Kernel based identification of systems with multiple outputs using nuclear norm regularization

Tillmann Falck, Bart De Moor and Johan A. K. Suykens

KU Leuven, ESAT-SCD and iMinds Future Health Department

International Workshop on Advances in Regularization, Optimization, Kernel Methods and Support Vector Machines: theory and applications

July 8 - 10, 2013; Leuven, Belgium

Outline

Introduction

Model derivation

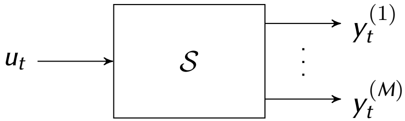
Example

Optimization aspects

Conclusions and outlook

Nonlinear system identification with multiple outputs

Overview

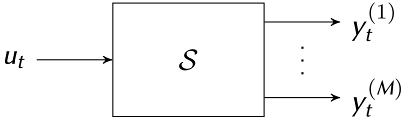


- Given input-output data $(u_t, \mathrm{y}_t)_{t=1}$ with $\mathrm{y}_t = [y_t^{(1)}, \ldots, y_t^{(M)}]^T$
- Estimate a mathematical model for $\mathcal{S}$
- Often solved via regression:
  $(y_{t-1}, \ldots, y_{t-P}, u_t, \ldots, u_{t-Q}) \rightarrow y_t$
- Kernel based methods and support vector techniques in particular quite successful

- Applications: load/demand forecasting, virtual sensors, ...

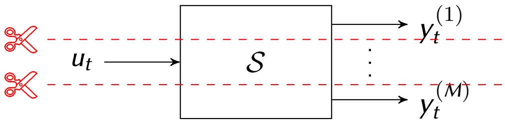
Nonlinear system identification with multiple outputs

Traditional versus proposed approach



Nonlinear system identification with multiple outputs

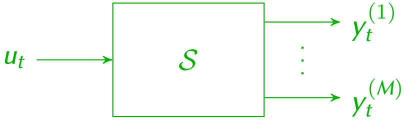
Traditional versus proposed approach



Traditional approach

Nonlinear system identification with multiple outputs

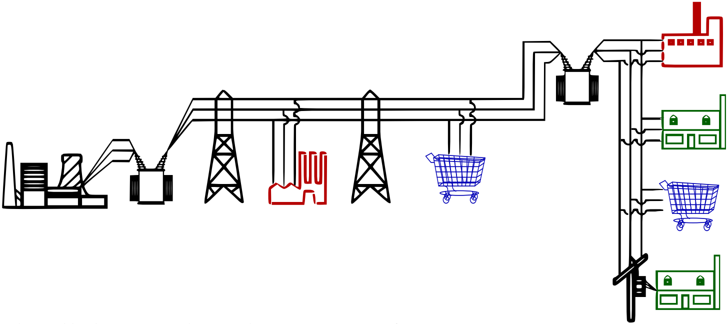
Traditional versus proposed approach



Traditional approach

Proposed approach

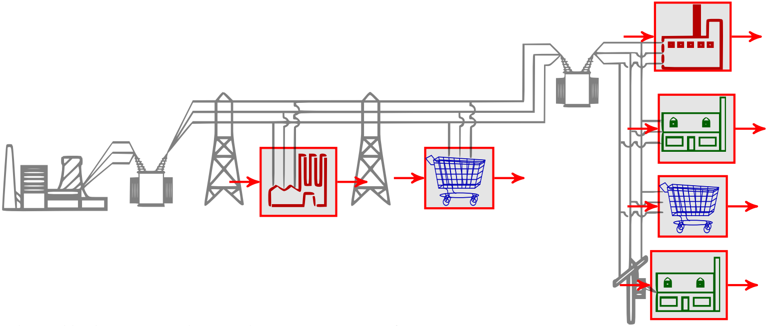
Example for multiple output system



Based on public domain image by United States Department of Energy

C. Alzate, M. Espinoza, B. De Moor, and J.A.K. Suykens (2009). *“Identifying customer profiles in power load time series using spectral clustering”*. In: Proc. of the 19th International Conference on Artificial Neural Networks.

Example for multiple output system



Based on public domain image by United States Department of Energy

C. Alzate, M. Espinoza, B. De Moor, and J.A.K. Suykens (2009). *"Identifying customer profiles in power load time series using spectral clustering"*. In: Proc. of the 19th International Conference on Artificial Neural Networks.

Key contributions and challenges

Contributions

- Kernel based model for nonlinear systems with multiple related outputs
- New primal-dual derivation of kernel based model with nuclear norm regularization

Challenges

- Finding a kernel based problem formulation
- Connecting dual, kernel based solution, to original model
- Numerical solution

Nuclear or trace norm

$$\|W\|_*$$

Basic properties

- matrix norm
- sum of singular values
- induces sparsity
  - can be interpreted as ℓ_1-norm of singular values
  - promotes low-rank solutions

As regularization term in a support vector model

- columns w_i model parameters, i.e. $y_t^{(i)} = w_i^T \varphi(x_t) + b$
- promotes relations between models in feature space

M. Fazel (2002). *"Matrix Rank Minimization with Applications"*. PhD thesis, Stanford.

Model formulation in a primal-dual setting

Primal model

$$\min_{\mathrm{W},\mathrm{b},\mathrm{e}_t} \quad \eta \|\mathrm{W}\|_* + \sum_{t=1}^{N} \mathrm{e}_t^T \mathrm{e}_t$$

$$\text{subject to} \quad y_t^{(i)} = \mathrm{w}_i^T \varphi(\mathrm{x}_t) + b_i + e_t^{(i)},$$
$$t = 1, \ldots, N, i = 1, \ldots, M$$

Derivation in primal-dual setting

(a) Write down Lagrangian for primal problem

(b) Take derivatives with respect to optimization variables

(c) Formulate KKT conditions for optimality

(d) Write down dual optimization problem

(e) Substitute dual solution into primal model

A. Argyriou, C.A. Micchelli, and M.A. Pontil (2009). *"When Is There a Representer Theorem? Vector Versus Matrix Regularizers"*. In: Journal of Machine Learning Research.

Model formulation in a primal-dual setting

Primal model

$$\min_{\mathrm{W},\mathrm{b},\mathrm{e}_t} \quad \eta\|\mathrm{W}\|_* + \sum_{t=1}^{N} \mathrm{e}_t^T \mathrm{e}_t$$

$$\text{subject to} \quad y_t^{(i)} = \mathrm{w}_i^T \varphi(\mathrm{x}_t) + b_i + e_t^{(i)},$$
$$t = 1, \ldots, N, i = 1, \ldots, M$$

Derivation in primal-dual setting

(a) Write down Lagrangian for primal problem

(b) Take derivatives with respect to optimization variables

(c) Formulate KKT conditions for optimality

(d) Write down dual optimization problem

(e) Substitute dual solution into primal model

A. Argyriou, C.A. Micchelli, and M.A. Pontil (2009). *"When Is There a Representer Theorem? Vector Versus Matrix Regularizers"*. In: Journal of Machine Learning Research.

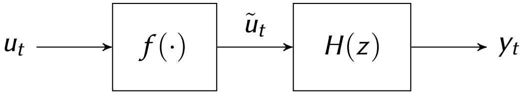
Advantage of primal-dual approach

Straightforward to incorporate additional structure

Example: simple constraints

- $\mathrm{w}_1^T \varphi(\mathrm{x}) = -\mathrm{w}_2^T \varphi(\mathrm{x})$
- $\mathrm{w}_i^T \varphi(\mathrm{x}) \geq y_0$

Example: Hammerstein systems



- Static nonlinearity $f(x) = \mathrm{w}^T \varphi(x) + c$
- Linear dynamical system $H : \hat{y}_t = \sum_{p=1}^{P} a_p y_{t-p} + \sum_{q=0}^{Q} b_q \tilde{u}_{t-q}$
- Approximate joint model
  $\hat{y}_t = \sum_{p=1}^{P} a_p y_{t-p} + \sum_{q=0}^{Q} \mathrm{w}_q^T \varphi(u_{t-q}) + \tilde{c}$

J.A.K. Suykens, C. Alzate, and K. Pelckmans (2010). *"Primal and dual model representations in kernel-based learning"*. In: Statistics Surveys.

Derivation of kernel based model

From Lagrangian to KKT conditions

Lagrangian

$$\mathcal{L} = \eta\|\mathrm{W}\|_* + \sum_{t=1}^{N} \mathrm{e}_t^T \mathrm{e}_t - \sum_{t=1}^{N} \boldsymbol{\alpha}^T(\mathrm{W}^T\boldsymbol{\varphi}(\mathrm{x}_t) + \mathrm{b} + \mathrm{e}_t - \mathrm{y}_t)$$

Nuclear norm is not differentiable!

Possible reformulations

- Dual norm, where $\|\cdot\|_2$ is dual norm of $\|\cdot\|_*$

  $$\|\mathrm{W}\|_* = \max_{\|\mathrm{C}\|_2 \leq 1} \mathrm{tr}(\mathrm{C}^T\mathrm{W})$$

- Conic duality, where $\mathcal{K}$ is convex cone $\{(\mathrm{X}, s) | \|\mathrm{X}\|_* \leq s\}$

  $$\|\mathrm{W}\|_* = t \text{ where } (\mathrm{W}, t) \in \mathcal{K}$$

Derivation of kernel based model

Kernel based optimization problem

$$\begin{aligned} \max_{\mathrm{A}} \quad & \mathrm{tr}(\mathrm{A}^T\mathrm{Y}) - \frac{1}{2}\,\mathrm{tr}(\mathrm{A}^T\mathrm{A}) \\ \text{subject to} \quad & \mathrm{A}^T\boldsymbol{\Omega}\mathrm{A} \preceq \eta \mathrm{I}_M, \\ & \mathrm{A}^T\mathbf{1}_N = \mathbf{0}_M \end{aligned}$$

- $\mathrm{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N]^T, \mathrm{Y} = [\mathrm{y}_1, \ldots, \mathrm{y}_N]^T$
- $\Omega_{ij} = K(\mathrm{x}_i, \mathrm{x}_j) = \boldsymbol{\varphi}(\mathrm{x}_i)^T\boldsymbol{\varphi}(\mathrm{x}_j)$

KKT condition for W

$$\mathrm{C} = \sum_{t=1}^{N} \boldsymbol{\varphi}(\mathrm{x}_t)\boldsymbol{\alpha}_t^T$$

- $\mathrm{W}, \mathrm{C} \in \mathbb{R}^{n_h \times M}$, $\boldsymbol{\varphi}(\mathrm{x}) \in \mathbb{R}^{n_h}$ and $\boldsymbol{\alpha}_t \in \mathbb{R}^M$
- No expansion of primal variables W!

Formulation of model in terms of dual solution

Characterization of solution set

$$\{W : \mathrm{tr}(W^T C) = \xi, \|W\|_* = \xi\}$$
$$= \{U_\eta H_\eta V_\eta^T : \mathrm{tr}(H_\eta) = \xi, 0 \preceq H_\eta \in \mathbb{R}^{r\times r}\}$$

- Here C and ξ are fixed
- $C = \mathbf{\Phi} A^T$
- U_η, V_η contain left and right singular vectors corresponding to largest singular value of C respectively

Connecting W and A

$$\begin{aligned}
\text{find} \quad & H_\eta \\
\text{subject to} \quad & H_\eta \succeq 0, \mathrm{tr}(H_\eta) = \xi \\
& y^{(i)} = [\mathbf{\Omega}_{i,1}\boldsymbol{\alpha}_1, \ldots, \mathbf{\Omega}_{i,M}\boldsymbol{\alpha}_M] V_\eta H_\eta V_\eta^T \boldsymbol{\epsilon}_i \\
& \quad + b_i \mathbf{1}_{N_i} + \boldsymbol{\alpha}_i, \quad i = 1, \ldots, M
\end{aligned}$$

Model representation

Primal model representation

$$\widehat{y}^{(i)} = f(\mathrm{z}) = \mathrm{w}_i^T \varphi(\mathrm{z}) + b_i$$

Dual model representation

$$\widehat{y}^{(i)} = f(\mathrm{z}) = \sum_{j=1}^{M} Q_{ji} \sum_{t=1}^{N} A_{tj} K(\mathrm{x}_t, \mathrm{z}) + b_i$$

with $\mathrm{Q} = \mathrm{V}_\eta \mathrm{H}_\eta \mathrm{V}_\eta^T$

One dimensional SVM

$$\widehat{y} = f(\mathrm{z}) = \sum_{t=1}^{N} \alpha_t K(\mathrm{x}_t, \mathrm{z}) + b$$

Example

Description

Toy example

- number of data: training 50, validation 100, test 150
- number of outputs: 20
- number of independent contributions: 3
- data generation: $\mathrm{Y} = \mathrm{W}_0^T \boldsymbol{\Phi} + \text{noise}$
- $\boldsymbol{\Phi} \in \mathbb{R}^{50\times 50}$, $\mathrm{W}_0 = \sum_{i=1}^{3} \mathrm{g}_i \mathrm{r}_i^T = \mathrm{G}\mathrm{R}^T$, $\mathrm{G} \in \mathbb{R}^{50\times 3}$, $\mathrm{R} \in \mathbb{R}^{20\times 3}$

Evaluated models

MIMO proposed nuclear norm regularized model

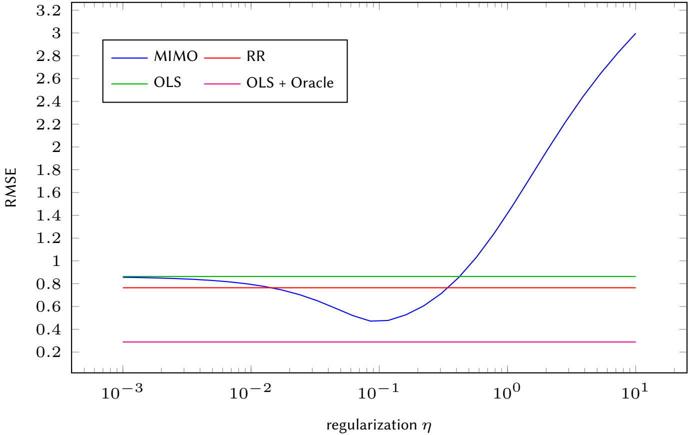
RR ridge regression model (LS-SVM model in primal with given feature map) with independent LS-SVM models for each output

OLS ordinary least squares model

OLS + oracle OLS given the true structure of problem

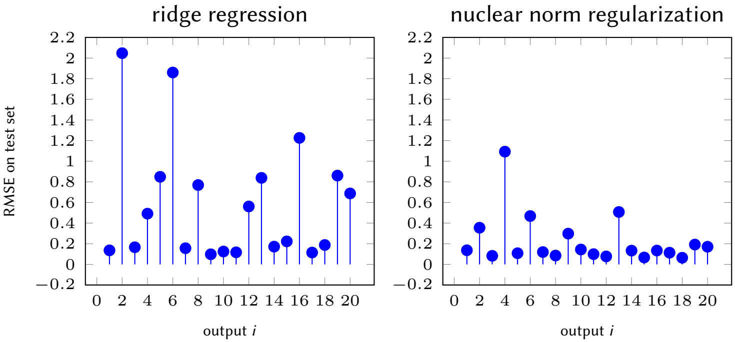
Example

Cross validation



Example

Results



Comparison of optimization problems

Primal formulation

$$\min_{\mathrm{W},\mathrm{b},\mathrm{e}_t} \quad \eta\|\mathrm{W}\|_* + \sum_{t=1}^{N} \mathrm{e}_t^T \mathrm{e}_t$$

$$\text{subject to} \quad y_t^{(i)} = \mathrm{w}_i^T \varphi(\mathrm{x}_t) + b_i + e_t^{(i)}$$

Dual formulation

Optimization problem

$$\max_{\mathrm{A}} \quad \mathrm{tr}(\mathrm{A}^T\mathrm{Y}) - \frac{1}{2}\mathrm{tr}(\mathrm{A}^T\mathrm{A})$$

$$\text{subject to} \quad \mathrm{A}^T\boldsymbol{\Omega}\mathrm{A} \preceq \eta \mathrm{I}_M, \quad \mathrm{A}^T\mathbf{1}_N = \mathbf{0}_M$$

Reconstruction of model

$$\text{find} \quad \mathrm{H}_\eta$$

subject to

$$\mathrm{H}_\eta \succeq 0, \mathrm{tr}(\mathrm{H}_\eta) = \xi$$

$$\mathrm{y}^{(i)} = [\boldsymbol{\Omega}_{i,1}\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\Omega}_{i,M}\boldsymbol{\alpha}_M] \cdot \mathrm{V}_\eta \mathrm{H}_\eta \mathrm{V}_\eta^T \boldsymbol{\epsilon}_i + b_i \mathbf{1}_{N_i} + \boldsymbol{\alpha}_i$$

Dimensionalities of optimization problems

	One input	L inputs
Primal		
Feature map	$n_h \times N$	$L \cdot (n_h \times N)$
Unknown W	$n_h \times M$	$L \cdot (n_h \times M)$
Dual		
Kernel matrix	$N \times N$	$(L \cdot N) \times (L \cdot N)$
Unknown A	$N \times M$	$(L \cdot N) \times M$

- N: number of data
- M: number of outputs
- n_h: dimension of feature map

Solution strategies

SDP formulation

- o Problems are SDP representable
- o Can be solved with general purpose SDP solvers
- \+ Small implementation effort
- \+ High accuracy
- \- High runtime costs, memory & times $\Rightarrow$ Limited to small problem sizes

First order techniques

- o (Accelerated) gradient projection can be applied
- o Higher implementation effort
- \+ Structure can be exploited
- \+ Scale to larger problem sizes
- \- Lower accuracy (crucial for reconstruction of dual model)

Conclusions

- Proposed novel identification for nonlinear systems with multiple outputs
  - Exploits relations between output variables
  - Illustrated improvement on small toy example
- Presented derivation of a nuclear norm regularized model in primal-dual setting
  - Allows straightforward integration of additional information

Challenges and Outlook

- Application on real world datasets
- Numerical solution on larger datasets
  - New algorithms are already being developed
  - Computational power increases exponentially
- Same primal-dual derivation can be applied to other nonquadratic regularization schemes
- Conjecture
  - Promising applications in system identification
  - Kernel based models can be used for many applications besides regression, these might also benefit from advanced regularization schemes