Statistica Sinica **26** (2016), 1021-1035
doi:http://dx.doi.org/10.5705/ss.2014.264

NUMERICAL ALGEBRAIC FAN OF A DESIGN FOR STATISTICAL MODEL BUILDING

Nikolaus Rudak, Sonja Kuhnt and Eva Riccomagno

TU Dortmund University, Dortmund University of Applied Sciences and Arts and University of Genova

Abstract: An important issue in the design of experiments is the question of identifiability of models. This paper deals with a modelling process, where linear modeling goes beyond the simple relationship between input and output variables. Observations or predictions from the chosen experimental design are themselves input variables for an eventual output. Tools developed to analyze designs from algebraic statistics are extended to noisy, irregular designs. They enable an advanced study of model identifiability. Model building is opened towards higher order interactions rather than restricting the class of considered models to main effects or two-way interactions only. The new approach is compared to classical model building strategies in an application to a thermal spraying process.

Key words and phrases: Algebraic statistics, aliasing, identifiability, noisy design, vanishing ideal.

1. Introduction

In this article we develop methods to determine polynomial models that are identifiable from non-standard, noisy experimental designs. In algebraic statistics an experimental design, defined as a finite set of distinct experimental settings, is expressed as solution of a system of polynomial equations. Thereby the design is described by a polynomial ideal and features and properties of the ideal provide insight into the structures of models identifiable by the design (Pistone, Riccomagno, and Wynn (2001)). Holliday et al. (1999) apply these ideas to a problem from the automotive industry with an incomplete standard factorial design.

Our work is motivated by a thermal spraying process used to produce a particle coating on a surface. Controllable process parameters (X variables) are varied according to standard experimental designs. Properties of the coating particles in flight (Y variables) are measured during the process. The observed values y or their predictions $\hat{y}$ are inputs in models describing coating properties (Z variables) as final output. The left diagram in Figure 1 summarises this.

The coating properties are very time-consuming and expensive to measure as the specimen has to be destroyed. It is thus desirable to predict coating

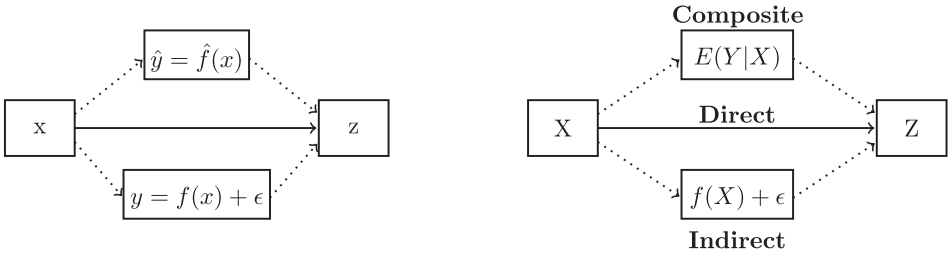

Figure 1. Left: occurence of noisy design points as observed y or predicted input values $\hat{y}$ for a final output z. Right: modelling approaches.

properties on the basis of particle properties. The present article treats the question of identifiable models from the Y to the Z variables. Basic tools from algebraic statistics for the analysis and design of experiments can be used to determine classes of identifiable models. However, some of the models returned from this basic theory are identifiable only due to small deviations of the design from more regular points. This leads to unwanted unstable models. As data or predictions on the Y variables are noisy, the analysis is very much affected by this problem. We extend existing theory by switching from symbolic, exact computations to numerical computations. Specifically, instead of polynomials whose corresponding polynomial equations have as solutions the design points, we identify a design with a set of polynomials which "almost vanish" at the design points. We do so using theoretical results and algorithms from Fassino (2010).

Section 2 reviews the notions of statistical and algebraic fans. In Section 3 the question of identifiable models from the designs occurring in the thermal spraying application is discussed and algorithms are developed to derive numerical fans for noisy designs in Section 4. In Section 5 theoretical properties and relationships between models potentially identifiable from the designs are analysed. Section 6 is the case study itself.

2. Background of Algebraic and Statistical Fans

A design or a set of observations can be seen as the zeros of a system of polynomial equations. This simple observation is the entry key for algebraic geometry to the design and analysis of experiments. For a generalization to designs with replicated points see Notari and Riccomagno (2010). A model with support $[f_1(x), \ldots, f_r(x)]$ is identified by the design with distinct points $d_1, \ldots, d_s$ if the design matrix $[f_j(d_i)]_{i=1,\ldots,s;j=1,\ldots,r}$ is full rank. It is saturated if the rank is $r = s$. The order ideal (or hierarchical) property states that any lower order term of an interaction term in the model is in the model as well.

Example 1. The design $D = \{(\pm 1, \pm 1), (0, 0)\}$ is the solution set of $p_1 = p_2 = p_3 = 0$ where $p_1 = x_1^3 - x_1$, $p_2 = x_2^3 - x_2$ and $p_3 = (x_2 - x_1)(x_1 + x_2)$. From

classical theory we know that only two saturated polynomial models with the hierarchical property are identifiable by $\mathcal{D}$. They have support $\{1, x_1, x_2, x_1x_2, x_1^2\}$ or $\{1, x_1, x_2, x_1x_2, x_2^2\}$. The corresponding design matrices coincide and are

$$X = \begin{pmatrix} 1 & x_1 & x_2 & x_1x_2 & x_1^2/x_2^2 \\ \hline 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{2.1}$$

Clearly X is invertible and hence the two models are identifiable. The power product x_1^2 and x_2^2 cannot be part of the same models because they are aliased, indeed $p_3 = (x_2 - x_1)(x_1 + x_2) = 0$ is equivalent to $x_1^2 = x_2^2$. Statistically this means that both effects are not distinguishable by data observed from $\mathcal{D}$. Substitution of $x_2^2 = x_1^2$ into p_2 gives a new polynomial $p_4 = x_2x_1^2 - x_2$ and the solution set of $p_1 = p_4 = p_3 = 0$ is still $\mathcal{D}$.

The set of hierarchical models identified by a design $\mathcal{D}$ with as many terms as distinct points in $\mathcal{D}$ is called the *statistical fan* of $\mathcal{D}$. It is finite and each of its elements, called leaves, is formed by as many power products (or monomials) as distinct points in $\mathcal{D}$. Each leaf is an order ideal, and hence it contains 1, and the design matrix for each leaf of the fan is invertible. The intersection of all leaves satisfies the hierarchical property and forms the support of polynomial models identifiable by $\mathcal{D}$. Subsets of the fan provide lists of saturated hierarchical models each of which can be input to a selection procedure for determination of a well-fitting parsimonious submodel.

In designs with a less regular structure, the statistical fan might not be easy to determine. We provide a systematic method to investigate at least an interesting part of it. The technical tools at the basis of the computation are a term order τ on the set of power products and the associated reduced Gröbner basis. A good reference for these algebraic notions is Cox, Little, and O-Shea (1996). Generally, the set of polynomials in the variables $x_1, \ldots, x_n$ and with real coefficients is indicated by $\mathbb{R}[x_1, \ldots, x_n]$ and the set of power products by T^n. A power product $x^\alpha = x_1^{\alpha_1} \ldots x_n^{\alpha_n}$ is represented by the vector of its exponents $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{Z}^n_{\geq 0}$. Hence ordering power products corresponds to ordering vectors with non-negative integer entries, more precisely a term order τ on T^n is a well-order relation on $\mathbb{Z}^n_{\geq 0}$. A saturated hierarchical model identifiable by the design is determined from a τ-Gröbner basis G as those power products in T^n which are not divisible by any of the leading terms with respect to τ of the elements of G. The obtained set, called a quotient basis or $\mathcal{O}_\tau(D)$

(Pistone, Riccomagno, and Wynn (2001)), is an element of the statistical fan of D and hence the design matrix

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

for D and $\mathcal{O}_{\tau}(D)$ is invertible. The set we are interested in is $F_D = \{\mathcal{O}_{\tau}(D) : \tau\}$, called the *algebraic fan* of D.

Example 2 (Example 1 contd.)**.** The design D is solution to $p_1 = p_2 = p_3 = 0$. However, its points satisfy also the equation $s_1(x_1^3 - x_1) + s_2(x_1^2 x_2 - x_2) + s_3(x_2 - x_1)(x_1 + x_2) = 0$ 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, written as

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

and referred to as the design ideal of D or the vanishing ideal at D. For any term order τ for which x_1 is smaller than x_2, the three polynomials $p_1 = \underline{x_1^3} - x_1$, $p_4 = \underline{x_1^2 x_2} - x_2$, $p_3 = (x_2 - x_1)(x_1 + x_2) = \underline{x_2^2} - x_1^2$ 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_2 is smaller than x_1. By a symmetry argument it is seen to be $\{x_2^3 - x_2, x_2^2 x_1 - x_1, (x_1 - x_2)(x_1 + x_2)\}$. The full algebraic fan of D is thus $\left\{\{1, x_1, x_2, x_2^2, x_1 x_2\}, \{1, x_1, x_2, x_1^2, x_1 x_2\}\right\}$.

Generally, the necessary computations to obtain the algebraic fan cannot be done by hand even for designs whose points exhibit regular geometric configuration. Usually, the algebraic fan is very large and can be much smaller than the statistical fan (Maruri-Aguilar (2007)). By Theorem 30 in Pistone, Riccomagno, and Wynn (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. Berstein et al. (2010) compare algebraic and statistical fans for some classes of designs, including Latin hypercube and orthogonal fractions, and derive a notion of design aberration. But 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 case study driving our work.

3. Motivation for a Numerical Fan of a Design

With the thermal spraying process in mind we are interested in a comparison of different models for the final response Z, which are either based on the initial inputs X, the intermediate outcomes Y or a prediction $\hat{Y}$. See the right diagram of Figure 1. To fix notation assume that X has q components, Y has p components, and Z has m components. Model building is based on an initial design

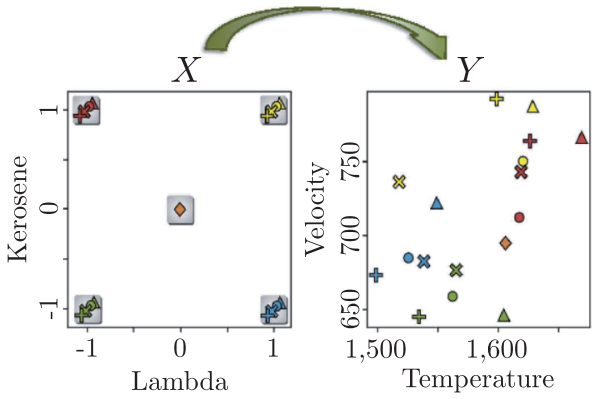

Figure 2. Settings of the controllable X variables Lambda versus Kerosene and the resulting observed values for the Y variables Temperature versus Velocity.

D_x for X and observed values D_y from Y and D_z from Z are available. To build models from Y to Z it is not immediately obvious which effects are identifiable, because compared to the controllable design parameters X the observed particle properties in D_y scatter strongly, see Figure 2 whose caption has to be read in the light of Section 6. In particular we are interested in finding out if information is lost or models are missed by considering any of the possible input types. For the model selection procedure the knowledge of possible maximal models is extremely useful as an all-subset selection is usually not feasible.

In polynomial models 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 θ_α real numbers and

$$L = \{0_q, (1,0,\dots,0), (0,1,0,\dots,0), \dots, (0,\dots,0,1)\}.$$

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 a statistical linear model of Z given X might be expressed as

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

where $E(\cdot|X=x)$ indicates expected conditional mean to $X=x$. The support of the model is indicated with $x_z = [x^\alpha]_{\alpha\in L}$ and is identified with the exponents of the power products in the set L. The parameter vector is $\gamma^* = [\gamma^*_\alpha]_{\alpha\in L}$.

The initial design D_x in the case study in Section 6 is a full factorial design with central point in four factors k, l, d, f. By generalization of Example 1 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 D_x. For any term order τ on the set of power products based on the letters f, d, l, k and for which f is lowest, the saturated model is

$$O_\tau(D_x) = \left\{ 1, f, f^2, d, l, k, \quad df, lf, ld, kf, kd, kl, \quad ldf, kdf, klf, kld, \quad kldf \right\}.$$

The $\mathcal{O}_\tau(D_x)$ above 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 power products and for which the design matrix is invertible. This statement follows by observing that (1) a power product in $\mathcal{O}_\tau(D_x)$ cannot have degree three or more in any variable because the four factors have three levels each and (2) d^2, k^2 and l^2 are aliased with f^2, indeed the two vectors of evaluations $[f^2(d)]_{d\in\mathcal{D}}$ and $[l^2(d)]_{d\in\mathcal{D}}$ 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 fan gives a hierarchical model with all 16 square free interactions up to order four.

The design in the X variables has a nice regular structure and we easily computed its fans. However, the "designs" D_y and $D_{\hat{y}}$ in the thermal spraying application look, although are not, random and have a fairly complex geometrical structure. The complexity of D_y and $D_{\hat{y}}$ carries over to their ideals and to their fans. Standard statistical techniques go only so far in their analysis and do not provide information on the aliasing structure imposed by the designs on the space of polynomial models for the final output variables Z. This is where, we believe, the algebraic method adopted in this paper becomes especially worthwhile.

Example 3 below shows another reason why it might be desirable to consider only a subset of the fans by excluding numerically unstable leaves. It shows some of the issues we encounter and overcome by using an approximated version of the design ideal and of its algebraic fan. 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 where the symbol $||\cdot||$ indicates matrix norm. The condition number is at least 1 and if it is much larger (depending on matrix size) than 1 then the matrix is ill-conditioned. In case of an ill-conditioned matrix, the solution of $Ax = b$ will be particularly sensitive to errors in A or in b.

Example 3. Let D be the 2^2 full factorial design with levels ± 1. The algebraic and statistical fans have only one leaf $\{1, x_1, x_2, x_1x_2\}$. If we substitute the point $(1, -1)$ with $(1, -1.001)$, the algebraic and statistical fans are equal and are given by the two leaves $\{1, x_1, x_2, x_1x_2\}$ and $\{1, x_1, x_2, x_2^2\}$ with corresponding design matrices X_1 and X_2 given below. The condition number of X_1 is almost 1.000707180 and of X_2 is $4 \cdot 10^3$.

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

The X_1 matrix is well-conditioned and so are problems relying on it, e.g. the stability of 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.

4. NBM Algorithm and Numerical Fan

We next consider designs whose points' coordinates are known up to a certain precision. We might think that there are measurement errors for D_y and prediction errors in $D_{\hat{y}}$. We seek a set of polynomials which *almost vanish* at the design points, namely evaluated at the design points they are close enough to zero. To do that, we use the numerical Buchberger-Möller (NBM) algorithm presented in Fassino (2010). It is based on a least square approximation and is a variant of the purely symbolic Buchberger-Möller algorithm (Möller and Buchberger (1982)) which is the computational tool at the basis of the application of algebraic geometry to design of experiments as described in Section 2.

The inputs to the NBM algorithm are a finite set of distinct points in n dimensions, say $\mathcal{D} \subset \mathbb{R}^n$, a term order τ and a precision vector $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \in \mathbb{R}^n_{\geq 0}$, where ϵ_i, $i = 1, \ldots, n$, gives allowed perturbation for the ith component. The outputs are a set of polynomials $\mathcal{G}$ and a hierarchical set of power products $\mathcal{O}$. It also includes a flag stating whether $\mathcal{O}$ has the same number of elements as there are points in $\mathcal{D}$. By construction, the matrix X whose columns are the evaluations of the power products in $\mathcal{O}$ at $\mathcal{D}$ is well-conditioned. Furthermore, X is full rank for all admissible perturbations of $\mathcal{D}$, although it might not be well-conditioned for all such perturbations.

We recall the basic notions from Fassino (2010). 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 $\mathcal{D}$. 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 $||X||_2/||c||_2 < O(\epsilon_M)$, where $X = [g(d)]_{d \in \mathcal{D}}$ is the evaluation vector of g at $\mathcal{D}$, $||\cdot||_2$ the Euclidean norm, and $\epsilon_M = \max\{\epsilon_i : i = 1, \ldots, n\}$. The set of polynomials almost vanishing at $\mathcal{D}$ is called the *approximated ideal* of tolerance ϵ.

A polynomial with small enough coefficients almost vanishes at many points and the zero set of a polynomial does not change when multiplied by a constant. Hence, when needed, we consider the unitary version of the approximating sets obtained by multiplying each polynomial in the generating sets of the approximated ideals with the inverse of the Euclidean norm of its coefficient vector.

The tail of the first polynomial returned by NBM is a linear combination of the smallest power products with respect to τ whose design matrix is full-rank for every ϵ-perturbed design. This can be interpreted as a high-dimensional surface

of a shape which is as simple as possible in τ and to which $\mathcal{D}$ is close in a least square sense. Indeed the NBM algorithm returns an implicit representation of $\mathcal{D}$ depending on the term order τ as input.

The polynomials returned by the NBM algorithm usually do not generate a proper polynomial ideal because they might have non-common zeros, unless $\epsilon_M = 0$. 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 a generating set of the exact design ideal.

The key technical step in the NBM algorithm is to start constructing (almost) vanishing polynomials by adding the lowest possible power product in τ. In short it is as follows: 1. start with $M := \{1\}$, 2. consider the smallest w.r.t. τ power product not in M, 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, 5. if yes, the obtained polynomial is almost vanishing while if not, add x^α to M and repeat from Step 2. The main computational cost at each iteration is the cost of computing the pseudo-inverse of an evaluation matrix in Step 3, involved in the solution of the least square problem, and of the estimation of the termination criterion in Step 4. Efficient algorithms are available for both.

To fully adapt this to the construction of a (numerical) fan, in Step 2 one needs to consider each possible x^α that preserves the order ideal structure. The fan is finite because the underlying variety is zero-dimensional and the procedure stops when no power product can be added. It returns the (numerical) statistical fan. In high dimension computing it is no trivial task, indeed it has not been implemented efficiently yet. In Section 6 we choose to approximate the numerical fan by computing a subset of the algebraic fan by running the NBM algorithm for some significant term orders.

5. Direct, Indirect and Composite Models

We next go back to the considered models in the right diagram of Figure 1 and compare them on a theoretical level. The *direct model* approach assumes $Z = h(X) + \delta$ with $E(\delta|X) = E(\delta) = 0$ and $Var(\delta|X) = Var(\delta)$ constant. Hence it holds that $E(Z|X = x) = h(x)$. For η, ϵ, and $\tilde{\eta}$ satisfying the same distributional assumptions of δ, the *composite model* is based on the assumptions $Z = g(f(X)) + \eta$ and $Y = f(X) + \epsilon$, thus $E(Z|X = x) = g(E(Y|X = x))$; while the *indirect model* takes $Z = g(Y) + \tilde{\eta}$ and $Y = f(X) + \epsilon$, hence $E(Z|X = x) = E(g(f(X) + \epsilon)|X = x)$. If g is a linear function then the indirect model becomes $E(Z|X) = g(f(x))$ by linearity of expectation, and the indirect and composite models coincide.

A straightforward full theoretical comparison of the three approaches is possible for the special case of linear models and main effects in going from Y to Z, and also for models beyond main effects in the direct strategy as well as from X to Y for the other two strategies. Without loss of generality set $m = 1$, hence $Z \in \mathbb{R}$. 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 power products of the original X-variables to model intercept, main effects, interactions, quadratic terms and so on, as required. It follows from these assumptions that

$$E(Z|X = x) = X_z^T \gamma^*. \tag{5.1}$$

For a main effect linear model between Y and Z, the composite model simplifies to

$$E(Z|X = x) = E(\gamma_0 + f(X)^T \gamma + \eta | X = x) = \gamma_0 + f(x)^T \gamma \tag{5.2}$$

with $\gamma_0 \in \mathbb{R}$ and $\gamma \in \mathbb{R}^p$ unknown parameters, for some suitable $p \in \mathbb{Z}_{\geq 0}$. Similarly when g gives a main effect linear model the indirect model reads

$$E(Z|X = x) = E(\tilde{\gamma_0} + (f(X) + \epsilon)^T \tilde{\gamma} + \tilde{\eta} | X) = \tilde{\gamma_0} + f(x)^T \tilde{\gamma}, \tag{5.3}$$

with $\tilde{\gamma_0} \in \mathbb{R}$ and $\tilde{\gamma} \in \mathbb{R}^q$ for some integer q. From (5.2) and (5.3) 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 (5.2) and (5.3) by a multivariate linear model. So for $i = 1, \ldots, p$ let the i-th component of f be written as

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

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 $\beta^i = [\beta_\alpha^i]_{\alpha \in L_i}$ gives the unknown parameter vector. Hence (5.2) becomes

$$E(Z|X = x) = \gamma_0 + f(x)^T \gamma = \gamma_0 + \left[\sum_{\alpha \in L_i} x^\alpha \beta_\alpha^i \right]_{i=1,\ldots,p}^T \gamma = \gamma_0 + \sum_{i=1}^p \sum_{\alpha \in L_i} x^\alpha \beta_\alpha^i \gamma_i. \tag{5.4}$$

By assuming equality of $E(Z|X = x)$ in all modeling approaches, from (5.1) and (5.4) we obtain an equality of polynomials in the x^α's,

$$\sum_{\alpha \in L} \gamma_\alpha^* x^\alpha = \gamma_0 + \sum_{i=1}^p \sum_{\alpha \in L_i} x^\alpha \beta_\alpha^i \gamma_i.$$

Table 1. Controllable and measured variables in the spraying process.

process parameters X	particles in-flight properties Y	coating properties Z
Kerosene (k)	Temperature (t)	
Lambda (l)	Velocity (v)	Hardness (Ha)
Stand-off Distance (d)	Flame width (w)	Deposition rate (Dr)
Feeder Disc Velocity (f)	Flame intensity (i)	

This holds true if and only if coefficients of the same power product 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 an intercept, so that a vector of suitable dimension and with all entries equal to zero belongs to L_i for all i from 1 to p. We also define L_i^* to be the set L_i without the zero vector. The above becomes $\gamma^*_{0_q} + \sum_{\alpha \in L^*} \gamma^*_\alpha x^\alpha = \gamma_0 + \sum_{i=1}^p \beta^i_{0_q}\gamma_i + \sum_{i=1}^p \sum_{\alpha \in L_i} x^\alpha \beta^i_\alpha \gamma_i$. 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 $0 = \sum_{i=1}^p \beta^i_\alpha \gamma_i$. This can be understood as *theoretical aliasing relationships* among the parameters for the indirect/composite case and the direct case.

6. Application to a HVOF Experiment

In the considered High Velocity Oxygen Fuel Flame (HVOF) spraying process considered, a spraying gun acts as a source of heat and spray material in powder form is fed to the gun from outside. Inside the gun the spray material is melted by the flame partly or completely and a gas stream accelerates the heated particles towards the substrate. The particles deposit as a coating after cooling down. The final coating can be analyzed but only at a cost of destroying it. A problem with this technology is a lack of reproducibility of the coatings caused by non-controllable effects which are, however, presumably visible in the particle properties. Our aim is to predict a coating with desired properties Z by means of controllable design parameters X on the spraying gun using information from observed particle properties Y.

In Table 1 the design parameters, particle properties, and coating properties identified by previous experiments (e.g., Tillmann et al. (2010)) are summarized. Experiments based on a full factorial design for the X variables with a center point (see Table 2) were carried out and particle properties together with coating properties measured.

Models for the Y variables depending on the X variables were obtained with an all-subset selection based on the AIC criterion. The maximal model for the submodel selection procedure is the intersection of the four leaves in the fans of

Table 2. Coded design parameters.

Coded values	-1	0	1
Kerosene level in $\frac{l}{h}$ (k)	17.5	20	22.5
Lambda (l)	1.075	1.15	1.225
Stand-off distance in mm (d)	225	250	275
Feeder disc velocity in $\frac{g}{min}$ (f)	7.5	10	12.5

$\mathcal{D}_x$. The fitted regression models, with estimates rounded to two digits, are

$$\begin{aligned}
\text{Temperature: } t = {} & 1581.36 - 20.09\,l + 32.76\,k - 17.93\,d + 9.46\,f + 2.63\,lf \\
& +12.95\,kf - 3.44\,df - 2.63\,lkd + 8.64\,lkf + 10.53\,ldf \\
& +7.40\,kdf + 3.89\,lkd, \\
\text{Velocity: } v = {} & 713.75 + 13.53\,l + 41.08\,k - 14.59\,d - 3.44\,f - 3.58\,tlk \\
& -2.69\,ld - 6.16\,kd + 8.56\,kf + 5.65\,df, \\
\text{Flame Width: } w = {} & 7.95 - 0.19\,l + 0.09\,k + 0.21\,d + 0.56\,f - 0.12\,lk + 0.15\,ld \\
& -0.20\,lf - 0.18\,kd + 0.08\,kf + 0.10\,df + 0.19\,ldf \\
& -0.14kdf - 0.05\,lkdf, \\
\text{Flame Intensity: } i = {} & 21.41 - 1.94\,l + 2.58\,k - 1.20\,d + 5.16\,f + 0.50\,lk + 0.45\,ld \\
& +2.23\,kf - 0.90\,df - 0.56lkd + 1.40\,lkf + 1.70\,ldf + 0.86\,kdf.
\end{aligned}$$

The predicted values at $\mathcal{D}_x$ give 17 distinct points in $\mathbb{R}^4$ collected in $\mathcal{D}_{\hat{y}}$. The generating process of the designs $\mathcal{D}_{\hat{y}}$ and $\mathcal{D}_y$ destroys the symmetries of the $\mathcal{D}_x$ design (e.g., Figure 2). Their irregularity comes from different sources, traceable back to the measurement errors of the observed Y-values, to an inherent complexity of the generating process, and to modeling approximation. The adjusted R^2 values, namely 0.92 for Temperature, 0.97 for Velocity, 0.86 Flame Width, and 0.96 for Flame Intensity are a measure of this.

6.1. Approximated vanishing ideals for the Y-designs

The polynomials in the exact design ideals $I(\mathcal{D}_y)$ and $I(\mathcal{D}_{\hat{y}})$ vanish at the points of the design $\mathcal{D}_y$ and $\mathcal{D}_{\hat{y}}$ by definition. Even when $\mathcal{D}_{\hat{y}}$ is an ϵ-perturbation of $\mathcal{D}_y$, their exact ideals may be very different. More informative is to consider approximated versions of the design ideals and compare them. Let $\mathcal{G}(\mathcal{D}_y)$ and $\mathcal{G}(\mathcal{D}_{\hat{y}})$ be the output of the NBM algorithm for $\mathcal{D}_y$ and $\mathcal{D}_{\hat{y}}$, $\tau = degrevlex(t, v, w, i)$ and $\epsilon = (5, 2, 0.01, 0.01)$, where ϵ is chosen together with engineers. The first component of ϵ refers to temperature, the second one to velocity, the third one to flame width, and the last one to flame intensity.

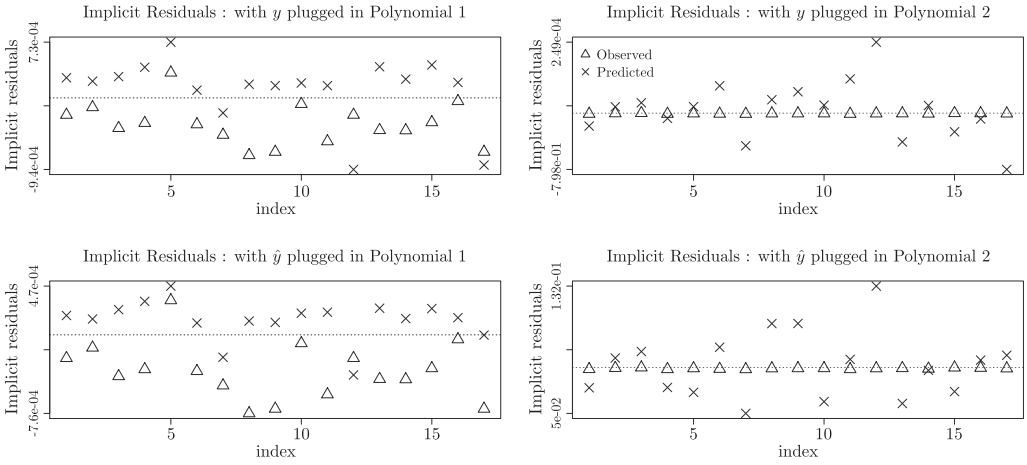

Figure 3. Implicit Residuals.

The number of polynomials in $\mathcal{G}(\mathcal{D}_y)$ is 17 and 15 in $\mathcal{G}(\mathcal{D}_{\hat{y}})$, this is a simple indication of the difference between the two almost vanishing ideals of, $\mathcal{D}_y$ and $\mathcal{D}_{\hat{y}}$. However the polynomials of the almost vanishing sets of $\mathcal{D}_{\hat{y}}$ have to almost vanish at the observed values $\mathcal{D}_y$ when $\mathcal{D}_{\hat{y}}$ is a good approximation of $\mathcal{D}_y$, even though the two designs are not an ϵ-perturbation of each other. A measure of this are the adjusted R^2 values.

In order to check this further and in order to check whether the first polynomials of the almost vanishing ideals are sufficiently informative to compute the fans, in Figures 3 we show what we call the **implicit residuals**. They are obtained by substituting $\mathcal{D}_y$ and $\mathcal{D}_{\hat{y}}$ in the first and second polynomials of $\mathcal{G}(\mathcal{D}_y)$ and $\mathcal{G}(\mathcal{D}_{\hat{y}})$. As expected, the implicit residuals are very small with absolute maximal value smaller than 10^{-3}.

6.2. Computation of the algebraic fan

In the previous section we considered one possible $\mathcal{O}$ set for $\mathcal{D}_{\hat{y}}$ and one $\mathcal{O}$ set for $\mathcal{D}_y$; thus for each design we computed one set of power products from which to start a model search. Here, we consider many $\mathcal{O}$ sets by varying the term order according to the strategy indicated in Section 4. We used the StableBBasisNBM5 function of CoCoA4.7.5 (CoCoATeam (2009)), which implements the numerical Buchberger-Möller algorithm. Ideally one would like to compute the full algebraic fan of the approximated ideals but this is not implemented in CoCoA4 or elsewhere, yet. Therefore, we chose three standard term orders: lexicographical, degree lexicographical, and reverse degree lexicographical ordering, and all possible permutations of the variables. These are quite extreme term orders: lexicographic orderings tend to include in $\mathcal{O}$ all powers of the smallest

Table 3. PRESS statistic values for different model building strategies.

	Hardness	Deposition rate
algebraic approach	**169268.96**	**164.75**
classical approach	172747.90	466.59
direct	198472.99	461.50

variable first; degree compatible term orders favor the inclusion of the first suitable power products with lowest total degree (sum of exponents). Thus, to each design there is associated a subset of its fan comprising 72 leaves a.k.a. as $\mathcal{O}$ sets. To compare the leaves within each subfan we counted the number of the 20 most frequent power products. Unsurprisingly they coincide, as $\mathcal{D}_{\hat{y}}$ is generated by the best fitting models of $\mathcal{D}_y$.

6.3. Comparison based on PRESS residuals

As the prediction of Z based on X is of major interest for the application, we compared the goodness of fit of the composite and the direct model. For a further comparison see (Rudak, Kuhnt, and Riccomagno (2013)). For reaching a good trade-off between parsimonious models and goodness of fit, the final model may be not hierarchical, even if the search spaces are hierarchical sets. For the direct model, a combination of backward and forward selection where the saturated model contains all main effects and interactions, led to $1370.45 + 78.15\,k - 46.82\,l - 134.35\,l^2 - 30.47\,k$ for Hardness and $47.12 + 2.75\,d$ for Deposition rate.

For composite models, we compared two model building methods. The new *algebraic approach* employs results from the algebraic analysis of $\mathcal{D}_{\hat{y}}$, whereas the *classical approach* relies on a common statistical model selection approach. For both approaches, particle properties were coded to values within $[-2, 2]$ (Temperature [1,300, 1,700], Velocity $[375, 825]$, Flame width $[5, 20]$, Intensity $[10, 30]$). In the algebraic approach, each leaf of the fan of $\mathcal{D}_{\hat{y}}$ is scope for a forward backward selection based on the AIC criterion. Models returned for different leaves are discriminated by AIC again resulted in $988.41 + 206.82\,t - 20.4\,v^5 + 148.85\,v$ for Hardness and in $63.42 + 58.34\,t^2v - 34.11\,t^2 - 5.38\,t^3 - 16.38\,v - 18.64\,t^2v^2$ for Deposition rate. In the classical approach, a forward-backward selection based on the AIC criterion and with maximal model all main effects and interactions, returned simpler models whose supports have only linear terms. For Hardness it returned $1094.71 + 183.5\,t$ and for Deposition rate $51.37 - 5.23\,t$.

To compare the approaches further we performed a leave-one-out cross validation analysis. Table 3 contains the PRESS statistic values (see Myers, Montgomery, and Anderson-Cook (2009)) defined as $\text{PRESS} = \sum_{i=1}^{17}(z_i - \hat{z}_{-i})^2$ where

$\hat{z}_{-i}$ is the predicted value for z_i based on the whole dataset except the i-th observation. The PRESS statistic values were smaller in the algebraic approach, up to the ratio of $1:3$ for Deposition rate.

This shows the worthwhileness of the computational effort required to compute (a part of) the fan of a design for model search. The final models for Deposition rate and Hardness obtained with the algebraic approach would not have been easily devised without the semi-automatic algebraic methods presented here. This is mainly because they include higher order terms and because of the very scattered configurations of design points for which they are computed. As often occurs, there is no pretence nor need to give an interpretation to the various terms in those models: they are models for the second part of the two stage modelling process and the engineering and physical knowledge on the Y-to-Z process is at the moment very limited.

7. Conclusion

Motivated by a thermal spraying process, we treated the question of identifiable models from noisy, irregular designs. We used almost vanishing ideals and an algorithm from Fassino (2010) for dealing with the instability in the observed or predicted designs. Models for the final response were analysed on a theoretical level with respect to their structural differences as well as aliasing relationships. Comparison with standard linear models showed that the use of algebraic statistics led to considering a wider choice of models and eventually a better fit, and thus demonstrated the efficacy of the proposed method. We introduced the notion of implicit residuals which measure the goodness-of-fit of the predicted design points.

More elaborate models like generalized linear models, non-linear models, or measurement error models might be more appropriate for the case study. However, the algebraic treatment would be very much the same and thus we chose the easier to handle linear models. The much improved model selection is due to an enhanced knowledge of the space of identifiable models achieved thanks to the proposed algebraic statistics method.

Acknowledgement

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

References

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

CoCoATeam (2009). *CoCoA: A System For Doing Computations in Commutative Algebra.* Available at `http://cocoa.dima.unige.it`.

Cox, D., Little, J. and O-Shea, D. (1996). *Ideals, Varieties, and Algorithms.* Springer-Verlag, New York.

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

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

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

Möller, H. M. and Buchberger, B. (1982). The construction of multivariate polynomials with preassigned zeros. *Computer Algebra.* Volume **144** of Lecture Notes in Comput. Sci., 24-31. Springer, Berlin.

Myers, R. H., Montgomery, D. C. and Anderson-Cook, C. M. (2009). *Response Surface Methodology.* Wiley , New Jersey.

Notari, R. and Riccomagno, E. (2010). Replicated measurements and algebraic statistics. *Algebraic and Geometric Methods in Statistics*, 187-202. Cambridge Univ. Press, Cambridge.

Pistone, G., Riccomagno, E. and Wynn, H.P. (2001). Algebraic statistics. Volume **89** of *Monographs on Statistics and Applied Probability.* Chapman & Hall/CRC, Boca Raton.

Rudak, N., Kuhnt, S. and Riccomagno, E. (2013). Numerical algebraic fan of a design for statistical model building. *SFB* 823 *Discussion Paper* 4/13, TU Dortmund University, Dortmund, Germany.

Tillmann, W., Vogli, E., Hussong, B., Kuhnt, S. and Rudak, N. (2010). Relations between in flight particle characteristics and coating properties by HVOF spraying. *Proceedings of ITSC* 2010 *Conference*, **264** of DVS-Berichte.

Faculty of Statistics, TU Dortmund University, 44221 Dortmund, Germany.

E-mail: rudak@statistik.tu-dortmund.de

Dortmund University of Applied Sciences and Arts, 44227 Dortmund, Germany.

E-mail: sonja.kuhnt@fh-dortmund.de

Department of Mathematics, University of Genova, 16146 Genova, Italy.

E-mail: riccomagno@dima.unige.it

(Received July 2014; accepted September 2015)