INTRODUCTION (1)

- Advantages of nonlinear polynomial predictors:
  - Linear predictors constitute a subset of polynomial predictors
  - Linear with respect to their parameters
- Disadvantages
  - Non parsimonious
  - Their estimation may involve ill-conditioned matrices
  - Not stable for generation

GENERALITIES (1)

- In the most general way, a causal polynomial model (also called NARMA, nonlinear ARMA) is expressed by [1]:

$$x_n = \sum_{i=0}^{d} f_i[\varepsilon_n, \varepsilon_{n-1}, \cdots, \varepsilon_{n-s}, x_{n-1}, \cdots, x_{n-t}]$$

Where functions $f_i(.)$ are ith-order polynomials with respect to the variables.

GENERALITIES (2)

- If the model is stable (in the sense that a bounded input $\{\varepsilon_n\}$ induces a bounded output $\{x_n\}$), it admits a convergent Volterra expansion expressed by:

$$x_n = h_0 + \sum_{s_1=0}^{S_1-1} h_1(s_1)\varepsilon_{n-s_1} + \sum_{s_1=0}^{S_1-1}\sum_{s_2=0}^{S_2-1} h_2(s_1,s_2)\varepsilon_{n-s_1}\varepsilon_{n-s_2}$$

$$+\cdots+ \sum_{s_1=0}^{S_1-1}\cdots\sum_{s_k=0}^{S_k-1} h_2(s_1,\cdots,s_k)\varepsilon_{n-s_1}\cdots\varepsilon_{n-s_k} +\cdots$$

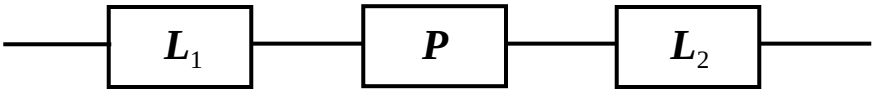
GENERALITIES (3)

- As for linear models, there are recursive realizations (NARMA) and non-recursive ones (truncated Volterra expansions).

- Polynomial models are generally used for prediction only, because they are unstable for inputs with non-bounded support (such as Gaussian ones). However, for the subset of bilinear models, stability conditions have been derived in some specific cases.

GENERALITIES (4)

- Another subset of simpler models is the LNL (linear-nonlinear-linear), which is a cascade:



with $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$ llinear filters and $\boldsymbol{P}$ a memoryless polynomial function

- Two sub-subsets are the Wiener (LN) and the Hammerstein (NL) ones.

GENERALITIES (5)

- Of course, we will focus here on polynomial NAR models expressed by:

$$x_n = \sum_{i=0}^{d} f_i[x_{n-1},\cdots,x_{n-t}] + \varepsilon_n$$

- In a MMSE setting, we will use the fact that polynomials have the property of universal approximation of continuous functions to use them to approximate the conditional expectation $\mathsf{E}[x_n|\boldsymbol{X}_{n-1}]$.

ESTIMATION (1)

- It is of course possible to derive a statistical expression of the estimation of a polynomial model, but this presents little interest. We shall deal only with the estimation of a predictor from N samples $\{x_n\}$, $n = 1, \ldots, N$.

- We suppose the maximum degree d, and the maximum delay t settled beforehand.

ESTIMATION (2)

- Since the polynomial tems are linear with repect to their parameters, least-square estimation amounts to write the model equation for all possible values of n, (i.e. $n = t+1$ to $n = N$). This lead to a matrix equation:

$$\mathbf{X}\boldsymbol{g} = \boldsymbol{x} + \boldsymbol{e}$$

- Let us illustrate that with a very simple example:

$$d = 2, t = 1$$

ESTIMATION (3)

Possible terms are 1, x_n, x_{n-1}, $x_n x_{n-1}$, x_n^2, x_{n-1}^2, and the equation to be solved in the LS sense is:

$$\begin{bmatrix} 1 & x_2 & x_1 & x_2x_1 & x_2^2 & x_1^2 \\ 1 & x_3 & x_2 & x_3x_2 & x_3^2 & x_2^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & x_{N-1} & x_{N-2} & x_{N-1}x_{N-2} & x_{N-1}^2 & x_{N-2}^2 \end{bmatrix} \begin{bmatrix} g_0 \\ g_1(1) \\ g_1(2) \\ g_2(1,2) \\ g_2(1,1) \\ g_2(2,2) \end{bmatrix}$$

ESTIMATION (4)

$$= \begin{bmatrix} x_3 \\ x_4 \\ \vdots \\ x_N \end{bmatrix} + \begin{bmatrix} e_3 \\ e_4 \\ \vdots \\ e_N \end{bmatrix}$$

- Two problems:
  - It is the simplest model but it contains 6 terms already. Some selection must take place.
  - Some of the columns of **X** risk to be almost linearly dependent.

ESTIMATION (5)

- The classical least-square solution:

$$\boldsymbol{g} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\boldsymbol{x}$$

may cause trouble because may be ill-conditioned, i.e., its eigenvalue spread may be large due to the presence of small eigenvalues.

- It is highly advised to use a robust numerical method such as Singular Value Decomposition (SVD) to solve the polynomial least-square problem.

ESTIMATION (6)

- SVD [2]: Any matrix $\mathbf{X}$ of size $K \times M$ and rank r can be written as:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}}$$

with $\mathbf{U}$ an orthogonal $K \times K$ matrix, $\mathbf{V}$ an orthogonal $M \times M$ matrix, and $\mathbf{S}$ a $K \times M$ matrix with only r non-zero elements $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$ (singular values) on its diagonal. The LS solution can be shown to be:

$$\boldsymbol{g} = \sum_{i=1}^{r} \frac{\mathbf{u}_i^{\mathsf{T}}\boldsymbol{x}}{\sigma_i}\mathbf{v}_i$$

ESTIMATION (7)

- In practical problems **X** is full rank and the M singular values are non zero. However, some are quite small.
- It can be shown that an efficient solution consists in considering an *effective rank* r_e instead of r, chosen as the smallest index i such that:

$$\frac{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_i^2}{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_M^2} > \theta \qquad \theta = 0.95 \text{ typically}$$

MODEL SELECTION (1)

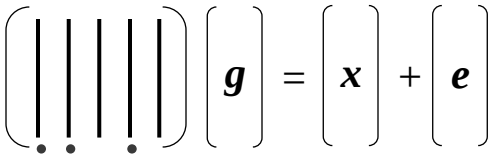
- An appropriate model can be selected using the MDL criterion :

$$\text{MDL} = N\ln\left(\hat{\sigma}_e^2\right) + M\ln(N)$$

with M the number of terms retained and $\hat{\sigma}_e^2$ the estimate of the error variance. Note it is a simplified criterion since encoding of the terms retained should also be included.

MODEL SELECTION (2)

- From the matrix equation viewpoint, model selection amounts to selecting columns of **X**:



One must obviously change the dimension of $\boldsymbol{g}$ also. Theoretically, one must apply MDL on all possible models (exhautstive search) or use sub-optimal empirical procedures.

MODEL SELECTION (3)

- For instance, one can use a genetic algorithm [3]: a population of models (coded with a bit string) evolves along the principle of *survival of the fittest*.
- The best individuals (with respect to MDL) are allowed to perpetuate, reproduce and mutate in the next generation.
- Typically, good solutions appear fast, but optimality cannot be guaranteed.

SEQUENTIAL MODEL SELECTION (1)

- Another approach, that gives good results in practice, is to select the terms sequentially [4].
- At each stage, the term producing maximum error variance reduction is included. This is the principle of *Matching Pursuit*, which is used for instance in some time-frequency techniques.
- This approach is quite attractive from a computational viewpoint, but is sub-optimal.

SEQUENTIAL MODEL SELECTION (2)

- The equation describing the model can be presented as:

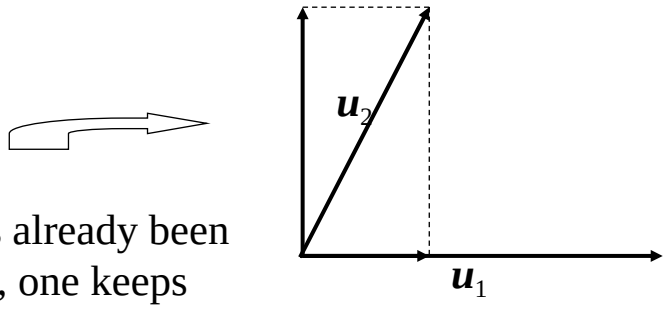
$$x_n = \sum_{m=1}^{M} c_m p_m(n) + e(n)$$

where each $p_m(.)$ is a polynomial term (such as $x_n x_{n-1}$ for instance), and corresponds this to a column of matrix **X**.

- These columns are not orthogonal so sequential selection imposes *orthogonalization*.

SEQUENTIAL MODEL SELECTION (3)

- Gram-Schmidt orthogonalization



If $\boldsymbol{u}_1$ has already been selected, one keeps from $\boldsymbol{u}_2$ the part orthogonal to $\boldsymbol{u}_1$ only. This procedure is easily generalized to a succession of vectors.

SEQUENTIAL MODEL SELECTION (4)

- Gram-schmidt procedure transforms an arbitrary basis $\{\boldsymbol{u}_i\}$ into an orthogonal basis $\{\boldsymbol{v}_i\}$ with:

$$\boldsymbol{v}_1 = \boldsymbol{u}_1$$

$$\alpha_{ki} = \frac{\langle \boldsymbol{u}_k, \boldsymbol{v}_i \rangle}{\langle \boldsymbol{v}_i, \boldsymbol{v}_i \rangle}, \quad 1 \le i < k$$

$$\boldsymbol{v}_k = \boldsymbol{u}_k - \sum_{i=1}^{k-1} \alpha_{ki} \boldsymbol{v}_i$$

SEQUENTIAL MODEL SELECTION (5)

- Building an orthogonal basis avoids to re-compute the coefficients already obtained. Projection of vector $\boldsymbol{x}$ on the orthogonal basis means for i^{th} coefficient :

$$a_i = \frac{\langle \boldsymbol{x}, \mathbf{v}_i \rangle}{\langle \mathbf{v}_i, \mathbf{v}_i \rangle}$$

that does not depend on the other ones.

SEQUENTIAL MODEL SELECTION (6)

- As a matter of fact solving

$$\mathbf{X}\boldsymbol{g} = \boldsymbol{x} + \boldsymbol{e}$$

In the least-square sense amounts to an orthogonal projection of $\boldsymbol{x}$ on the subspace generated by the columns of $\mathbf{X}$, with $\boldsymbol{e}$ the projection error.

- Thus one orthogonalizes these columns, so as to optimize the projection on a smaller subspace spanned by the most important orthogonal basis elements.

SEQUENTIAL MODEL SELECTION (7)

- Summary

1. Determine vector $\boldsymbol{x}$ of all samples x_n for which prediction is possible. Typically

$$\boldsymbol{x} = [x_{t+1}, x_{t+2}, \ldots, x_N]^{\mathsf{T}}$$

2. Determine all columns $\boldsymbol{p}_m$ of $\mathbf{X}$, i.e. all possible polynomial terms $p_m(.)$.
3. Find the column k of $\mathbf{X}$ minimizing the sum of squared errors:

SEQUENTIAL MODEL SELECTION (8)

$$\left\| \boldsymbol{x} - \alpha_k \boldsymbol{p}_k \right\|^2 = \sum_{n=t+1}^{N} \left[x_n - \alpha_k p_k(n) \right]^2 = \sum_{n=t+1}^{N} \left[e_n \right]^2$$

with

$$\alpha_k = \frac{\langle \boldsymbol{y}, \boldsymbol{p}_k \rangle}{\langle \boldsymbol{p}_k, \boldsymbol{p}_k \rangle} = \frac{\sum_{n=n_0}^{N} y(n)\, p_k(n)}{\sum_{n=n_0}^{N} \left[p_k(n) \right]^2}$$

SEQUENTIAL MODEL SELECTION (9)

3. One sets $q_1(.) = p_k(.)$, $\boldsymbol{q}_1 = \boldsymbol{p}_k$, et $a_1 = \alpha_k$. The reduction in least-square error is:

$$a_1^2 \|\boldsymbol{g}_1\|^2 = a_1^2 \sum_{n=t+1}^{N} [g_1(n)]^2$$

4. For the i^{th} term, one finds column $\boldsymbol{p}_k$ of $\mathbf{X}$ such that:

$$\boldsymbol{q}_i = \boldsymbol{p}_k - \sum_{l=1}^{i-1} \gamma_l \boldsymbol{q}_l \qquad \gamma_l = \frac{\langle \boldsymbol{p}_k, \boldsymbol{q}_l \rangle}{\langle \boldsymbol{q}_l, \boldsymbol{q}_l \rangle}$$

SEQUENTIAL MODEL SELECTION (10)

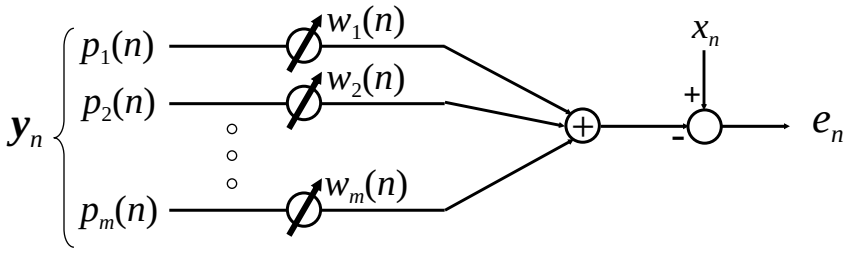
Produces maximum LS error reduction:

$$\alpha_i^2 \|\boldsymbol{q}_i\|^2 \quad \text{with} \quad \alpha_i = \frac{\langle \mathbf{y}, \boldsymbol{q}_i \rangle}{\langle \boldsymbol{q}_i, \boldsymbol{q}_i \rangle}$$

4. One increments i to $i + 1$ and back to step 3. Recursion may be stopped if the error norm stops decreasing or, preferably, if MDL starts to increase.

ADAPTIVE POLYNOMIAL PREDICTION (1)

- Since polynomial predictors are linear with respect to their coefficients, one may be interested to perform an adaptive prediction:



ADAPTIVE POLYNOMIAL PREDICTION (2)

- The most popular algorithm is the LMS, which is a gradient algorithm on $\mathsf{E}[e_n^2]$ with an instantaneous estimate e_n^2:

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n - \mu\hat{\nabla}_n$$
$$= \boldsymbol{w}_n + 2\mu e(n)\boldsymbol{y}_n$$

- But LMS convergence is highly influenced by the eigenvalue spread of the input vector covariance matrix (i.e. correlation between inputs).

ADAPTIVE POLYNOMIAL PREDICTION (3)

- Unfortunately polynomial terms are often correlated, so other approaches:
  - Recursive least-squares (RLS)
  - Lattice LMS

Are generally more efficient [5].

REFERENCES

1. V. J. Mathews and G. L. Sicuranza, *Polynomial Signal Processing*, Wiley, NY, 2000.
2. G. H. Golub and C. F. Van Loan, *Matrix Computations*, John Hopkins Univ. Press, 3rd Ed., 1996.
3. J.-M. Vesin and R. Grueter, "Model selection using a simplex reproduction genetic algorithm," *Signal Processing*, vol. 7, no. 3, June 1999.
4. M. J. Korenberg and L. D. Paarmann, « Orthogonal approaches to time-series analysis and system identification," *IEEE Sig. Proc. Mag.*, vol. 8, pp. 29-43, July 1991.
5. V. J. Mathews, "Adaptive polynomial filters, ," *IEEE Sig. Proc. Mag.*, vol. 8, pp. 10-26, July 1991.