Regularized switched system identification: a statistical learning perspective

Université de Lorraine, CNRS, CRAN, LORIA

Louis MASSUCCI, Fabien LAUER, Marion GILSON

Massucci et al., Structural risk minimization for hybrid system identification. *CDC2020*
Massucci et al., How statistical learning can help to estimate the number of modes in switched system identification? *SYSID21*
Massucci et al., Regularized switched system identification: a statistical learning perspective. *ADHS21*

Aim of this talk

Use **AI** (Statistical learning theory) and **system identification techniques** to produce new solutions for **estimating hybrid systems**

Outline

Hybrid system identification

Estimating the number of modes

Regularization

Conclusions

Outline

Hybrid system identification

SISO arbitrarily switched ARX system:

$$\underbrace{y_i}_{\text{output}} = f_{q_i}(\boldsymbol{x}_i) + \underbrace{\nu_i}_{\text{noise}} \tag{1}$$

- $\boldsymbol{x}_i = [\underbrace{y_{i-1}, \ \dots, y_{i-n_a}}_{\text{past outputs}}, \underbrace{u_i}_{\text{input}}, \underbrace{\dots, u_{i-n_b}}_{\text{past inputs}}]^T$
- f_j: the j-th submodel
- $q_i \in \{1 \dots C\}$: active mode at time i

Problem:

Given a data set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and a set of possible submodels $\mathcal{F}$, estimate the number of submodels C, the submodels f_j in $\mathcal{F}$, and the switching sequence $(q_i)_{1 \le i \le n}$.

Literature for switched system identification

Methods for a fixed number of modes C

- K-LinReg [Lauer, 2013]
- Algebraic Methods [Vidal et al., 2003, Ozay et al., 2015]
- Others...

Methods that estimate C from a threshold on the prediction error:

- Sparse Optimization [Bako, 2011]
- Sum-of-norm regularization [Ohlsson and Ljung, 2013]
- Bounded-error approach [Bemporad et al., 2005]

Challenge: Estimate the number of modes using techniques from statistical learning

Literature for switched system identification

Methods for a fixed number of modes C

- K-LinReg [Lauer, 2013]
- Algebraic Methods [Vidal et al., 2003, Ozay et al., 2015]
- Others...

Methods that estimate C from a threshold on the prediction error:

- Sparse Optimization [Bako, 2011]
- Sum-of-norm regularization [Ohlsson and Ljung, 2013]
- Bounded-error approach [Bemporad et al., 2005]

Challenge: Estimate the number of modes using techniques from statistical learning

Literature for switched system identification

Methods for a fixed number of modes C

- K-LinReg [Lauer, 2013]
- Algebraic Methods [Vidal et al., 2003, Ozay et al., 2015]
- Others...

Methods that estimate C from a threshold on the prediction error:

- Sparse Optimization [Bako, 2011]
- Sum-of-norm regularization [Ohlsson and Ljung, 2013]
- Bounded-error approach [Bemporad et al., 2005]

Challenge: Estimate the number of modes using techniques from statistical learning

Outline

Estimating the number of modes

Structural Risk Minimization:

- Model selection method from statistical learning
- Derive statistical guarantees on the prediction error
- Select the model with the best guarantees

→ Choose the number of modes C that minimizes an upper bound on the prediction error

Learning theory

Setting:

- A pair of random variables (X, Y) of unknown distribution
- A training set $((x_i, y_i))_{1\leq i\leq N}$: a sample realization of N **independent** copies (X_i, Y_i) of (X, Y)
- $\mathcal{F}$ a set of possible models

Typical form of distribution free risk bounds:

With probability at least $1-\delta$, for all $f \in \mathcal{F}$:

$$L(f) \leq \hat{L}_n(f) + \epsilon(n, \mathcal{F}, \delta) \tag{2}$$

- $L(f) = \mathbb{E}_{X,Y}\ell(f, X, Y)$: the risk or prediction error
- $\hat{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(f, X_i, Y_i)$: the empirical risk
- $\epsilon(n, \mathcal{F}, \delta)$: a confidence interval to be defined

Typical loss for regression: $\ell(f, X, Y) = (Y - f(X))^2$

for classification: $\ell(f, X, Y) = \mathbb{1}_{(X)\neq Y}$

Learning theory

Setting:

- A pair of random variables (X, Y) of unknown distribution
- A training set $((x_i, y_i))_{1\leq i\leq N}$: a sample realization of N **independent** copies (X_i, Y_i) of (X, Y)
- $\mathcal{F}$ a set of possible models

Typical form of distribution-free risk bounds:

With probability at least $1 - \delta$, for all $f \in \mathcal{F}$:

$$L(f) \leq \hat{L}_n(f) + \epsilon(\delta, n, \mathcal{F}) \tag{2}$$

- $L(f) = \mathbb{E}_{X,Y}\ell(f, X, Y)$: the risk or prediction error
- $\hat{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(f, X_i, Y_i)$: the empirical risk
- $\epsilon(n, \mathcal{F}, \delta)$: a confidence interval to be defined

Typical loss for regression: $\ell(f, X, Y) = (Y - f(X))^2$

for classification: $\ell(f, X, Y) = \mathbb{1}_{(X)\neq Y}$

Learning theory

Setting:

- A pair of random variables (X, Y) of unknown distribution
- A training set $((x_i, y_i))_{1 \leq i \leq N}$: a sample realization of N **independent** copies (X_i, Y_i) of (X, Y)
- $\mathcal{F}$ a set of possible models

Typical form of distribution-free risk bounds:

With probability at least $1 - \delta$, for all $f \in \mathcal{F}$:

$$L(f) \leq \hat{L}_n(f) + \epsilon(\delta, n, \mathcal{F}) \tag{2}$$

- $L(f) = \mathbb{E}_{X,Y} \ell(f, X, Y)$: the risk or prediction error
- $\hat{L}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f, X_i, Y_i)$: the empirical risk
- $\epsilon(n, \mathcal{F}, \delta)$: a confidence interval to be defined

Typical loss for regression: $\ell(f, X, Y) = (Y - f(X))^2$

for classification: $\ell(f, X, Y) = \mathbb{1}_{f(X) \neq Y}$

Confidence interval

- The confidence interval $\epsilon(n, \mathcal{F}, \delta)$ depends on a measure of complexity of the model
- Common complexity measures: VC-dimension, Rademacher Complexity,...
- Computed using statistical learning theory for i.i.d samples, depending on $\mathcal{L}$:

$$\mathcal{L} = \{\ell_f : \ell_f(z) = \ell(f, x, y),\ f \in \mathcal{F}\} \tag{3}$$

Rademacher complexity:

Empirical Rademacher complexity $$\hat{\mathcal{R}}_{\boldsymbol{Z}_n}(\mathcal{L}) = \mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{\ell \in \mathcal{L}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(Z_i) \,\middle|\, \boldsymbol{Z}_n\right], \tag{4}$$

with $\boldsymbol{Z}_n = (Z_i)_{1 \leq i \leq n} = ((X_i, Y_i))_{1 \leq i \leq n}$, and $\boldsymbol{\sigma}_n = (\sigma_i)_{1 \leq i \leq n}$ is a sequence of Rademacher variables, i.e., random variables uniformly distributed in $\{-1, +1\}$.

Rademacher complexity bound

Rademacher complexity:

$$\text{Empirical Rademacher complexity} \quad \hat{\mathcal{R}}_{\boldsymbol{Z}_n}(\mathcal{L}) = \mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{\ell \in \mathcal{L}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(Z_i) \,\middle|\, \boldsymbol{Z}_n\right], \tag{5}$$

with $\boldsymbol{Z}_n = (Z_i)_{1 \leq i \leq n} = ((X_i, Y_i))_{1 \leq i \leq n}$, and $\boldsymbol{\sigma}_n = (\sigma_i)_{1 \leq i \leq n}$ is a sequence of Rademacher variables, i.e., random variables uniformly distributed in $\{-1, +1\}$.

Theorem (Theorem 1 in [Mohri et al., 2018])

Let $\mathcal{L}$ be a class of functions from $\mathcal{Z}$ into $[0, B]$ and $\boldsymbol{Z}_n = (Z_i)_{1 \leq i \leq n}$ be a sequence of independent copies of the random variable $Z \in \mathcal{Z}$. Then, for any fixed $\delta \in (0, 1)$, with probability at least $1 - \delta$, uniformly over all $\ell \in \mathcal{L}$,

$$\underbrace{\mathbb{E}_Z \ell(Z)}_{\text{Risk}} \leq \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(Z_i)}_{\text{Empirical risk}} + \underbrace{2\hat{\mathcal{R}}_{\boldsymbol{Z}_n}(\mathcal{L}) + 3B\sqrt{\frac{\log \frac{2}{\delta}}{2n}}}_{\text{Confidence interval}}. \tag{6}$$

Example in linear regression

Consider the model class:

$$\mathcal{F} = \{f : f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x},\ \boldsymbol{w} \in \mathbb{R}^d,\ \|\boldsymbol{w}\| \leq R_w\}, \tag{7}$$

And the loss function

$$\mathcal{L} = \{\ell \in [0, 4M^2]^{\mathcal{Z}} : \ell(f, x, y) = |y - f(x)|^p,\ f \in \mathcal{F}\}, \tag{8}$$

with $Y \in [-M, M]$.
Using a contraction argument ([Ledoux and Talagrand, 1991]),

$$\hat{\mathcal{R}}_{\boldsymbol{Z}_n}(\mathcal{L}) \leq p(2M)^{p-1} \hat{\mathcal{R}}_{\boldsymbol{X}_n}(\mathcal{F}) \tag{9}$$

Where, using standard computation of Rademacher complexity we have

$$\hat{\mathcal{R}}_{\boldsymbol{X}_n}(\mathcal{F}) \leq \frac{R_w \sqrt{\sum_{i=1}^n \|\boldsymbol{X}_i\|^2}}{n}. \tag{10}$$

Example in switching linear regression

Consider the model class:

$$\mathcal{F} = \{f : f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x},\ \boldsymbol{w} \in \mathbb{R}^d,\ \|\boldsymbol{w}\| \leq R_w\}, \tag{11}$$

And the loss function

$$\mathcal{L} = \{\ell \in [0, 4M^2]^{\mathcal{Z}} : \ell(\boldsymbol{f}, x, y) = \min_{j\in\{1,\dots,C\}} |y - f_j(x)|^p,\ f_j \in \mathcal{F}\}, \tag{12}$$

with $Y \in [-M, M]$.

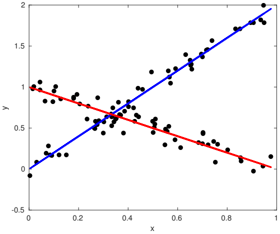


Figure: Switching regression

Example in switching linear regression 2

Consider the model class:

$$\mathcal{F} = \{f : f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x},\ \boldsymbol{w} \in \mathbb{R}^d,\ \|\boldsymbol{w}\| \leq R_w\}, \tag{13}$$

And the loss function

$$\mathcal{L} = \{\ell \in [0, 4M^2]^{\mathcal{Z}} : \ell(\boldsymbol{f}, x, y) = \min_{j \in \{1,\dots,C\}} |y - f_j(x)|^p,\ f_j \in \mathcal{F}\}, \tag{14}$$

with $Y \in [-M, M]$.
Using Rademacher calculus [Lauer, 2020],

$$\hat{\mathcal{R}}_{\boldsymbol{Z}_n}(\mathcal{L}) \leq p(2M)^{p-1} C\, \hat{\mathcal{R}}_{\boldsymbol{X}_n}(\mathcal{F}) \tag{15}$$

Examples

Final prediction error bounds (in case $p = 2$):

- For linear regression

$$\mathbb{E}_{X,Y}(Y - f(X))^2 \leq \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2 + \frac{8MR_w\sqrt{\sum_{i=1}^{n}\|\boldsymbol{X}_i\|^2}}{n} + 12M^2\sqrt{\frac{\log\frac{2}{\delta}}{2n}}. \quad (16)$$

- For switching linear regression

$$\mathbb{E}_{X,Y}\min_{j\in\{1,\dots,C\}}(Y - f_j(X))^2 \leq \frac{1}{n}\sum_{i=1}^{n}\min_{j\in\{1,\dots,C\}}(Y_i - f_j(X_i))^2 + \frac{8MCR_w\sqrt{\sum_{i=1}^{n}\|\boldsymbol{X}_i\|^2}}{n} + 12M^2\sqrt{\frac{\log\frac{2}{\delta}}{2n}}, \quad (17)$$

with $\boldsymbol{f} = (f_1, \dots, f_C)$.

Bound only valid in static case.

How can we adapt it for dynamical systems?

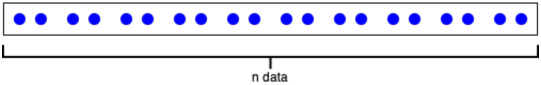
Error bounds for dynamical system

Problem:

- For dynamical systems, i.i.d. assumption doesn't hold

Proposed solution:

- Assume data are β-mixing
- Dependence between two data points decreases with the time interval between them



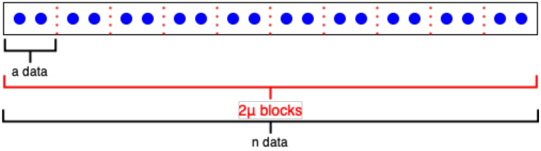
Error bounds for dynamical system

Problem:

- For dynamical systems, i.i.d. assumption doesn't hold

Proposed solution:

- Assume data are β-mixing
- Dependence between two data points decreases with the time interval between them
- Independent blocks method [Yu, 1994]



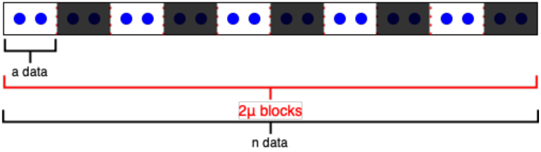
Error bounds for dynamical system

Problem:

- For dynamical systems, i.i.d. assumption doesn't hold

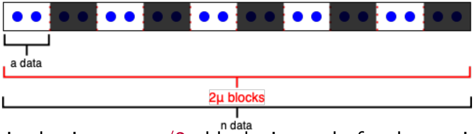
Proposed solution:

- Assume data are β-mixing
- Dependence between two data points decreases with the time interval between them
- Independent blocks method [Yu, 1994]
- Dependency between odd blocks weakens with the size of blocks



Error bounds for dynamical system

Independent Blocks Method:



- Bound is derived using $\mu = n/2a$ blocks instead of n data points [Mohri and Rostamizadeh, 2009]
- The confidence interval depends on a mixing coefficient β(a)

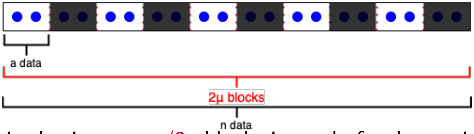
With probability at least $1-\delta$, for all $f \in \mathcal{F}$:

$$L(f) \leq \hat{L}_n(f) + \epsilon(n, \mathcal{F}, \delta) \quad \text{(i.i.d case)} \tag{18}$$

$$L(f) \leq \hat{L}_n(f) + \epsilon(\mu, \beta(a), \mathcal{F}, \delta) \quad \text{(non i.i.d case)} \tag{19}$$

Error bounds for dynamical system

Independent Blocks Method:



- Bound is derived using $\mu = n/2a$ blocks instead of n data points [Mohri and Rostamizadeh, 2009]
- The confidence interval depends on a mixing coefficient β(a)

With probability at least $1-\delta$, for all $f \in \mathcal{F}$:

$$L(f) \leq \hat{L}_n(f) + \epsilon(n, \mathcal{F}, \delta) \quad \text{(i.i.d case)} \tag{18}$$

$$L(f) \leq \hat{L}_n(f) + \epsilon(\mu, \beta(a), \mathcal{F}, \delta) \quad \text{(non i.i.d case)} \tag{19}$$

→ Using the previous results on the Rademacher complexity for switching regression, we obtain:

$$L(f) \leq \hat{L}_n(f) + \epsilon(C, \mu, \beta(a), \mathcal{F}, \delta) \tag{20}$$

Proposed method to estimate C

Require: The data set $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ and a maximum number of modes $\overline{C}$

1: **for** $C = 1$ to $\overline{C}$ **do**
2: Run a generic algorithm to estimate a model $\boldsymbol{f}$ with C modes
3: Compute the error bound $J(C)$
4: **end for**
5: Select the "best" number of modes

$$\hat{C} = \underset{C \in \{1 \dots \overline{C}\}}{\arg\min} \; J(C)$$

6: **return** the selected model with $\hat{C}$ modes

With $J(C) = \hat{L}_n(f) + \epsilon(C, \mu, \beta(a), \mathcal{F}, \delta)$

Numerical Experiment

Case study:

- switched ARX system with $C = 3$ modes, orders $n_a = n_b = 2$
- $n = 10^5$ points
- Gaussian noise with $SNR = 10dB$

Numerical Experiment

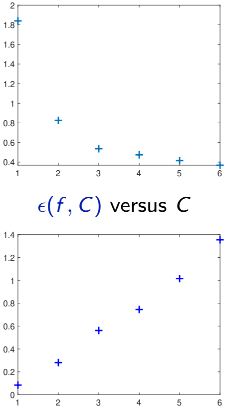
Case study:

- switched ARX system with $C = 3$ modes, orders $n_a = n_b = 2$
- $n = 10^5$ points
- Gaussian noise with $SNR = 10dB$

Results with L1-loss and a block size of a=2

$\hat{L}_n(f)$ versus C

$\epsilon(f, C)$ versus C



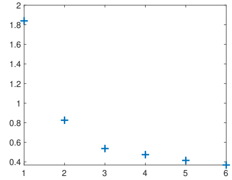
Numerical Experiment

Results with L1-loss and a block size of a=2

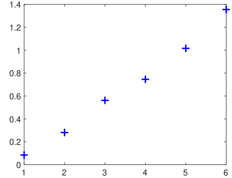
Case study:

- switched ARX system with $C = 3$ modes, orders $n_a = n_b = 2$
- $n = 10^5$ points
- Gaussian noise with $SNR = 10dB$

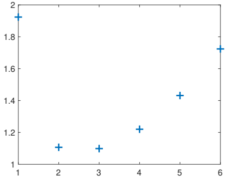
$\hat{L}_n(f)$ versus C



$\epsilon(f, C)$ versus C



$J(C)$



Minimum achieved at $C = 3$

Numerical Experiment

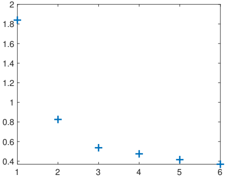
Case study:

- switched ARX system with $C = 3$ modes, orders $n_a = n_b = 2$
- $n = 10^5$ points
- Gaussian noise with $SNR = 10dB$

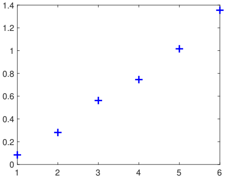
Evaluation of the method over 100 trials with colored noise shows promising results [Massucci et al., 2020]
Comparison with other methods in [Massucci et al., 2021]

Results with L1-loss and a block size of a=2

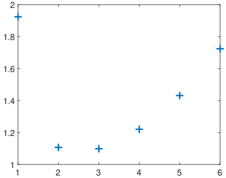
$\hat{L}_n(f)$ versus C



$\epsilon(f, C)$ versus C



$J(C)$



Minimum achieved at $C = 3$

Regularization

What could be the benefits of regularization ?

Outline

Regularization

A standard technique to control model complexity while learning from data by minimizing a trade-off:

$$\min_{\boldsymbol{w}\in\mathbb{R}^D} \underbrace{\mathcal{E}(\boldsymbol{w},\boldsymbol{X},\boldsymbol{y})}_{\text{error term}} \quad + \quad \underbrace{\lambda\Gamma(\boldsymbol{w})}_{\text{Regularization term}} \quad ,$$

For switching systems:

$$\mathcal{E}(\boldsymbol{w},\mathrm{X},\boldsymbol{y}) = \sum_{i=1}^{n} \ell(\boldsymbol{w},\boldsymbol{x}_i,y_i)$$

with $\ell(\boldsymbol{w},\boldsymbol{x}_i,y_i) = \min_{j\in\{1,\dots,C\}} |y_i - \boldsymbol{w}_j^T\boldsymbol{x}_i|^p$ for $p \in \{1,2\}$

$$\Gamma(\boldsymbol{w}) = \|\Omega(\boldsymbol{w})\|_q$$

where $\Omega(\boldsymbol{w}) = [\|\boldsymbol{w}_1\|_2,\dots,\|\boldsymbol{w}_C\|_2]^T$, $q \in \{1,2,\infty\}$

$$\lambda > 0$$

Regularization

A more fine-grained measure of complexity $\|\Omega(\boldsymbol{w})\|_q$, where

$$\forall q \in (0, \infty], \quad \|\Omega(\boldsymbol{w})\|_q \leq C \max_{j \in \{1,\dots,C\}} \|\boldsymbol{w}_j\|_2 = \|\Omega(\boldsymbol{w})\|_\infty \quad (21)$$

Consequence of $\|\Omega(\boldsymbol{w})\|_q$:

- Consider the number of submodels
- And the complexity of each submodels

Corresponding model class:

$$\mathcal{F}(R_w) = \left\{ \boldsymbol{f} \in \mathcal{F}_0(R_w)^C : \|\Omega(\boldsymbol{w})\|_q \leq R_w \right\}, \quad (22)$$

Bound for regularized switching models

Use of [Lauer, 2020] leads to the following complexity term:

$$\hat{\mathcal{R}}_{Z_\mu}(\mathcal{L}) \leq p(2M)^{p-1}\alpha(C,q)\frac{R_w\sqrt{\sum_{i=1}^{\mu}\|\boldsymbol{X}_{2a(i-1)+1}\|^2}}{\mu}, \qquad (23)$$

where the dependence on C is now characterized by

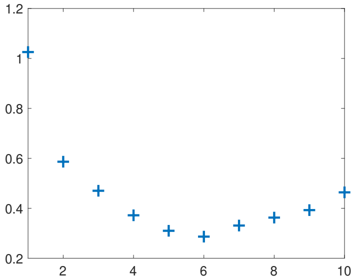
$$\alpha(C,q) = \begin{cases} C, & \text{if } q=\infty \quad \text{(Previous case, independent submodels)} \\ 2\sqrt{C}, & \text{if } q=2 \quad \text{(Classic case)} \\ 1+\log C, & \text{if } q=1. \quad \text{(Sparse case)} \end{cases} \qquad (24)$$

Numerical Experiment

Case study:

- $q = 2$, switched ARX system with $C = 6$ modes, orders $n_a = n_b = 2$
- $n = 3.10^5$ points
- Gaussian noise with $SNR = 30dB$

Regularized J(C) versus C



Massucci et al., "Regularized switched system identification: a statistical learning perspective." *ADHS21* for more details on regularization

Outline

Hybrid system identification

Estimating the number of modes

Regularization

Conclusions

Conclusions

To summarize

- New error bounds for switched systems in the non-I.I.D. case
- New model selection method to estimate the number of modes
- Refined analysis with regularized model

Open issues

- Estimating the mixing coefficient $\beta(a)$
- Tighten the bounds to make the method more efficient with less data

Take-home message

Statistical learning theory can be used to produce **non-asymptotic error bounds** for **hybrid system** identification and a method to estimate the number of modes

References I

Bako, L. (2011).
Identification of switched linear systems via sparse optimization.
Automatica, 47(4):668–677.

Bemporad, A., Garulli, A., Paoletti, S., and Vicino, A. (2005).
A bounded-error approach to piecewise affine system identification.
IEEE Transactions on Automatic Control, 50(10):1567–1580.

Lauer, F. (2013).
Estimating the probability of success of a simple algorithm for switched linear regression.
Nonlinear Analysis: Hybrid Systems, 8:31–47.

Lauer, F. (2020).
Risk bounds for learning multiple components with permutation-invariant losses.
In *International Conference on Artificial Intelligence and Statistics*, pages 1178–1187. PMLR.

Ledoux, M. and Talagrand, M. (1991).
Probability in Banach Spaces: Isoperimetry and Processes.
Springer-Verlag, Berlin.

References II

Massucci, L., Lauer, F., and Gilson, M. (2020).
Structural risk minimization for switched system identification.
In *Proc. of the 59th IEEE Conference on Decision and Control, CDC 2020*.

Massucci, L., Lauer, F., and Gilson, M. (2021).
How statistical learning can help to estimate the number of modes in switched system identification?
In *Proc. of the 19th IFAC Symposium on system identification, Virtual Event*.

Mohri, M. and Rostamizadeh, A. (2009).
Rademacher complexity bounds for non-i.i.d. processes.
In *Advances in Neural Information Processing Systems 21*, pages 1097–1104.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018).
Foundations of Machine Learning.
The MIT Press, second edition.

Ohlsson, H. and Ljung, L. (2013).
Identification of switched linear regression models using sum-of-norms regularization.
Automatica, 49(4):1045–1050.

References III

Ozay, N., Lagoa, C., and Sznaier, M. (2015).
Set membership identification of switched linear systems with known number of subsystems.
Automatica, 51:180–191.

Vidal, R., Soatto, S., Ma, Y., and Sastry, S. (2003).
An algebraic geometric approach to the identification of a class of linear hybrid systems.
In *Proc. of the 42nd IEEE Conference on Decision and Control (CDC), Maui, HI, USA*, pages 167–172.

Yu, B. (1994).
Rates of convergence for empirical processes of stationary mixing sequences.
The Annals of Probability, 22(1):94–116.

Comparison with algebraic methods

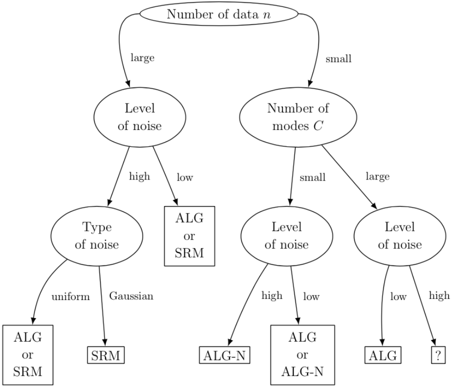


Figure: Guide to select a suitable method to estimate C.

- ALG is Algebraic method for noiseless data [Vidal et al., 2003]
- ALG-N is Algebraic method for noisy data
- SRM is Structural risk minimization method