

Experiments with neural network for modeling of nonlinear dynamical systems: Design problems

Ewa Skubalska-Rafajłowicz

**Wrocław University of Technology, Wrocław,
Wrocław, Poland**

Summary

Introduction

Preliminaries

Motivations for using random projections

Model 1 NFIR with random projections

Information matrix for Model 1

D-optimal inputs for Model 1

Simulation studies for Model 1

Brief outline of Model 2 – NARX with random projections

Experiment design for Model 2

Conclusions

Our aims in this lecture include:

- ▶ providing a brief introduction to modeling dynamical systems by neural networks
- ▶ focusing on neural networks with **random projections**,
- ▶ on their estimation and optimal input signals.

Preliminaries

We start with a time-invariant continuous time dynamical system (MISO)

$$\mathbf{x}'(t) = \mathbf{F}(\mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{y}(t) = \mathbf{H}(\mathbf{x}(t)), \quad \mathbf{x} \in \mathbb{R}^D, \quad \mathbf{u} \in$$

or discrete time system:

$$\mathbf{x}_{n+1} = \mathbf{f}(\mathbf{x}_n, \mathbf{u}_n), \quad \mathbf{y}_n = \mathbf{h}(\mathbf{x}_n).$$

Non-autonomous dynamical system ==
system of differential equations with
exogenous (external) inputs.

Observations with (usually white) noise:

$$\mathbf{y}_n = \mathbf{h}(\mathbf{x}_n) + \boldsymbol{\epsilon}_n.$$

Artificial neural networks at glance:

Neural networks are (general) nonlinear black-box structures.

- ▶ Classical neural network architectures (**feed-forward structures**):
 - ▶ multi-layer perceptron with one hidden layer and sigmoid activation functions (MLP),
 - ▶ radial bases function networks (RBF),
 - ▶ orthogonal (activation functions) neural networks (wavelets networks, fourier networks)
 - ▶ spline networks
- ▶ Other classes of NN:
 - ▶ support vector machines,
 - ▶ Kohonen nets (based on vector quantization),
 - ▶ based on Kolmogorov Representation Thm. (Sprecher networks and others)

Dynamical neural networks:

1. **networks with lateral or feedback connections,**
 - ▶ Hopfield networks (associative memory),
 - ▶ Ellman nets (context sensitive)
2. **networks with external dynamics – NARMAX, NARX, NFIR (see below).**

Aims of modeling dynamical systems:

- ▶ **simulation – estimate outputs when only inputs are available,**
- ▶ **prediction – also last observations of outputs y_n, \dots, y_{n-p} are given.**

NFIR and NARX models:

We focus our attention on two stable models
(external dynamics approach):

NFIR (nonlinear finite impulse response)

$$y_{n+1} = g(u_n, \dots, u_{n-r}),$$

where g is a continuous (smooth) function.

NARX (nonlinear autoregressive model with external inputs),

$$y_{n+1} = G(y_n, \dots, y_{n-p}, u_n, \dots, u_{n-r}),$$

where G is a continuous (smooth) function.

Further remarks on NN:

NN are usually **non-linear-in-the-parameters**.

Parametric vs nonparametric approach ??

If we allow a net to grow with the number of observations, they are sufficiently rich to approximate smooth functions. In practice, **finite structures** are considered, leading to a parametric (weights) optimizing approach.

Learning (training) == selecting weights, using nonlinear LMS (using Levenberg - Marquardt local optimization algorithm).

Our idea 1: some simplifications may lead to linear-in-the-parameter network structures.

Further remarks on NN 2:

By a feed-forward NN we mean a real valued function on \mathbb{R}^d :

$$g_M(x; \bar{w}, \bar{\theta}) = \theta_0 + \sum_{j=1}^M \theta_j \varphi(\langle w_j, x \rangle),$$

where $x \in \mathcal{K} \subset \mathbb{R}^d$, \mathcal{K} is a compact set.

Activation functions $\varphi(t)$ is a sigmoid function, e.g, logistic – $\frac{1}{1+\exp(-t)}$

hiperbolic tan – $\tanh(t) = \frac{1-\exp(-2t)}{1+\exp(-2t)}$

$\arctan \frac{2}{\pi} \arctan(t)$

Further remarks on NN 3:

Universal approximation property (UAP):

Certain classes of neural network models are shown to be universal approximators, in the sense that:

- for every continuous function $f : \mathbb{R}^d \rightarrow \mathbb{R}$,
- for every compact $\mathcal{K} \subset \mathbb{R}^d$, $\forall \epsilon > 0$,
a network of appropriate size M and a corresponding set of parameters (weights) $\bar{\mathbf{w}}, \bar{\boldsymbol{\theta}}$ exist, such that

$$\sup_{\mathbf{x} \in \mathcal{K}} ||f(\mathbf{x}) - g_M(\bar{\mathbf{w}}, \bar{\boldsymbol{\theta}}; \mathbf{x})|| < \epsilon.$$

Further remarks on NN 4:

Having a learning sequence (x_n, y_n) , $n = 1, 2, \dots, N$, the weights $\bar{w}, \bar{\theta}$ are usually selected by minimization of

$$\sum_{n=1}^N [y_n - g_M(\bar{w}, \bar{\theta}; x_n)]^2 \quad (1)$$

w.r.t. $\bar{w}, \bar{\theta}$. This is frequently complicated – by spurious local minima – iterative optimization process of searching global minimum. It can be unreliable when dimensions of $\bar{w}, \bar{\theta}$ are large, as in modeling dynamical systems.

Further remarks on NN 5:

Our idea is to **replace \bar{w}** in

$$g_M(x; \bar{w}, \bar{\theta}) = \theta_0 + \sum_{j=1}^M \theta_j \varphi(< w_j, x >), \quad (2)$$

by **randomly selected vectors \bar{s}_j 's**, say, and replace x past inputs $u_n, u_{n-1} \dots, u_{n-r}$ and (or) by past outputs $y_n, y_{n-1} \dots, y_{n-r}$, which converts (2) into dynamic models with outer dynamics.

M is frequently large and a testing procedure for **selecting essentially non zero θ_j 's** will be necessary. Later we skip θ_0 parameter, which is usually not necessary for modeling dynamics.

Motivations for using random projections

To motivate our further discussion, consider the well known, simple finite impulse response (FIR) model:

$$y_n = \sum_{j=1}^J \alpha_j u_{n-j} + \epsilon_n, \quad n = 1, 2, \dots, N, \quad (3)$$

where y_n are outputs observed with the noise ϵ_n 's, while u_n 's form an input signal, which is observed (or even designed) in order to estimate α_j 's. Let us suppose that our system has a long memory – needs $J \approx 10^3$ for adequate modeling, e.g., chaotic systems. **Is it reasonable to estimate $\sim 10^3$ parameters ?**, even if N is very large ?

The alternative idea: project vector of past u_n 's onto random directions and select only those projections that are relevant for a proper modeling.

Random projections

Projection: $x \mapsto v = Sx$, $R^d \rightarrow R^k$, $k \ll d$ is defined by projection matrix S :

$$\begin{bmatrix} s_{11} & s_{12} & \dots & \dots & \dots & s_{1d} \\ s_{21} & s_{22} & \dots & \dots & \dots & s_{2d} \\ \vdots & \vdots & \dots & \dots & \dots & \vdots \\ s_{k1} & s_{k2} & \dots & \dots & \dots & s_{kd} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{bmatrix}$$

Random projections are closely related to the Johnson-Lindenstrauss lemma (Johnson-Lindenstrauss 1984), which states that any set A , say, of N points in an Euclidean space can be embedded in an Euclidean space of lower dimension ($\sim O(\log N)$) with relatively small distortion of the distances between any pair of points from A .

Model 1 – with internal projections

For T denoting the transposition, define:

$$\bar{\mathbf{u}}_n = [\mathbf{u}_{n-1}, \mathbf{u}_{n-2}, \dots, \mathbf{u}_{(n-r)}]^T.$$

Above, $r \geq 1$ is treated as large (hundreds, say) – we discuss models with long memory.

Model 1. For $n=(r+1), (r+2), \dots, N$,

$$\mathbf{y}_n = \sum_{k=1}^K \theta_k \varphi(\underbrace{\bar{\mathbf{s}}_k^T \bar{\mathbf{u}}_n}_{\text{int.proj.}}) + \epsilon_n, \quad (4)$$

where ϵ_n 's are i.i.d. random errors, having Gaussian distr. with zero mean and finite variance; $E(\epsilon_n) = 0$, $\sigma^2 = E(\epsilon_n)^2 < \infty$,

Model 1 – assumptions

- ▶ $\bar{\theta} = [\theta_1, \theta_2, \dots, \theta_K]^T$ – vector of unknown parameters, to be estimated from (u_n, y_n) , $n = 1, 2, \dots, N$ – observations of inputs u_n and outputs y_n .
- ▶ $\varphi : \mathbb{R} \rightarrow [-1, 1]$ a sigmoidal function, $\lim_{t \rightarrow \infty} \varphi(t) = 1$, $\lim_{t \rightarrow -\infty} \varphi(t) = -1$, nondecreasing, e.g. $\varphi(t) = 2 \arctg(t)/\pi$.
We **admit also**: $\varphi(t) = t$. When observations properly scaled, we approximate $\varphi(t) \approx t$.
- ▶ $\bar{s}_k = s_k / ||s_k||$, where $r \times 1$ i.i.d. random vectors s_k : $E(s_k) = 0$, $\text{cov}(s_k) = I_r$. s_k 's mutually independent from ϵ_n 's.

Model 1:
$$y_n = \sum_{k=1}^K \theta_k \underbrace{\varphi(\bar{s}_k^T \bar{u}_n)}_{\text{int.proj.}} + \epsilon_n.$$

- ▶ K is large, but $K \ll r = \dim(\bar{u}_n)$
- ▶ Model 1 looks similarly as the projection pursuit regression (PPR), but there are important differences:

1. directions of projections \bar{s}_k 's are **drawn at random** uniformly from the unit sphere, (instead of estimated)
2. φ is given (instead of estimated)

Idea: project \bar{u}_n onto many random directions, estimate θ_k 's by LSQ, test $\theta_k \neq 0$, reject the terms $\theta_k \approx 0$ and re-estimate.

We derive Fisher's information matrix (FIM):

Case A) FIM exact, if $\varphi(\mathbf{t}) = \mathbf{t}$,

Case B) FIM approximate if $\varphi(\mathbf{t}) \approx \mathbf{t}$.

Rewrite Model 1 as:

$$\mathbf{y}_n = \bar{\boldsymbol{\theta}}^T \bar{\mathbf{x}}_n + \epsilon_n, \quad (5)$$

$$\begin{aligned} \bar{\mathbf{x}}_n^T &\stackrel{\text{def}}{=} [\varphi(\bar{\mathbf{s}}_1^T \bar{\mathbf{u}}_n), \varphi(\bar{\mathbf{s}}_2^T \bar{\mathbf{u}}_n), \dots, \varphi(\bar{\mathbf{s}}_K^T \bar{\mathbf{u}}_n)], \\ \bar{\mathbf{x}}_n &\underbrace{=}_{\approx \text{ if B)}} \mathbf{S}^T \bar{\mathbf{u}}_n, \end{aligned} \quad (6)$$

$\mathbf{S} \stackrel{\text{def}}{=} [\bar{\mathbf{s}}_1, \bar{\mathbf{s}}_2, \dots, \bar{\mathbf{s}}_K]$. **Then, for $\sigma = 1$, FIM:**

$$\mathbf{M}_N(\mathbf{S}) = \sum_{n=r+1}^N \bar{\mathbf{x}}_n \bar{\mathbf{x}}_n^T = \mathbf{S}^T \left[\sum_{n=r+1}^N \bar{\mathbf{u}}_n \bar{\mathbf{u}}_n^T \right] \mathbf{S}. \quad (7)$$

Averaged FIM (AFIM):

Define the correlation matrix of lagged input vectors:

$$\mathbf{R}_u = \lim_{N \rightarrow \infty} (N - r)^{-1} \sum_{n=r+1}^N \bar{\mathbf{u}}_n \bar{\mathbf{u}}_n^T, \quad (8)$$

which is well defined for stationary and ergodic sequences. AFIM is defined as:

$$\mathcal{M}(\mathbf{R}_u) = \mathbf{E}_S \left[\lim_{N \rightarrow \infty} (N - r)^{-1} \mathbf{M}_N(\mathbf{S}) \right] \quad (9)$$

AFIM – final form:

$$\mathcal{M}(\mathbf{R}_u) = \mathbf{E}_S [\mathbf{S}^T \mathbf{R}_u \mathbf{S}] \quad (10)$$

Problem statement:

The constraint on input signal power:

$$\text{diag. el. } [R_u] = \lim_{N \rightarrow \infty} N^{-1} \sum_{i=1}^N u_i^2 \leq 1. \quad (11)$$

Problem statement – D optimal experiment design for Model 1

Assume that the class of Models 1 is sufficiently rich to include unknown system. Motivated by relationships between the estimation accuracy and FIM, find R_u^* , under constraints (11) such that

$$\max_{R_u} \text{Det}[\mathcal{M}(R_u)] = \text{Det}[\mathcal{M}(R_u^*)] \quad (12)$$

Remarks on averaging w.r.t. S

As usual, when averages are involved, one can consider different ways of combining averaging with the matrix inversion and calculating determinants. **The way selected above is manageable.** Other possibilities: the minimization, w.r.t. R_u , either

$$\text{Det} \left\{ \mathbf{E}_S \left[N \mathbf{M}_N^{-1}(S) \right] \right\}, \text{ as } N \rightarrow \infty \quad (13)$$

$$\text{or } \mathbf{E}_S \text{ Det} \left[N \mathbf{M}_N^{-1}(S) \right], \text{ as } N \rightarrow \infty. \quad (14)$$

The above dilemma is formally very similar to the one that arises when we consider the bayesian prior on unknown parameters.

D-optimal experiment design for Model 1

Result 1

$\max_{R_u} \text{Det}[\mathcal{M}(R_u)]$, under (11), is attained when all off-diagonal elements of R_u^* are zero.

It follows from known fact that for symmetric A :

$\text{Det}A \leq \prod_{k=1}^r a_{kk}$ with $=$ iff A is a diagonal matrix.

Selecting $R_u^* = I_r - r \times r$ identity matrix, we obtain:

$$\mathcal{M}(R_u^*) = E_S[S^T S] = I_K,$$

because: $E(\bar{s}_j^T \bar{s}_K) = 0$ for $k \neq j$ and 1, for $k = j$.

Remark

For $N \rightarrow \infty$ sequences u_n 's with $R_u = I_r$ can be generated as i.i.d. $\mathcal{N}(0, 1)$. For finite N pseudorandom binary signals are known, for which $R_u \approx I_r$.

Summary – $y_n = \sum_{k=1}^K \theta_k \varphi(\underbrace{\bar{s}_k^T \bar{u}_n}_{\text{int.proj.}}) + \epsilon_n.$

1. Select $r \geq 1$ – expected length of the system memory.
2. Generate input signal u_n 's with $R_u^* = I_r$.
3. Observe pairs (u_n, y_n) , $n = 1, 2, \dots, N$.
4. Select $K \ll r$ and generate \bar{s}_k 's as $\mathcal{N}(0, I_r)$ and normalize their lengths to 1.
5. Select φ and calculate $\bar{x}_n = \text{vec}[\varphi(\bar{s}_k \bar{u}_n)]$.
6. Estimate $\bar{\theta}$ by LSQ:

$$\hat{\theta} = \arg \min_{\bar{\theta}} \sum_{n=r+1}^N [y_n - \bar{\theta}^T \bar{x}_n]^2 \quad (15)$$

Summary – $y_n = \sum_{k=1}^K \theta_k \underbrace{\varphi(\bar{s}_k^T \bar{u}_n)}_{\text{int.proj.}} + \epsilon_n.$

- ▶ Test $H_0 : \hat{\theta}_k = 0$, for all components of $\hat{\theta}$.
- ▶ Form \mathcal{K} – the set of those k that $H_0 : \hat{\theta}_k = 0$ is rejected.
- ▶ Re-estimate θ_k , $k \in \mathcal{K}$ by LSQ – denote them by $\tilde{\theta}_k$.
- ▶ **Form the final model for prediction:**

$$\hat{y}_n = \sum_{k \in \mathcal{K}} \tilde{\theta}_k \varphi(\bar{s}_k^T \bar{u}_n) \quad (16)$$

and validate it on data that were not used for its estimation.

Remarks on estimating Model 1

1. The nice feature of the above method is **its computational simplicity.**
2. It can be used also when the experiment is not active, i.e., **u_n 's are only observed.**
3. The method can be applied also **for estimating a regression function with a large number of regressors** – replace \bar{u}_n by them.
4. For testing $H_0 : \hat{\theta}_k = 0$ we can use the standard t-Student test.

Preliminary simulation studies – Model 1

The following system, with zero IC, was simulated: $t \in (0, H_{or})$, $H_{or} = 100$,

$$\dot{x}(t) = -0.2 x(t) + 3 u(t) \quad (17)$$

$$\dot{y}(t) = -0.25 y(t) + 0.5 x(t), \quad (18)$$

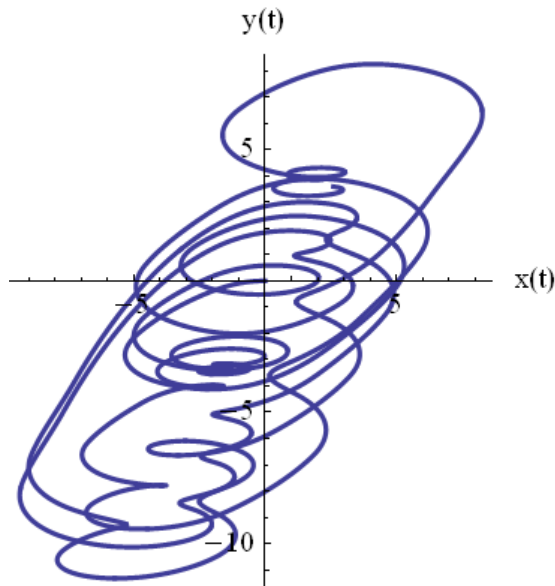
where $u(t)$ – interpolated PRBS. The observations:

$$y_n = y(n \tau) + \epsilon_n, \quad \tau = 0.01,$$

$\epsilon_n \sim \mathcal{N}(0, 1)$, $n = 1, 2, \dots, 10^4$. The first 5900 observations used for learning (estimation), the rest – for testing.

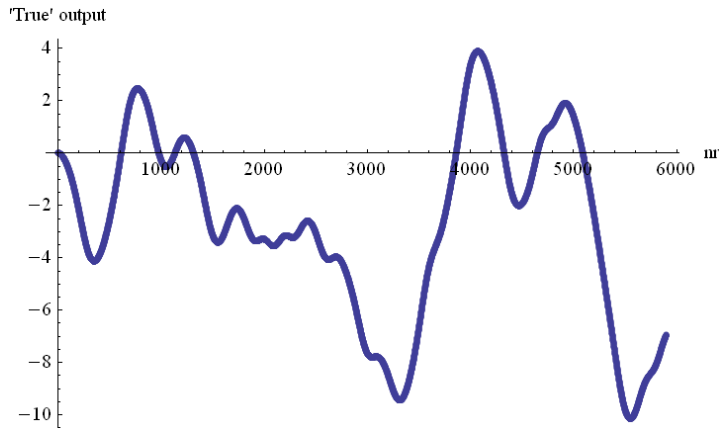
Simulation studies – the system behaviour

Complicated dynamics caused by PRBS input



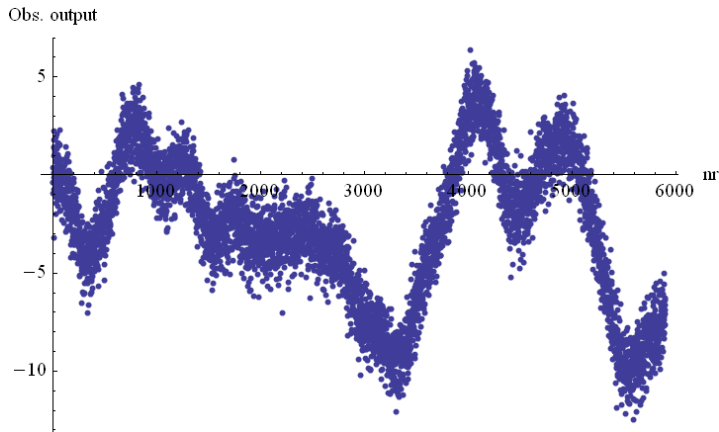
Simulation studies – the system behaviour

"True" output $y(t)$



Simulation studies – the system behaviour

Sampled output $y(n\tau)$ + noise



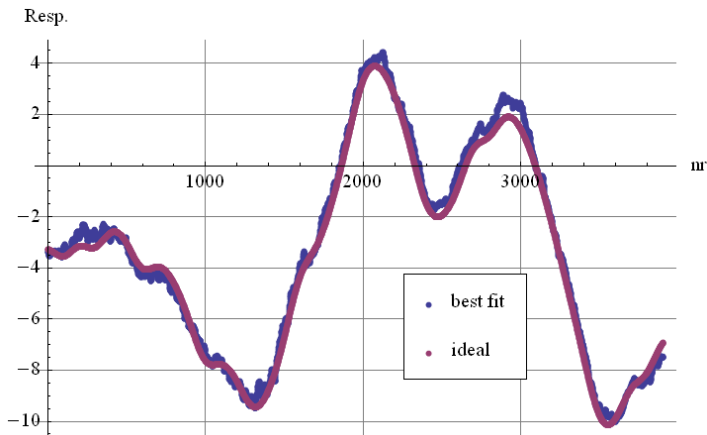
Simulation studies – estimated model:

$$\mathbf{y}_n = \sum_{k=1}^K \theta_k (\bar{\mathbf{s}}_k^T \bar{\mathbf{u}}_n) + \epsilon_n, \quad (19)$$

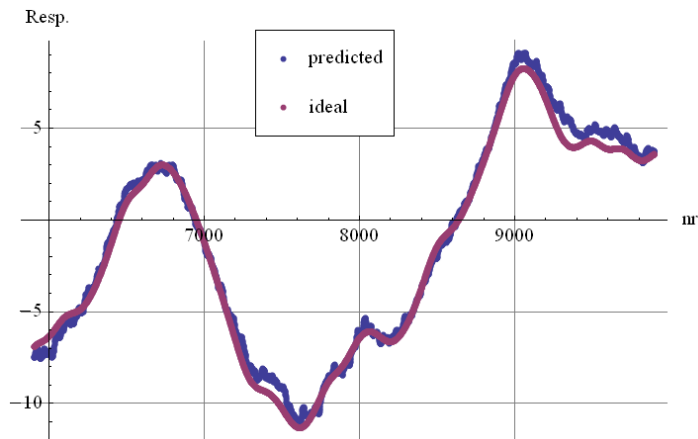
i.e., $\phi(\nu) = \nu$ ($\phi(\nu) = \arctan(0.1 \nu)$ provides very similar results), where

- ▶ $K = 50$ – the number of random projections,
- ▶ $r = 2000$ – the number of past inputs ($r = \dim(\bar{\mathbf{u}}_n)$) that are projected by
- ▶ $\bar{\mathbf{s}}_k \sim \mathcal{N}(0, \mathbf{I}_r)$ – normalized to 1.

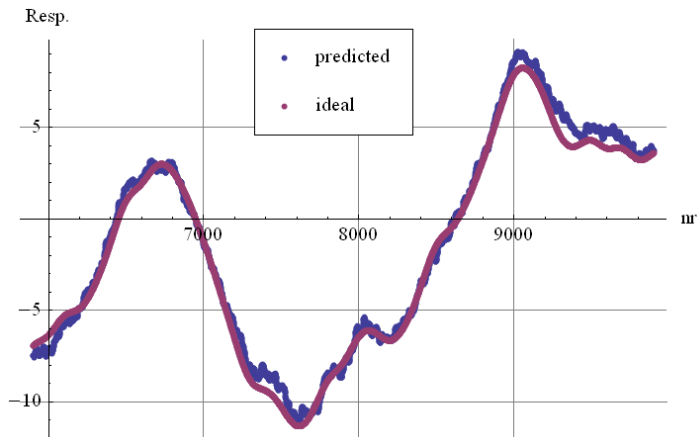
Estimated model – response vs learning data



Estimated model – one step ahead prediction vs testing data



Estimated model – one step ahead prediction vs testing data,
after rejecting 21 terms with parameters having $p\text{-val.} > 0.05$
in t-Student test (29 terms remain);



What if a time series is generated from a nonlinear and chaotic ODE's ?

Consider the well known **chaotic Lorenz system, perturbed** by (interpolated) PRBS $u(t)$:

$$\dot{x}(t) = 100 u(t) - 5 (x(t) - y(t)) \quad (20)$$

$$\dot{y}(t) = x(t) (-z(t) + 26.5) - y(t) \quad (21)$$

$$\dot{z}(t) = x(t)y(t) - z(t) \quad (22)$$

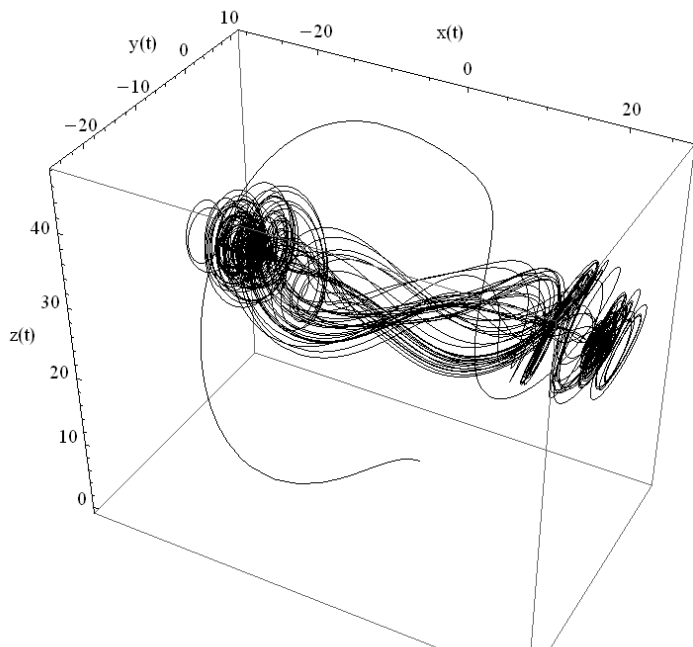
Our aim: select and estimate models in order:

A) to predict $x(t_n)$ from $\chi_n = x(t_n) + \epsilon_n$,

B) to predict $y(t_n)$ from $\eta_n = y(t_n) + \epsilon_n$,

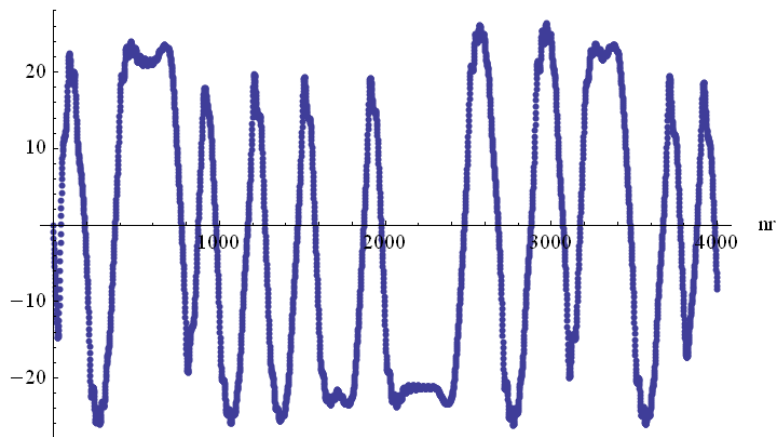
without using the knowledge about (20)-(22).

The system behaviour – phase plot



The system behaviour – $x(t)$ component – a part used for the estimation (learning):

'True' output



+ noise $\mathcal{N}(0, 0.1)$, sampled $\tau = 1$.

Estimated model – Case A):

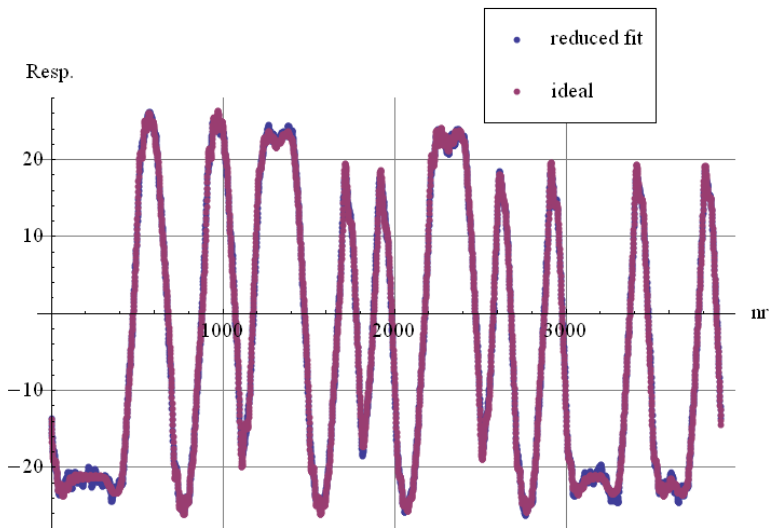
$$\mathbf{y}_n = \sum_{k=1}^K \theta_k (\bar{\mathbf{s}}_k^T \bar{\mathbf{u}}_n) + \epsilon_n, \quad (23)$$

where

- ▶ $K = 150$ – the number of random projections,
- ▶ $r = 2000$ – the number of past inputs ($r = \dim(\bar{\mathbf{u}}_n)$) that are projected by
- ▶ $\bar{\mathbf{s}}_k \sim \mathcal{N}(0, \mathbf{I}_r)$ – normalized to 1.

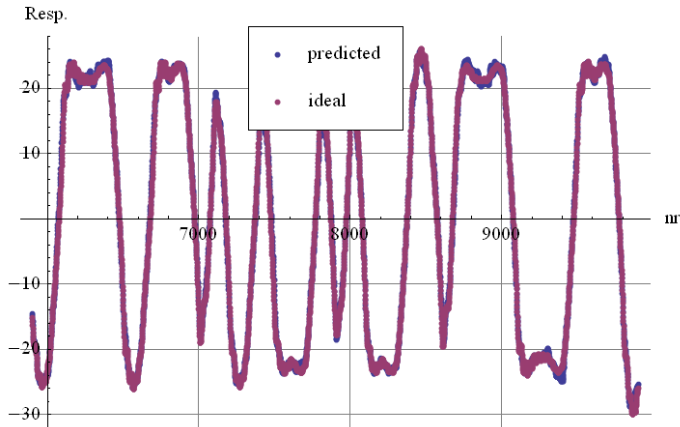
Note that we use 3 times larger number of projections than in the earlier example, while r remains the same.

Estimated model output vs learning data



obtained after selecting 73 terms out of $K = 150$.

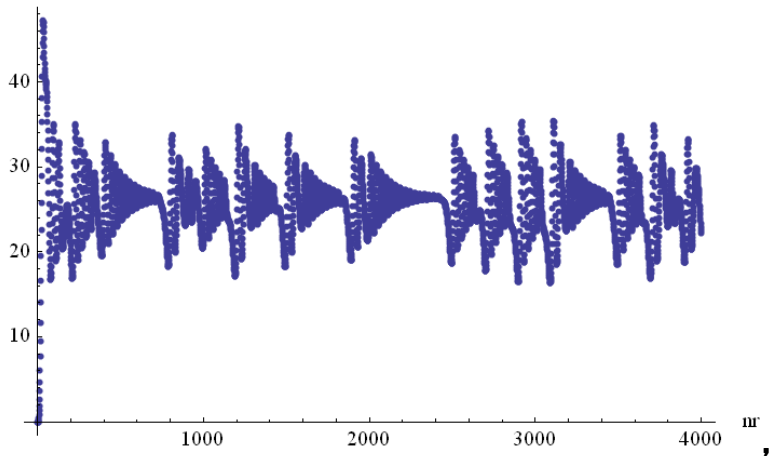
One step ahead prediction from the estimated model output vs testing data (2000 recorded)



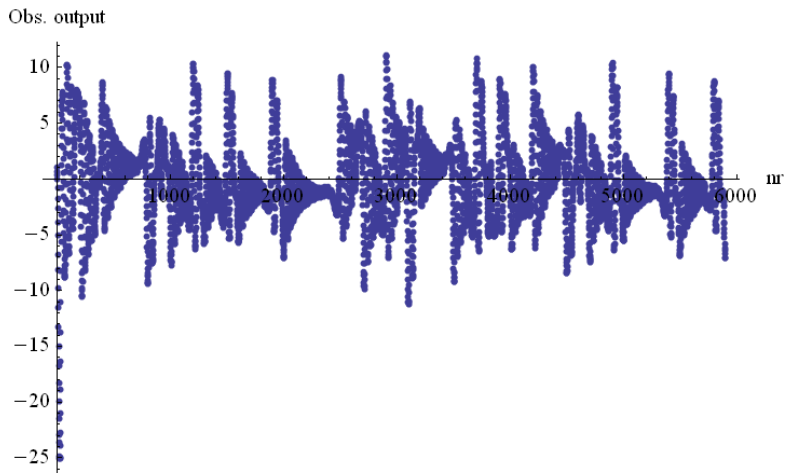
Very accurate prediction ! So far, so good ?
Let's consider a real challenge: prediction of $y(t)$ component.

Even noiseless $y(t)$ is really chaotic:

'True' output



Simulated 10^4 noisy observations: ~ 8000 used for learning,
 ~ 2000 for testing prediction abilities.



Estimated model – Case B):

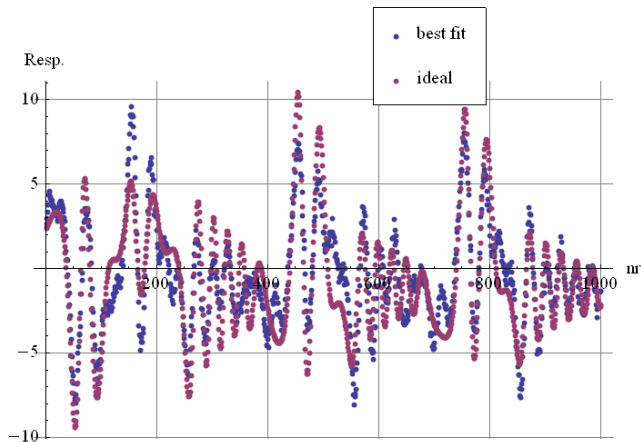
$$y_n = \sum_{k=1}^K \theta_k (\bar{\mathbf{s}}_k^T \bar{\mathbf{u}}_n) + \epsilon_n, \quad (24)$$

where

- ▶ $K = 200$ – the number of random projections,
- ▶ $r = 750$ – the number of past inputs ($r = \dim(\bar{\mathbf{u}}_n)$) that are projected by
- ▶ $\bar{\mathbf{s}}_k \sim \mathcal{N}(0, \mathbf{I}_r)$ – normalized to 1.

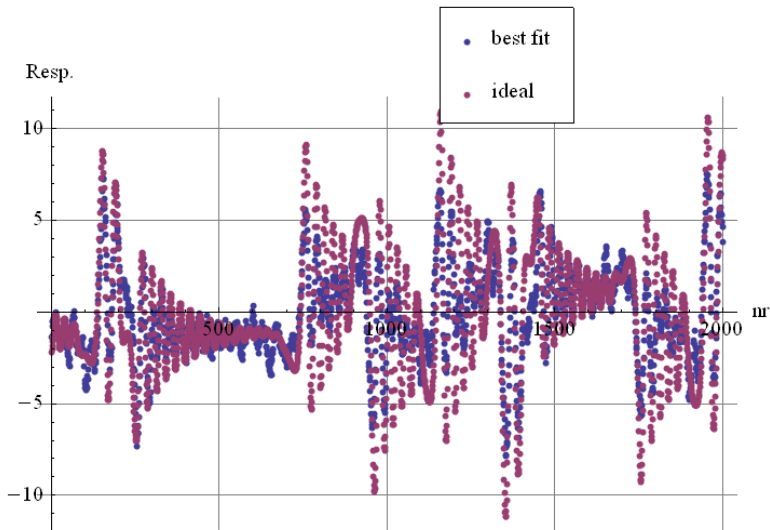
Note that we use 2.66 times smaller number of projections than in case B), while r is 30% larger.

Estimated model – response vs learning data. Case B), the first 1000 observations



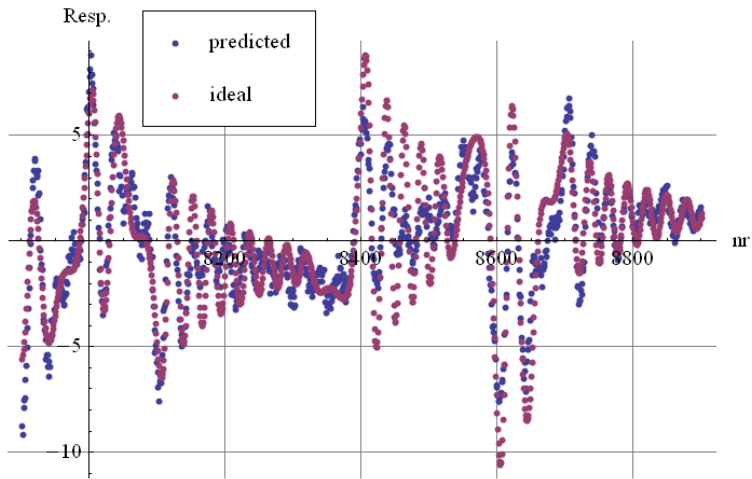
The fit is not perfect, but retains almost all oscillations.

Estimated model – response vs learning data. Case B), next 2000 observations

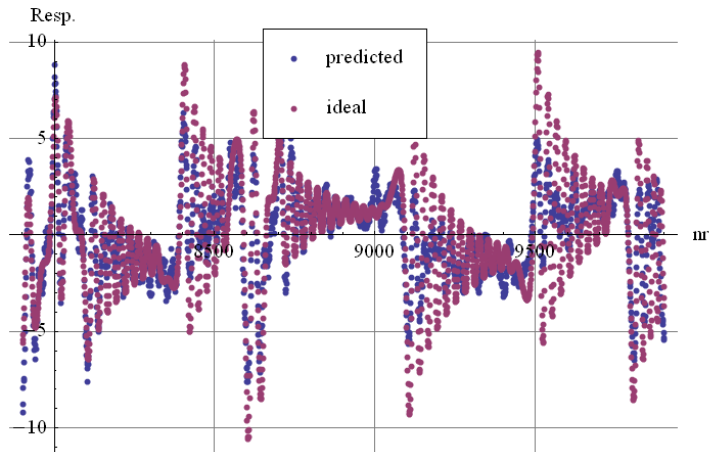


Estimated model – one step ahead prediction

Case B), the first 1000 testing data



Estimated model – one step ahead prediction Case B), all testing data



The prediction is far from precision, but still retains a general shape of the highly chaotic sequence.

Conclusions for Model 1

- ▶ Projections of long sequence of past inputs plus LSQ provide easy to use method of predicting sequences having complicated behaviours.
- ▶ The choice of $r = \dim(\bar{u}_n)$ is important. Here it was done by try and error, but one can expect that Akaike's or Rissanen's criterions will be useful.
- ▶ The choice of K – the number of projections is less difficult, because too large number of projections is corrected at the stage of rejecting terms with small parameters.
- ▶ One can also consider a mixture of projections having different lengths.

Brief outline of Model 2

Based on **projections of past outputs**. Define:

$$\bar{y}_n = [y_{n-1}, y_{n-2}, \dots, y_{(n-p)}]^T.$$

Above, $p \geq 1$ is treated as large.

Model 2. For $n=(p+1), (p+2), \dots, N$,

$$y_n = \sum_{l=1}^L \beta_l \varphi\left(\underbrace{\bar{s}_l^T \bar{y}_n}_{\text{past outputs}}\right) + \beta_0 \underbrace{u_n}_{\text{input}} + \epsilon_n, \quad (25)$$

where $\epsilon_n \sim \mathcal{N}(0, \sigma)$, u_l – external inputs (to be selected), β_l 's unknown parameters (to be estimated). $\phi(\nu) = \nu$ or $\phi(\nu) \approx \nu$ – sigmoid.

Does not have a counterpart in PPR.

Model 2 – estimation

Important: we assume that ϵ_n 's are uncorrelated. Note that σ is also unknown and estimated.

The estimation algorithm for model 2:
(formally the same as for Model 1)

- ▶ replace \bar{u}_n in Model 1 by \bar{y}_n concatenated with u_n ,
- ▶ use LSQ plus rejection of spurious terms (by testing $H_0 : \beta_l = 0$),
- ▶ and re-estimate.

Mimicking the proof from Goodwin Payne (Thm. 6.4.9), we obtain the following

Model 2 – experiment design

averaged and normalized FIM, when $N \rightarrow \infty$:

$$\bar{\mathbf{M}} = \left[\begin{array}{c|c|c} \mathbf{E}_S (\mathbf{S}^T \mathbf{A} \mathbf{S}) & \mathbf{B} & \mathbf{0} \\ \hline \mathbf{B}^T & \mathbf{C} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & 1/(2\sigma^2) \end{array} \right].$$

where \mathbf{S} is $p \times K$ is the random projection matrix. Define: $\rho(j - k) = \mathbf{E}(y_{i-j} y_{i-k})$ and

$$\mathbf{V}^T = [\rho(1), \rho(2), \dots, \rho(p)].$$

\mathbf{A} is $p \times p$ Toeplitz matrix, build on

$$[\rho(0), \rho(1), \dots, \rho(p - 1)]^T$$

Model 2 – experiment design 2

Then, the rest of blocks in $\bar{\mathbf{M}}$ are given by:

$$\mathbf{B} = (\mathbf{V} - \mathbf{A} \bar{\boldsymbol{\beta}}) / \beta_0,$$

$$\mathbf{C} = (\rho(0) - 2 \bar{\boldsymbol{\beta}}^T \mathbf{V} + \bar{\boldsymbol{\beta}}^T \mathbf{A} \bar{\boldsymbol{\beta}} - \sigma) / \beta_0^2$$

Task: find an input sequence u_1, u_2, \dots such that $\text{Det} [\bar{\mathbf{M}}]$ is maximized, under the constraint:

$$\rho(0) = \lim_{N \rightarrow \infty} \frac{1}{(N - p)} \sum_{n=p}^N y_n^2 \leq 1, \quad (26)$$

that is interpreted as the constraint on the power of the output signal.

Model 2 – experiment design 2

Mimicking the proof from Goodwin Payne (Thm. 6.4.9) and using the assumed properties of S , we obtain:

Theorem 2.

Assume $\rho(0) > \sigma$ and that unknown system can be adequately described by (25). Select $\phi(\nu) = \nu$. Then, $\text{Det}(\bar{M})$ is maximized when A is the unit matrix, which holds if $\rho(0) = 1$, $\rho(k) = 0$ for $k > 0$, i.e., **when the system output is an uncorrelated sequence.**

Model 2 – experiment design 3 – Remarks

Condition $\rho(k) = 0$ for $k > 0$ can formally be ensured by the following **minimum variance control law** (negative feedback from past y_n):

$$u_n = -\beta_0^{-1} \sum_{l=1}^L \beta_l \phi(\bar{s}_l^T \bar{y}_n) + \eta_n, \quad (27)$$

where η_n 's is a set point sequence, which should be i.i.d. sequence with zero mean and the variance $(\rho(0) - \sigma)/\beta_0^2$. In practice, realization of (27) can cause troubles (not quite adequate model, unknown β etc.), but random projections are expected to reduce them, **due to their smoothing effect**.

Model 3 = Model 1 + Model 2

The estimation and rejection procedure for the combined model:

$$y_n = \sum_{l=1}^L \beta_l \varphi(\bar{s}_l^T \bar{y}_n) + \sum_{l=L+1}^{L+K} \beta_l \phi(\bar{s}_l^T \bar{u}_n) + \epsilon_n. \quad (28)$$

is the as above.

Open problem: design D-optimal input signal for the estimation of β_l 's in (28), under input and/or output power constraint.

One can expect that it is possible to derive the equivalence thm., assuming $\phi(\nu) = \nu$.






Concluding remarks






1. Random projections of past inputs and/or outputs occurred to be a powerful tool for modeling systems with long memory.
2. The proposed Model 1 provides a very good prediction for linear dynamic systems, while for quickly changing chaotic systems its is able to predict a general shape of the output signal.
3. D-optimal input signal can be designed for Model 1 and 2, mimicking the proofs of the classical results for linear systems without projections.






Concluding remarks 2






- Despite the similarities of Model 1 to PPR, random projections + rejections of spurious terms leads to **much simpler estimation procedure**.
- A similar procedure **can be used for a regression function estimation**, when we have a large number of candidate terms, while the number of observations is not enough for their estimation – projections allow for estimating a common impact of several terms.
- Model 2 without an input signal **can be used for predicting time series** such as sun spots.

PARTIAL BIBLIOGRAPHY

-  D. Aeyels, Generic observability of differentiable systems, SIAM J. Control and Optimization, vol. 19, pp. 595-603, 1981.
-  D. Aeyels, On the number of samples necessary to achieve observability, Systems and Control Letters, vol. 1, no. 2, pp. 92-94, August 1981.
-  Casdagli, M., Nonlinear Prediction of Chaotic Time Series, Physical Review D, 35(3):35-356, 1989.
-  Leontaritis, I.J. and S.A. Billings, Input-Output Parametric Models for Nonlinear Systems part I: Deterministic Nonlinear Systems, International Journal of Control, 41(2):303-328, 1985.
-  A.U. Levin and K.S. Narendra. Control of non-linear dynamical systems using neural networks. controllability and stabilization. IEEE Transactions on Neural Networks, 4:192-206, March 1993.

-  A.U. Levin, K.S. Narendra, Recursive identification using feedforward neural networks, International Journal of Control 61 (3) (1995) 533-547.
-  Juditzky, H. Hjalmarsson, A. Benveniste, B.Delyon, L. Ljung, J. Sjöberg and Q. Zhang, Nonlinear Black-box Models in System Identification: Mathematical Foundations, Automatica 31(12): 1725–1750, 1995. Ljung, L. System Identification - Theory for the User. Prentice-Hall, N.J. 2nd edition, 1999.
-  Pintelon R. and Schoukens, J. System Identification. A Frequency Domain Approach, IEEE Press, New York, 2001.
-  Söderström, T. and P. Stoica, System Identification, Prentice Hall, Englewood Cliffs, NJ. 1989.
-  Sjöberg, J., Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.-Y. Glorennec, H. Hjalmarsson, and A. Juditsky: "Non-linear Black-box Modeling in System Identification: a Unified Overview" , Automatica, 31:1691-1724, 1995.

-  G. Cybenko, Approximation by superpositions of a sigmoidal function, Mathematics of Control, Signals, and Systems 2 (1989) 303–314.
-  K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Networks 2 (1989) 359–366.
-  Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). A Distribution-Free Theory of Nonparametric Regression. Springer: New York. MR1987657
-  Jones, Lee K. (1992), A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training, Annals of Statistics, 20, 608-613
-  W.B. Johnson, J. Lindenstrauss, Extensions of Lipschitz mapping into Hilbert space, Contemporary Mathematics 26 (1984) 189–206.

-  Matousek, J.: On variants of the JohnsonLindenstrauss lemma. Random Structures and Algorithms, 33(2): 142-156, 2008.
-  E. J. Candès and Terence Tao, Near-optimal signal recovery from random projections: Universal encoding strategies?, IEEE Transactions on Information Theory, vol. 52, no. 12, pp. 5406-5425, 2006.
-  Takens F. Detecting strange attractors in fluid turbulence. In: Rand D, Young LS, editors. Dynamical systems and turbulence. Berlin: Springer; 1981.
-  J. Stark, Delay embeddings for forced systems. I. Deterministic forcing, Journal of Nonlinear Science 9 (1999) 255-332.
-  Stark, J., D.S. Broomhead, M.E. Davies, and J. Huke, Takens Embedding Theorems for Forced and Stochastic Systems, Nonlinear Analysis: Theory, Methods and Applications, 30(8):5303-5314, 1997.



J. Stark, D.S. Broomhead, M.E. Davies, J. Huke, Delay embeddings for forced systems. II. Stochastic forcing, Journal of Nonlinear Science 13 (2003) 519577.



Ucinski D. Optimal Measurement Methods for Distributed Parameter System Identification CRC Press, Boca Raton, FL, 2005.