 \cdot l + 41.075 \cdot k - 14.588 \cdot d - 3.437 \cdot f - 3.575 \cdot l \cdot k \\ & -2.688 \cdot l \cdot d - 6.162 \cdot k \cdot d + 8.562 \cdot k \cdot f \\ & +5.650 \cdot d \cdot f \end{aligned}$$

Flame Width:

$$\begin{aligned} w =\,& 7.9471 - 0.1875 \cdot l + 0.0875 \cdot k + 0.2125 \cdot d + 0.5625 \cdot f - 0.1250 \cdot l \cdot k \\ & +0.1500 \cdot l \cdot d - 0.2000 \cdot l \cdot f - 0.1750 \cdot k \cdot d + 0.0750 \cdot k \cdot f \\ & +0.1000 \cdot d \cdot f + 0.1875 \cdot l \cdot d \cdot f \\ & -0.1375 \cdot k \cdot d \cdot f - 0.0500 \cdot l \cdot k \cdot d \cdot f \end{aligned}$$

Flame Intensity:

$$\begin{aligned} i =\,& 21.4059 - 1.9375 \cdot l + 2.5750 \cdot k - 1.2000 \cdot d + 5.1625 \cdot f + 0.5000 \cdot l \cdot k \\ & +0.4500 \cdot l \cdot d + 2.2250 \cdot k \cdot f - 0.9000 \cdot d \cdot f \\ & -0.5625 \cdot l \cdot k \cdot d + 1.4000 \cdot l \cdot k \cdot f + 1.7000 \cdot l \cdot d \cdot f \\ & +0.8625 \cdot k \cdot d \cdot f \end{aligned}$$

Method Simple

Temperature: $t = 1581.359 - 20.088 \cdot l + 32.763 \cdot k - 17.925 \cdot d + 9.463 \cdot f$

Velocity: $v = 713.753 + 13.525 \cdot l + 41.075 \cdot k - 14.588 \cdot d - 3.437 \cdot f$

Flame Width: $w = 7.9471 - 0.1875 \cdot l + 0.0875 \cdot k + 0.2125 \cdot d + 0.5625 \cdot f$

Flame Intensity: $i = 21.406 - 1.938 \cdot l + 2.575 \cdot k - 1.200 \cdot d + 5.163 \cdot f$