Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 48-6 (2015) 154–158

Selecting Transients Automatically for the Identification of Models for an Oil Well

Antônio H. Ribeiro & **Luis A. Aguirre**

*Department of Electronic Engineering, Universidade Federal de Minas Gerais (UFMG). Av. Antônio Carlos 6627, Belo Horizonte, MG, Brazil (*antonio-ribeiro@ufmg.com.br, aguirre@cpdee.ufmg.br*)*

Abstract: This paper proposes a procedure to automatically select transient windows for system identification from routine operation data. To this end two metrics are proposed. One quantifies the transient content in a given window and the other provides an overall measure of correlation between such transients and the chosen model input. The procedure is illustrated using data from an oil well that operates in deep waters.

Keywords: Automatic transient selection, system identification, soft sensors, intelligent oil fields

1. INTRODUCTION

The set of techniques for constructing models from observed data is known, in the control theory context, as *System Identification*.

To construct models from a data set it is necessary to have a set of data that contains relevant information about the system. Some times only "routine operation" data is available and, other times, it will be necessary to create tests for extract dynamic information about the system.

There are tests that excite a wide range of system frequencies (e.g. pseudorandom binary signal (PRBS)) and, therefore, are proper for linear system identification since the output data obtained will contain significant information about the system dynamics.

For nonlinear models there is a need to drive the system over a wider range of amplitudes and PRBS may not be the best choice as shown by Leontaritis and Billings (1987). Classical system identification textbooks offer some practical guidance in what concerns testing (Ljung, 1987).

Sometimes it will not be possible to perform experiments on the system and historical data is used. Since the data is recorded during "routine operation", the system will probably be in steady state most of the time and there will be few data windows containing relevant dynamical information.

Therefore in practical problems of system identification from routine operation data the choice of informative windows of data are both subjective and greatly time consuming. Hence this paper puts forward a criterion to aid in the choice of informative windows of data from a large data set. The method was developed for an oil well (Teixeira et al., 2014), but should also be useful in other applications.

In this paper, two metrics are proposed to classify windows from the oil well recorded data with the identification in view. Each metric addresses one of the following points:

(1) The transient is appropriate for identification only if it contains relevant information about the system dynamics;
(2) The output should be well correlated with the input, otherwise the transient is caused by an unmeasured disturbance and the window are not appropriate for identification of an input-output model.

Using such metrics it is possible to create automatic routines that choose good transients for identification.

The remainder of the paper is organized as follows. In Section 2 some mathematical concepts are quickly revisited. These concepts will be used on Sections 3 and 4 to define the metrics. In Section 5 the use of the metrics is illustrated considering a real numerical problem, and finally, some concluding remarks are provided in Section 6

2. MATHEMATICAL BACKGROUND

2.1 Singular Value Decomposition (SVD)

Theorem 1. Any $m \times n$ matrix $\mathbf{A}$ with rank r can be factored as (Strang, 1988):

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T,$$

where $\mathbf{U}$ $(m \times m)$ and $\mathbf{V}$ $(n \times n)$ are both ortogonal [1]. And Σ is a $m \times n$ matrix containing r elements on its main diagonal:

$$\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r).$$

The scalars $\{\sigma_1, \sigma_2, \ldots, \sigma_r\}$ are the singular values of $\mathbf{A}$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.

Proof. This is shown in (Strang, 1988, pp.450-451).

The key ideia behind the SVD factorization is to take into consideration the diagonalization of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$:

Corollary 2. The columns of $\mathbf{U}$ are the eigenvectors of the matrix $\mathbf{A}\mathbf{A}^T$ and the columns of $\mathbf{V}$ are the eigenvectors of

[1] The matrix U is ortogonal if and only if $UU^T = I$

10.1016/j.ifacol.2015.08.024

the matrix $\mathbf{A}^T\mathbf{A}$. The singular values $\{\sigma_1, \sigma_2, \ldots, \sigma_r\}$ are the square roots of the nonzero eigenvalues of both $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$.

This can be easily understood considering:

$$\mathbf{A}\mathbf{A}^T = (\mathbf{U}\Sigma\mathbf{V}^T)(\mathbf{V}\Sigma^T\mathbf{U}^T) = \mathbf{U}\Sigma\Sigma^T\mathbf{U}^T,$$

and interpreting $\mathbf{U}\Sigma\Sigma^T\mathbf{U}^T$ as the diagonalization of the symmetric matrix $\mathbf{A}\mathbf{A}^T$. The diagonal matrix containing the eigenvalues of $\mathbf{A}\mathbf{A}^T$ is:

$$\Sigma\Sigma^T = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2),$$

and the columns of $\mathbf{U}$ are its eigenvectors.

Analogously, the columns of $\mathbf{V}$ are eigenvectors of $\mathbf{A}^T\mathbf{A}$ and its eigenvalues are the squared singular values.

2.2 Cross Correlation Function

Given two stationary signals $y(k)$ and $u(k)$, the cross correlation function measures the similarity between u and copies of y shifted (lagged) by τ. It is defined as the expected value ($\text{E}\{\cdot\}$) of u times a shifted copy of y:

$$r_{uy}(\tau) = \text{E}\{u(t)y(t+\tau)\}, \tag{1}$$

The cross correlation function can be estimated for a finite time series with m samples by:

$$\hat{r}_{uy}(\tau) = \frac{1}{m}\sum_{k=1}^{m} u(k)y(k+\tau). \tag{2}$$

An important practical consideration is how large the cross correlation should be so it indicate statistically significant correlation between two signals. And, in fact, for two uncorrelated signals $u(t)$ and $y(t)$ (where the expected cross correlation is zero) there is a 95% probability that the *estimated* normalized cross correlation falls within the *confidence interval*:

$$-\frac{1.96}{\sqrt{m}} \leq \rho_{uy}(\tau) \leq \frac{1.96}{\sqrt{m}}, \tag{3}$$

where $\rho_{uy} = \hat{r}_{uy}/\sigma_u\sigma_y$ is the normalized cross correlation (σ stands for standard deviation).

Hence if the cross correlation is outside the confidence interval at some lag (τ) it is fair to say that the two signals have a high probability to be correlated at lag τ.

The correlation between a stationary signal $y(t)$ and a shifted version of itself is known as autocorrelation and is denoted by:

$$r_{yy}(\tau) = \text{E}\{y(t)y(t+\tau)\}. \tag{4}$$

All previous considerations are equally valid for the autocorrelation function.

3. DYNAMIC BASED METRICS

More informative signals are better suited for identification and will yield better parameter estimation. The next section will discuss how to adjust an autoregressive model to the signal y and a clear way to evaluate this signal activity will arise as consequence.

3.1 Autoregressive (AR) Models and Regressor Matrix

A linear *autoregressive* (AR) model is defined as

$$y(k) = a_1y(k-1)+a_2y(k-2)+\cdots+a_ny(k-n)+e(k), \tag{5}$$

where $e(k)$ is white noise and the scalar parameters $\{a_1, a_2, \ldots, a_n\}$ may be estimated from recorded data. If the signal y is known from the instant $k = 1$ to the instant $k = m$, then

$$\begin{aligned}
y(1) &= a_1y(0) + a_2y(-1) + \cdots + a_ny(-n+1) + e(1)\\
y(2) &= a_1y(1) + a_2y(0) + \cdots + a_ny(-n+2) + e(2)\\
&\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots\\
y(m) &= a_1y(m-1) + a_2y(m-2) + \cdots + a_ny(m-n) + e(m),
\end{aligned} \tag{6}$$

which can be rewritten in the matrix form as:

$$\mathbf{A}\mathbf{x} + \mathbf{e} = \mathbf{y}, \tag{7}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ is called the *AR regressor matrix*. Such matrix and vectors have the following structure:

$$\mathbf{A} = \begin{bmatrix} y(0) & y(-1) & \ldots & y(-n+1) \\ y(1) & y(0) & \ldots & y(-n+2) \\ \vdots & \vdots & \ddots & \vdots \\ y(m-1) & y(m-2) & \ldots & y(m-n) \end{bmatrix};$$

$$\mathbf{y} = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(m) \end{bmatrix}; \quad \mathbf{e} = \begin{bmatrix} e(1) \\ e(2) \\ \vdots \\ e(m) \end{bmatrix}; \quad \mathbf{x} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}.$$

The AR regressor matrix may be written as:

$$\mathbf{A} = [\mathbf{A}_1 \quad \mathbf{A}_2 \quad \ldots \quad \mathbf{A}_n].$$

where $\mathbf{A}_i \in \mathbb{R}^m$ is the i-th column of matrix $\mathbf{A}$. Equation 7 may be rewritten as:

$$[\mathbf{A}_1 \quad \mathbf{A}_2 \quad \ldots \quad \mathbf{A}_n] \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} + \mathbf{e} = \mathbf{y},$$

or, equivalently as $a_1\mathbf{A}_1 + a_2\mathbf{A}_2 + \cdots + a_n\mathbf{A}_n + \mathbf{e} = \mathbf{y}$.

The *range* of matrix a $\mathbf{A}$ is defined as the vector space that contais all possible results of $\mathbf{A}\mathbf{x}$ or, in other words, all vectors that may be written as linear combinations of the columns of $\mathbf{A}$: $a_1\mathbf{A}_1 + a_2\mathbf{A}_2 + \cdots + a_n\mathbf{A}_n$.

Because of the noise vector $\mathbf{e}$, the output vector $\mathbf{y}$ does not belong to the *range* space of the matrix $\mathbf{A}$ and there is no general exact solution to (7). Its is generally possible to find a solution $\hat{\mathbf{x}}$ in the *least square* sense, which corresponds to the orthogonal projection $\hat{\mathbf{y}}$ of $\mathbf{y}$ onto the *range* space of $\mathbf{A}$. In this case the solution $\hat{\mathbf{x}}$ is such that:

$$\mathbf{A}\hat{\mathbf{x}} = \hat{\mathbf{y}},$$

or, equivalently $\hat{a}_1\mathbf{A}_1 + \hat{a}_2\mathbf{A}_2 + \cdots + \hat{a}_n\mathbf{A}_n = \hat{\mathbf{y}}$.

The next theorem states the importance of the rank r of the matrix $\mathbf{A}$ on the parameter estimates.