Smoothness priors, Shrinkage and Sparsity in System Identification: Bayesian procedures from a classical perspective

Alessandro Chiuso

Department of Information Engineering
University of Padova

Control Day - Padova
September 20th, 2013

Joint work with: G. Pillonetto, G. De Nicolao, A. Aravkin, J. Burke, T. Chen, L. Ljung

Main Goal of this lecture

Discuss *Nonparametric Bayesian Procedures* for System Identification

Main Goal of this lecture

Discuss *Nonparametric Bayesian Procedures* for System Identification

- **Regularization for Sparsity and Shrinking**

Main Goal of this lecture

Discuss *Nonparametric Bayesian Procedures* for System Identification

- **Regularization for Sparsity and Shrinking**
  - Link with Multiple Kernel Learning (MKL) and Group Lasso

Main Goal of this lecture

Discuss *Nonparametric Bayesian Procedures* for System Identification

- **Regularization for Sparsity and Shrinking**
  - Link with Multiple Kernel Learning (MKL) and Group Lasso
  - Sparsity vs. Shrinking (Example)

Main Goal of this lecture

Discuss *Nonparametric Bayesian Procedures* for System Identification

- **Regularization for Sparsity and Shrinking**
  - Link with Multiple Kernel Learning (MKL) and Group Lasso
  - Sparsity vs. Shrinking (Example)
  - MSE Properties (Classical View)

Main Goal of this lecture

Discuss *Nonparametric Bayesian Procedures* for System Identification

- **Regularization for Sparsity and Shrinking**
  - Link with Multiple Kernel Learning (MKL) and Group Lasso
  - Sparsity vs. Shrinking (Example)
  - MSE Properties (Classical View)
- **Kernels for SI**
  - Exponentially weighted Kernels

Main Goal of this lecture

Discuss *Nonparametric Bayesian Procedures* for System Identification

- **Regularization for Sparsity and Shrinking**
  - Link with Multiple Kernel Learning (MKL) and Group Lasso
  - Sparsity vs. Shrinking (Example)
  - MSE Properties (Classical View)
- **Kernels for SI**
  - Exponentially weighted Kernels
  - Empirical Bayes estimators

Main Goal of this lecture

Discuss *Nonparametric Bayesian Procedures* for System Identification

- **Regularization for Sparsity and Shrinking**
  - Link with Multiple Kernel Learning (MKL) and Group Lasso
  - Sparsity vs. Shrinking (Example)
  - MSE Properties (Classical View)
- **Kernels for SI**
  - Exponentially weighted Kernels
  - Empirical Bayes estimators
  - MSE Properties (white inputs)
- **Ongoing work**
  - Kernel Design

Main Goal of this lecture

Discuss *Nonparametric Bayesian Procedures* for System Identification

- **Regularization for Sparsity and Shrinking**
  - Link with Multiple Kernel Learning (MKL) and Group Lasso
  - Sparsity vs. Shrinking (Example)
  - MSE Properties (Classical View)
- **Kernels for SI**
  - Exponentially weighted Kernels
  - Empirical Bayes estimators
  - MSE Properties (white inputs)
- **Ongoing work**
  - Kernel Design
  - Multi-output Systems: Nuclear Norm and/or Vector Kernels

Main Goal of this lecture

Discuss *Nonparametric Bayesian Procedures* for System Identification

- **Regularization for Sparsity and Shrinking**
  - Link with Multiple Kernel Learning (MKL) and Group Lasso
  - Sparsity vs. Shrinking (Example)
  - MSE Properties (Classical View)
- **Kernels for SI**
  - Exponentially weighted Kernels
  - Empirical Bayes estimators
  - MSE Properties (white inputs)
- **Ongoing work**
  - Kernel Design
  - Multi-output Systems: Nuclear Norm and/or Vector Kernels
  - MSE Properties (general Inputs)

Motivation

- **Data sets with large cross-sectional dimension**

Motivation

- **Data sets with large cross-sectional dimension**
  - Spatially distributed sensor networks

Motivation

- **Data sets with large cross-sectional dimension**
  - Spatially distributed sensor networks
  - Econometrics/Finance

Motivation

- **Data sets with large cross-sectional dimension**
  - Spatially distributed sensor networks
  - Econometrics/Finance
  - High dimensional sensor processing (Vision - Tactile)

Motivation

- **Data sets with large cross-sectional dimension**
  - Spatially distributed sensor networks
  - Econometrics/Finance
  - High dimensional sensor processing (Vision - Tactile)
  - Social Networks

Motivation

- **Data sets with large cross-sectional dimension**
  - Spatially distributed sensor networks
  - Econometrics/Finance
  - High dimensional sensor processing (Vision - Tactile)
  - Social Networks
  - ....

Motivation

- **Data sets with large cross-sectional dimension**
  - Spatially distributed sensor networks
  - Econometrics/Finance
  - High dimensional sensor processing (Vision - Tactile)
  - Social Networks
  - ....
- **Parsimonious Models**

Motivation

- **Data sets with large cross-sectional dimension**
  - Spatially distributed sensor networks
  - Econometrics/Finance
  - High dimensional sensor processing (Vision - Tactile)
  - Social Networks
  - ....
- **Parsimonious Models**
  - Tradeoffs needed in high dimension

Motivation

- **Data sets with large cross-sectional dimension**
  - Spatially distributed sensor networks
  - Econometrics/Finance
  - High dimensional sensor processing (Vision - Tactile)
  - Social Networks
  - ....
- **Parsimonious Models**
  - Tradeoffs needed in high dimension
  - Interpretable models (who influences/is influenced by who?): Emphasis on dynamic interactions

Thermodynamic modeling

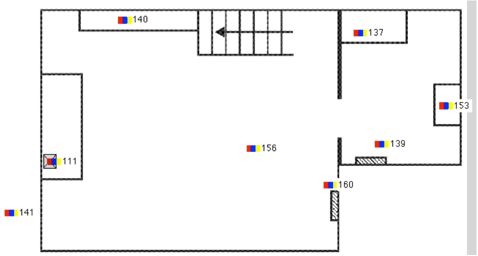


Figure : Sensors

Thermodynamic modeling

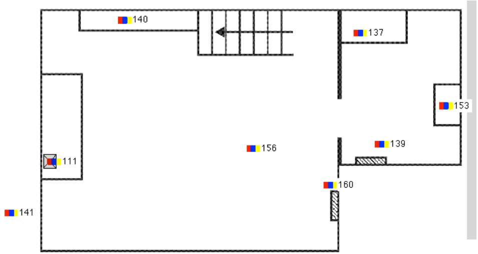
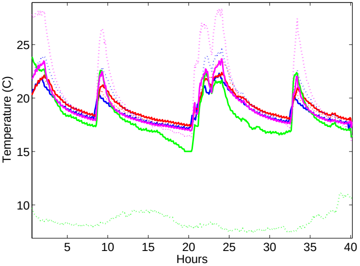


Figure : Sensors



Example 1

Thermodynamic modeling

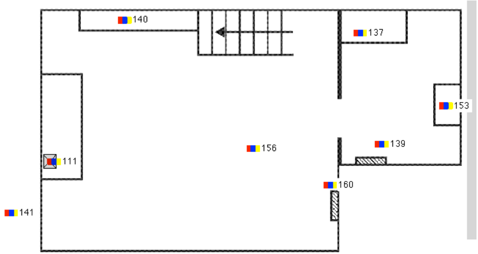
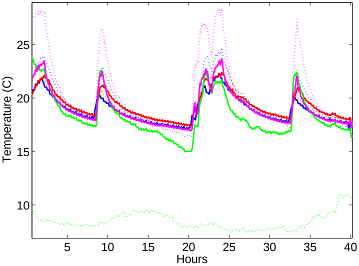


Figure : Sensors

- Fine prediction/modeling with few sensors



Thermodynamic modeling

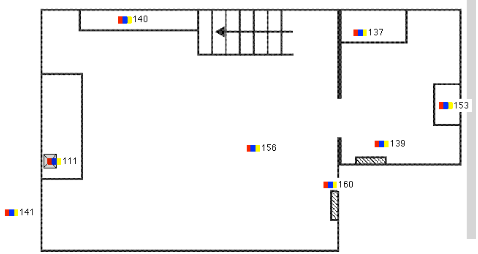
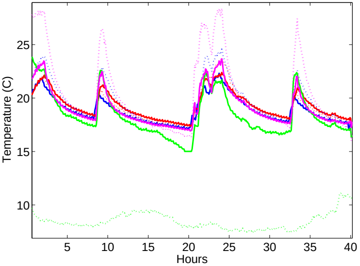


Figure : Sensors

- Fine prediction/modeling with few sensors
- Optimal sensor placement



Thermodynamic modeling

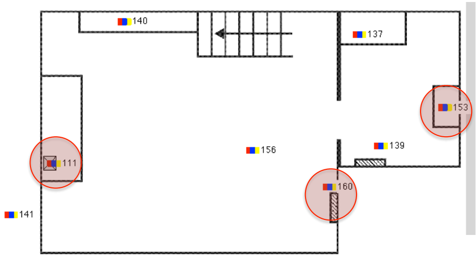
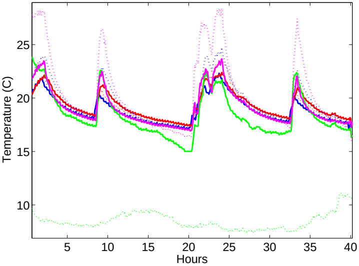


Figure : Sensors

- Fine prediction/modeling with few sensors
- Optimal sensor placement



Example 2

Video Sequence Processing

Figure : Video Courtesy of Mario Sznaier

Dynamic Bayesian Network

Vector Process z_t, $t \in \mathbb{Z}$

$$z(t) := \begin{pmatrix} z_t^{(1)} \\ \vdots \\ z_t^{(m)} \end{pmatrix}$$

Dynamic Bayesian Network

Vector Process z_t, $t \in \mathbb{Z}$

$$z(t) := \begin{pmatrix} z_t^{(1)} \\ \vdots \\ z_t^{(m)} \end{pmatrix}$$

$$\begin{aligned} \hat{z}^{(1)}_{t|t-1} &:= \hat{\mathbb{E}}\left[z_t^{(1)} | z_{t-1}, z_{t-2}, \ldots\right] \\ &= \sum_{i=1}^{m} \left[h^{(i)} * z^{(i)}\right](t) \end{aligned}$$

Dynamic Bayesian Network

Vector Process z_t, $t \in \mathbb{Z}$

$$z(t) := \begin{pmatrix} z_t^{(1)} \\ \vdots \\ z_t^{(m)} \end{pmatrix}$$

$$\begin{aligned} \hat{z}^{(1)}_{t|t-1} &:= \hat{\mathbb{E}}\left[z_t^{(1)} | z_{t-1}, z_{t-2}, \dots\right] \\ &= \textstyle\sum_{i=1}^m \left[h^{(i)} * z^{(i)}\right](t) \end{aligned}$$

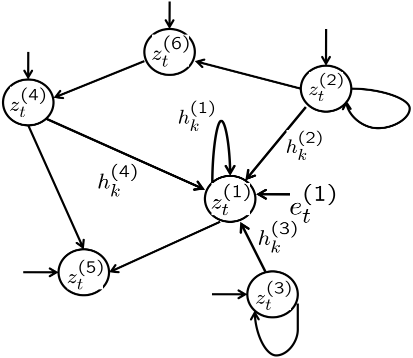


Figure : Granger causality graph

Dynamic Bayesian Network

Vector Process z_t, $t \in \mathbb{Z}$

$$z(t) := \begin{pmatrix} z_t^{(1)} \\ \vdots \\ z_t^{(m)} \end{pmatrix}$$

$$\begin{aligned} \hat{z}^{(1)}_{t|t-1} &:= \hat{\mathbb{E}}\left[z_t^{(1)} | z_{t-1}, z_{t-2}, \dots\right] \\ &= \sum_{i=1}^{m} \left[h^{(i)} * z^{(i)}\right](t) \\ &= \sum_{i=1}^{4} \left[h^{(i)} * z_t^{(i)}\right](t) \end{aligned}$$

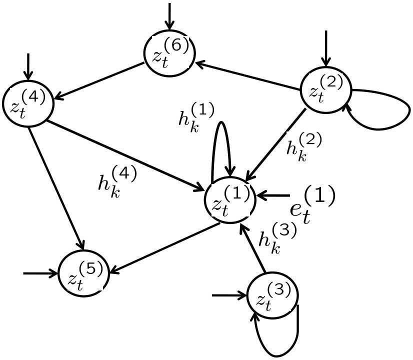


Figure : Granger causality graph

Dynamic Bayesian Network

Vector Process z_t, $t \in \mathbb{Z}$

$$z(t) := \begin{pmatrix} z_t^{(1)} \\ \vdots \\ z_t^{(m)} \end{pmatrix}$$

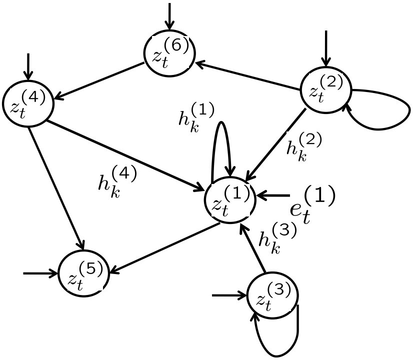


Figure : Granger causality graph

Dynamic Bayesian Network

Vector Process z_t, $t \in \mathbb{Z}$

$$z(t) := \begin{pmatrix} z_t^{(1)} \\ \vdots \\ z_t^{(m)} \end{pmatrix}$$

Identification

$\hat{h}^{(i)} = ?$

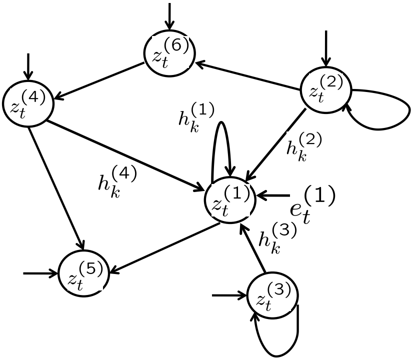


Figure : Granger causality graph

Dynamic Bayesian Network

Vector Process z_t, $t \in \mathbb{Z}$

$$z(t) := \begin{pmatrix} z_t^{(1)} \\ \vdots \\ z_t^{(m)} \end{pmatrix}$$

Identification

$\hat{h}^{(i)} = ?$

Variable Selection

$i \;\; s.t. \;\; \hat{h}^{(i)} = 0$

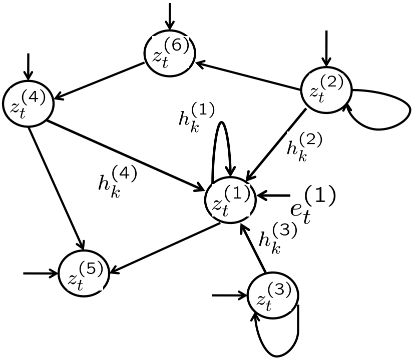


Figure : Granger causality graph

Case Study: ARMAX systems with variable selection

200 Sparse Randomly generated ARMAX systems

$$y_t = \sum_{i=1}^{20} \left[q^{(i)} * u^{(i)} \right](t) + [\ell * e](t)$$

Case Study: ARMAX systems with variable selection

Classical Parametric Perspective

Parametric Models (ARMAX/SS/Rational Basis etc...)

$$\hat{y}_{t|t-1}(\theta) = \left[h^{(1)}(\theta) * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)}(\theta) * u^{(i)}\right](t) \quad \theta \in \Theta \subseteq \mathbb{R}^n$$

Case Study: ARMAX systems with variable selection

Classical Parametric Perspective

Parametric Models (ARMAX/SS/Rational Basis etc...)

$$\hat{y}_{t|t-1}(\theta) = \left[h^{(1)}(\theta) * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)}(\theta) * u^{(i)}\right](t) \quad \theta \in \Theta \subseteq \mathbb{R}^n$$

Prediction Error Minimization (PEM)

$$\hat{\theta} := \arg\min_{\theta} \sum_{t} \left(y_t - \hat{y}_{t|t-1}(\theta)\right)^2$$

Case Study: ARMAX systems with variable selection

Classical Parametric Perspective

Parametric Models (ARMAX/SS/Rational Basis etc...)

$$\hat{y}_{t|t-1}(\theta) = \left[h^{(1)}(\theta) * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)}(\theta) * u^{(i)}\right](t) \quad \theta \in \Theta \subseteq \mathbb{R}^n$$

Prediction Error Minimization (PEM)

$$\hat{\theta} := \arg\min_{\theta} \sum_t \left(y_t - \hat{y}_{t|t-1}(\theta)\right)^2$$

1. **Model order estimation** (=Mc Millan Degree)

Case Study: ARMAX systems with variable selection

Classical Parametric Perspective

Parametric Models (ARMAX/SS/Rational Basis etc...)

$$\hat{y}_{t|t-1}(\theta) = \left[h^{(1)}(\theta) * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)}(\theta) * u^{(i)}\right](t) \quad \theta \in \Theta \subseteq \mathbb{R}^n$$

Prediction Error Minimization (PEM)

$$\hat{\theta} := \arg\min_{\theta} \sum_t \left(y_t - \hat{y}_{t|t-1}(\theta)\right)^2$$

1. **Model order estimation** (=Mc Millan Degree)
2. **Variable selection** (= test which “inputs” and/or “lags” are significant)

Parametric Models (ARMAX/SS/Rational Basis etc...)

$$\hat{y}_{t|t-1}(\theta) = \left[h^{(1)}(\theta) * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)}(\theta) * u^{(i)}\right](t) \quad \theta \in \Theta \subseteq \mathbb{R}^n$$

Prediction Error Minimization (PEM)

$$\hat{\theta} := \arg\min_{\theta} \sum_t \left(y_t - \hat{y}_{t|t-1}(\theta)\right)^2$$

1. **Model order estimation** (=Mc Millan Degree)
2. **Variable selection** (= test which “inputs” and/or “lags” are significant)
   - multiple tests: unfeasible for lager m (combinatorial)

Case Study: ARMAX systems with variable selection

Classical Parametric Perspective

Parametric Models (ARMAX/SS/Rational Basis etc...)

$$\hat{y}_{t|t-1}(\theta) = \left[h^{(1)}(\theta) * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)}(\theta) * u^{(i)}\right](t) \quad \theta \in \Theta \subseteq \mathbb{R}^n$$

Prediction Error Minimization (PEM)

$$\hat{\theta} := \arg\min_{\theta} \sum_t \left(y_t - \hat{y}_{t|t-1}(\theta)\right)^2$$

1. **Model order estimation** (=Mc Millan Degree)
2. **Variable selection** (= test which “inputs” and/or “lags” are significant)
   - multiple tests: unfeasible for lager m (combinatorial)
   - greedy procedures: stagewise methods... (may take advantage of *submodularity*)

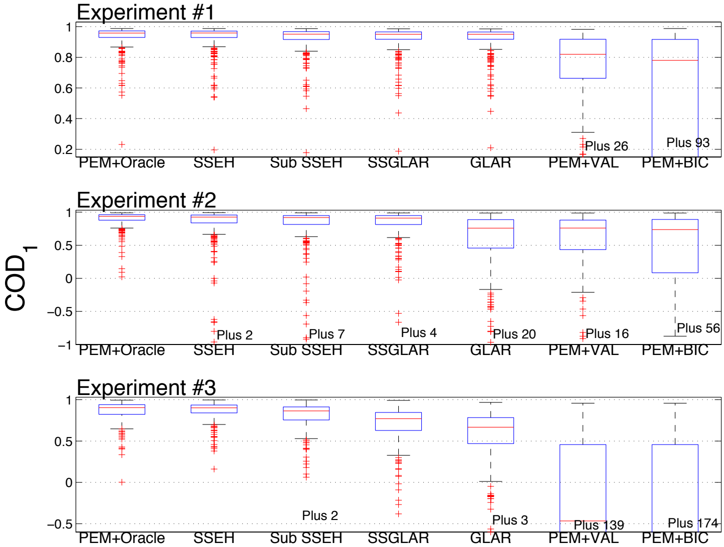
Case Study: ARMAX systems

Coefficient of determination

$$COD_k := 1 - \frac{Var(y_t - \hat{y}_{t|t-k})}{Var(y_t)}$$

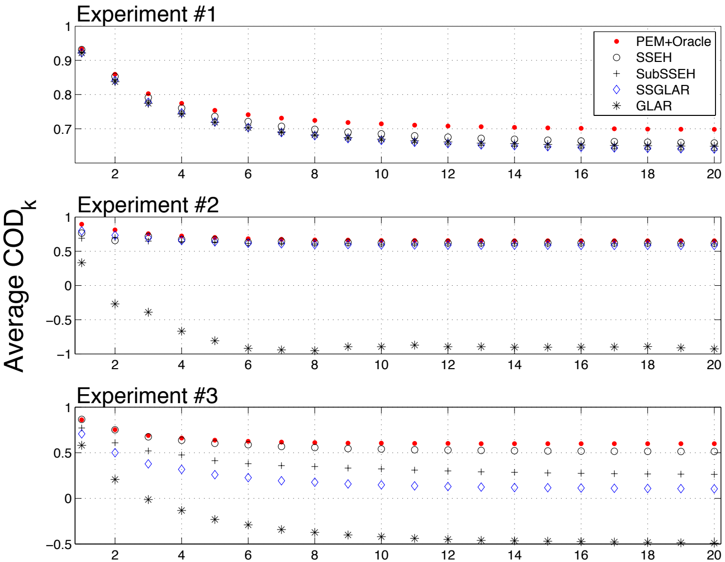
Case Study: ARMAX systems

Boxplots: coefficient of determination



Case Study: ARMAX systems

Average coefficient of determination



Case Study: ARMAX systems
Sparsity

Exp. #	Bayesian	SS-GLAR	GLAR + ARX
#1	98.8%	45.93%	63.41%
#2	98.64%	49.76%	70.09%
#3	95.05%	56.58%	67.16%

Table : Percentage of the $h^{(i)}$ correctly set to zero

Case Study: ARMAX systems

Sparsity

Exp. #	Bayesian	SS-GLAR	GLAR + ARX
#1	98.8%	45.93%	**63.41%**
#2	98.64%	49.76%	**70.09%**
#3	95.05%	56.58%	**67.16%**

Table : Percentage of the $h^{(i)}$ correctly set to zero

Case Study: ARMAX systems
Sparsity

Exp. #	Bayesian	SS-GLAR	GLAR + ARX
#1	**98.8%**	45.93%	**63.41%**
#2	**98.64%**	49.76%	**70.09%**
#3	**95.05%**	56.58%	**67.16%**

Table : Percentage of the $h^{(i)}$ correctly set to zero

Critical Aspect of Classical Parametric Methods

MODEL SELECTION

$$\hat{y}_{t|t-1}(\theta) = \left[h^{(1)}(\theta) * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)}(\theta) * u^{(i)}\right](t) \quad \theta \in \Theta \subseteq \mathbb{R}^n$$

Critical Aspect of Classical Parametric Methods
MODEL SELECTION

$$\hat{y}_{t|t-1}(\theta) = \left[h^{(1)}(\theta) * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)}(\theta) * u^{(i)}\right](t) \quad \theta \in \Theta \subseteq \mathbb{R}^n$$

1. Need to estimate several models (*local minima*)

Critical Aspect of Classical Parametric Methods

MODEL SELECTION

$$\hat{y}_{t|t-1}(\theta) = \left[h^{(1)}(\theta) * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)}(\theta) * u^{(i)}\right](t) \quad \theta \in \Theta \subseteq \mathbb{R}^n$$

1. Need to estimate several models (*local minima*)
2. Order estimation (AIC, BIC, FPE etc.) based on asymptotic arguments

Critical Aspect of Classical Parametric Methods

MODEL SELECTION

$$\hat{y}_{t|t-1}(\theta) = \left[h^{(1)}(\theta) * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)}(\theta) * u^{(i)}\right](t) \quad \theta \in \Theta \subseteq \mathbb{R}^n$$

1. Need to estimate several models (*local minima*)
2. Order estimation (AIC, BIC, FPE etc.) based on asymptotic arguments
3. Statistical properties of **PMSE** (**P**oct **M**odel **S**election **E**stimators) hard to obtain (Leeb and Pötscher (2005)) + unreliable results in some cases (experimental evidence)

Bayesian viepoint

Sparse Bayesian Learning (SBL)

$$\hat{y}_{t|t-1} = \left[h^{(1)} * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)} * u^{(i)}\right](t)$$

$$\hat{y}_{t|t-1} = \left[h^{(1)} * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)} * u^{(i)}\right](t)$$

1. **Identification**: Gaussian Processes $h^{(i)} \sim \mathcal{N}(0, \lambda_i K_i)$

$$\hat{y}_{t|t-1} = \left[h^{(1)} * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)} * u^{(i)}\right](t)$$

1. **Identification**: Gaussian Processes $h^{(i)} \sim \mathcal{N}(0, \lambda_i K_i)$
   - Convexify the problem for given λ_i and K_i (closed form solution)
   - No order estimation: the *Kernels* K_i control the “complexity”

$$\hat{y}_{t|t-1} = \left[h^{(1)} * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)} * u^{(i)}\right](t)$$

1. **Identification**: Gaussian Processes $h^{(i)} \sim \mathcal{N}(0, \lambda_i K_i)$
   - Convexify the problem for given λ_i and K_i (closed form solution)
   - No order estimation: the *Kernels* K_i control the "complexity"
2. **Variable Selection**: the *hyperparameter* λ_i performs selection (SBL)

$$\hat{y}_{t|t-1} = \left[h^{(1)} * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)} * u^{(i)}\right](t)$$

1. **Identification**: Gaussian Processes $h^{(i)} \sim \mathcal{N}(0, \lambda_i K_i)$
   - Convexify the problem for given λ_i and K_i (closed form solution)
   - No order estimation: the *Kernels* K_i control the "complexity"
2. **Variable Selection**: the *hyperparameter* λ_i performs selection **(SBL)**

Key observation of SBL

$$\lambda_i = 0 \Leftrightarrow \hat{h}^{(i)} = 0$$

Linear (Infinite Dimensional) Model

$$y_t = \hat{y}_{t|t-1} + e_t$$

Linear (Infinite Dimensional) Model

$$
\begin{aligned}
y_t &= \hat{y}_{t|t-1} + e_t \\
&= \left[h^{(1)} * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)} u^{(i)}\right](t) + e_t
\end{aligned}
$$

Linear (Infinite Dimensional) Model

$$
\begin{aligned}
y_t &= \hat{y}_{t|t-1} + e_t \\
&= \left[h^{(1)} * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)} u^{(i)}\right](t) + e_t \\
&= \sum_{i=1}^{m} G_t^{(i)} \theta^{(i)} + e_t
\end{aligned}
$$

Linear (Infinite Dimensional) Model

$$
\begin{aligned}
y_t &= \hat{y}_{t|t-1} + e_t \\
&= \left[h^{(1)} * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)} u^{(i)}\right](t) + e_t \\
&= \sum_{i=1}^{m} G_t^{(i)} \theta^{(i)} + e_t \\
&= G_t \theta + e_t
\end{aligned}
$$

Linear (Infinite Dimensional) Model

$$
\begin{aligned}
y_t &= \hat{y}_{t|t-1} + e_t \\
&= \left[h^{(1)} * y\right](t) + \sum_{i=2}^{m} \left[h^{(i)} u^{(i)}\right](t) + e_t \\
&= \sum_{i=1}^{m} G_t^{(i)} \theta^{(i)} + e_t \\
&= G_t \theta + e_t
\end{aligned}
$$

$$\Downarrow$$

Linear Model with Grouped Variables

$$Y = G\theta + E$$

Linear (Infinite Dimensional) Model

$$
\begin{aligned}
y_t &= \hat{y}_{t|t-1} + e_t \\
&= \left[h^{(1)} * y\right](t) + \textstyle\sum_{i=2}^{m} \left[h^{(i)} u^{(i)}\right](t) + e_t \\
&= \textstyle\sum_{i=1}^{m} G_t^{(i)} \theta^{(i)} + e_t \\
&= G_t \theta + e_t
\end{aligned}
$$

$$\Downarrow$$

Linear Model with Grouped Variables

$$Y = G\theta + E$$

$$
Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \quad G = [G^{(1)}, \ldots, G^{(m)}] \quad \theta = \begin{bmatrix} \theta^{(1)} \\ \vdots \\ \theta^{(m)} \end{bmatrix}
$$

Bayesian Model

Sparse Bayesian Learning (SBL)/Penalized ADR (PARD)

Aravkin et al. (2011, 2013)

$$Y = \sum_i G^{(i)}\theta^{(i)} + E$$

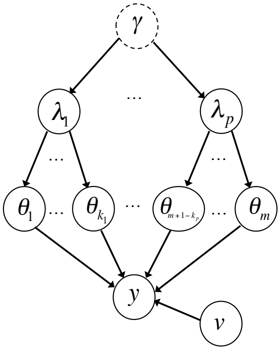


Figure : Bayesian Model.

Bayesian Model

Sparse Bayesian Learning (SBL)/Penalized ADR (PARD)

Aravkin et al. (2011, 2013)

$$Y = \sum_i G^{(i)}\theta^{(i)} + E$$

- $\theta^{(i)}|\lambda_i \sim \mathcal{N}(0, \lambda_i K_i)$

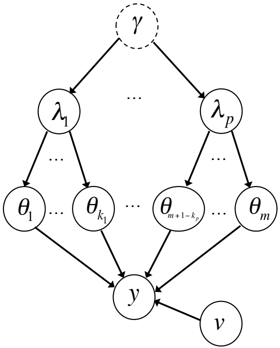


Figure : Bayesian Model.

Bayesian Model

Sparse Bayesian Learning (SBL)/Penalized ADR (PARD)

Aravkin et al. (2011, 2013)

$$Y = \sum_i G^{(i)}\theta^{(i)} + E$$

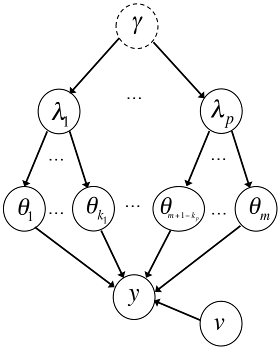


Figure : Bayesian Model.

- $\theta^{(i)}|\lambda_i \sim \mathcal{N}(0, \lambda_i K_i)$
- $Y|\theta \sim \mathcal{N}\left(\sum_i G^{(i)}\theta^{(i)}, \sigma^2 I\right)$

Bayesian Model

Sparse Bayesian Learning (SBL)/Penalized ADR (PARD)

Aravkin et al. (2011, 2013)

$$Y = \sum_i G^{(i)}\theta^{(i)} + E$$

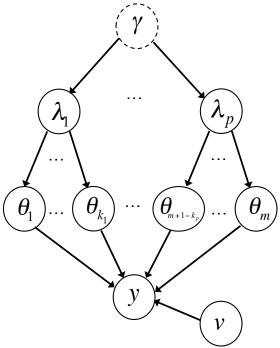


Figure : Bayesian Model.

- $\theta^{(i)}|\lambda_i \sim \mathcal{N}(0, \lambda_i K_i)$
- $Y|\theta \sim \mathcal{N}\left(\sum_i G^{(i)}\theta^{(i)}, \sigma^2 I\right)$
- $(\lambda_1, .., \lambda_p) \sim \gamma e^{-\gamma \sum_i \lambda_i} \quad \lambda_i \geq 0$

Bayesian Model

Sparse Bayesian Learning (SBL)/Penalized ADR (PARD)

Aravkin et al. (2011, 2013)

$$Y = \sum_i G^{(i)}\theta^{(i)} + E$$

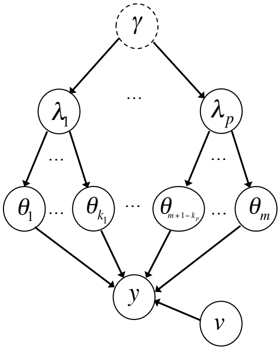


Figure : Bayesian Model.

- $\theta^{(i)}|\lambda_i \sim \mathcal{N}(0, \lambda_i K_i)$
- $Y|\theta \sim \mathcal{N}\left(\sum_i G^{(i)}\theta^{(i)}, \sigma^2 I\right)$
- $(\lambda_1, .., \lambda_p) \sim \gamma e^{-\gamma \sum_i \lambda_i} \quad \lambda_i \geq 0$

SBL/PARD

$$\hat{\lambda} = \arg\max_{\lambda} p(\lambda|Y)$$

Bayesian Model

Sparse Bayesian Learning (SBL)/Penalized ADR (PARD)

Aravkin et al. (2011, 2013)

$$Y = \sum_i G^{(i)}\theta^{(i)} + E$$

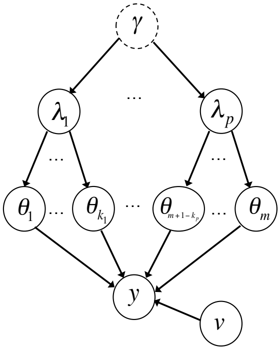


Figure : Bayesian Model.

- $\theta^{(i)}|\lambda_i \sim \mathcal{N}(0, \lambda_i K_i)$
- $Y|\theta \sim \mathcal{N}\left(\sum_i G^{(i)}\theta^{(i)}, \sigma^2 I\right)$
- $(\lambda_1, .., \lambda_p) \sim \gamma e^{-\gamma \sum_i \lambda_i} \quad \lambda_i \geq 0$

SBL/PARD

$$\begin{aligned} \hat\lambda &= \arg\max_\lambda p(\lambda|Y) \\ \hat\theta^{(i)} &:= \mathbb{E}[\theta^{(i)}|Y, \hat\lambda] \\ &= \hat\lambda_i K_i \left(G^{(i)}\right)^\top \Sigma^{-1}(\hat\lambda) Y \\ \Sigma(\hat\lambda) &:= \sum_i \hat\lambda_i G^{(i)} K_i \left(G^{(i)}\right)^\top + \sigma^2 I \end{aligned}$$

SBL/PARD vs. MKL/GLASSO

Objectives

Define:

$$\Sigma(\lambda) := \sum_i \lambda_i G^{(i)} K_i \left(G^{(i)} \right)^\top + \sigma^2 I$$

SBL/PARD vs. MKL/GLASSO

Objectives

Define:

$$\Sigma(\lambda) := \sum_i \lambda_i G^{(i)} K_i \left(G^{(i)}\right)^\top + \sigma^2 I$$

SBL/PARD (Difference of Convex)

$$\hat{\lambda} = \arg\min_\lambda \; \log(\det(\Sigma(\lambda)) + Y^\top \Sigma^{-1}(\lambda) Y + \gamma \sum_i \lambda_i$$

SBL/PARD vs. MKL/GLASSO

Objectives

Define:

$$\Sigma(\lambda) := \sum_i \lambda_i G^{(i)} K_i \left(G^{(i)}\right)^\top + \sigma^2 I$$

SBL/PARD (Difference of Convex)

$$\hat{\lambda} = \arg\min_\lambda \ \log(\det(\Sigma(\lambda)) + Y^\top \Sigma^{-1}(\lambda) Y + \gamma \sum_i \lambda_i$$

MKL/GLASSO (convex)

$$\hat{\lambda} = \arg\min_\lambda \ Y^\top \Sigma^{-1}(\lambda) Y + \gamma_{MKL} \sum_i \lambda_i$$

Model:

$$
\begin{aligned}
y &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \theta^{(1)} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \theta^{(2)} + e \\
\theta^{(1)} &= 0, \theta^{(2)} = 1, \quad e \sim \mathcal{N}(0, \sigma^2 I)
\end{aligned}
$$

Sparsity vs. Shrinking

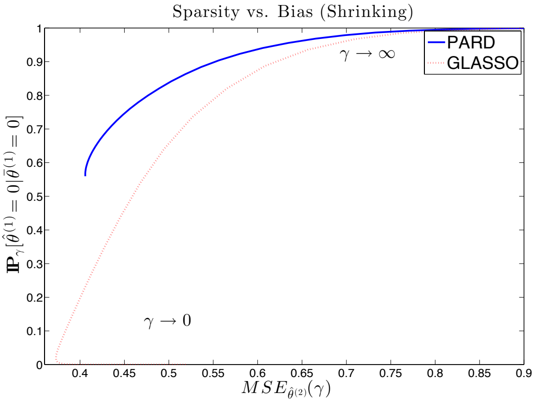
Case study: MSE Analysis

MSE

$$MSE(\theta) := \mathrm{Trace}\,\mathbb{E}\left[\left(\hat{\theta} - \theta_0\right)\left(\hat{\theta} - \theta_0\right)^\top\right]$$

MSE

$$MSE(\theta) := \mathrm{Trace}\,\mathbb{E}\left[\left(\hat{\theta}-\theta_0\right)\left(\hat{\theta}-\theta_0\right)^\top\right]$$

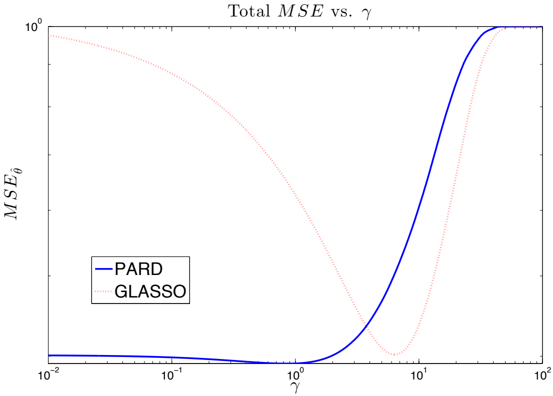


Sparsity vs. Shrinking

Case study: MSE Analysis

MSE

$$MSE(\theta) := \mathrm{Trace}\, \mathbb{E}\left[\left(\hat{\theta} - \theta_0\right)\left(\hat{\theta} - \theta_0\right)^\top\right]$$



Empirical Bayes Estimator (Marginal Likelihood)

$$\hat{\lambda}^N = \arg\min_{\lambda} \frac{1}{2} \log\det(\Sigma(\lambda)) + \frac{1}{2} Y^\top \Sigma^{-1}(\lambda) Y + \gamma \sum_{i=1}^{m} \lambda_i, \tag{1}$$

Empirical Bayes Estimator (Marginal Likelihood)

$$\hat{\lambda}^N = \arg\min_{\lambda} \frac{1}{2}\log\det(\Sigma(\lambda)) + \frac{1}{2}Y^\top \Sigma^{-1}(\lambda) Y + \gamma \sum_{i=1}^{m} \lambda_i, \qquad (1)$$

Mean Squared Error

$$\begin{aligned} MSE_N^{(i)}(\lambda_i) &:= \mathrm{Trace}\left[\mathbb{E}\left[(\hat{\theta}_N^{(i)}(\lambda) - \theta^{(i)})(\hat{\theta}_N^{(i)}(\lambda) - \theta^{(i)})^\top\right]\right] \\ &= \mathrm{Trace}\left[\mathbb{E}\left[(\hat{\beta}_N^{(i)}(\lambda) - \beta_N^{(i)})(\hat{\beta}_N^{(i)}(\lambda) - \beta_N^{(i)})^\top\right]\right] \end{aligned}$$

Asymptotics of PARD

MSE and WEIGHTED MSE

Empirical Bayes Estimator (Marginal Likelihood)

$$\hat{\lambda}^N = \arg\min_{\lambda} \frac{1}{2} \log\det(\Sigma(\lambda)) + \frac{1}{2} Y^\top \Sigma^{-1}(\lambda) Y + \gamma \sum_{i=1}^{m} \lambda_i, \qquad (1)$$

Mean Squared Error

$$\begin{aligned} MSE_N^{(i)}(\lambda_i) &:= \mathrm{Trace}\left[\mathbb{E}\left[(\hat{\theta}_N^{(i)}(\lambda) - \theta^{(i)})(\hat{\theta}_N^{(i)}(\lambda) - \theta^{(i)})^\top\right]\right] \\ &= \mathrm{Trace}\left[\mathbb{E}\left[(\hat{\beta}_N^{(i)}(\lambda) - \beta_N^{(i)})(\hat{\beta}_N^{(i)}(\lambda) - \beta_N^{(i)})^\top\right]\right] \end{aligned}$$

Weighted Mean Squared Error

$$WMSE_N^{(i)}(\lambda_i) := \mathrm{Trace}\left[D_N^4\, \mathbb{E}\left[(\hat{\beta}_N^{(i)}(\lambda) - \beta_N^{(i)})(\hat{\beta}_N^{(i)}(\lambda) - \beta_N^{(i)})^\top\right]\right]$$

$$\check{\lambda}_i^N := \arg\min_{\lambda_i} WMSE_N^{(i)}(\lambda_i) \quad \text{with} \quad \lambda_j = \bar{\lambda}_j^n \quad \text{for} \quad j \neq i$$

$$\hat{\lambda}^N = \arg\min_{\lambda} \frac{1}{2} \log\det(\Sigma(\lambda)) + \frac{1}{2} Y^\top \Sigma^{-1}(\lambda) Y + \gamma \sum_{i=1}^{m} \lambda_i$$

Asymptotics of PARD

MSE and WEIGHTED MSE: MAIN RESULT

$$\check{\lambda}_i^N := \arg\min_{\lambda_i} WMSE_N^{(i)}(\lambda_i) \quad \text{with} \quad \lambda_j = \bar{\lambda}_j^n \quad \text{for} \quad j \neq i$$

$$\hat{\lambda}^N = \arg\min_{\lambda} \frac{1}{2}\log\det(\Sigma(\lambda)) + \frac{1}{2} Y^\top \Sigma^{-1}(\lambda) Y + \gamma \sum_{i=1}^{m} \lambda_i$$

THEOREM

(ML optimization vs. WMSE optimization, Aravkin et al. (2013))

For $\gamma = 0$,

$$\lim_{N\to\infty} \check{\lambda}_i^N = \lim_{N\to\infty} \hat{\lambda}_i^N = \frac{\|\theta^{(i)}\|_{K_i}^2}{T_i}$$

Kernels for Dynamical Systems

Simplest model:

$$\theta^{(i)} \simeq \mathcal{N}(0, \lambda_i K(\rho_i))$$

Kernels for Dynamical Systems

Simplest model:

$$\theta^{(i)} \simeq \mathcal{N}(0, \lambda_i K(\rho_i))$$

Exponentially decaying kernels

$$K_i = K(\rho_i) := \mathrm{diag}\{\rho_i, \rho_i^2, ..., \rho_i^{T_i}\}$$

Kernels for Dynamical Systems

Simplest model:

$$\theta^{(i)} \simeq \mathcal{N}(0, \lambda_i K(\rho_i))$$

Exponentially decaying kernels

$$K_i = K(\rho_i) := \mathrm{diag}\{\rho_i, \rho_i^2, ..., \rho_i^{T_i}\}$$

where ρ_i is an hyperparameter which describes the Kernels' shape

Define:

$$\Sigma(\lambda,\rho) := \sum_i \lambda_i G^{(i)} K(\rho_i) \left(G^{(i)}\right)^\top + \sigma^2 I$$

Define:

$$\Sigma(\lambda, \rho) := \sum_i \lambda_i G^{(i)} K(\rho_i) \left(G^{(i)}\right)^\top + \sigma^2 I$$

Empirical Bayes Estimator of hyperparametrs

$$\begin{aligned}(\hat{\rho}^N, \hat{\lambda}^N) &= \arg\max_{\lambda,\rho} \int p_\gamma(\lambda, \rho, \theta | Y) d\theta \\ &= \arg\min_{\lambda,\rho} \ \log(\det(\Sigma(\lambda, \rho)) + Y^\top \Sigma^{-1}(\lambda, \rho) Y + \gamma \textstyle\sum_i \lambda_i\end{aligned}$$

Define:

$$\Sigma(\lambda,\rho) := \sum_i \lambda_i G^{(i)} K(\rho_i) \left(G^{(i)}\right)^\top + \sigma^2 I$$

Empirical Bayes Estimator of hyperparametrs

$$\begin{aligned}(\hat{\rho}^N, \hat{\lambda}^N) &= \arg\max_{\lambda,\rho} \int p_\gamma(\lambda,\rho,\theta|Y)d\theta \\ &= \arg\min_{\lambda,\rho} \ \log(\det(\Sigma(\lambda,\rho)) + Y^\top \Sigma^{-1}(\lambda,\rho) Y + \gamma \sum_i \lambda_i\end{aligned}$$

QUESTION: where do the Empirical Bayes Estimators $\hat{\rho}_i^N$, $\hat{\lambda}_i^N$converge to?

Exponentially decaying kernels

White noise inputs: asymptotic analysis

Exponentially decaying kernels

$$\lambda K = \lambda K(\rho) := \lambda \, \mathrm{diag}\{\rho, \rho^2, ..., \rho^T\}$$

+ white noise inputs

Exponentially decaying kernels

White noise inputs: asymptotic analysis

Exponentially decaying kernels

$$\lambda K = \lambda K(\rho) := \lambda\,\mathrm{diag}\{\rho, \rho^2, \ldots, \rho^T\}$$

+ white noise inputs

Theorem, Carli et al. (2012)

If $\gamma = 0$ the Empirical Bayes Estimators $\hat{\rho}^N$ and $\hat{\lambda}^N$ converge, as $N \to \infty$, to $\hat{\rho}$ and $\hat{\lambda}$ which satisfy

$$\hat{\lambda} = \frac{1}{T}\sum_{k=1}^{T} \frac{(\theta_k)^2}{\hat{\rho}^k} = \frac{1}{T}\|\theta\|^2_{K(\hat{\rho})}$$

$$\sum_{k=1}^{T} \frac{\theta_k^2}{\hat{\rho}^k}\left(\frac{T+1}{2} - k\right) = 0$$

Exponentially decaying kernels

$$\lambda K = \lambda K(\rho) := \lambda\,\mathrm{diag}\{\rho, \rho^2, ..., \rho^T\}$$

+ white noise inputs

REMARK

If the truncation index $T \to \infty$ and $\gamma = 0$ the Empirical Bayes Estimators $\hat{\rho}^N$ and $\hat{\lambda}^N$ converge to

$$\hat{\lambda} = \lim_{T\to\infty} \frac{1}{T} \|\theta\|^2_{K(\hat{\rho})}$$

$$\hat{\rho} = \max_j |p_j|^2$$

p_j = poles of the system: $\max_j |p_j|^2$ = modulus of the dominant mode.

Exponentially decaying kernels:

White noise inputs: numerical results

Exponentially decaying kernels

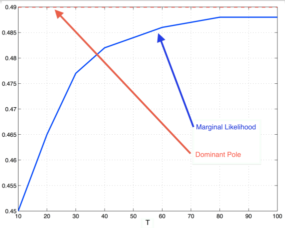
$$\lambda K = \lambda K(\rho) := \lambda\,\mathrm{diag}\{\rho, \rho^2, ..., \rho^T\}$$

$+$ white noise inputs

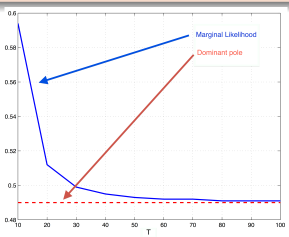
Exponentially decaying kernels

$$\lambda K = \lambda K(\rho) := \lambda\, \mathrm{diag}\{\rho, \rho^2, \ldots, \rho^T\}$$

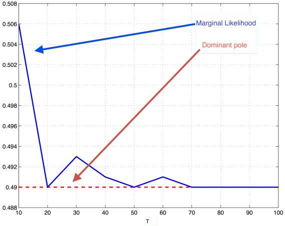
$+$ white noise inputs



$$h_k = (.7)^k + 3(.3)^k$$



$$h_k = (.7)^k - 3(.3)^k$$



$$h_k = (.7)^k \cos(\pi k/3)$$

Exponentially decaying kernels:

White noise inputs: MSE Analysis

WMSE (σ^2 = noise variance)

$$
\begin{aligned}
WMSE_N(\rho,\lambda) &= \text{Trace}\left[K(\bar{\rho})\,\mathbb{E}\left[(\hat{\theta}_N(\rho,\lambda)-\theta)(\hat{\theta}_N(\rho,\lambda)-\theta)^\top\right]\right] \\
&\propto \sum_{k=1}^{T} \bar{\rho}^k \frac{\lambda^2\rho^{2k}+\frac{\sigma^2}{N}\theta_k^2}{\left(\lambda\rho^k+\frac{\sigma^2}{N}\right)^2}
\end{aligned}
\tag{2}
$$

Exponentially decaying kernels:

White noise inputs: Weighted MSE optimization

$$\begin{aligned}(\check{\rho}^N, \check{\lambda}^N) &= \arg\min_{\rho,\lambda} WMSE_N(\rho, \lambda) \\ (\hat{\rho}^N, \hat{\lambda}^N) &= \arg\max_{\lambda,\rho} \int p_\gamma(\lambda, \rho, \theta | Y) d\theta\end{aligned}$$

Exponentially decaying kernels:

White noise inputs: Weighted MSE optimization

$$
\begin{aligned}
(\check{\rho}^N, \check{\lambda}^N) &= \arg\min_{\rho,\lambda} WMSE_N(\rho, \lambda) \\
(\hat{\rho}^N, \hat{\lambda}^N) &= \arg\max_{\lambda,\rho} \int p_\gamma(\lambda, \rho, \theta | Y) d\theta
\end{aligned}
$$

Theorem, Carli et al. (2012)

If $\bar{\rho} = \hat{\rho}^N$ then

$$
\lim_{N\to\infty} \hat{\rho}^N = \lim_{N\to\infty} \check{\rho}^N
$$

$$
\lim_{N\to\infty} \hat{\lambda}^N = \lim_{N\to\infty} \check{\lambda}^N
$$

Current/Future work

1. Kernel design, Conic Combination of Kernels, Optimization Algorithms

Current/Future work

1. Kernel design, Conic Combination of Kernels, Optimization Algorithms
2. Multi-Input-Multi-Output systems: vector Kernels-Nuclear Norm penalties etc.

Current/Future work

1. Kernel design, Conic Combination of Kernels, Optimization Algorithms
2. Multi-Input-Multi-Output systems: vector Kernels-Nuclear Norm penalties etc.
3. Properties of Local Minima, criteria for “smart” initialization

Current/Future work

1. Kernel design, Conic Combination of Kernels, Optimization Algorithms
2. Multi-Input-Multi-Output systems: vector Kernels-Nuclear Norm penalties etc.
3. Properties of Local Minima, criteria for "smart" initialization
4. Recursive algorithms for on-line adaptation

Current/Future work

1. Kernel design, Conic Combination of Kernels, Optimization Algorithms
2. Multi-Input-Multi-Output systems: vector Kernels-Nuclear Norm penalties etc.
3. Properties of Local Minima, criteria for "smart" initialization
4. Recursive algorithms for on-line adaptation
5. MSE properties for more general input/Kernel design *e.g. Stable Spline Kernels*

Thanks!

Thanks for your attention

http://automatica.dei.unipd.it/people/chiuso.html

chiuso@dei.unipd.it

References

A. Aravkin, J. Burke, A. Chiuso, and G. Pillonetto. Convex vs nonconvex approaches for sparse estimation: Glasso, multiple kernel learning and hyperparameter glasso. In *IEEE Conf. on Dec. and Control*, 2011.

A. Aravkin, J. Burke, A. Chiuso, and G. Pillonetto. Convex vs non-convex estimators for regression and sparse estimation: the mean squared error properties of ARD and GLasso. Technical report, University of Padova, 2013. submitted to Journal of Machine Learning Research.

F. Carli, T. Chen, A. Chiuso, L. Ljung, and G. Pillonetto. On the estimation of hyperparameters for bayesian system identification with exponentially decaying kernels. In *CDC*, 2012.

A. Chiuso and G. Pillonetto. A bayesian approach to sparse dynamic network identification. *Automatica*, 48(8):1553 – 1565, 2012. ISSN 0005-1098. doi: 10.1016/j.automatica.2012.05.054. URL `http://www.sciencedirect.com/science/article/pii/S0005109812002270`.

H. Leeb and B. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21:2159, 2005.