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 \((u_t, y_t)_{t=1}^T\) with \(y_t = [y_t^{(1)}, \ldots, y_t^{(M)}]^T\)

Estimate a mathematical model for \(S\)

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

\[ u_t \rightarrow S \rightarrow y_t^{(1)} \rightarrow \cdots \rightarrow y_t^{(M)} \]
Nonlinear system identification with multiple outputs

Traditional versus proposed approach

Traditional approach

Proposed 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

Example for multiple output system

Based on public domain image by United States Department of Energy

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 \( \ell_1 \)-norm of singular values
  - promotes low-rank solutions

As regularization term in a support vector model
- columns \( w_i \) model parameters, i.e. \( y_{(i)}^t = w_i^T \varphi(x_t) + b \)
- promotes relations between models in feature space

Model formulation in a primal-dual setting

Primal model

\[
\begin{align*}
\min_{W,b,e_t} & \quad \eta \|W\|_* + \sum_{t=1}^{N} e_t e_t \\
\text{subject to} & \quad y_t^{(i)} = w_i^T \varphi(x_t) + b_i + e_t^{(i)}, \\
& \quad t = 1, \ldots, N, i = 1, \ldots, M
\end{align*}
\]

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


Model formulation in a primal-dual setting

Primal model

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

subject to

$$y_t^{(i)} = w_i^T \varphi(x_t) + b_i + e_t^{(i)},$$

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

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

Advantage of primal-dual approach

Straightforward to incorporate additional structure

Example: simple constraints

- \( w_1^T \varphi(x) = -w_2^T \varphi(x) \)
- \( w_i^T \varphi(x) \geq y_0 \)

Example: Hammerstein systems

- Static nonlinearity \( f(x) = w^T \varphi(x) + c \)
- 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 \( \hat{y}_t = \sum_{p=1}^{P} a_p y_{t-p} + \sum_{q=0}^{Q} w_q^T \varphi(u_{t-q}) + \tilde{c} \)

Derivation of kernel based model

From Lagrangian to KKT conditions

Lagrangian

\[
\mathcal{L} = \eta \|W\|_* + \sum_{t=1}^{N} e_t^T e_t - \sum_{t=1}^{N} \alpha^T (W^T \varphi(x_t) + b + e_t - y_t)
\]

Nuclear norm is not differentiable!

Possible reformulations

- Dual norm, where \(\| \cdot \|_2\) is dual norm of \(\| \cdot \|_*\)
  \[
  \|W\|_* = \max_{\|C\|_2 \leq 1} \text{tr}(C^T W)
  \]

- Conic duality, where \(\mathcal{K}\) is convex cone \(\{(X, s) | \|X\|_* \leq s\}\)
  \[
  \|W\|_* = t \text{ where } (W, t) \in \mathcal{K}
  \]
Derivation of kernel based model

Kernel based optimization problem

\[
\max_A \quad \text{tr}(A^T Y) - \frac{1}{2} \text{tr}(A^T A) \\
\text{subject to} \quad A^T \Omega A \preceq \eta I_M, \\
A^T 1_N = 0_M
\]

\[\begin{align*}
\text{A} &= [\alpha_1, \ldots, \alpha_N]^T, \quad Y = [y_1, \ldots, y_N]^T \\
\Omega_{ij} &= K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)
\end{align*}\]

KKT condition for \( W \)

\[
C = \sum_{t=1}^{N} \varphi(x_t) \alpha_t^T
\]

\[\begin{align*}
W, C &\in \mathbb{R}^{n_h \times M}, \quad \varphi(x) \in \mathbb{R}^{n_h} \text{ and } \alpha_t \in \mathbb{R}^M \\
\text{No expansion of primal variables W!}
\end{align*}\]
Formulation of model in terms of dual solution

Characterization of solution set

\[ \{ W : \text{tr}(W^T C) = \xi, \|W\|_* = \xi \} \]

\[ = \{ U_\eta H_\eta V_\eta^T : \text{tr}(H_\eta) = \xi, 0 \preceq H_\eta \in \mathbb{R}^{r \times r} \} \]

- Here \( C \) and \( \xi \) are fixed
- \( C = \Phi A^T \)
- \( U_\eta, V_\eta \) contain left and right singular vectors corresponding to largest singular value of \( C \) respectively

Connecting \( W \) and \( A \)

find \( H_\eta \)

subject to \( H_\eta \succeq 0, \text{tr}(H_\eta) = \xi \)

\[ y^{(i)} = [\Omega_{i,1} \alpha_1, \ldots, \Omega_{i,M} \alpha_M] V_\eta H_\eta V_\eta^T \epsilon_i \]

\[ + b_i 1_{N_i} + \alpha_i, \quad i = 1, \ldots, M \]
Model representation

Primal model representation

\[ \hat{y}^{(i)} = f(z) = w_i^T \varphi(z) + b_i \]

Dual model representation

\[ \hat{y}^{(i)} = f(z) = \sum_{j=1}^{M} Q_{ji} \sum_{t=1}^{N} A_{tj} K(x_t, z) + b_i \]

with \( Q = V_\eta H_\eta V_\eta^T \)

One dimensional SVM

\[ \hat{y} = f(z) = \sum_{t=1}^{N} \alpha_t K(x_t, z) + b \]
Example

Description

Toy example

- number of data: training 50, validation 100, test 150
- number of outputs: 20
- number of independent contributions: 3
- data generation: \( Y = W_0^T \Phi + \text{noise} \)
- \( \Phi \in \mathbb{R}^{50 \times 50}, \ W_0 = \sum_{i=1}^{3} g_i r_i^T = GR^T, \ G \in \mathbb{R}^{50 \times 3}, \ R \in \mathbb{R}^{20 \times 3} \)

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

\begin{figure}[h]
\centering
\begin{subfigure}{0.4\textwidth}
\centering
\includegraphics[width=\textwidth]{ridge_regression}
\caption{ridge regression}
\end{subfigure}\hfill
\begin{subfigure}{0.4\textwidth}
\centering
\includegraphics[width=\textwidth]{nuclear_norm_regularization}
\caption{nuclear norm regularization}
\end{subfigure}
\end{figure}
Comparison of optimization problems

Primal formulation

\[
\begin{align*}
\min_{W,b,e_t} & \quad \eta \|W\|_* + \sum_{t=1}^{N} e_t^T e_t \\
\text{subject to} & \quad y_t^{(i)} = w_i^T \varphi(x_t) + b_i + e_t^{(i)}
\end{align*}
\]

Dual formulation

Optimization problem

\[
\max_{A} \quad \text{tr}(A^T Y) - \frac{1}{2} \text{tr}(A^T A)
\]
\[
\text{subject to} \quad A^T \Omega A \preceq \eta I_M,
\]
\[
A^T 1_N = 0_M
\]

Reconstruction of model

\[
\text{find} \quad H_{\eta}
\]
\[
\text{subject to} \quad H_{\eta} \succeq 0, \text{tr}(H_{\eta}) = \xi
\]
\[
y^{(i)} = [\Omega_{i,1} \alpha_1, \ldots, \Omega_{i,M} \alpha_M]
\]
\[
\cdot V_{\eta} H_{\eta} V_{\eta}^T \epsilon_i + b_i 1_N + \alpha_i
\]
## Dimensionalities of optimization problems

<table>
<thead>
<tr>
<th></th>
<th>One input</th>
<th>( L ) inputs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primal</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feature map</td>
<td>( n_h \times N )</td>
<td>( L \cdot (n_h \times N) )</td>
</tr>
<tr>
<td>Unknown ( W )</td>
<td>( n_h \times M )</td>
<td>( L \cdot (n_h \times M) )</td>
</tr>
<tr>
<td><strong>Dual</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kernel matrix</td>
<td>( N \times N )</td>
<td>( (L \cdot N) \times (L \cdot N) )</td>
</tr>
<tr>
<td>Unknown ( A )</td>
<td>( N \times M )</td>
<td>( (L \cdot N) \times M )</td>
</tr>
</tbody>
</table>

- \( N \): number of data
- \( M \): number of outputs
- \( n_h \): dimension of feature map
Solution strategies

SDP formulation

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

First order techniques

- (Accelerated) gradient projection can be applied
- 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