Pseudo-measurement simulations and bootstrap for the experimental cross-section covariances estimation with quality quantification

S. Varet[1]

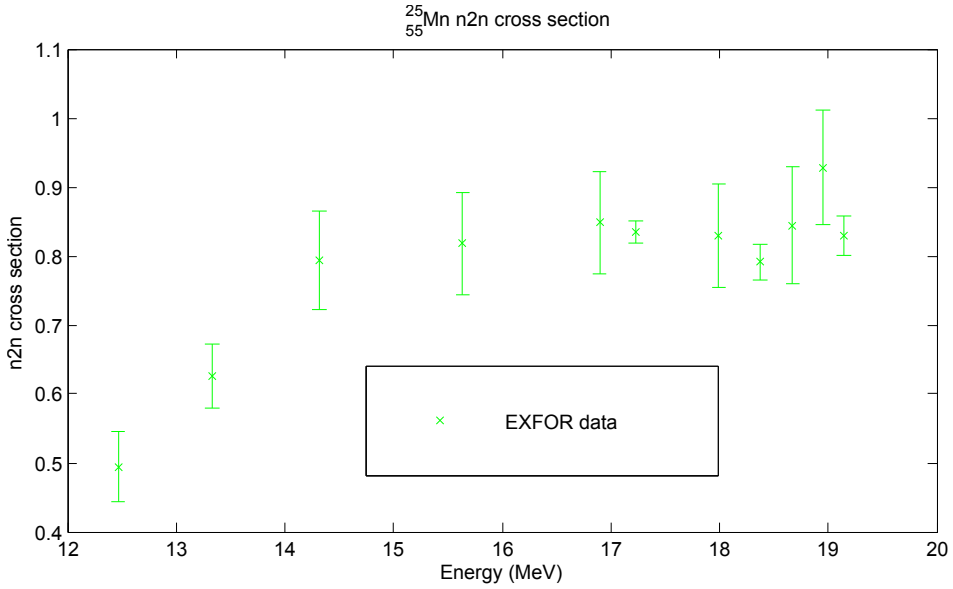
P. Dossantos-Uzarralde[1] N. Vayatis[2] E. Bauge[1]

[1]CEA-DAM-DIF

[2]ENS Cachan

WONDER-2012 September 25-28

Motivations: experimental cross-sections



Available informations:

- the measurements
- their uncertainty

$\longrightarrow$ covariance matrix estimation and its inverse

Example

Evaluated cross sections uncertainty: generalized χ^2 ([VDUVB11])

Motivations: experimental cross-sections covariances

HOW?

- Empirical estimator: only one measurement per energy

$$\mathrm{cov}(F_i, F_j) \approx \frac{1}{n} \sum_{k=1}^{n} (F_i^{(k)} - \bar{F}_i)(F_j^{(k)} - \bar{F}_j)$$

- Conventionnal approach: propagation error formula [IBC04],[Kes08]

Motivations: experimental cross-sections covariances

HOW?

- Empirical estimator: only one measurement per energy
- Conventionnal approach: propagation error formula [IBC04],[Kes08]

$$\mathrm{cov}(F_i, F_j) \approx \sum_{\text{param. exp. } (q_k,\, q_l)} \frac{\partial F_i}{\partial q_k} . \frac{\partial F_j}{\partial q_l} . \mathrm{cov}(q_k, q_l)$$

ex. : fission fragments yield, recoil protons yield,

$\longrightarrow$ the needed informations are rarely available
$\longrightarrow$ linearity assumption

Motivations: experimental cross-sections covariances

HOW?

- Empirical estimator: only one measurement per energy
- Conventionnal approach: propagation error formula [IBC04],[Kes08]

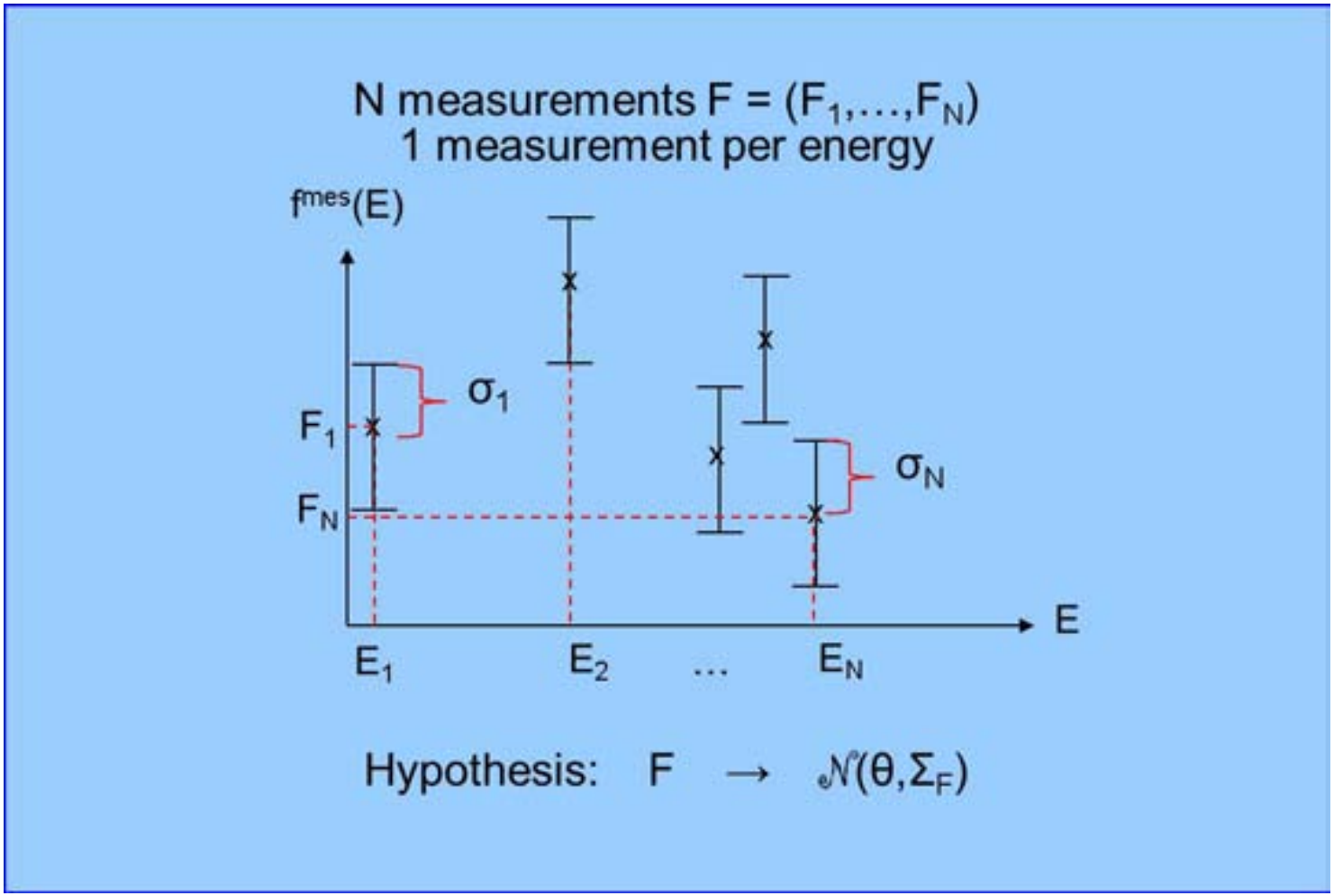
QUALITY OF THE COVARIANCES ESTIMATION?

Outline

Outline

Notations

Schematic representation of the experimental cross sections



Outline

Pseudo-measurements method

Given the available information (the measurements and their uncertainty)

The idea:

Repeat the available information in order to artificially increase the number of measurements

The method:

1. Construction of a regression model h (SVM, polynomial,...)
2. Generation of r pseudo-measurements $(S^{(1)}, ..., S^{(r)})$: gaussian noise centered on h: $\mathcal{N}((h(E_1), ..., h(E_N))^t, diag(\sigma_1, ..., \sigma_N))$
3. $\widehat{\Sigma_F}$: empirical estimator of Σ_F
4. Diagonal terms are imposed

Pseudo-measurements method

Given the available information (the measurements and their uncertainty)

The idea:

Repeat the available information in order to artificially increase the number of measurements

The method:

1. Construction of a regression model h (SVM, polynomial,...)
2. Generation of r pseudo-measurements $(S^{(1)}, ..., S^{(r)})$: gaussian noise centered on h: $\mathcal{N}((h(E_1), ..., h(E_N))^t, diag(\sigma_1, ..., \sigma_N))$
3. $\widehat{\Sigma_F}$: empirical estimator of Σ_F
4. Diagonal terms are imposed

The SVM regression principle [SS04], [VGS97]

Linear case:
Assume $F_i = h(E_i) = w.E_i + b$ with $E_i \in \mathbb{R}^s$ and $w \in \mathbb{R}^s$ (here $s = 1$).
The svm regression model is a solution of the constraint optimisation problem:

$$\text{flat function: minimize} \quad \frac{1}{2}\|w\|^2$$

$$\text{errors lower than } \varepsilon\text{: subject to} \quad \begin{cases} F_i - (w.E_i + b) \leq \varepsilon \\ (w.E_i + b) - F_i \leq \varepsilon \end{cases}$$

Non linear case:
Find a map of the initial space of energy, (into a higher dimensionnal space) such as the problem becomes linear in the new space (mapping via Kernel).

Pseudo-measurements method

Given the available information (the measurements and their uncertainty)

The idea:

Repeat the available information in order to artificially increase the number of measurements

The method:

1. Construction of a regression model h (SVM, polynomial,...)
2. Generation of r pseudo-measurements $(S^{(1)}, ..., S^{(r)})$: gaussian noise centered on h: $\mathcal{N}((h(E_1), ..., h(E_N))^t, diag(\sigma_1, ..., \sigma_N))$
3. $\widehat{\Sigma_F}$: empirical estimator of Σ_F
4. Diagonal terms are imposed

Pseudo-measurements method

$\widehat{\Sigma_F}$ term at the ith line and j^{th} column, for i and $j \in \{1, ..., N\}$:

$$C_{ij} = \frac{\sigma_i \sigma_j}{n+r} \sum_{k=1}^{n} \frac{(F_i^{(k)} - h(E_i))(F_j^{(k)} - h(E_j))}{\sqrt{\widehat{V}_i}\sqrt{\widehat{V}_j}} + \frac{\sigma_i \sigma_j}{n+r} \sum_{k=1}^{r} \frac{(S_i^{(k)} - h(E_i))(S_j^{(k)} - h(E_j))}{\sqrt{\widehat{V}_i}\sqrt{\widehat{V}_j}}$$

and

$$C_{ii} = \sigma_i^2$$

where $n = 1$.

Number r of pseudo-measurements?

Constraint on r

r too small:
$\widehat{\Sigma_F}$ non invertible

r too large:
$\widehat{\Sigma_F}$ diagonal matrix

$$r = 0$$

$\widehat{\Sigma_F}$ invertible
$\downarrow$
no pseudo-measurements

$\widehat{\Sigma_F}$ non invertible
$\downarrow$
r such as $\widehat{\Sigma_F}$ invertible

Outline

Toy model

Goal:

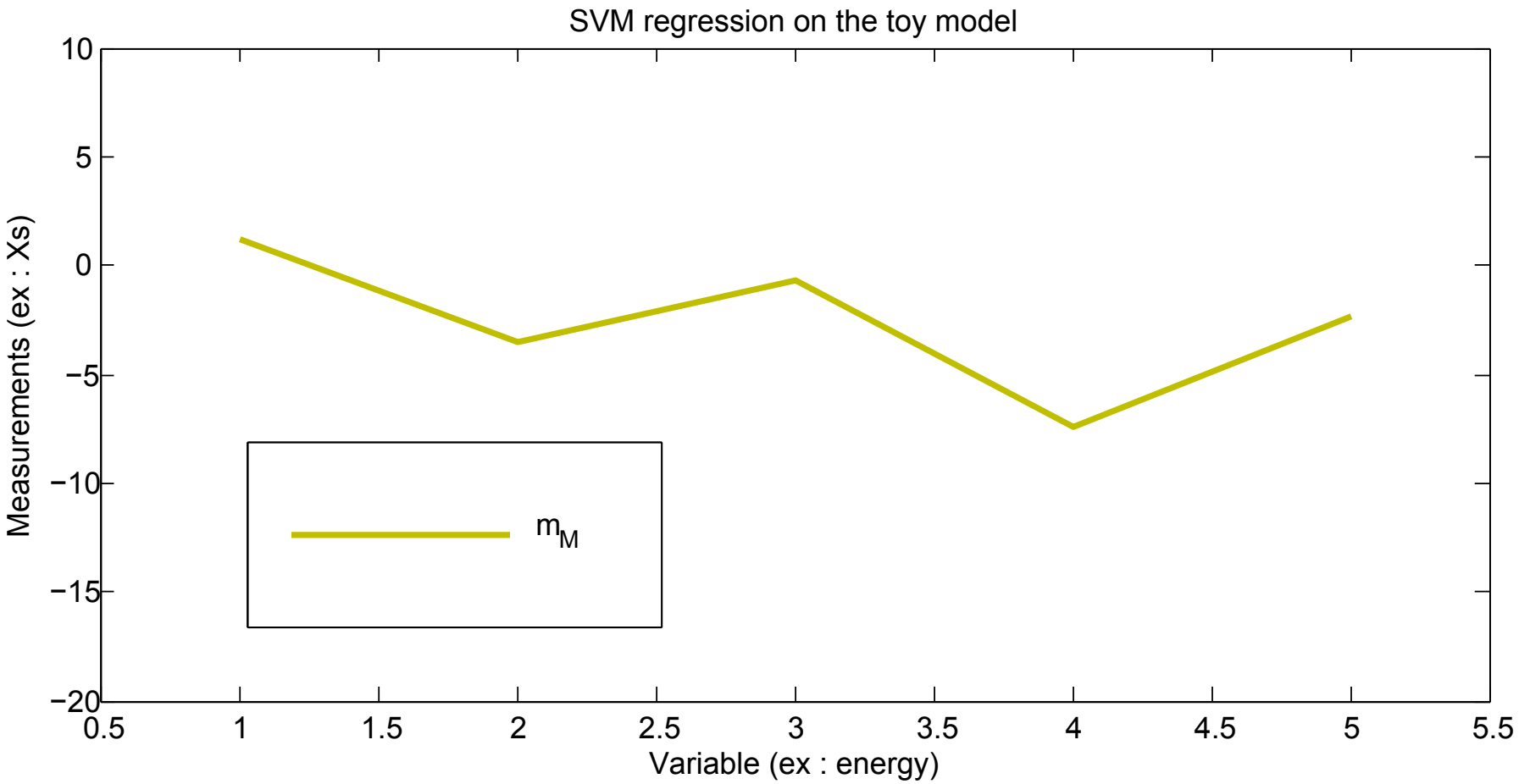
Compare $\widehat{\Sigma_F}$ with Σ_F: real case: Σ_F is unknown
$\Rightarrow$ we build a toy model with Σ_F fixed $\Rightarrow$ numerical validation

Toy Model:

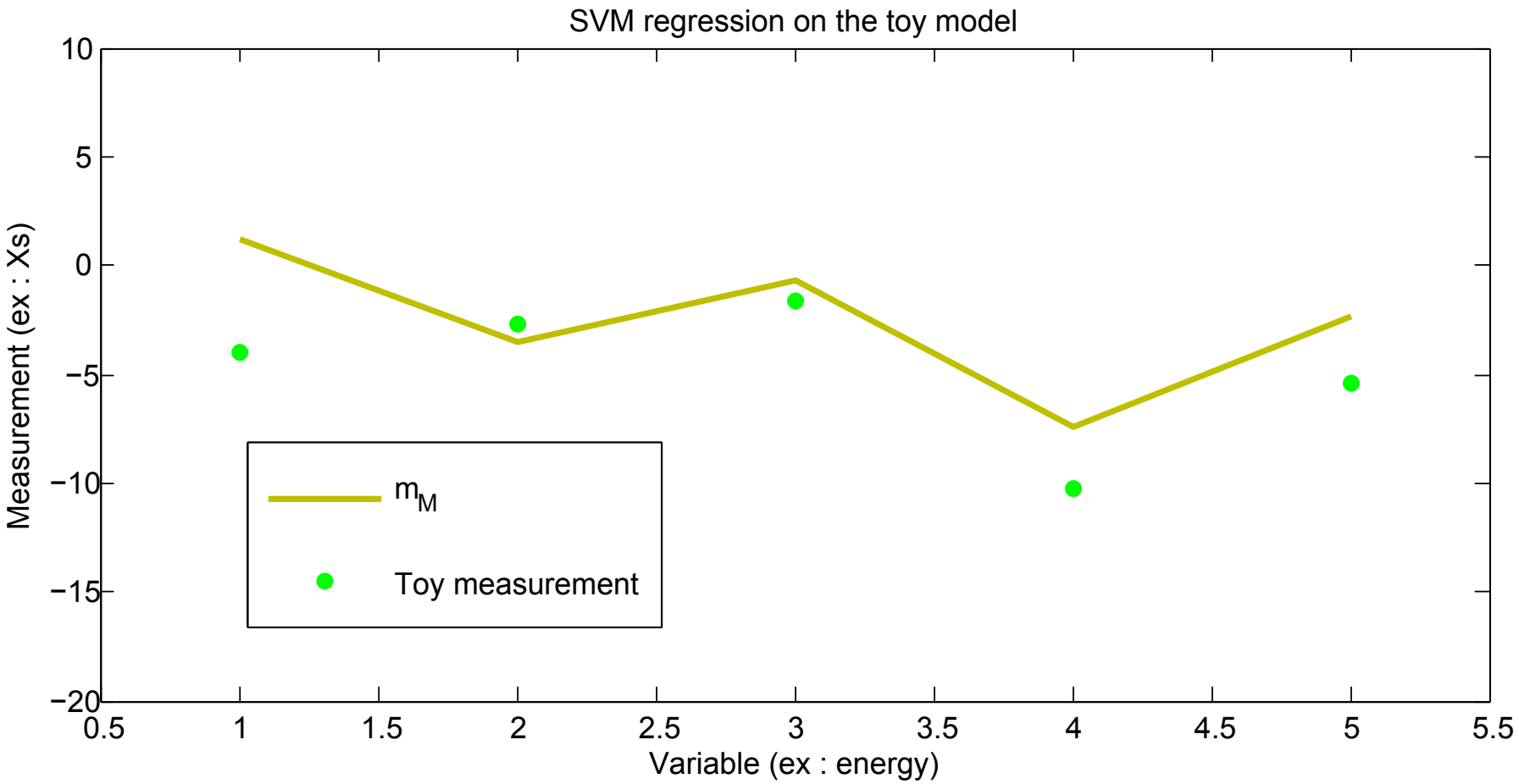
- $M := \begin{pmatrix} M^1 \\ \vdots \\ M^N \end{pmatrix} (\equiv F) \hookrightarrow \mathcal{N}_N(m_M, \Sigma_M)$
- number of M realisations $n = 1$, $N = 5$

$$\theta = \begin{pmatrix} 1.25 \\ -3.49 \\ -0.67 \\ -7.41 \\ -2.29 \end{pmatrix} \quad \Sigma_F = \begin{pmatrix} 9.49 & -1.20 & 2.39 & -0.03 & 0.07 \\ -1.20 & 3.64 & -3.21 & 0.03 & -0.09 \\ 2.39 & -3.21 & 5.27 & 0.48 & 0.10 \\ -0.03 & 0.03 & 0.48 & 6.79 & 0.01 \\ 0.07 & -0.09 & 0.10 & 0.01 & 5.72 \end{pmatrix}$$

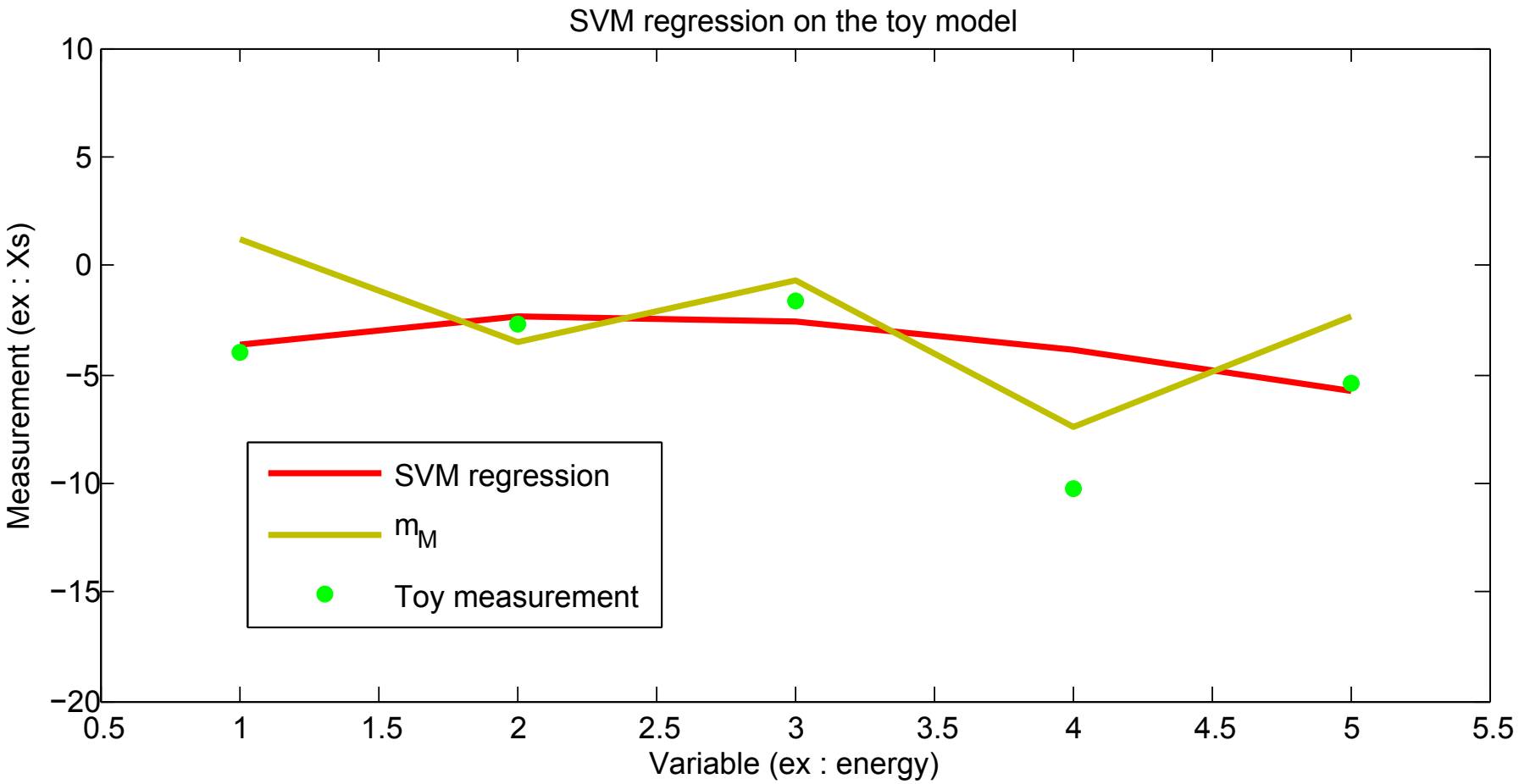
Pseudo-measurements with the 'Toy Model'



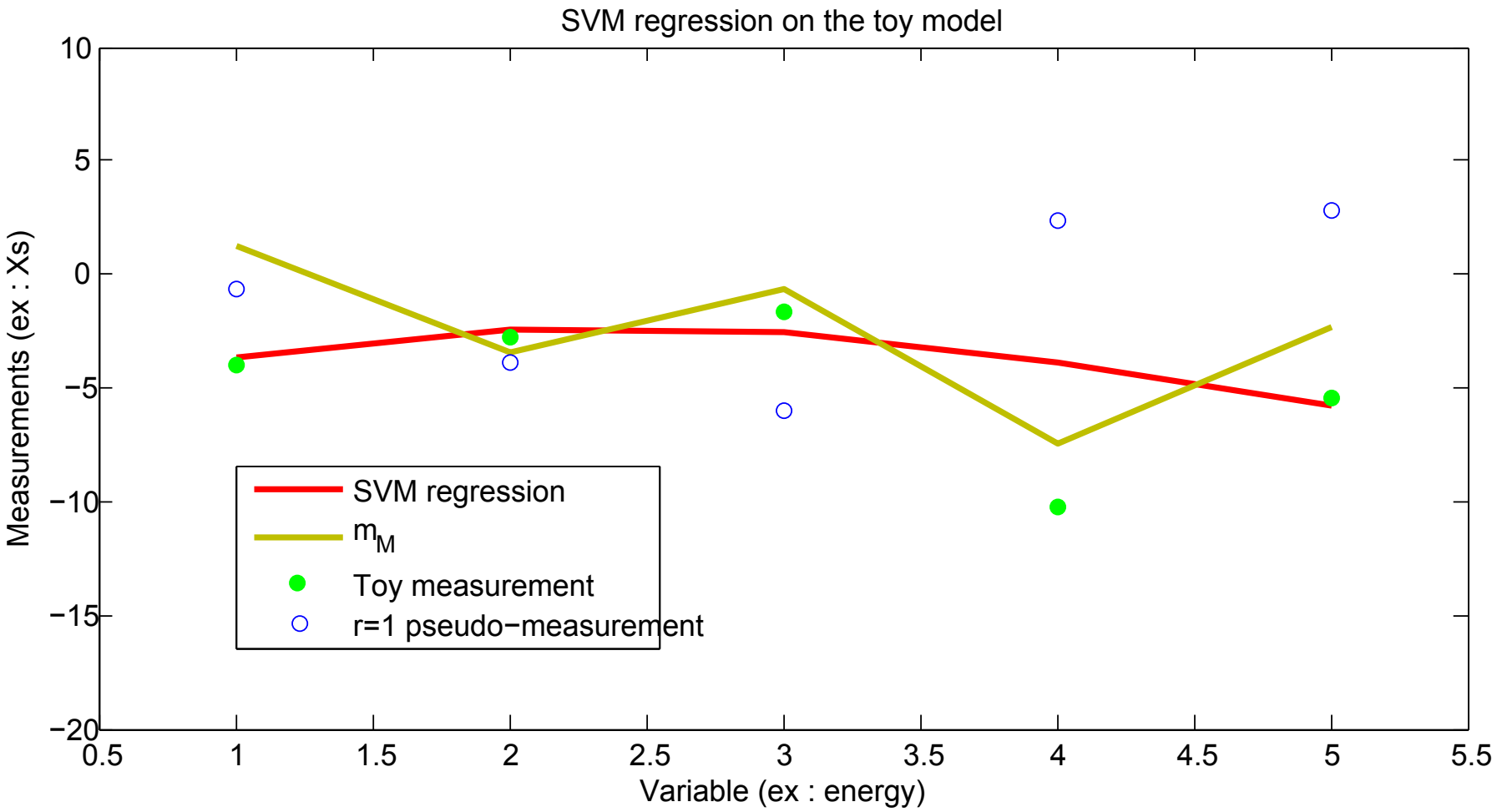
Pseudo-measurements with the 'Toy Model'



Pseudo-measurements with the 'Toy Model'

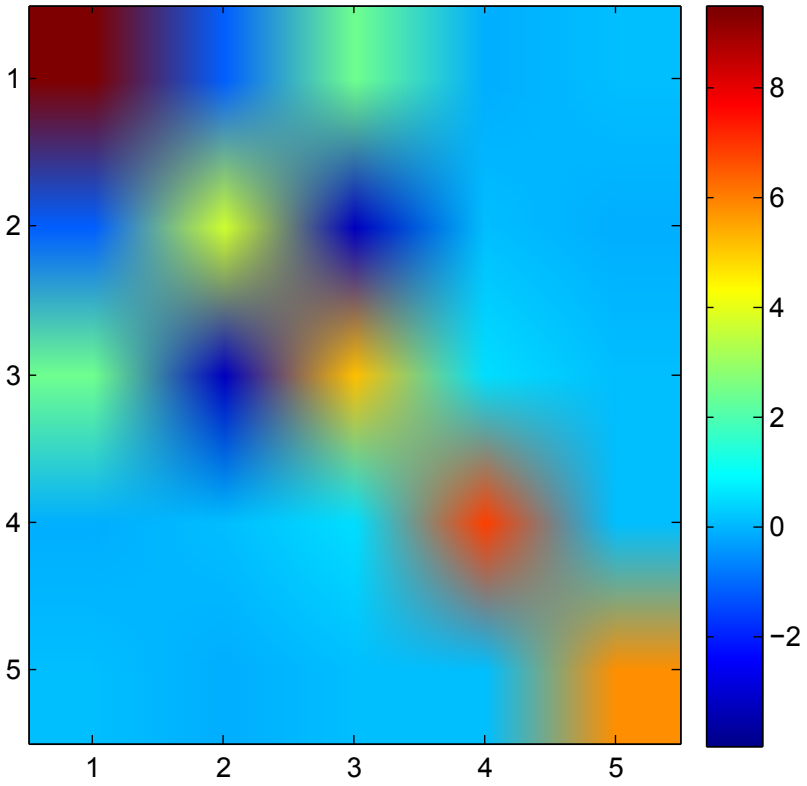


Pseudo-measurements with the 'Toy Model'

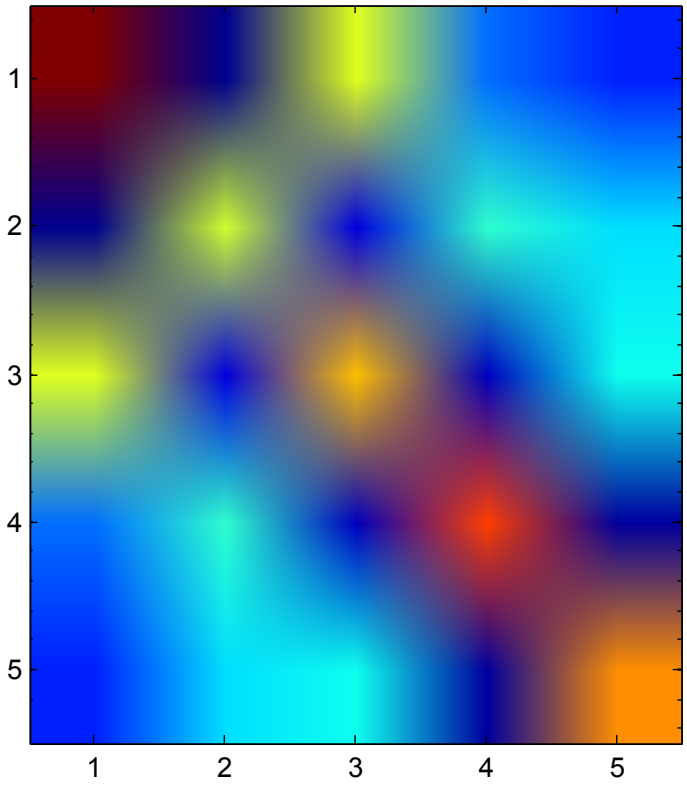


Pseudo-measurements with the 'Toy Model'

N=5, $r = 1$ sample of pseudo-measurements



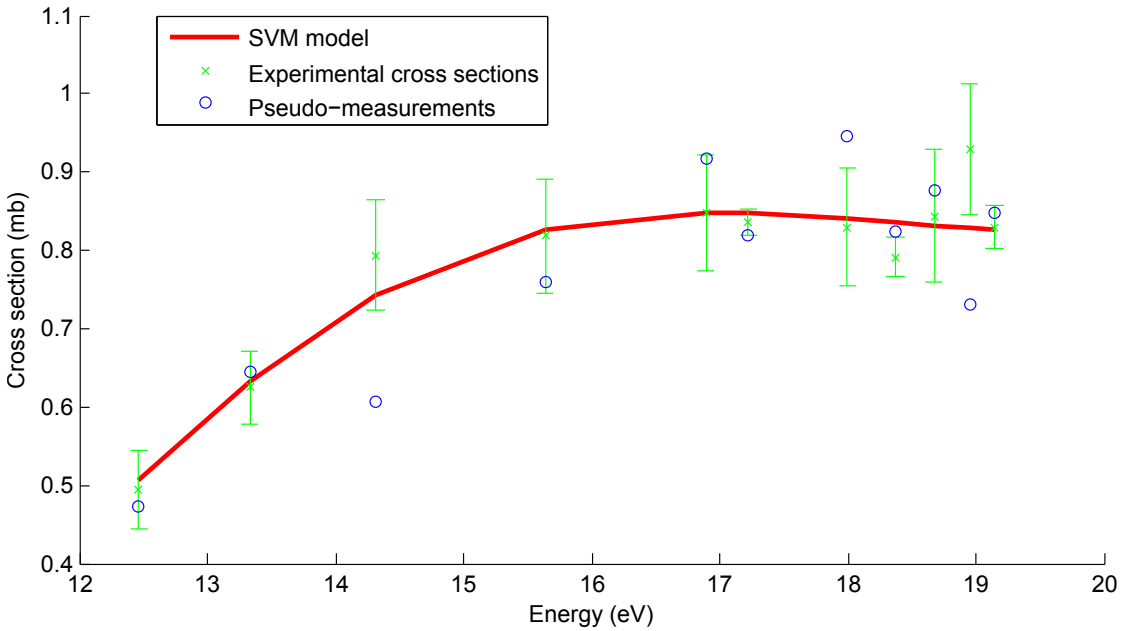
Real matrix

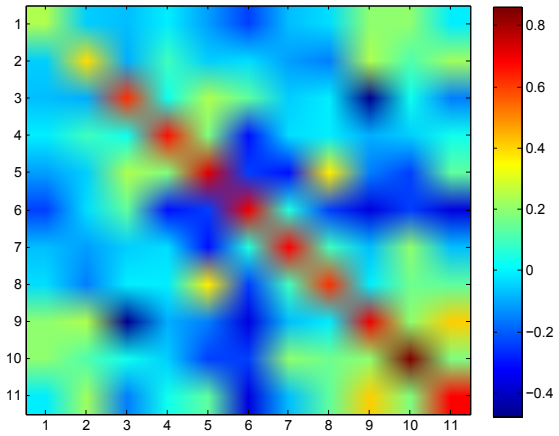


Estimation

$^{55}_{25}Mn$: (n,2n) cross section

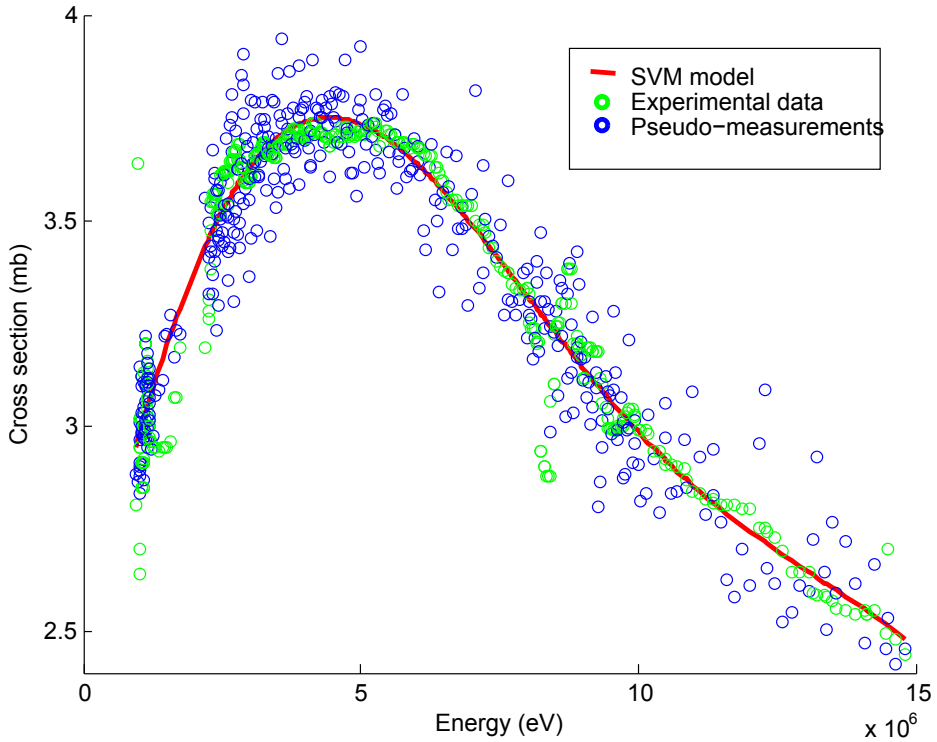
Values extracted from [MTN11]
N=11, r=1, $\widehat{\Sigma_F}$ is a 11x11 matrix

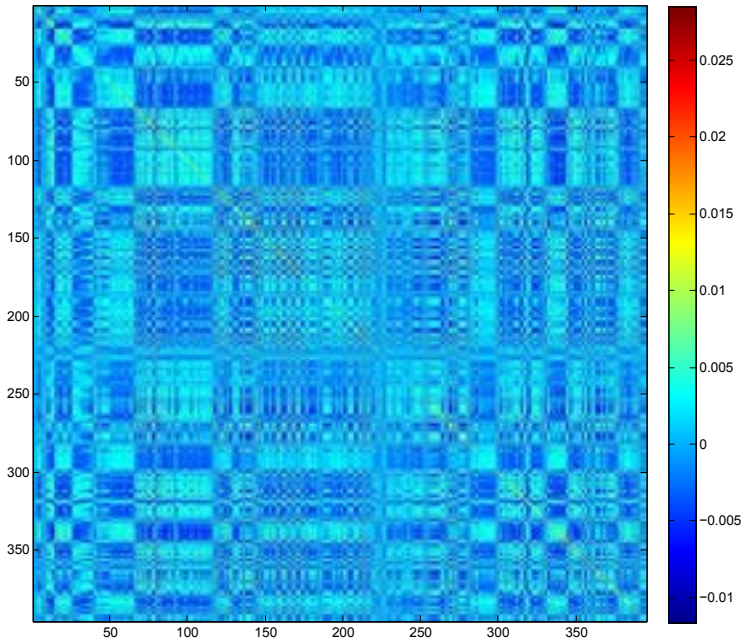




$^{55}_{25}Mn$: total cross section

Values extracted from EXFOR [EXF]
N=399, r=1, $\widehat{\Sigma_F}$ is a 399x399 matrix





How far is the estimation from the real matrix?

Outline

Estimator quality

Criterion choice: quality of $\widehat{\Sigma_F}$ and/or quality of $\widehat{\Sigma_F}^{-1}$?
$\Rightarrow$ distance criterion

1. Distance between $\widehat{\Sigma_F}$ and Σ_F: Frobenius norm

$$\boldsymbol{E}(\|\widehat{\Sigma_F} - \Sigma_F\|_{fro}) = \boldsymbol{E}\left(\left(\sum_{i=1}^{N}\sum_{j=1}^{N}|\widehat{C_{ij}} - C_{ij}|^2\right)^{\frac{1}{2}}\right)$$

2. Distance between $\widehat{\Sigma_F}^{-1}$ and Σ_F^{-1}: Kullback-Leibler distance [LRZ08]

$$\mathrm{KL}(\widehat{\Sigma_F}, \Sigma_F) = \mathrm{trace}((\widehat{\Sigma_F})^{-1}.\Sigma_F) - \log(\det(\widehat{\Sigma_F}^{-1}.\Sigma_F)) - N$$

Criterion estimation $\rightarrow$ parametric bootstrap

Bootstrap principle

$$F : \theta_F,\ \Sigma_F: \quad \text{unknown}$$

$$\downarrow$$

$$(F^{(1)}, ..., F^{(n)}) : h,\ \widehat{\Sigma}_F: \quad \text{sample measurements}$$

$$(F'^{(1)}, ..., F'^{(n)})_1, \cdots (F'^{(1)}, ..., F'^{(n)})_B \quad \text{resample measurements}$$

$$\downarrow \qquad\qquad \downarrow$$

$$\widehat{\widehat{\Sigma}_F}^{1} \qquad\qquad \widehat{\widehat{\Sigma}_F}^{B}$$

Resampling allows to make statistics on $\widehat{\Sigma_F}$:

$$\widehat{\boldsymbol{E}}(\widehat{\Sigma_F}),\ \widehat{\mathrm{Var}}(\widehat{\Sigma_F}),\ \widehat{\mathrm{Bias}}(\widehat{\Sigma_F}),...$$

Resample strategy:

- classical: draw with replacement in the initial sample $(F^{(1)}, ..., F^{(n)})$
- parametric: simulations with a fixed law (here $\mathcal{N}(h, \widehat{\Sigma_F})$)

Estimation of $\mathbb{E}(\|\widehat{\Sigma_F} - \Sigma_F\|_{fro})$ and $\mathrm{KL}(\widehat{\Sigma_F}, \Sigma_F)$

<u>1st case</u>: with $r = 0$, $\widehat{\Sigma_F}$ invertible

Usual parametric bootstrap:

$(F'^{(1)}, ..., F'^{(n)})_1, ..., (F'^{(1)}, ..., F'^{(n)})_B$ where $F'^{(i)} \hookrightarrow \mathcal{N}(H, \widehat{\Sigma_F})$ with $H = (h(E_1), ..., h(E_N))^t$.

$\widehat{\widehat{\Sigma_F}}^b$ term at the ith line and j^{th} column, for i and $j \in \{1, ..., N\}$ and for $b = 1, ..., B$:

$$C_{ij}^b = \frac{\sigma_i \sigma_j}{n} \sum_{k=1}^{n} \frac{(F_i'^{(k)b} - h(E_i))(F_j'^{(k)b} - h(E_j))}{\sqrt{\widehat{V_i'}}\sqrt{\widehat{V_j'}}} \qquad \text{and } C_{ii}^b = \sigma_i^2$$

$$\boldsymbol{E}(\|\widehat{\Sigma_F} - \Sigma_F\|) \approx \frac{1}{B} \sum_{b=1}^{B} \|\widehat{\widehat{\Sigma_F}}^b - \widehat{\Sigma_F}\|$$

$$\mathrm{KL}(\widehat{\Sigma_F}, \Sigma_F) \approx \frac{1}{B} \sum_{b=1}^{B} \mathrm{KL}(\widehat{\widehat{\Sigma_F}}^b, \widehat{\Sigma_F})$$

Estimation of $\mathbb{E}(\|\widehat{\Sigma_F} - \Sigma_F\|_{fro})$ and $\mathrm{KL}(\widehat{\Sigma_F}, \Sigma_F)$

<u>2nd case</u>: with $r = 0$, $\widehat{\Sigma_F}$ non invertible

2 simultaneous parametric bootstraps:

$(F'^{(1)}, ..., F'^{(n)})_1, ..., (F'^{(1)}, ..., F'^{(n)})_B$ where $F'^{(i)} \hookrightarrow \mathcal{N}(H, \widehat{\Sigma_F})$ et

$(S'^{(1)}, ..., S'^{(r)})_1, ..., (S'^{(1)}, ..., S'^{(r)})_B$ where $S'^{(i)} \hookrightarrow \mathcal{N}(H, \mathrm{diag}(\widehat{\Sigma_F}))$

$\widehat{\widehat{\Sigma_F}}^b$ term at the ith line and j^{th} column, for i and $j \in \{1, ..., N\}$ and for $b = 1, ..., B$:

$$C_{ij}^b = \frac{\sigma_i \sigma_j}{n+r} \sum_{k=1}^{n} \frac{(F_i'^{(k)b} - h(E_i))(F_j'^{(k)b} - h(E_j))}{\sqrt{\widehat{V_i'}}\sqrt{\widehat{V_j'}}} + \frac{\sigma_i \sigma_j}{n+r} \sum_{k=1}^{r} \frac{(S_i'^{(k)b} - h(E_i))(S_j'^{(k)b} - h(E_j))}{\sqrt{\widehat{V_i'}}\sqrt{\widehat{V_j'}}}$$

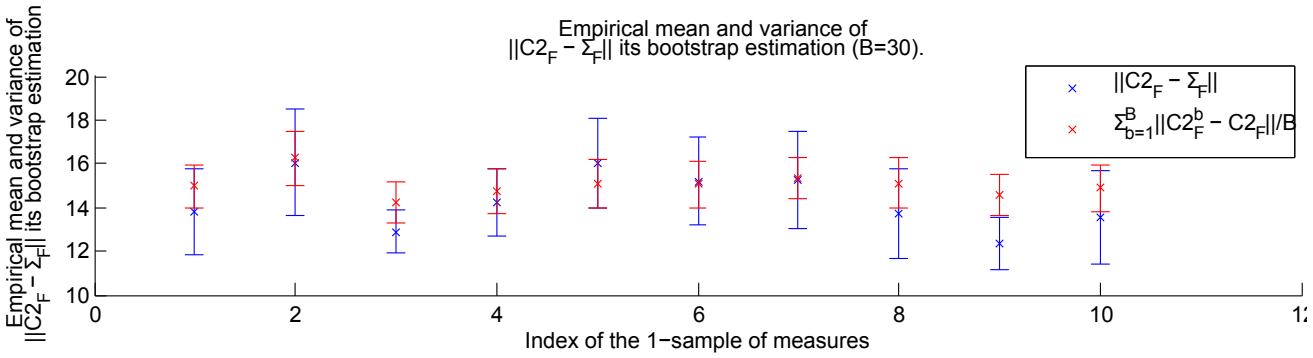
and $C_{ii}^b = \sigma_i^2$

$$\boldsymbol{E}(\|\widehat{\Sigma_F} - \Sigma_F\|) \approx \frac{1}{B} \sum_{b=1}^{B} \|\widehat{\widehat{\Sigma_F}}^b - \widehat{\Sigma_F}\|$$

$$\mathrm{KL}(\widehat{\Sigma_F}, \Sigma_F) \approx \frac{1}{B} \sum_{b=1}^{B} \mathrm{KL}(\widehat{\widehat{\Sigma_F}}^b, \widehat{\Sigma_F})$$

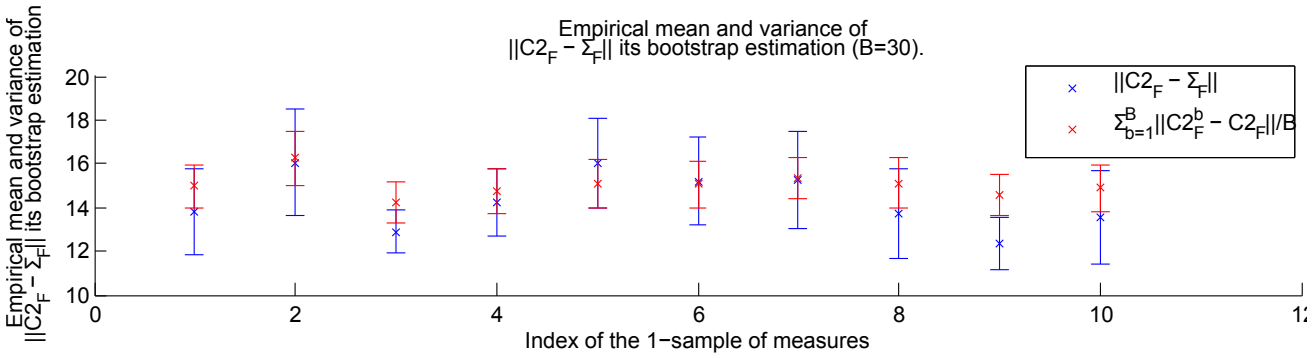
Bootstrap on the 'Toy Model'

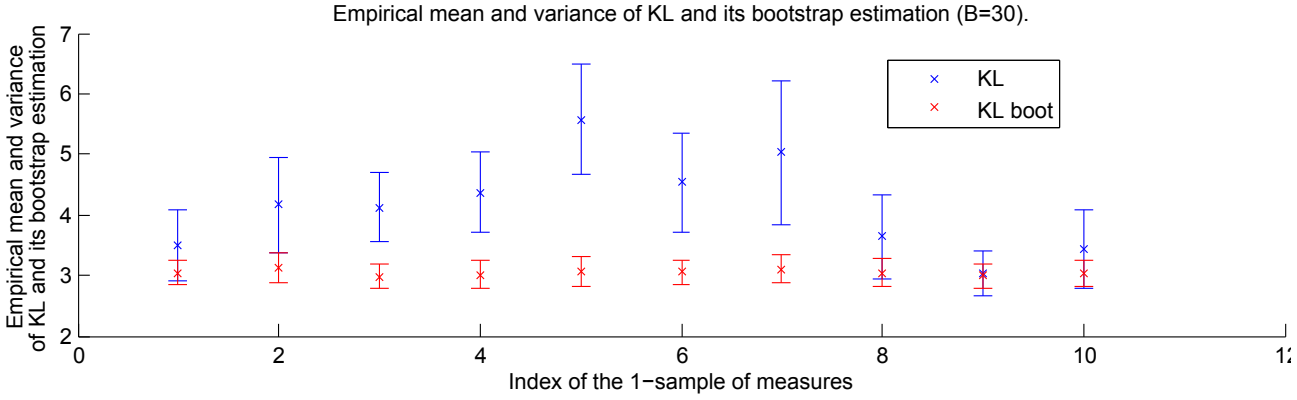
$N = 5$, $n = 1$, $B = 30$, $\|\Sigma_F\| = 15.6680$, $r = 1$,



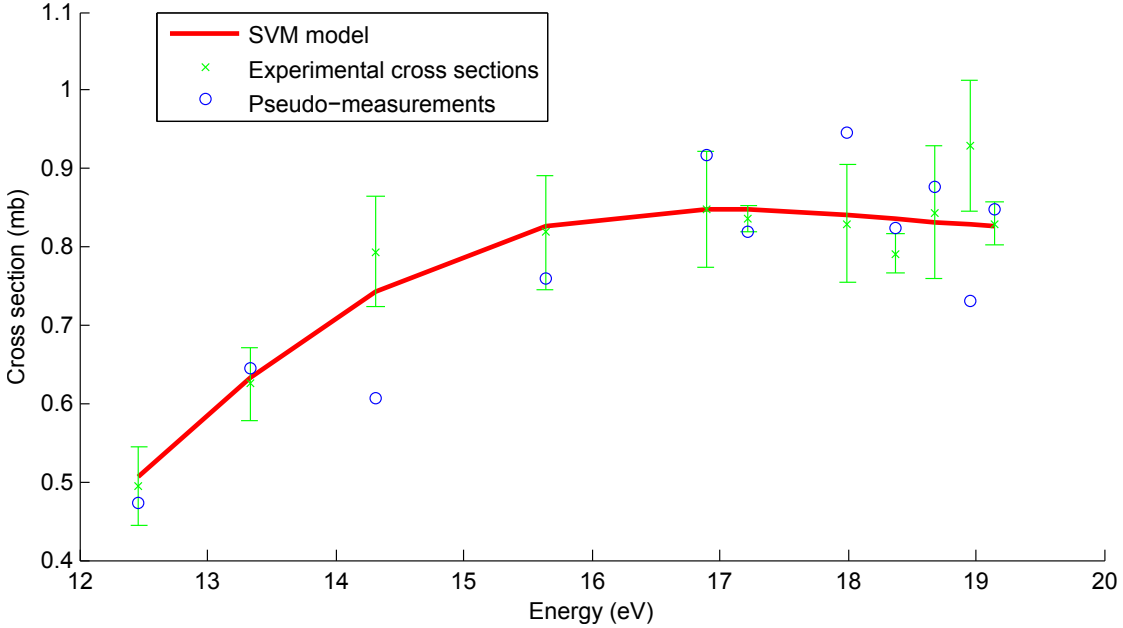
Bootstrap on the 'Toy Model'

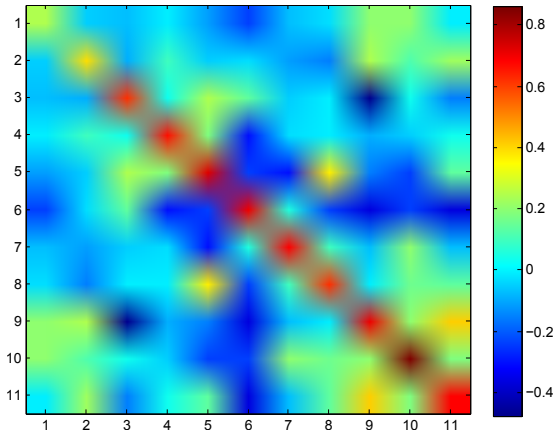
$N = 5$, $n = 1$, $B = 30$, $\|\Sigma_F\| = 15.6680$, $r = 1$,





$^{55}_{25}Mn$: (n,2n) cross section [MTN11]

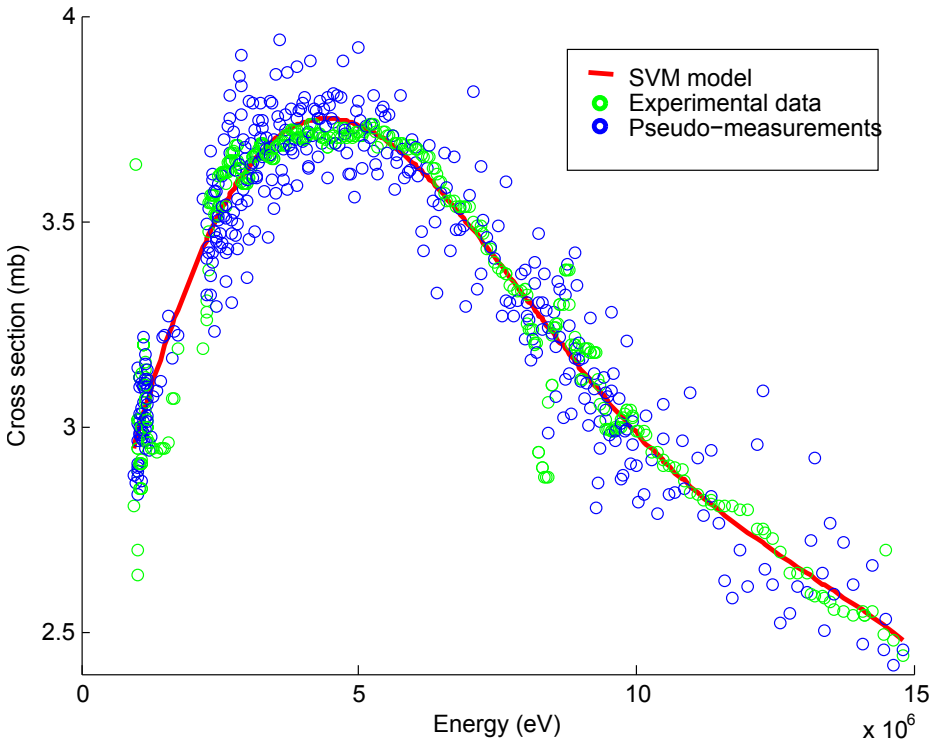


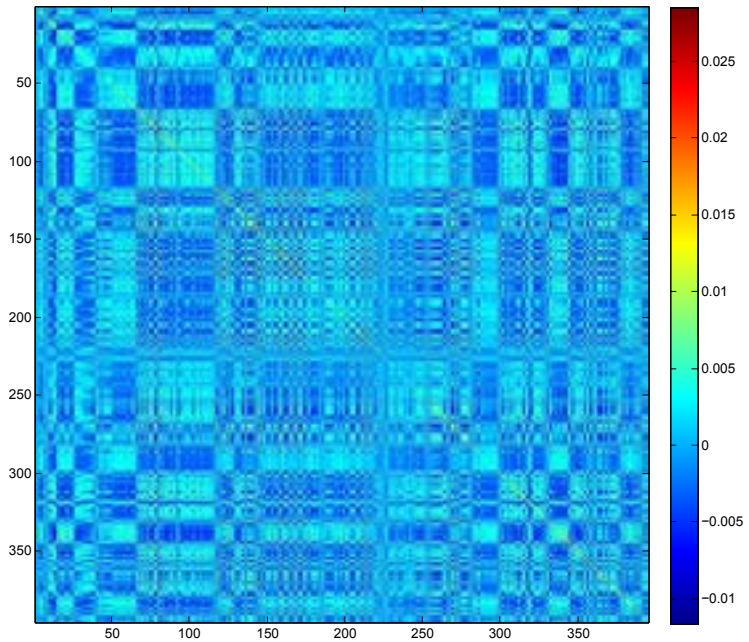


$$\boldsymbol{E}(\|\widehat{\Sigma_F} - \Sigma_F\|) \approx 0.019 \qquad \boldsymbol{E}(KL(\widehat{\Sigma_F})) \approx 6.834$$

$$\mathrm{Var}(\|\widehat{\Sigma_F} - \Sigma_F\|) \approx 1.2\,10^{-5} \qquad \mathrm{Var}(KL(\widehat{\Sigma_F})) \approx 2.351$$

$^{55}_{25}Mn$: total cross section





$$\boldsymbol{E}(\|\widehat{\Sigma_F} - \Sigma_F\|) \approx 2.385 \qquad \boldsymbol{E}(KL(\widehat{\Sigma_F})) \approx 386.36$$

$$\mathrm{Var}(\|\widehat{\Sigma_F} - \Sigma_F\|) \approx 4.3\,10^{-5} \qquad \mathrm{Var}(KL(\widehat{\Sigma_F})) \approx 6.19$$

Outline

Conclusion

Σ_F estimation

1. Classical approaches
   - ✓ one measurement per energy
   - ✓ experimental informations rarely available
   - ✓ quite unrealistic assumptions
2. Our approach:
   - ✓ only measurements are needed
   - ✓ quality measure of the estimation

Estimator quality

Numerical validation

Application

Conclusion

Σ_F estimation

Estimator quality

- Estimation of $\|\widehat{\Sigma}_F - \Sigma_F\|$ with Bootstrap
- Estimation of $KL(\widehat{\Sigma}_F)$ with Bootstrap

Numerical validation

Application

Conclusion

Σ_F estimation

Estimator quality

Numerical validation

- Toy model

Application

Conclusion

Σ_F estimation

Estimator quality

Numerical validation

Application

- Application to $^{55}_{25}Mn$

Future work

New approach

- Optimisation: shrinkage approach (in progress)

Other approaches

- Kriging: cf poster session

References

- Experimental Nuclear Reaction Data EXFOR, *http://www.oecd-nea.org/dbdata/x4/*.
- M. Ionescu-Bujor and D. G. Cacuci, *A comparative review of sensitivity and uncertainty analysis of large-scale systems $-i$: Deterministic methods*, Nuclear Science and Engineering **147** (2004), 189–203.
- Grégoire Kessedjian, *Mesures de sections efficaces d'actinides mineurs d'intérêt pour la transmutation*, Ph.D. thesis, Bordeaux 1, 2008.
- E. Levina, A. Rothman, and J. Zhu, *Sparse estimation of large covariance matrices via a nested lasso penalty'*, The Annals of Applied Statistics **2** (2008), no. 1, 245–263.
- A. Milocco, A. Trkov, and R. Capote Noy, *Nuclear data evaluation of 55 mn by the empire code with emphasis on the capture cross-section*, Nuclear Engineering and Design **241** (2011), no. 4, 1071–1077.
- A.J. Smola and B. Schölkopf, *A tutorial on support vector regression*, Statistics and Computing **14** (2004), 199–222.
- S. Varet, P. Dossantos-Uzarralde, N. Vayatis, and E. Bauge, *Experimental covariances contribution to cross-section uncertainty determination*, NCSC2 Vienna, 2011.
- V. Vapnik, S. Golowich, and A. Smola, *Support vector method for function approximation, regression estimation, and signal processing*, Advances in Neural Information Processing Systems **9** (1997), 281–287.