NONLINEAR STRUCTURE IDENTIFICATION WITH LINEAR LEAST SQUARES AND ANOVA

Ingela Lind

Division of Automatic Control, Department of Electrical Engineering, Linköping University, SE-581 83 Linköping, Sweden. E-mail: ingela@isy.liu.se

Abstract: The objective of this paper is to find the structure of a nonlinear system from measurement data, as a prior step to model estimation. Applying ANOVA directly on a dataset is compared to applying ANOVA on residuals from a linear model. The distributions of the involved test variables are computed and used to show that ANOVA is effective in finding which regressors give linear effects and what regressors produce nonlinear effects. The ability to find nonlinear substructures depending on only subsets of regressors is an ANOVA feature which is shown not to be affected by subtracting a linear model.

Keywords: System identification, Nonlinear systems, Structural properties, Analysis of variance, Linear estimation

1. INTRODUCTION

The aim of this paper is to quantify the difference between two different ways to use Analysis of Variance (ANOVA) as a tool for nonlinear system identification.

In general, system identification is data centred. Simple things are tried first; Is a linear model sufficient to describe the data? To invalidate a linear model, the residuals are examined with whiteness tests and the fit of the model on validation data is used to form an opinion of how good the model is. Thus, a linear model is often available, or easily computed.

ANOVA (Miller, 1997; Montgomery, 1991) can be used for finding proper regressors and model structure for a nonlinear model by fitting a locally constant model to the response surface of the data (Lind, 2000; Lind and Ljung, 2005, 2003). A clever parameterisation of a locally constant model makes it possible to perform hypothesis tests in a balanced and computationally very effective way. Let

$$\begin{aligned} y(t) &= g(u(t), u(t-T), \dots, u(t-kT)) + e(t) \\ &= \theta_1^T \varphi_1(t) + g_2(\varphi_2(t)) + e(t) \end{aligned}$$

be a general nonlinear finite impulse response model with input $u(t)$ and output $y(t)$, sampled with sampling time T. Let $\varphi_1(t)$ be a vector containing the regressors that affect the output linearly (with parameters θ_1) and $\varphi_2(t)$ the regressors that affect $y(t)$ nonlinearly through the function $g_2(\cdot)$. Three main questions can be answered by both running ANOVA directly on identification data and running ANOVA on the residuals from a linear model:

- Should the regressor $u(t-k_iT)$ be included in the model at all, and should it be included in $\varphi_1(t)$ or $\varphi_2(t)$?
- What interaction pattern is present? Can $g(\cdot)$ be divided into additive parts containing only subsets of the regressors? What subsets?
- Are there nonlinear effects in the residuals from a linear model?

There are much to be gained by the division into a linear and a nonlinear subset of the regressors (Ljung et al., 2004) instead of assuming a full nonlinear model. The complexity of any black-box type of model depends heavily on the size of $\varphi_2(t)$.

An idealised case is examined to quantify the difference between running ANOVA directly on identification data and first estimate an affine (linear with constant offset) model and then running ANOVA on its residuals. The input signal is chosen to keep computations simple, while being sufficiently exciting to make a nonlinear black-box identification possible.

The structure of the paper is as follows: First, the true data model is stated. In section 3, the linear model is estimated and the residuals are formed. In section 4, ANOVA is run directly on the estimation data, and in section 5 ANOVA is run on the residuals from the linear model. Section 6 explores the differences between the two approaches and give examples. The conclusions are made in section 7.

2. TRUE DATA MODEL

The system is a finite impulse response model:

$$y(t) = g(u(t), u(t-T)) + e(t), \tag{1}$$

where $e(t) \sim N(0, \sigma^2)$ is independent identically distributed Gaussian noise with mean zero and variance σ^2. The input signal $u(t)$ is a pseudo-random multi-level signal with mean zero, in which each level combination of $u(t)$ and $u(t-T)$ occur equally many times. The last condition defines a balanced dataset and will give independence between sums of squares. This type of signal can be given a nearly white spectra, see Godfrey (1993). The number of levels the signal assumes is m and the levels are denoted u_i, where $i = 1, \ldots, m$. An integer number, n, of periods from the input/output data are collected and denoted by $\mathbf{Z}^N$ ($N = np$, where p is the length of the period). $g(\cdot)$ is a function of two variables.

The results extend to more regressors, but since the equations do not fit in this short format, only two regressors are considered below.

3. ESTIMATION OF THE AFFINE MODEL

A linear FIR model with an extra parameter for the mean level of the signal,

$$\begin{aligned}\hat{y}(t) &= \hat{a}u(t) + \hat{b}u(t-T) + \hat{c} \\ &= \begin{bmatrix}\hat{a} & \hat{b} & \hat{c}\end{bmatrix} \begin{bmatrix} u(t) \\ u(t-T) \\ 1 \end{bmatrix} = \hat{\theta}^T \varphi(t),\end{aligned}$$

is estimated using linear least squares (Ljung, 1999, page 203). The loss function

$$V_N(\theta, \mathbf{Z}^N) = \frac{1}{N}\sum_{t=1}^{N} \frac{1}{2}(y(t) - \hat{y}(t))^2,$$

is minimised by the estimate

$$\begin{aligned}\hat{\theta}_N^{LS} &= \begin{bmatrix}\hat{a} & \hat{b} & \hat{c}\end{bmatrix}^T = \arg\min V_N(\theta, \mathbf{Z}^N) \\ &= [\frac{1}{N}\sum_{t=1}^{N} \varphi(t)\varphi^T(t)]^{-1} \frac{1}{N}\sum_{t=1}^{N} \varphi(t)y(t).\end{aligned}$$

For the pseudo-random multi-level input signal we have that:

$$\frac{1}{N}\sum_{t=1}^{N} u(t) = 0, \quad \frac{1}{N}\sum_{t=1}^{N} u(t-T) = 0, \tag{2}$$

$$\frac{1}{N}\sum_{t=1}^{N} u(t)u(t-T) = 0, \tag{3}$$

and

$$\frac{1}{N}\sum_{t=1}^{N} u^2(t) = \frac{1}{m}\sum_{i=1}^{m} u_i^2 = R_u.$$

If assumption (3) is not valid, change variables to $\tilde{u}(t-T) = u(t-T) - \alpha u(t)$, such that $1/N\sum_{t=1}^{N} u(t)\tilde{u}(t-T) = 0$. Now,

$$\left[\frac{1}{N}\sum_{t=1}^{N} \varphi(t)\varphi^T(t)\right]^{-1} = \frac{1}{R_u}\begin{bmatrix}1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & R_u\end{bmatrix}$$

and

$$\begin{aligned}&\frac{1}{N}\sum_{t=1}^{N} \varphi(t)y(t) \\ &= \frac{1}{N}\sum_{t=1}^{N} \begin{bmatrix} u(t) \\ u(t-T) \\ 1 \end{bmatrix} (g(u(t), u(t-T)) + e(t)) \\ &= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} \begin{bmatrix} u_i g(u_i, u_j) \\ u_j g(u_i, u_j) \\ g(u_i, u_j) \end{bmatrix} \\ &+ \frac{1}{nm^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{n} \begin{bmatrix} u_i e_{ijk} \\ u_j e_{ijk} \\ e_{ijk} \end{bmatrix},\end{aligned}$$

where e_{ijk} is the value of $e(t)$ when $u(t) = u_i$ and $u(t-T) = u_j$ for the k:th time, that is, the k:th measurement in cell ij. With vector notation;

$$\mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_m \end{bmatrix}, \qquad \mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix},$$

$$\mathbf{f} = \begin{bmatrix} f_1 \\ \vdots \\ f_m \end{bmatrix} = \frac{1}{m}\sum_{j=1}^{m} \begin{bmatrix} g(u_1, u_j) \\ \vdots \\ g(u_m, u_j) \end{bmatrix} = \frac{\mathbf{G1}}{m},$$

$$\mathbf{h} = \begin{bmatrix} h_1 \\ \vdots \\ h_m \end{bmatrix} = \frac{1}{m}\sum_{i=1}^{m} \begin{bmatrix} g(u_i, u_1) \\ \vdots \\ g(u_i, u_m) \end{bmatrix} = \frac{\mathbf{G}^T\mathbf{1}}{m},$$

$$\mathbf{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} = \frac{1}{mn}\sum_{j=1}^{m}\sum_{k=1}^{n}\begin{bmatrix} e_{1jk} \\ \vdots \\ e_{mjk} \end{bmatrix} = \frac{\mathbf{E1}}{m},$$

$$\mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{k=1}^{n}\begin{bmatrix} e_{i1k} \\ \vdots \\ e_{imk} \end{bmatrix} = \frac{\mathbf{E}^T\mathbf{1}}{m},$$

where $\mathbf{G}_{ij} = g(u_i, u_j)$ and $\mathbf{E}_{ij} = 1/n\sum_{k=1}^{n} e_{ijk}$ is the noise average in cell ij, the parameter estimates can be written as:

$$\hat{\theta}_N^{LS} = \begin{bmatrix} \frac{1}{mR_u}\mathbf{u}^T(\mathbf{f}+\mathbf{v}) \\ \frac{1}{mR_u}\mathbf{u}^T(\mathbf{h}+\mathbf{w}) \\ \frac{1}{m^2}\mathbf{1}^T(\mathbf{G}+\mathbf{E})\mathbf{1} \end{bmatrix}. \quad (4)$$

The residuals from the affine model are denoted

$$\epsilon(t) = g(u(t), u(t-T)) + e(t) - (\hat{\theta}_N^{LS})^T\varphi(t)$$

4. ANOVA APPLIED TO THE DATA DIRECTLY

The statistical analysis method ANOVA (Miller (1997); Montgomery (1991), among many others) is a widely spread tool for finding out which factors contribute to given measurements. It has been used and discussed since the 1930's and is a common tool in, e.g., medicine and quality control applications.

The method is based on hypothesis tests with F-distributed test variables computed from the residual quadratic sum. There are several slightly different variants (Miller, 1997). Here the fixed effects model with two factors will be used.

Assume that the collected measurement data can be described by a linear statistical model,

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + w_{ijk}, \quad (5)$$

where the w_{ijk} are independent Gaussian distributed variables with zero mean and constant variance σ^2. The parameter μ is the overall mean. For each level $i = 1, \ldots, m$ of the first regressor $(u(t))$ there is a corresponding effect τ_i, and for each level $j = 1, \ldots, m$ of the second regressor$(u(t-T))$ the corresponding effect is β_j. The interaction between the regressors is described by the parameters $(\tau\beta)_{ij}$. The sum of a batch of indexed parameters over any of its index is zero.

For a linear $(y(t) = au(t) + bu(t-T) + e(t))$ or a non-linear additive system $(y(t) = g_1(u(t)) + g_2(u(t-T)) + e(t))$, the interaction parameters $(\tau\beta)_{ij}$ are zero. These are needed when the non-linearities have a non-additive nature, i.e., $y(t) = g(u(t), u(t-T)) + e(t)$.

Since the regressors are quantised, it is a very simple procedure to estimate the model parameters by the computation of means. For example, the constant μ would correspond to $\overline{y}_{...}$, while the effects from the first regressor are computed as $\tau_i = \overline{y}_{i..} - \overline{y}_{...}$. In each cell ij, for the model (1),

$$y_{ijk} = g(u_i, u_j) + e_{ijk}.$$

This gives the total mean over all cells (all data)

$$\bar{y}_{...} = \frac{1}{m^2}\mathbf{1}^T(\mathbf{G}+\mathbf{E})\mathbf{1}.$$

The m different row means $\bar{y}_{i..}$, and column means $\bar{y}_{.j.}$, are given by

$$\bar{y}_{i..} = f_i + v_i = \frac{1}{m}\mathbf{i}^T(\mathbf{G}+\mathbf{E})\mathbf{1} \text{ and}$$

$$\bar{y}_{.j.} = h_j + w_j = \frac{1}{m}\mathbf{1}^T(\mathbf{G}+\mathbf{E})\mathbf{j}.$$

respectively. Here $\mathbf{i}$ and $\mathbf{j}$ are vectors with one nonzero element in row i and j respectively. $\|\mathbf{i}\| = \|\mathbf{j}\| = 1$. The m^2 different cell means are given by

$$\bar{y}_{ij.} = \mathbf{i}^T(\mathbf{G}+\mathbf{E})\mathbf{j},$$

ANOVA is used for testing which of the parameters that significantly differ from zero and for estimating the values of the parameters with standard errors, which makes it a tool for exploratory data analysis. The residual quadratic sum, SS_T, is used to design test variables for the different batches (e.g., the τ_i:s) of parameters. The total residual sum of squares is divided into the four parts

$$\begin{aligned} SS_A^d &= nm\sum_{i=1}^{m}\tau_i^2 = nm\sum_{i=1}^{m}(\bar{y}_{i..} - \bar{y}_{...})^2 \\ &= nm\sum_{i=1}^{m}(\mathbf{i}^T(\mathbf{f}+\mathbf{v}) - \frac{1}{m}\mathbf{1}^T(\mathbf{f}+\mathbf{v}))^2 \\ &= (\mathbf{f}+\mathbf{v})^T\mathbf{A}(\mathbf{f}+\mathbf{v}), \\ SS_B^d &= nm\sum_{j=1}^{m}\beta_j^2 = nm\sum_{j=1}^{m}(\bar{y}_{.j.} - \bar{y}_{...})^2 \\ &= nm\sum_{j=1}^{m}(\mathbf{j}^T(\mathbf{h}+\mathbf{w}) - \frac{1}{m}\mathbf{1}^T(\mathbf{h}+\mathbf{w}))^2 \\ &= (\mathbf{h}+\mathbf{w})^T\mathbf{A}(\mathbf{h}+\mathbf{w}), \\ SS_{AB}^d &= n\sum_{i=1}^{m}\sum_{j=1}^{m}(\tau\beta)_{ij}^2 \\ &= n\sum_{i=1}^{m}\sum_{j=1}^{m}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 \\ &= n\sum_{i=1}^{m}\sum_{j=1}^{m}((\mathbf{i} - \frac{\mathbf{1}}{m})^T\mathbf{Y}(\mathbf{j} - \frac{\mathbf{1}}{m}))^2 \\ &= \frac{1}{m}\text{trace}(\mathbf{Y}^T\mathbf{A}\mathbf{Y}) - \frac{1}{m^2}\mathbf{1}^T\mathbf{Y}^T\mathbf{A}\mathbf{Y}\mathbf{1}, \\ SS_E^d &= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{n}w_{ijk}^2 = \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2 \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{n}(e_{ijk} - \bar{e}_{ij.})^2, \end{aligned}$$

with $\mathbf{A} = nm(I - \frac{1}{m}\mathbf{1}\mathbf{1}^T)$, $\mathbf{Y} = \mathbf{G} + \mathbf{E}$ and where d stands for sums of squares computed directly from the dataset $\mathbf{Z}^N$. The index for each part is related to one batch of parameters in (5). If all the parameters in the batch are zero, the corresponding quadratic sum is χ^2-distributed if divided by the true variance σ^2 (see, e.g., Montgomery (1991, page 59)). Since the true variance is not available, the estimate $\hat{\sigma}^2 = \frac{SS_E}{m^2(n-1)}$ is used to form F-distributed test variables, e.g., for τ_i;

$$v_A^d = \frac{SS_A^d/(m-1)}{SS_E^d/(m^2(n-1))}, \quad H_{A,d}^0 : \ \tau_i = 0 \ \forall i.$$

If all the τ_i:s are zero, v_A belongs to an F-distribution with $m-1$ and $m^2(n-1)$ degrees of freedom. If any τ_i is nonzero it will give a large value of v_A^d, compared to an F-table. This is, of course, a test of the null hypothesis that all the τ_i:s are zero, which correspond to the case that the regressor $u(t)$ does not have any main effect on the measurements $y(t)$.

5. ANOVA APPLIED TO THE RESIDUALS FROM THE AFFINE MODEL

In each cell ij we have the residuals

$$\epsilon_{ijk} = g(u_i, u_j) + e_{ijk} - \hat{a}u_i - \hat{b}u_j - \hat{c},$$

where the parameters $\hat{a}$, $\hat{b}$ and $\hat{c}$ are computed according to (4). The total mean is now given by

$$\begin{aligned}\bar{\epsilon}_{...} &= \frac{1}{nm^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{n}\epsilon_{ijk} \\ &= \frac{1}{m^2}\mathbf{1}^T(\mathbf{G}+\mathbf{E})\mathbf{1} - \frac{\hat{a}}{m}\mathbf{u}^T\mathbf{1} - \frac{\hat{b}}{m}\mathbf{u}^T\mathbf{1} - \hat{c} = 0,\end{aligned}$$

where the last equality is due to (2) and (4). The row means changes to

$$\bar{\epsilon}_{i..} = \frac{1}{nm}\sum_{j=1}^{m}\sum_{k=1}^{n}\epsilon_{ijk} = \mathbf{i}^T(\mathbf{f} + \mathbf{v} - \hat{a}\mathbf{u}) - \hat{c},$$

since the sum over u_j is zero. The column means are given by

$$\bar{\epsilon}_{.j.} = \frac{1}{nm}\sum_{i=1}^{m}\sum_{k=1}^{n}\epsilon_{ijk} = \mathbf{j}^T(\mathbf{h} + \mathbf{w} - \hat{b}\mathbf{u}) - \hat{c},$$

and, finally, the cell means are given by

$$\bar{\epsilon}_{ij.} = \frac{1}{n}\sum_{k=1}^{n}\epsilon_{ijk} = \mathbf{i}^T(\mathbf{G}+\mathbf{E})\mathbf{j} - \hat{a}\mathbf{i}^T\mathbf{u} - \hat{b}\mathbf{u}^T\mathbf{j} - \hat{c}.$$

The sums of squares SS_A and SS_B are changed to

$$\begin{aligned}SS_A^r &= nm\sum_{i=1}^{m}(\bar{\epsilon}_{i..} - \bar{\epsilon}_{...})^2 \\ &= nm\sum_{i=1}^{m}((\mathbf{i}^T - \frac{1}{m}\mathbf{1}^T - \frac{\mathbf{i}^T\mathbf{u}}{mR_u}\mathbf{u}^T)(\mathbf{f}+\mathbf{v}))^2 \\ &= (\mathbf{f}+\mathbf{v})^T\mathbf{A}_1(\mathbf{f}+\mathbf{v}),\end{aligned}$$

with $\mathbf{A}_1 = nm(I - \frac{1}{m}\mathbf{1}\mathbf{1}^T - \frac{1}{mR_u}\mathbf{u}\mathbf{u}^T)$, and

$$\begin{aligned}SS_B^r &= nm\sum_{j=1}^{m}(\bar{\epsilon}_{.j.} - \bar{\epsilon}_{...})^2 \\ &= nm\sum_{j=1}^{m}((\mathbf{j}^T - \frac{1}{m}\mathbf{1}^T - \frac{\mathbf{j}^T\mathbf{u}}{mR_u}\mathbf{u}^T)(\mathbf{h}+\mathbf{w}))^2 \\ &= (\mathbf{h}+\mathbf{w})^T\mathbf{A}_1(\mathbf{h}+\mathbf{w}).\end{aligned}$$

It is easy to verify that

$$SS_{AB}^r = n\sum_{i=1}^{m}\sum_{j=1}^{m}(\bar{\epsilon}_{ij.} - \bar{\epsilon}_{i..} - \bar{\epsilon}_{.j.} + \bar{\epsilon}_{...})^2 = SS_{AB}^d \tag{6}$$

and that

$$\begin{aligned}SS_E^r &= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{n}(\epsilon_{ijk} - \bar{\epsilon}_{ij.})^2 \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{n}(e_{ijk} - \bar{e}_{ij.})^2 = SS_E^d,\end{aligned}$$

where r stands for sums of squares computed from the residuals from the affine model.

6. DIFFERENCES AND DISTRIBUTIONS

The sum of squares corresponding to the regressor $u(t)$, SS_A, changes with the following amount when an affine model is extracted from the data:

$$SS_A^d - SS_A^r = (\mathbf{f}+\mathbf{v})^T\mathbf{A}_2(\mathbf{f}+\mathbf{v}),$$

with $\mathbf{A}_2 = \mathbf{A} - \mathbf{A}_1 = \frac{n}{R_u}\mathbf{u}\mathbf{u}^T$. The change in sums of squares corresponding to the regressor $u(t-T)$ is

$$SS_B^d - SS_B^r = (\mathbf{h}+\mathbf{w})^T\mathbf{A}_2(\mathbf{h}+\mathbf{w}).$$

6.1 Distributions

From, e.g., Miller (1997), the following is known:

$$SS_A^d \sim \sigma^2\chi^2(m-1, \frac{\mathbf{f}^T\mathbf{A}\mathbf{f}}{\sigma^2}),$$

$$SS_B^d \sim \sigma^2\chi^2(m-1, \frac{\mathbf{h}^T\mathbf{A}\mathbf{h}}{\sigma^2}),$$

$SS_{AB}^d \sim$

$$\sigma^2\chi^2((m-1)^2, \frac{m\mathrm{trace}(\mathbf{G}^T\mathbf{A}\mathbf{G}) - \mathbf{1}^T\mathbf{G}^T\mathbf{A}\mathbf{G}\mathbf{1}}{m^2\sigma^2}),$$

and

$$SS_E^2 \sim \sigma^2\chi^2(m^2(n-1)),$$

where $\sim \chi^2(d, \omega)$ means distributed as a non-central Chi-square distribution with d degrees of freedom and non-centrality parameter ω. The sums of squares SS_A^d, SS_B^d, SS_{AB}^d and SS_E^d are independently distributed if the dataset is balanced, i.e., if all combinations of $u(t) = u_i, u(t-T) = u_j$ are present equally many times in the input.

To find the distributions of SS_A^r, SS_B^r, $SS_A^d - SS_A^r$ and $SS_B^d - SS_B^r$ the following theorems, numbered as in Khatri (1980), can be applied. Let $\mathbf{v} \sim N(\mu, \sigma_v^2\mathbf{V})$ and $\mathbf{q} = \mathbf{v}^T\mathbf{A}\mathbf{v} + 2\mathbf{l}^T\mathbf{v} + c$. The notation in Theorem 2 is changed from matrix valued to vector valued $\mathbf{v}$.

Theorem 2. $\mathbf{q} \sim \lambda\sigma_v^2\chi^2(d, \frac{\Omega}{\lambda^2\sigma_v^2})$ iff (i) λ is the nonzero eigenvalue of $\mathbf{VA}$ (or $\mathbf{AV}$) repeated d times, (ii) $(\mathbf{l}^T+\mu^T\mathbf{A})\mathbf{V} = \mathbf{k}^T\mathbf{VAV}$ for some vector $\mathbf{k}$ and (iii) $\Omega = (\mathbf{l}^T + \mu^T\mathbf{A})\mathbf{V}(\mathbf{l}^T + \mu^T\mathbf{A})^T$ and $\mu^T\mathbf{A}\mu+2\mathbf{l}^T\mu+c = (\mathbf{l}^T + \mu^T\mathbf{A})\mathbf{V}(\mathbf{l}^T + \mu^T\mathbf{A})^T/\lambda$. $\mathbf{q} \sim \lambda\sigma_v^2\chi^2(d)$ iff (i) $\mathbf{VAVAV} = \lambda\mathbf{VAV}$ and (ii) $(\mathbf{l}^T + \mu^T\mathbf{A})\mathbf{V} = 0 = \mu^T\mathbf{A}\mu + 2\mathbf{l}^T\mu + c$.

Theorem 4. Let $\mathbf{q}_i = \mathbf{v}^T\mathbf{A}_i\mathbf{v} + 2\mathbf{l}_i^T\mathbf{v} + c_i$, $i = 1, 2$, where $\mathbf{A}_1$ and $\mathbf{A}_2$ are symmetric matrices. Then $\mathbf{q}_1$ and $\mathbf{q}_2$ are independently distributed iff (i) $\mathbf{VA}_1\mathbf{VA}_2\mathbf{V} = 0$, (ii) $\mathbf{VA}_2\mathbf{V}(\mathbf{A}_1\mu + \mathbf{l}_1) = \mathbf{VA}_1\mathbf{V}(\mathbf{A}_2\mu+\mathbf{l}_2) = 0$ and (iii) $(\mathbf{l}_1+\mathbf{A}_1\mu)^T\mathbf{V}(\mathbf{l}_2 + \mathbf{A}_2\mu) = 0$.

Now set

$$\begin{aligned} \mathbf{q}_1 &= SS_A^r \text{ and} \quad & \mathbf{q}_2 &= SS_A^d - SS_A^r. \\ \text{Let } \mathbf{l}_1 &= \mathbf{A}_1\mathbf{f}, & c_1 &= \mathbf{f}^T\mathbf{A}_1\mathbf{f}, \\ \mathbf{l}_2 &= \mathbf{A}_2\mathbf{f}, & c_2 &= \mathbf{f}^T\mathbf{A}_2\mathbf{f} \text{ and} \\ \mathbf{v} &\sim N(0, \frac{\sigma^2}{nm}I). \end{aligned}$$

Then independence is shown by;

(i) $\mathbf{VA}_1\mathbf{VA}_2\mathbf{V} = \mathbf{A}_1\mathbf{A}_2$
$= \frac{n^2m}{R_u}(I - \frac{1}{m}\mathbf{1}\mathbf{1}^T - \frac{1}{mR_u}\mathbf{u}\mathbf{u}^T)\mathbf{u}\mathbf{u}^T = 0,$

since $\mathbf{u}^T\mathbf{u} = mR_u$ and $\mathbf{1}^T\mathbf{u} = 0$.

(ii) $\mathbf{VA}_2\mathbf{V}(\mathbf{A}_1\mu + \mathbf{l}_1) = \mathbf{A}_2\mathbf{A}_1\mathbf{f}$
$= \frac{n^2m}{R_u}\mathbf{u}\mathbf{u}^T(I - \frac{1}{m}\mathbf{1}\mathbf{1}^T - \frac{1}{mR_u}\mathbf{u}\mathbf{u}^T) = 0,$
$\mathbf{VA}_1\mathbf{V}(\mathbf{A}_2\mu + \mathbf{l}_2) = \mathbf{A}_1\mathbf{A}_2\mathbf{f} = 0,$
(iii) $(\mathbf{l}_1 + \mathbf{A}_1\mu)^T\mathbf{V}(\mathbf{l}_2 + \mathbf{A}_2\mu) = \mathbf{f}^T\mathbf{A}_1\mathbf{A}_2\mathbf{f} = 0$

for the same reasons as in (i).

Since conditions (i),(ii) and (iii) in Theorem 4 are fulfilled, $\mathbf{q}_1$ and $\mathbf{q}_2$ (that is, SS_A^r and $SS_A^d - SS_A^r$) are independently distributed. The same argument is valid for the independence of SS_B^r and $SS_B^d - SS_B^r$ if all occurrences of $\mathbf{f}$ are replaced with $\mathbf{h}$ and $\mathbf{v}$ with $\mathbf{w}$. If assumption (3) is not valid, the independence is lost.

To compute the distributions of $\mathbf{q}_1$ and $\mathbf{q}_2$ the conditions in Theorem 2 are checked:

(i) $\lambda_1 = eig(\mathbf{VA}_1) = nm$ with $d_1 = m - 2$.
$\lambda_2 = eig(\mathbf{VA}_2) = nm$ with $d_2 = 1$.

(ii) $(\mathbf{l}_1 + \mu\mathbf{A}_1)\mathbf{V} = \mathbf{f}^T\mathbf{A}_1 = \mathbf{k}^T\mathbf{VA}_1\mathbf{V}$ for $\mathbf{k} = \mathbf{f}$.
$(\mathbf{l}_2 + \mu\mathbf{A}_2)\mathbf{V} = \mathbf{f}^T\mathbf{A}_2 = \mathbf{k}^T\mathbf{VA}_2\mathbf{V}$ for $\mathbf{k} = \mathbf{f}$.

(iii) $\Omega_1 = (\mathbf{l}_1 + \mu\mathbf{A}_1)\mathbf{V}(\mathbf{l}_1 + \mu\mathbf{A}_1)^T$
$= \mathbf{f}^T\mathbf{A}_1\mathbf{A}_1\mathbf{f} = \lambda_1\mathbf{f}^T\mathbf{A}_1\mathbf{f}$ and
$\mu^T\mathbf{A}_1\mu + 2\mathbf{l}_1^T\mu + c_1 = c_1 = \mathbf{f}^T\mathbf{A}_1\mathbf{f} = \Omega_1/\lambda_1.$
$\Omega_2 = (\mathbf{l}_2 + \mu\mathbf{A}_2)\mathbf{V}(\mathbf{l}_2 + \mu\mathbf{A}_2)^T$
$= \mathbf{f}^T\mathbf{A}_2\mathbf{A}_2\mathbf{f} = \lambda_2\mathbf{f}^T\mathbf{A}_2\mathbf{f}$ and
$\mu^T\mathbf{A}_2\mu + 2\mathbf{l}_2^T\mu + c_2 = c_2 = \mathbf{f}^T\mathbf{A}_2\mathbf{f} = \Omega_2/\lambda_2.$

By Theorem 2,

$$SS_A^r = \mathbf{q}_1 \sim \sigma^2\chi^2(m-2, \frac{\mathbf{f}^T\mathbf{A}_1\mathbf{f}}{\sigma^2}),$$

and

$$SS_A^d - SS_A^r = \mathbf{q}_2 \sim \sigma^2\chi^2(1, \frac{\mathbf{f}^T\mathbf{A}_2\mathbf{f}}{\sigma^2}).$$

As before, all SS_A can be replaced by SS_B if $\mathbf{f}$ is replaced by $\mathbf{h}$.

6.2 Interpretation

There are five test variables of interest:

$$\begin{aligned} v_{AB} &= \frac{SS_{AB}/(m-1)^2}{SS_E/m^2(n-1)}, \quad & H_{AB}^0 &: (\tau\beta)_{ij} = 0 \ \forall i, \\ v_A^d &= \frac{SS_A^d/(m-1)}{SS_E/m^2(n-1)}, \quad & H_{A,d}^0 &: \tau_i = 0 \ \forall i, \\ v_B^d &= \frac{SS_B^d/(m-1)}{SS_E/m^2(n-1)}, \quad & H_{B,d}^0 &: \beta_j = 0 \ \forall j, \\ v_A^r &= \frac{SS_A^r/(m-2)}{SS_E/m^2(n-1)}, \quad & H_{A,r}^0 &: \tau_i = 0 \ \forall i, \\ v_B^r &= \frac{SS_B^r/(m-2)}{SS_E/m^2(n-1)}, \quad & H_{B,r}^0 &: \beta_j = 0 \ \forall j. \end{aligned}$$

All of these belong to F-distributions if the corresponding null hypotheses are true, that is, large values of the test variables (compared to an $F(d_1, d_2)$-table) are interpreted as that there are effects from the corresponding regressor. If v_{AB} is large an interaction effect between $u(t)$ and $u(t-T)$ is assumed. This means that the system can not be decomposed into additive subsystems. If v_{AB} is small, the null hypothesis can not be rejected, so it is assumed that the system can be decomposed into additive subsystems. For both v_A^d and v_A^r large, the interpretation is that the effect from $u(t)$ is nonlinear. If v_A^d is large, but v_A^r is small, the effect from $u(t)$ can be described by the linear model. If both v_A^d and v_A^r are small, $u(t)$ can not be shown to affect the output of the system. The same reasoning built on v_B^d and v_B^r is valid for the effects from $u(t-T)$.

Since SS_{AB} is not changed when the linear model is extracted from the data, see (6), we can make the conclusion that all information about the interactions in the system is left in the residuals. The interaction information is not destroyed by subtracting a linear model in a balanced dataset.

6.3 Linear example

Let the true system be given by

$$y(t) = au(t) + bu(t-T) + e(t).$$

Then $\mathbf{f} = a\mathbf{u}$ and $\mathbf{h} = b\mathbf{u}$. This gives

$$SS_{AB} \sim \sigma^2\chi^2((m-1)^2, 0),$$
$$SS_A^d \sim \sigma^2\chi^2(m-1, nm^2a^2R_u/\sigma^2), \quad (7)$$
$$SS_B^d \sim \sigma^2\chi^2(m-1, nm^2b^2R_u/\sigma^2), \quad (8)$$

and

$$SS_A^r, SS_B^r \sim \sigma^2\chi^2(m-2, 0). \quad (9)$$

The size of the non-centrality parameters in (7) and (8) depends on how many data are collected, the size of the true linear effects and the variance of the input and the noise. This is what effects the power of the F-tests. In the sums of squares computed from the residuals from the linear model all dependence on the true model is removed (9). Thus the conclusion can be made; that if SS_A^r or SS_B^r are found to be large by the F-tests, then the data are probably not collected from a true linear system.

6.4 Quadratic example

Let $y(t) = u^2(t) + e(t)$. Then $\mathbf{f} = [u_1^2, \ldots, u_m^2]^T$, $\mathbf{h} = \mathbf{0}$, and

$$SS_{AB} \sim \sigma^2\chi^2((m-1)^2, 0),$$
$$SS_A^d \sim \sigma^2\chi^2(m-1, nc_d),$$
$$SS_B^d \sim \sigma^2\chi^2(m-1, 0),$$
$$SS_A^r \sim \sigma^2\chi^2(m-2, nc_r),$$
$$SS_B^r \sim \sigma^2\chi^2(m-2, 0)$$

with

$$nc_d = \frac{n}{\sigma^2}(m\sum_{i=1}^{m} u_i^4 - R_u^2), \text{ and}$$
$$nc_r = \frac{n}{\sigma^2}(m\sum_{i=1}^{m} u_i^4 - R_u^2 - \frac{1}{R_u}(\sum_{i=1}^{m} u_i^3)^2)$$

Here, it is clear that it matters how the levels of the input signal are chosen. $\sum_{i=1}^{m} u_i^3$ can vary considerably while (2) is valid, since there are no constraints on the level distribution. Also $\hat{a}$ is proportional to $\sum_{i=1}^{m} u_i^3$, so a large difference between ANOVA directly and ANOVA applied to the residuals means that the nonlinear effect have been picked up by the affine model, due to irregular sampling of the function. If $u(t)$ is symmetric around its mean, it is clear that $\sum_{i=1}^{m} u_i^3 = 0$, so $nc_d = nc_r$.

7. CONCLUSIONS

Applying ANOVA directly on a dataset was compared to applying ANOVA on the residuals from a linear model estimated with linear least squares. The distributions for the sums of squares needed for the ANOVA analysis in the latter case were computed. These distributions were used to show that by combining the two approaches ANOVA is effective in finding what regressors give linear effects and what regressors give nonlinear effects. In section 6.2 it was shown how to divide the regressors into a linear and a nonlinear subset, depending on the outcome of the ANOVA tests.

The ability to structure the proposed nonlinear function into additive parts depending on only subsets of regressors is an ANOVA feature which is not affected by subtraction of a linear model. The results in the paper extend to more regressors, but an important limitation is that the dataset should be balanced, see section 2.

This work has been supported by the Swedish Research Council(VR) which is gratefully acknowledged.

REFERENCES

K. Godfrey. *Perturbation Signals for System Identification*. Prentice Hall, New York, 1993.

C. G. Khatri. Quadratic forms in normal variables. In P.R. Krishnaiah, editor, *Handbook of Statistics*, volume 1, pages 443–469, Amsterdam, 1980. North-Holland.

I. Lind. Model order selection of N-FIR models by the analysis of variance method. In *Proc IFAC Symposium SYSID 2000*, pages 367–372, Santa Barbara, Jun 2000.

I. Lind and L. Ljung. Structure selection with ANOVA: Local linear models. In P. van der Hof, B. Wahlberg, and S. Weiland, editors, *Proc. 13th IFAC Symposium on System Identification*, pages 51 – 56, Rotterdam, the Netherlands, aug 2003.

I. Lind and L. Ljung. Regressor selection with the analysis of variance method. *Automatica*, 41(4): 693–700, Apr 2005.

L. Ljung. *System Identification, Theory for the User*. Prentice Hall, New Jersey, 2nd edition, 1999.

L. Ljung, Q. Zhang, P. Lindskog, and A. Juditsky. Modeling a non-linear electric circuit with black box and grey box models. In *Proc. NOLCOS 2004 - IFAC Symposium on Nonlinear Control Systems*, pages 543–548, Stuttgart, Germany, Sep 2004.

R.G. Miller, Jr. *Beyond ANOVA*. Chapman and Hall, London, 1997.

D.C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, New York, 3rd edition, 1991.