# Structured Kernel Based Modeling and its Application To Electric Load Forecasting

## Doctoral Presentation

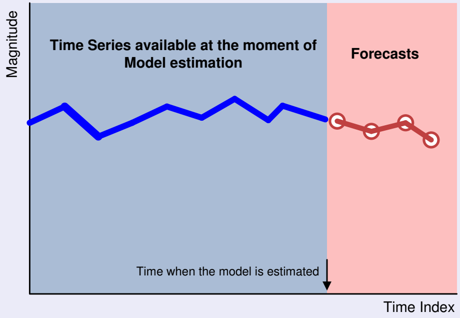
Marcelo Espinoza

Promotors
Prof. Dr. Ir. Bart De Moor
Prof. Dr. Ir. Ronnie Belmans

Department Electrical Engineering
Faculty of Engineering
Katholieke Universiteit Leuven

# Time Series: To Know Something Before it Happens

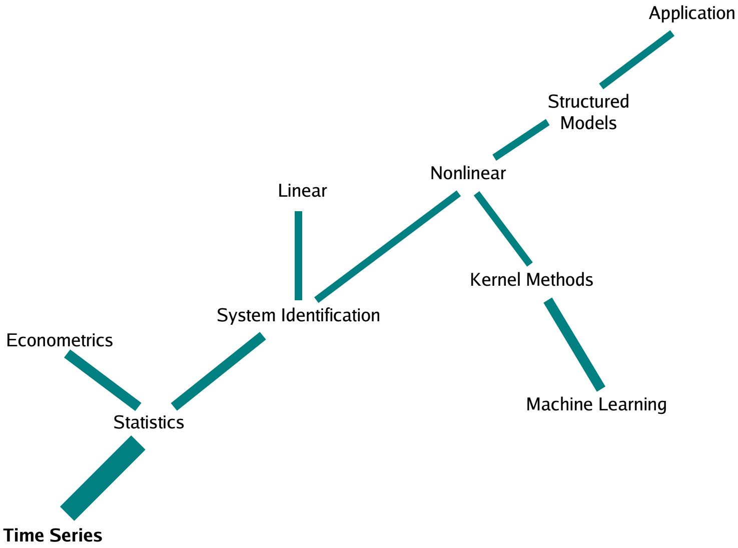
## Time Series Representation



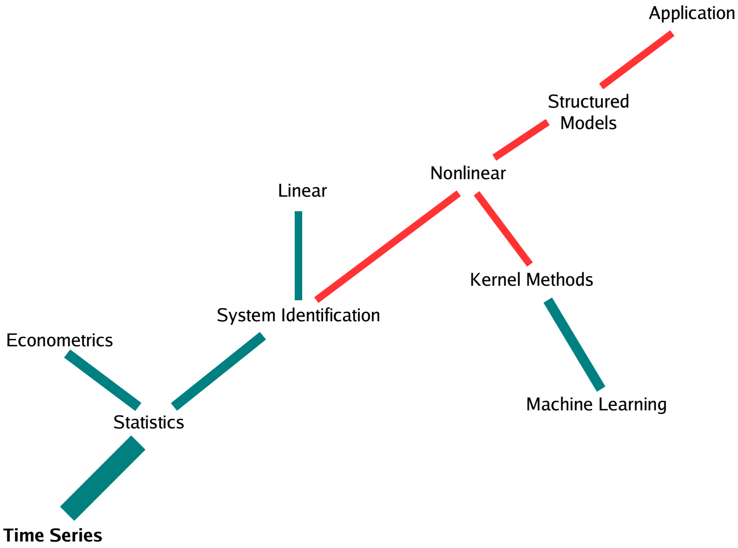

The Goal : a Model for Prediction

- Important in many domains and applications
- Economics, Process Industry, Traffic, Energy, (Bio)Medical, and many others
- Large datasets available

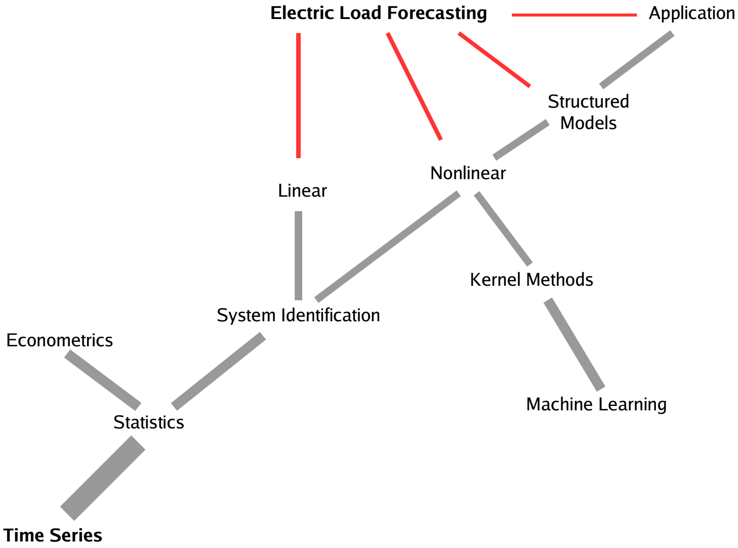
# Context and Outline



# Context and Outline



# Context and Outline



# The Importance of Short-Term Load Forecasting

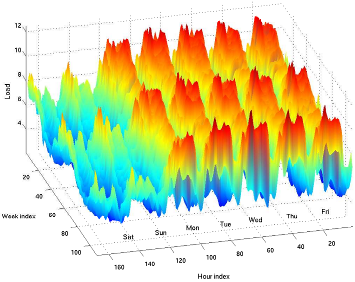
## Important for Generation

- Electricity cannot be efficiently stored
- Forecasts help to manage generation
- Everyday use of short-term forecasts

## Important for Energy Markets

- Liberalization
- Energy Market Exchanges
- Market players interaction leads to price setting

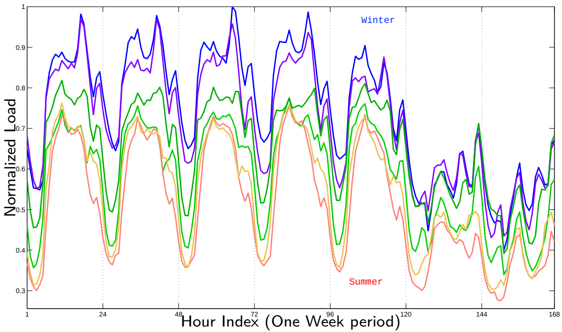
# Using Data from Substations




- Off-take point for local distribution
- Voltage reduction from transmission grid
- Datasets from ELIA (Transmission System Operator)
- Long time series (40,000 datapoints)
- Seasonal Behavior

# Seasonality in the Load Series

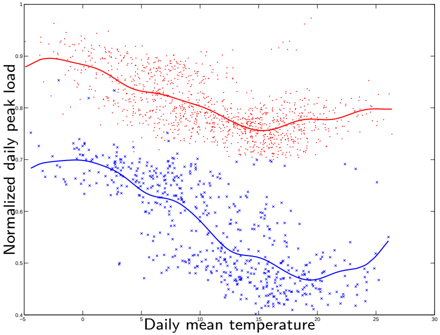
## Daily, Weekly and Yearly Patterns




- Seasonal patterns are present
- Combined effect at different levels

# The Nonlinear Relation between Weather and Load

## Daily Peak Load and Temperature




- Heating
- Cooling
- Nonlinear function

# Load Forecasting Models

Short-Term Load Forecasting:

- Predict 1 hour (or less) ahead, up to several days
- Accuracy, Interpretability
- Scenario Simulation, weather normalization

**Objective of this work**

Obtain the best possible predictions given all available information

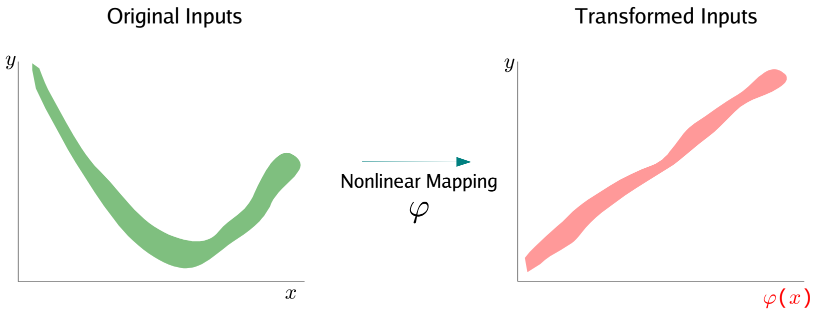
[Huang, IEEE TPS, 2003] [Hylleberg, 1992] [Hippert *et al.*, IJF, 2005] [Fay *et al.*, Neurocomputing, 2003]

# Kernel Methods for Function Estimation

# Nonlinear Function Estimation

Using kernel methods

- The data is mapped to a higher dimensional space
- The mapping is computed using kernel functions



# Least-Squares Support Vector Machines

LS-SVM: **Optimization Approach to Kernel-Based Modeling**

Basic Model + Prior Knowledge (constraints)

Solution in terms of
Equivalent Kernel

- Other kernel methods: gaussian processes, kriging, regularization networks, RKHS

# LS-SVM Regression Formulation

## Primal Space

**Model**

$$y_i = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i$$

**Problem**

$$\min_{\boldsymbol{w}, b, e_i} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma \frac{1}{2}\sum_{i=1}^{N} e_i^2$$

$$\text{s.t.} \quad y_i = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i, \ \ i = 1, \ldots, N$$

## Dual Space

**Model**

$$y(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

**Solution**

$$\left[\begin{array}{c|c} \boldsymbol{\Omega} + \gamma^{-1}\mathbf{I} & \mathbf{1} \\ \hline \mathbf{1}^T & 0 \end{array}\right] \left[\begin{array}{c} \boldsymbol{\alpha} \\ \hline b \end{array}\right] = \left[\begin{array}{c} \boldsymbol{y} \\ 0 \end{array}\right]$$

Use kernel trick $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\varphi}(\boldsymbol{x}_i)^T \boldsymbol{\varphi}(\boldsymbol{x}_j)$ with positive-definite kernel function $K$

# Using Kernel Functions

## Some common kernel functions

- Radial Basis Function (RBF) with parameter $\sigma$
  $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{\frac{-||\boldsymbol{x}_i - \boldsymbol{x}_j||^2}{\sigma^2}}$
- Polynomial of degree $d$, with $c > 0$
  $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T \boldsymbol{x}_j + c)^d$
- Linear
  $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T \boldsymbol{x}_j$

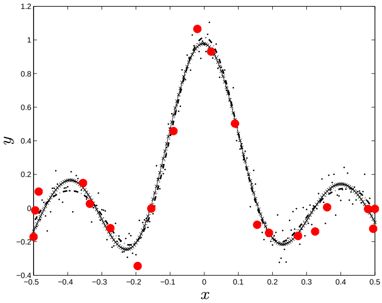
More kernels can be defined from existing kernels

[Suykens *et al.*, LS-SVM, 2002] [Vapnik, Statistical Learning Theory, 1998]

# Large Scale Problems and Primal Space

## Fixed Size LS-SVM

- Select a subsample of the training set
- Build finite-dimensional approximation $\hat{\varphi}$ using Nyström Methods
- Evaluate $\hat{\varphi}$ for all training set
- Solve the problem in Primal Space




### Advantages

- Practical for Large Scale problems
- Sparse Model representation

# LS-SVM: Living in Two Worlds

## Primal Space

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}) + b$$

- Parametric Model
- Estimate $\boldsymbol{w} \in \mathbb{R}^d$
- Large Scale Problems

**Prior Knowledge here**

Kernel Trick →

← Nyström Method

## Dual Space

$$y(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

- Nonparametric Model
- Estimate $\boldsymbol{\alpha} \in \mathbb{R}^N$
- Large Dimensional Inputs

# Imposing Prior Knowledge to Models

**Golden Rule**

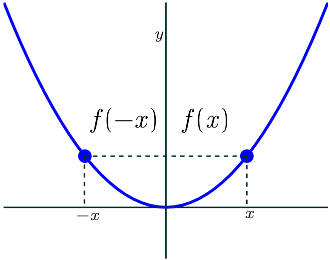
"Do not estimate what you already know"

- Black-Box regression can be too general
- Prior Knowledge is usual in real-life situations

Some forms of prior knowledge

- Symmetry
- Linear Terms
- Autocorrelated Residuals

# Imposing Symmetry




- Optimization Problem

$$\min_{\boldsymbol{w},b,e_i} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2$$

$$\text{s.t.} \quad \begin{cases} y_i = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i \\ \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_i) = \boldsymbol{w}^T\boldsymbol{\varphi}(-\boldsymbol{x}_i) \end{cases}$$

$$i = 1, \ldots, N$$

- Final Model $y(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i K_{eq}(\boldsymbol{x}_i, \boldsymbol{x}) + b$
- Equivalent Kernel $K_{eq}(\boldsymbol{x}_i, \boldsymbol{x}) = \frac{1}{2}[(K(\boldsymbol{x}_i, \boldsymbol{x}) + K(-\boldsymbol{x}_i, \boldsymbol{x})]$
- Also possible for odd functions

[Espinoza *et al.*, IEEE CDC 2005]

# Partially Linear Model with LS-SVM

## A Partially Linear Structure

$$y = \boldsymbol{\beta}^T \boldsymbol{z} + f(\boldsymbol{x})$$

can be estimated with LS-SVM

Final model is given by

$$\hat{y}(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{\beta}^T \boldsymbol{z} + \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b.$$

- Simultaneous estimation of the linear and nonlinear parts
- Model can be estimated in Primal Space

[Espinoza *et al.*, IEEE CDC 2004]

# Imposing Autocorrelated Residuals

## Regression with AR(1) Residuals

$$\begin{cases} y_i = f(\boldsymbol{x}_i) + e_i \\ e_i = \rho e_{i-1} + r_i \end{cases}$$

can be estimated with LS-SVM

The estimated $f$ is obtained as

$$\hat{f}(\boldsymbol{x}_i) = \sum_{j=2}^{N} \alpha_{j-1} K_{eq}(\boldsymbol{x}_j, \boldsymbol{x}_i) + b$$

Equivalent Kernel $K_{eq}(\boldsymbol{x}_j, \boldsymbol{x}_i) = K(\boldsymbol{x}_j, \boldsymbol{x}_i) - \rho K(\boldsymbol{x}_{j-1}, \boldsymbol{x}_i)$

- Extension to general AR($q$) case

[Espinoza *et al.*, SYSID 2006]

# Nonlinear System Identification with Structured Kernel Based Modeling

# Model Structures in System Identification

**System Identification...**

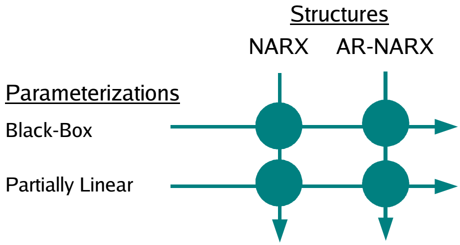
...is the discipline of making mathematical models of systems, starting from experimental data, measurements, observations

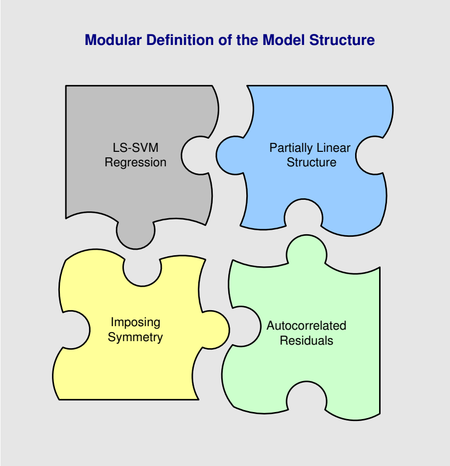
[Ljung, 1987] [Guidorzi, 2003]
[Sjöberg et al., Automatica, 1995]

**Model Structures**

| Linear | Nonlinear |
|---|---|
| FIR | NFIR |
| ARX | NARX |
| AR-ARX | AR-NARX |
| ARMAX | NARMAX |
| OE | NOE |
| BJ | NBJ |

# A Modular Approach




- Model Structure, Parameterization and Estimation

# Model Structures to Consider

Using the regression vector

$\boldsymbol{z}_t = [y_{t-1}; \cdots ; y_{t-p}; \boldsymbol{u}_t; \boldsymbol{u}_{t-1}; \cdots ; \boldsymbol{u}_{t-q}]$

## NARX Structure

$$y_t = g(\boldsymbol{z}_t) + e_t$$

$e_t$ i.i.d., zero mean and constant variance

## AR-NARX Structure

$$\begin{cases} y_t = g(\boldsymbol{z}_t) + e_t \\ A(z^{-1})e_t = r_t \end{cases}$$

with $A(z^{-1})$ a polynomial in the lag operator, $r_t$ i.i.d., zero mean and constant variance

# The Parameterizations for $g(\cdot)$

## Black-Box with LS-SVM

$$g(\boldsymbol{z}_t) = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_t) + b$$

## Partially Linear LS-SVM

$$g(\boldsymbol{z}_t) = \boldsymbol{\beta}^T \boldsymbol{z}_{A,t} + \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{z}_{B,t}) + b$$

with a partition $\boldsymbol{z}_t = [\boldsymbol{z}_{A,t}\ ,\ \boldsymbol{z}_{B,t}]$

- Final Models can be represented in Primal and Dual forms
- Large Scale implementations using Equivalent Kernels

# Examples of Model Representations

## NARX Model

Primal $\hat{y}_t = \boldsymbol{w}^T \varphi(\boldsymbol{z}_t) + b$

Dual $\hat{y}_t = \sum_{i=1}^{N} \alpha_i K(\boldsymbol{z}_i, \boldsymbol{z}_t) + b$

## AR(1)-NARX Model

Primal $\hat{y}_t = \rho y_{t-1} + \boldsymbol{w}^T \varphi(\boldsymbol{z}_t) - \rho \boldsymbol{w}^T \varphi(\boldsymbol{z}_{t-1}) + (1-\rho) b$

Dual $\hat{y}_t = \rho y_{t-1} + h(\boldsymbol{z}_t) - \rho h(\boldsymbol{z}_{t-1})$

with

$h(\boldsymbol{z}_t) = \sum_{i=2}^{N} \alpha_{i-1} \left[K(\boldsymbol{z}_i, \boldsymbol{z}_t) - \rho K(\boldsymbol{z}_{i-1}, \boldsymbol{z}_t)\right] + b$

## PL-NARX Model

Primal $\hat{y}_t = \boldsymbol{\beta}^T \boldsymbol{z}_{A,t} + \boldsymbol{w}^T \varphi(\boldsymbol{z}_{B,t}) + b$

Dual $\hat{y}_t = \boldsymbol{\beta}^T \boldsymbol{z}_{A,t} + \sum_{i=1}^{N} \alpha_i K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,t})$

# Equivalent Kernel Functions

For any given positive-definite kernel function $K$:

$$K^{\text{eq}}_{\text{narx}} = K(\boldsymbol{z}_i, \boldsymbol{z}_j)$$

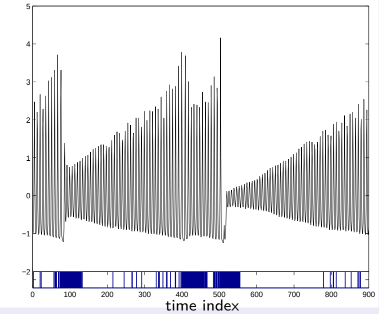
$$K^{\text{eq}}_{\text{ar-narx}} = K(\boldsymbol{z}_{i+1}, \boldsymbol{z}_{j+1}) - \rho K(\boldsymbol{z}_i, \boldsymbol{z}_{j+1}) - \rho K(\boldsymbol{z}_{i+1}, \boldsymbol{z}_j) + \rho^2 K(\boldsymbol{z}_i, \boldsymbol{z}_j)$$

$$K^{\text{eq}}_{\text{pl-narx}} = K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,j})$$

$$K^{\text{eq}}_{\text{pl-ar-narx}} = K(\boldsymbol{z}_{B,i+1}, \boldsymbol{z}_{B,j+1}) - \rho K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,j+1}) - \rho K(\boldsymbol{z}_{B,i+1}, \boldsymbol{z}_{B,j}) + \rho^2 K(\boldsymbol{z}_{B,i}, \boldsymbol{z}_{B,j})$$

# Example NARX Model

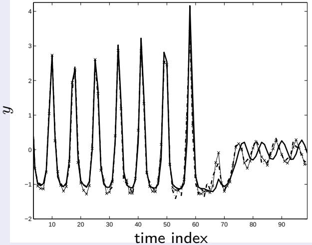
## Santa Fe Laser Data




- Subsample of 200 datapoints for Fixed-Size LS-SVM

- Hyperparameters σ, γ determined by cross-validation

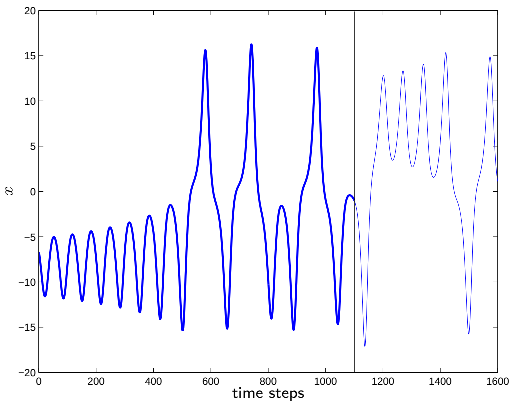
## Iterative Predictions

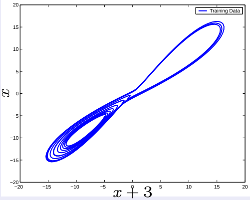


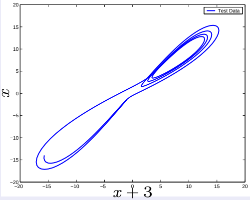

[Espinoza *et al.*, IEEE CDC 2003]

# Example of NARX Model with Symmetry

## Chaotic Time Series

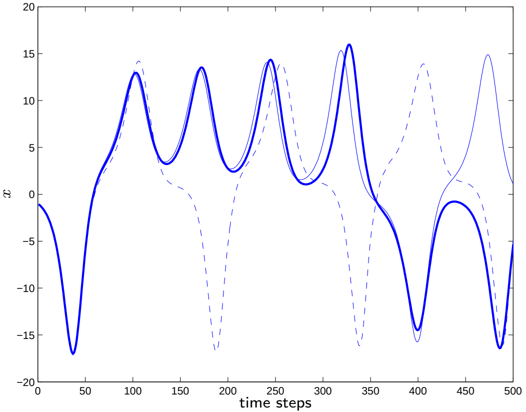


# Using Symmetry Improves Forecasting Performance

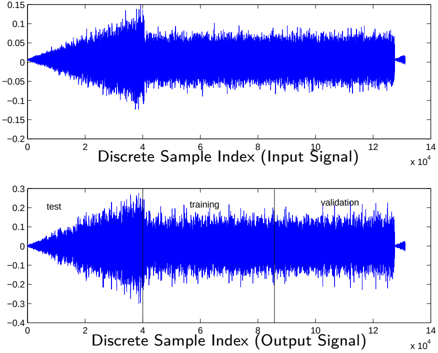
## Test Set Predictions

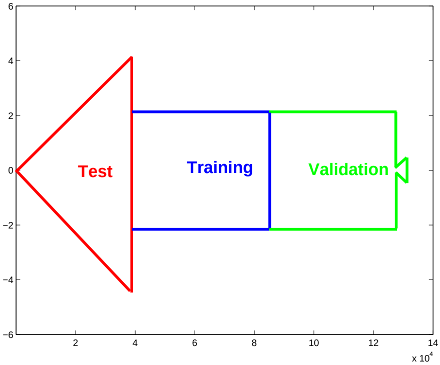



[Espinoza *et al.*, NOLTA 2005]

- Actual (thin)
- LS-SVM (dashed)
- **LS-SVM+S** (thick)

# The SilverBox Case Study



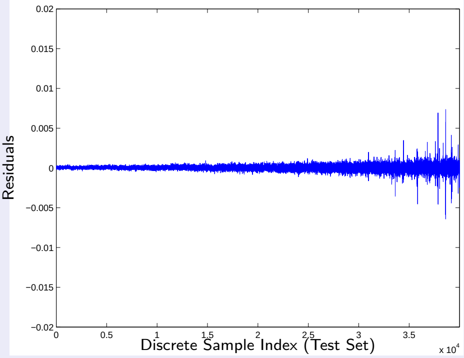




- Physical System, benchmark study
- Estimate model with 90,000 datapoints
- Simulate 40,000 test points

[Schoukens *et al.*, Automatica, 2003]

# SilverBox with NARX

## Test Set Residuals



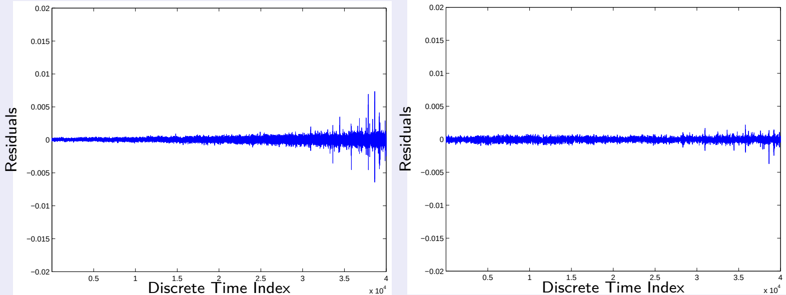

- Fixed-Size LS-SVM, Subsample of 1000 datapoints
- Polynomial kernel of degree $d = 3$
- RMSE$=3.2\times10^{-4}$
- Best Result in Special Session NOLCOS '04

[Espinoza *et al.*, NOLCOS 2004]

# SilverBox with PL-NARX

- Using prior knowledge about linear regressors
- RMSE$=2.7\times10^{-4}$

## Improvement over NARX

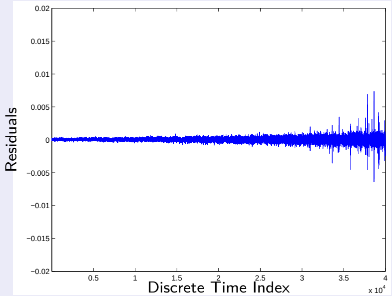


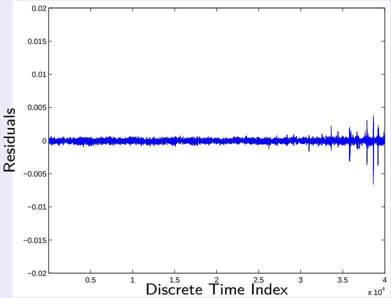

[Espinoza *et al.*, IEEE TAC Special Issue, 2005]

# SilverBox with NARX+Symmetry

- Using prior knowledge on symmetry of the nonlinear function
- RMSE=$2.8\times10^{-4}$

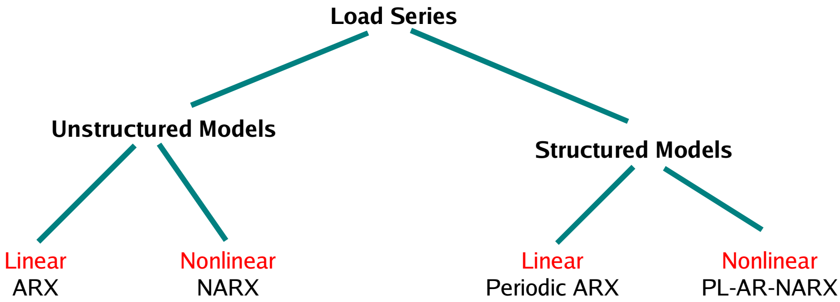
## Improvement over NARX



# Short-Term Load Forecasting

# Models Comparison

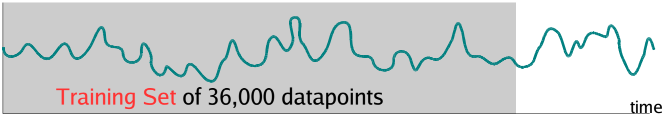



## Explanatory Variables

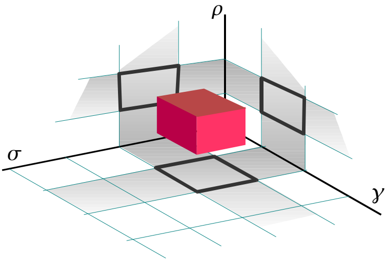
- Past values of the load
- Calendar Information
- Temperature
- Prior knowledge: Seasonal variations

# Implementation Issues

**Load Series**




**Hyperparameters** $\sigma$ $\gamma$ $\rho$
Determined by cross-validation




**Fixed-Size LS-SVM**

Using Equivalent Kernel
from each model structure

# (N)ARX Models

**Load Series**

**Unstructured Models**

Linear
ARX

Nonlinear
NARX

**LS-SVM in Dual Space**
Using the last 2,000 datapoints

**Fixed-Size LS-SVM**
Using the entire training set

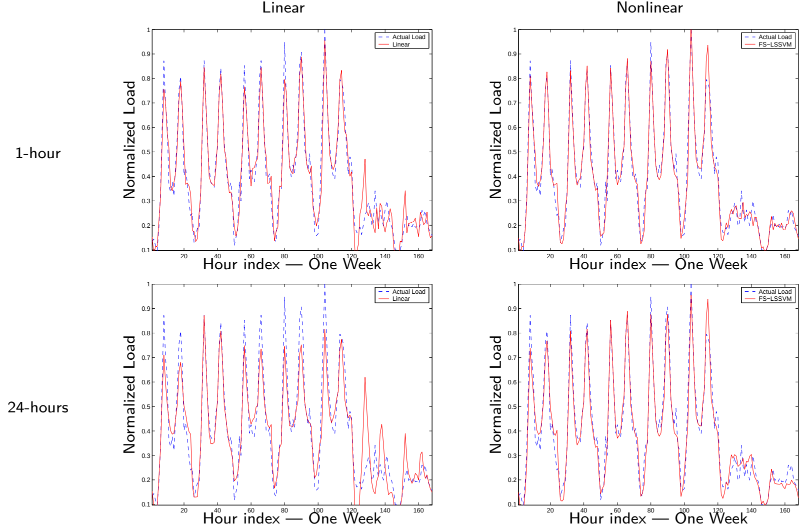
**Models Comparison**
1 hour ahead prediction, 24 hours ahead simulation
Test set of 1 week after training data
**10 load series**

[Espinoza *et al.*, CMS 2006]
[Espinoza *et al.*, IWANN 2005]

# (N)ARX Results for 3 series

| Series | Mode | Performance | LS-SVM | FS-LSSVM | Linear |
|---|---|---|---|---|---|
| Series 1 | 1-h-ahead | MSE | 2.2% | 0.6% | 1.4% |
| | | MAPE | 2.8% | 1.5% | 2.5% |
| | 24-h-ahead | MSE | 5.0% | 2.7% | 9.5% |
| | | MAPE | 4.3% | 3.1% | 5.9% |
| Series 2 | 1-h-ahead | MSE | 3.4% | 2.3% | 3.0% |
| | | MAPE | 4.3% | 3.4% | 3.9% |
| | 24-h-ahead | MSE | 20.2% | 11.5% | 11.9% |
| | | MAPE | 10.6% | 7.4% | 7.9% |
| Series 3 | 1-h-ahead | MSE | 9.7% | 6.7% | 10.2% |
| | | MAPE | 29.4% | 17.7% | 24.9% |
| | 24-h-ahead | MSE | 15.1% | 9.4% | 15.0% |
| | | MAPE | 30.1% | 23.1% | 29.7% |

# (N)ARX Forecasting Examples



# Structured Models

**Load Series**

**Structured Models**

Linear
Periodic ARX

Nonlinear
PL-AR-NARX

**Periodic ARX**
System of 24 equations
Very Interpretable
"Typical Daily Profile"
(Clustering)

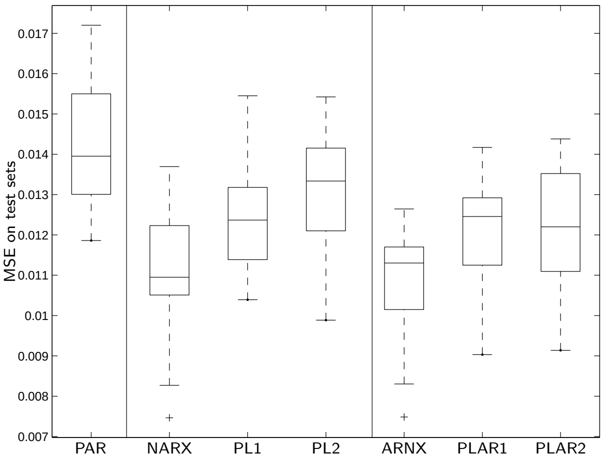
**Fixed-Size LS-SVM**
Single equation
Different variants
Residuals correlated past 24 hours

**Models Comparison**
1 hour ahead, 24 hours ahead
50 Test set of 1 week each
**4 load series**

[Espinoza *et al.*, IEEE TPS 2005]
[Espinoza *et al.*, PSCC 2005]

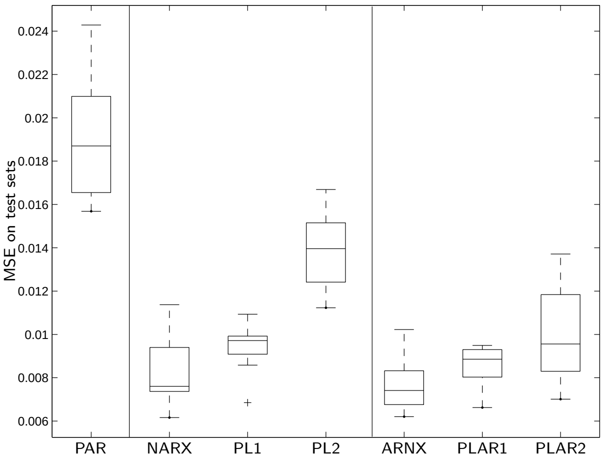
# Structured Models: Performance Comparison

## Predictions 1-h-ahead: Test Set Performance



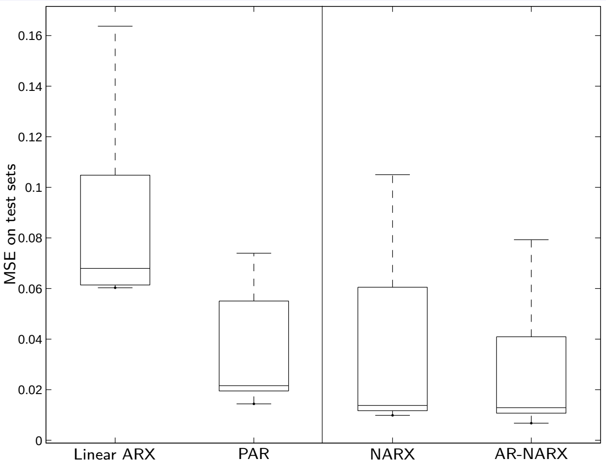
# Structured Models: Performance Comparison

## Predictions 1-h-ahead: Test Set Performance



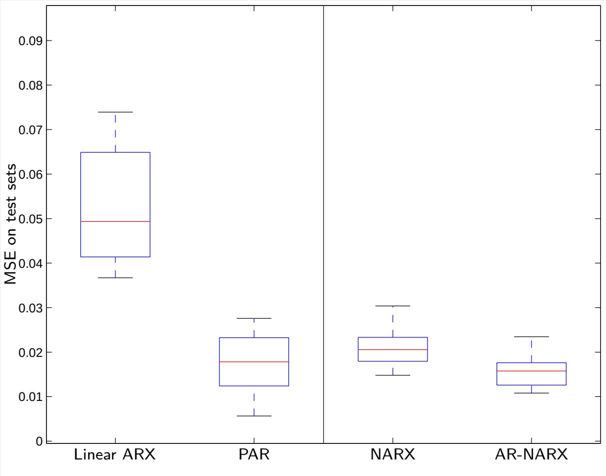
# The Benefit of using Structured Models for Simulation

## Simulations 24-h-ahead: Test Set Performance



# The Benefit of using Structured Models for Simulation

## Simulations 24-h-ahead: Test Set Performance

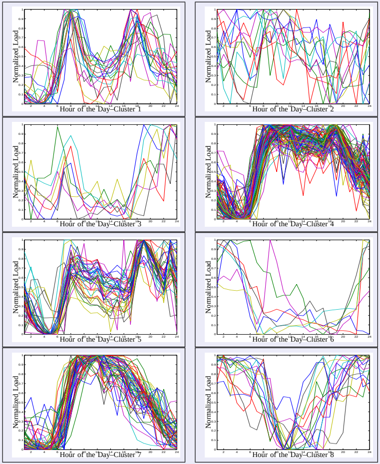


# Clustering using Periodic ARX Models

- Exploiting the vectorial representation
- Estimate models for 245 load series
- Define Typical Daily Profile
- Clean from seasonalities and weather
- Use the Profile for clustering

[Espinoza *et al.*, IEEE TPS 2005]

## 8 clusters found



# Conclusions and Future Research

# Conclusions

- Extend the LS-SVM regression to incorporate elements of prior knowledge
  - Symmetry
  - Partially Linear Structure
  - Autocorrelated Residuals
- Use extensions as building blocks for a modular approach to Nonlinear System Identification
  - (PL)-(AR)-NARX
  - Equivalent Kernels are very practical for large scale problems
  - Structured models with prior knowledge show better performance

# Conclusions

For the application of Load Forecasting

- It is good to build a model from a large dataset
- Structured models are preferable over unstructured models for the series under study
  - Linear PAR models
  - Nonlinear PL-AR-NARX
- PL-AR-NARX gives accurate predictions and interpretable coefficients

# Future Research

- Extend the framework to hypothesis testing
- Define new hyperparameters selection procedure
- Extend to other model structures (NARMAX, NBJ, NOE)
- Compute prediction errors in presence of additional constraints
- Further applications around and beyond Load Forecasting

# Acknowledgements


- Promotors
  Bart De Moor, Ronnie Belmans
- Kernel Methods
  Johan Suykens, Kristiaan Pelckmans, Luc Hoegaerts, Jos De Brabanter
- System Identification
  Tony Van Gestel, Jeroen Boets, Katrien De Cock
- Energy Islands
  Edwin Haesen, Vu Van Thong, Johan Driesen
- Glycemia Modeling
  Tom Van Herpe, Bert Pluymers, Ivan Goethals
- External Projects
  Caroline Joye (ELIA), Peter Maas (MasterCard)