Tensor Estimation with Multilinear Spectral Penalties

ICML Workshop on Machine Learning For System Identification

Marco Signoretto

ESAT-SCD/SISTA, KULeuven

Atlanta June 21, 2013

Acknowledgements

- **Andreas Argyriou**,
  Center for Learning and Visual Computing, Ecole Centrale Paris
- **Borbála Hunyadi**,
  ESAT-SCD/SISTA, KULeuven
- **Quoc Tran Din**,
  Laboratory for Information and Inference Systems, EPFL
- **Lieven De Lathauwer**,
  ESAT-SCD/SISTA, KULeuven
- **Johan A. K. Suykens**,
  ESAT-SCD/SISTA, KULeuven

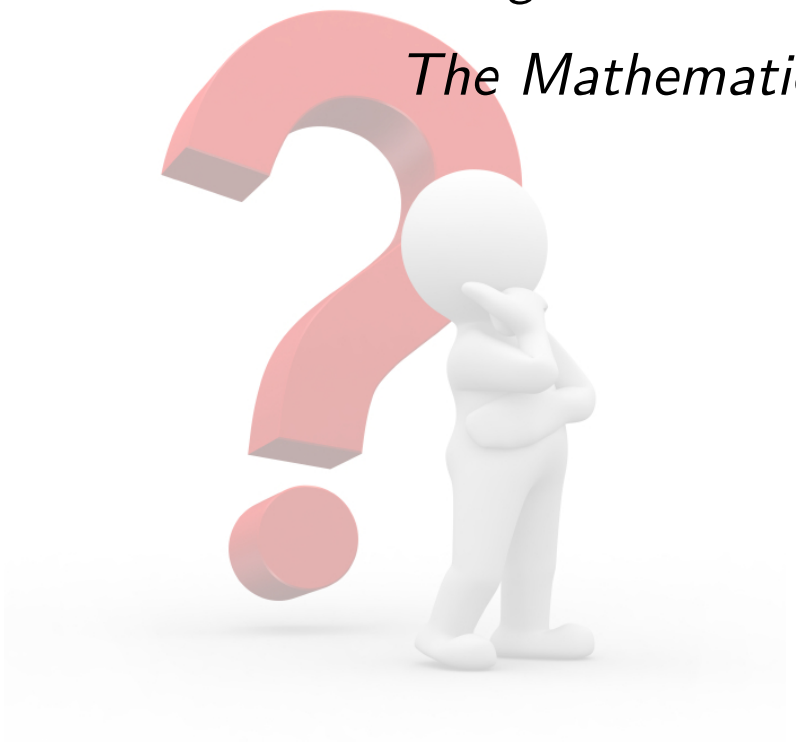
Outline

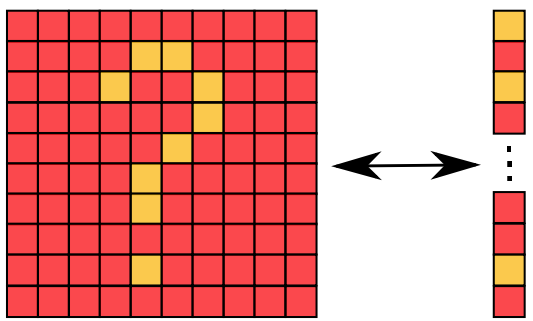
1. General Setting
   - Learning with Tensors
   - Tools for Tensors
2. Multilinear Spectral Learning
   - Inductive Learning
   - Transductive Learning
3. Tensor Estimation in RKHSs
   - Tools for Tensors (Revisited)
   - Inference Problem and Representer Theorem

Why Tensors?

One could argue that the most important aspect of intelligence [...] is the ability to learn from very few examples. [...] Evolution has probably done a part of the learning but so have we, when we choose for any given task an **appropriate input representation** *for our learning machines.*

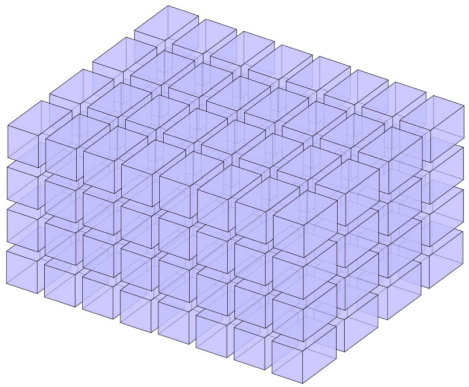
The Mathematics of Learning: Dealing with Data
T. Poggio and S. Smale





Why Tensors?

- natural for many type of data
- expressive multi-way structure
- tools from multilinear algebra
- tensor transforms



Higher Order Representations in Real-life Applications

neuroscience:	EEG data (time samples $\times$ frequency $\times$ electrodes)
computer vision:	image (/video) compression/completion/$\cdots$ (pixel $\times$ illumination $\times$ expression $\times \cdots$)
web mining:	analyze users behaviours (users $\times$ queries $\times$ webpages)

Tensor-based Data Analysis

A Supervised Learning Approach

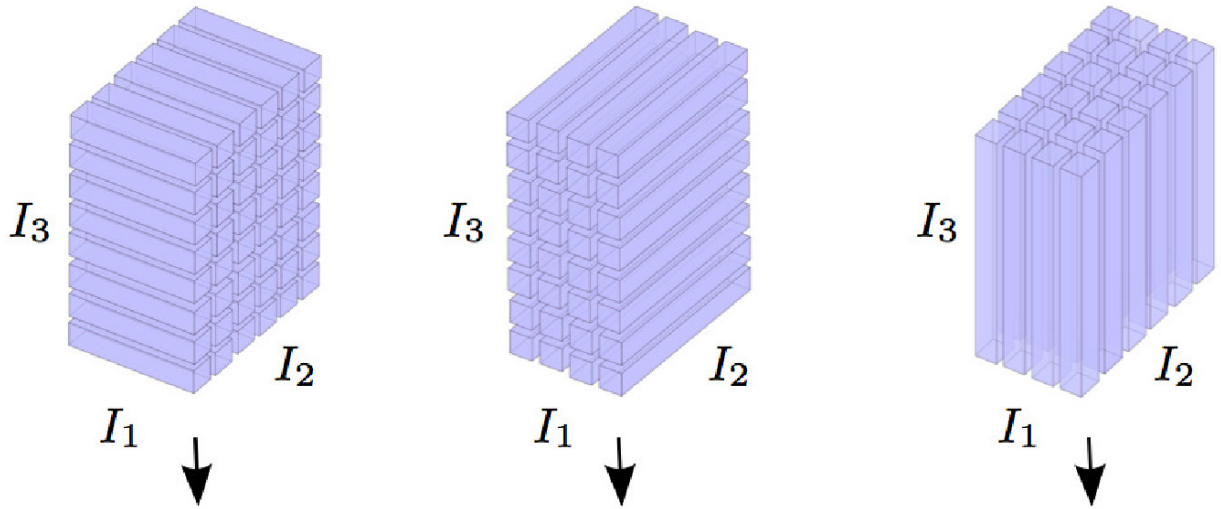
- generalization of spectral penalties to tensors
- unifying convex framework and template algorithm
- exploit the structure of data with tensorial representation
- multitask (/multiclass) learning

Marco Signoretto, Quoc Tran Dinh, Lieven De Lathauwer, Johan A. K. Suykens. *Learning with tensors: a framework based on convex optimization and spectral regularization*, Machine Learning (In press) 2013

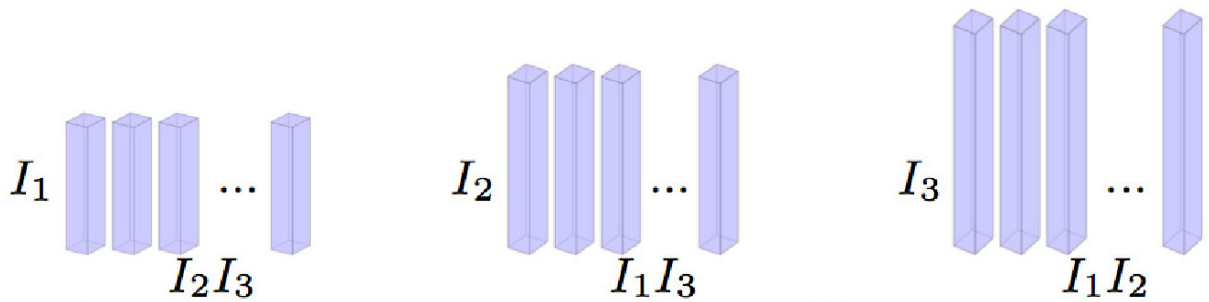
Tensor Unfolding

- $q-$mode *unfolding*: $M_q : \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_Q} \longrightarrow \mathbb{R}^{I_q \times I_1 I_2 \cdots I_{q-1} I_{q+1} \cdots I_Q}$
- $q-$mode *refolding*: $M_q^* : \mathbb{R}^{I_q \times I_1 I_2 \cdots I_{q-1} I_{q+1} \cdots I_Q} \longrightarrow \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_Q}$

tensor $\mathcal{X}$ $\longrightarrow$

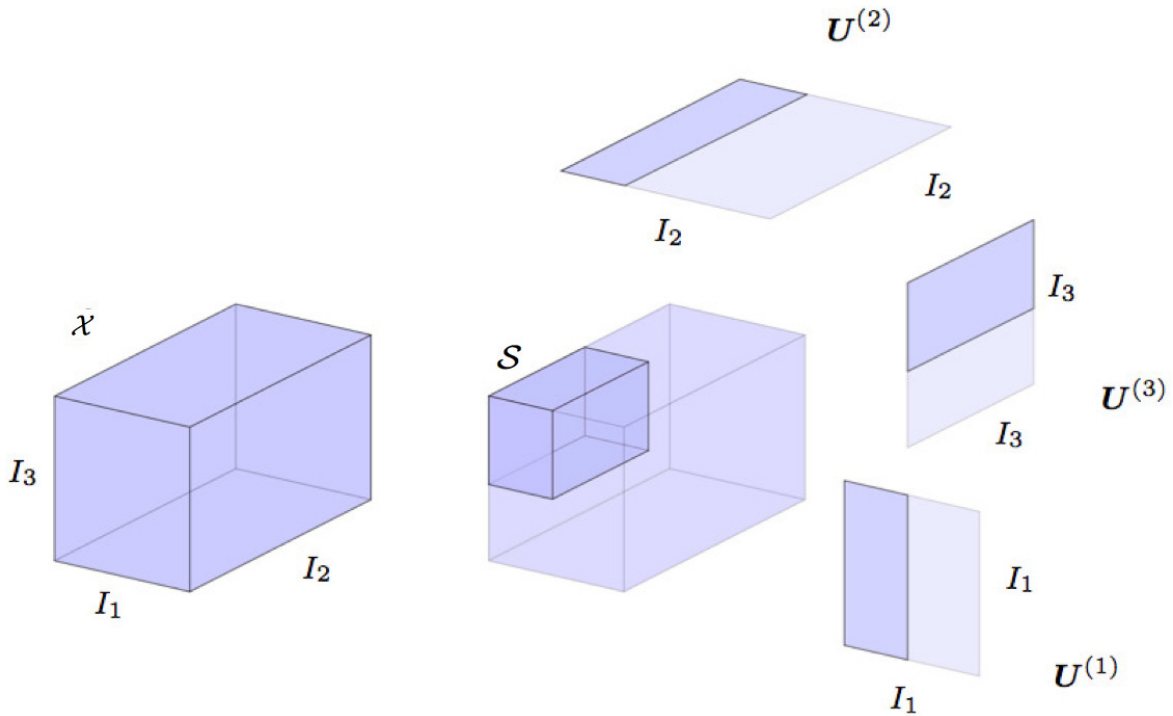


q-mode vectors $\longrightarrow$



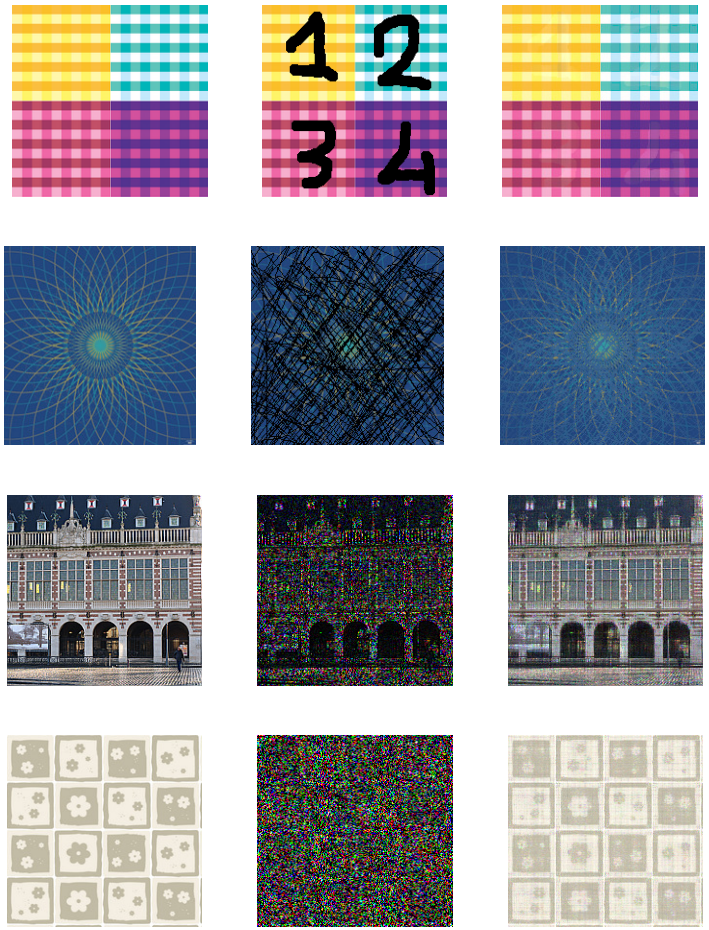
Multilinear Singular Value Decomposition (MLSVD)

- mlrank$(\mathcal{X})$:= $(R_1, R_2, \ldots, R_Q)$, $\quad R_q :=$ dim(span(q-mode vectors))
- q-mode product: $\qquad \mathcal{X} = \mathcal{S} \times_q \boldsymbol{U} \quad \Leftrightarrow \quad M_q(\mathcal{X}) = \boldsymbol{U} M_q(\mathcal{S})$
- MLSVD: $\qquad \mathcal{X} = \mathcal{S} \times_1 \boldsymbol{U}^{(1)} \times_2 \boldsymbol{U}^{(2)} \times_3 \cdots \times_Q \boldsymbol{U}^{(Q)}$



Multilinear Spectral Regularization

$$\text{mlrank}(\mathcal{X}) = (\text{rank}(M_1(\mathcal{X})), \text{rank}(M_2(\mathcal{X})), \ldots, \text{rank}(M_Q(\mathcal{X})))$$



- penalty on the mlrank:

$$\Omega(\mathcal{X}) = \frac{1}{Q} \sum_{q \in \mathbb{N}_Q} \text{rank}(M_q(\mathcal{X}))$$

- convex proxy:

$$\Omega(\mathcal{X}) = \frac{1}{Q} \sum_{q \in \mathbb{N}_Q} \|M_q(\mathcal{X})\|_1$$

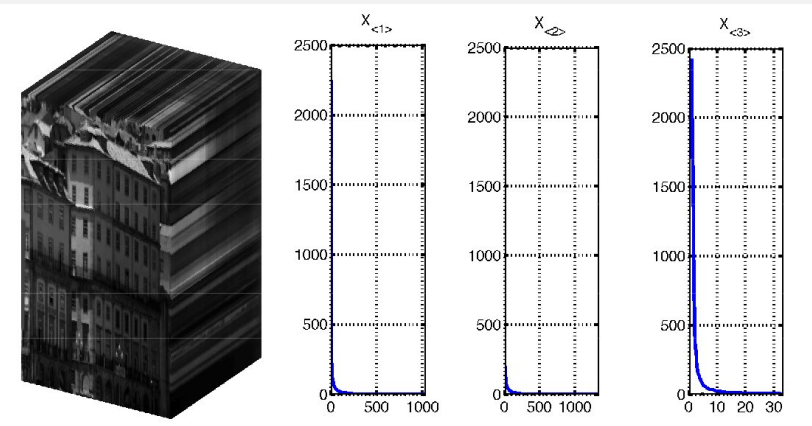
where $\|\boldsymbol{A}\|_1 := \sum_r \sigma_r(\boldsymbol{A})$

Taxonomy of Learning Problems

inductive learning with tensors

data: pairs of fully specified input features and vectors of target labels

output: models for out-of-sample evaluations of multiple tasks



transductive learning with tensors

soft-completion

data: partially specified input data tensor and matrix of target labels

output: latent features and missing labels

hard-completion

data: partially specified input data tensor

output: missing input data

Inductive Problems

inference problem: $\underset{\mathcal{W},b}{\text{minimize}} \; f_{\mathscr{D}_N}(\mathcal{W}, b) + \sum_{q \in \mathbb{N}_{Q+1}} \lambda_q \|M_q(\mathcal{W})\|_1$

Modeling Assumptions (Logistic Regression)

low mlrank discriminative features:

$$\mathcal{X}^{(n)} = \tilde{\mathcal{X}}^{(n)} + \mathcal{E}^{(n)} = \mathcal{S}^{(n)} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times \cdots \times_Q \boldsymbol{U}_Q + \mathcal{E}^{(n)}$$

probability model:

$$p(y_{tn}|\tilde{y}_{tn}) = 1/(1 + \exp(-y_{tn}\tilde{y}_{tn})), \quad \tilde{y}_{tn} = \langle \mathcal{S}_{\tilde{\mathcal{X}}^{(n)}}, \mathcal{S}_{\mathcal{W}^{(t)}} \rangle + b^{(t)}$$

task relatedness:

$$w_{d_1 d_2 \cdots d_{Q+1}} := w^{(t)}_{d_1 d_2 \cdots d_Q}, \qquad \text{rank}(M_{Q+1}(\mathcal{W})) \ll T$$

Classification of structured EEG data

Single-trial P300 classification for ERP-based BCIs

Task: detect the onset of positive and negative auditory stimuli

Challenge:
improving accuracy from mobile recordings (e.g., walking outdoors)



Problem Features:

- P300 signature spans multiple channels/frequencies
- few high-dimensional examples
- need robustness to noise

Classification of structured EEG data

Single-trial P300 classification for ERP-based BCIs

- **LDA:** features/electrode concatenated in a long vector
- **mNNL:** features/electrode constitute the row of a matrix
- **tNNL:** Hankel matrix per electrode stacked into a 3rd order tensor

B. Hunyadi, M. Signoretto, W. Van Paesschen, J. Suykens, S. Van Huffel, and M. De Vos, Incorporating structural information from the multichannel EEG improves patient-specific seizure detection, *Clinical Neurophysiology*, 2012

B. Hunyadi, M. Signoretto, S. Debener, S. Van Huffel, M. De Vos, Classification of structured EEG tensors using nuclear norm regularization: improving P300 classification , *Workshop on Pattern Recognition in NeuroImaging* (PRNI) 2013

Classification of structured EEG data

Single-trial P300 classification for ERP-based BCIs

- **LDA:** features/electrode concatenated in a long vector
- **mNNL:** features/electrode constitute the row of a matrix
- **tNNL:** Hankel matrix per electrode stacked into a 3rd order tensor

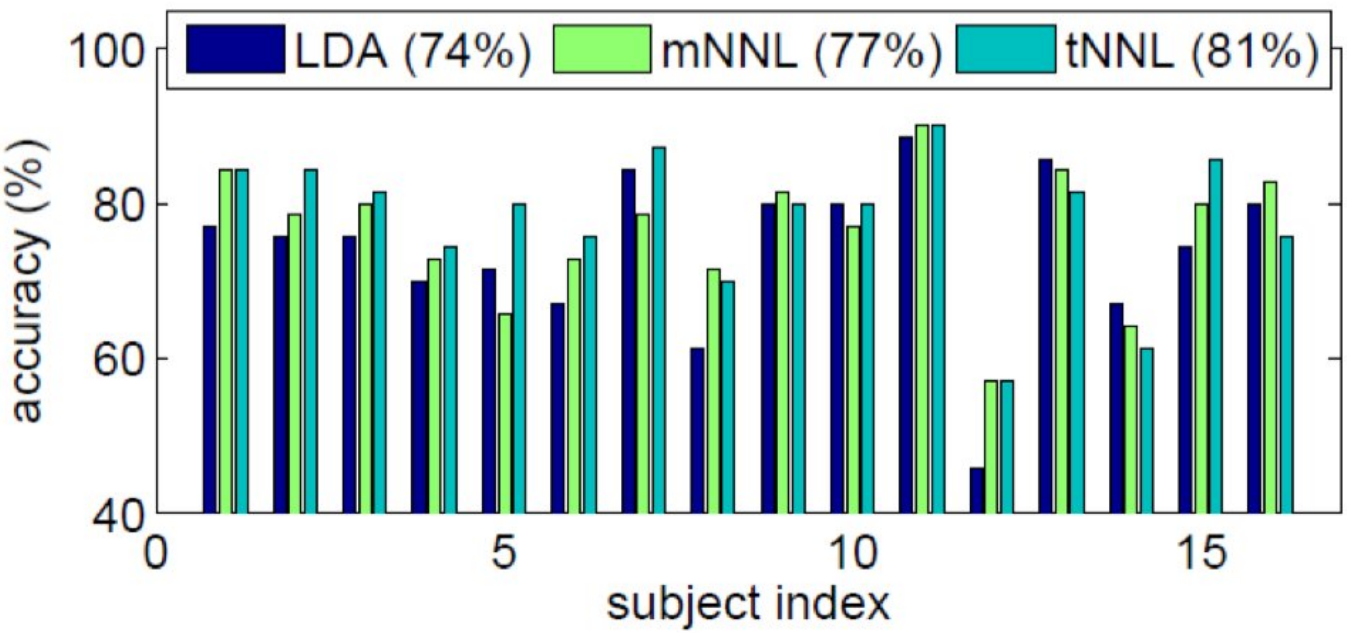


Figure: inside

Classification of structured EEG data

Single-trial P300 classification for ERP-based BCIs

- **LDA:** features/electrode concatenated in a long vector
- **mNNL:** features/electrode constitute the row of a matrix
- **tNNL:** Hankel matrix per electrode stacked into a 3rd order tensor

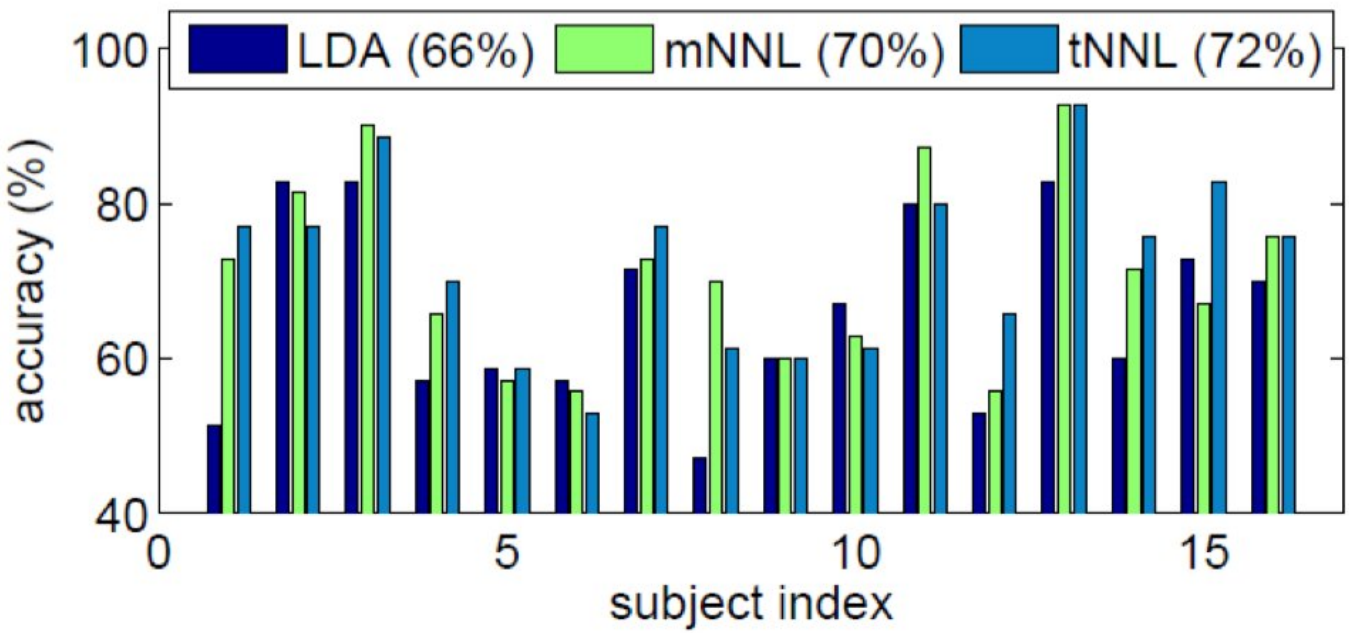


Figure: outside

Transductive Problems

goal: *jointly* fill in the missing entries (purple tones)

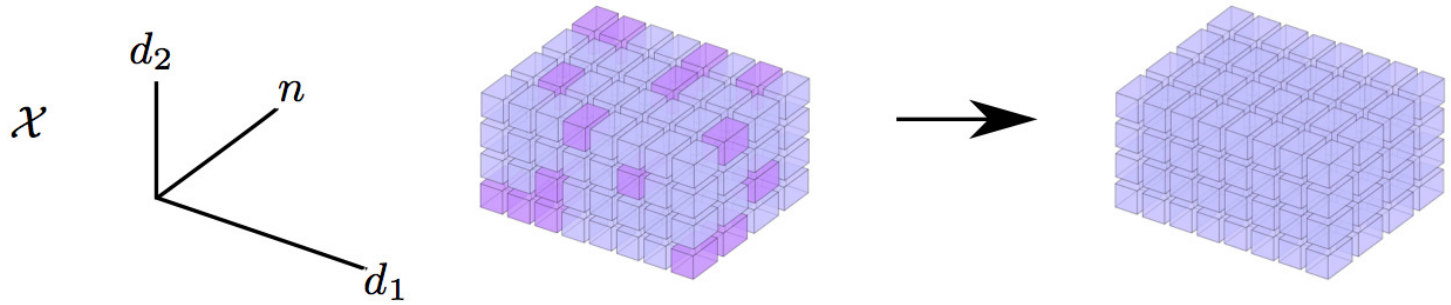


Figure: Input observations ($D_1 \times D_2$ matrices) stacked along the third-mode of the input dataset $\mathcal{X}$

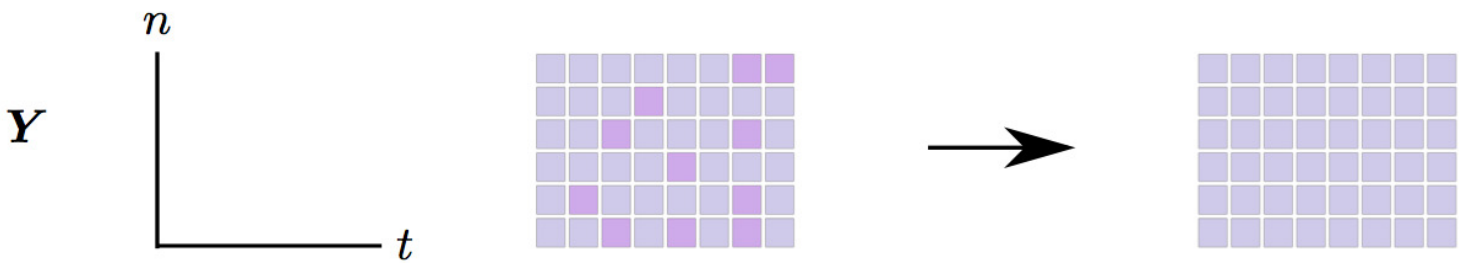


Figure: Target observations ($T-$dimensional vectors) stacked in the matrix $\boldsymbol{Y}$

Olivetti Faces

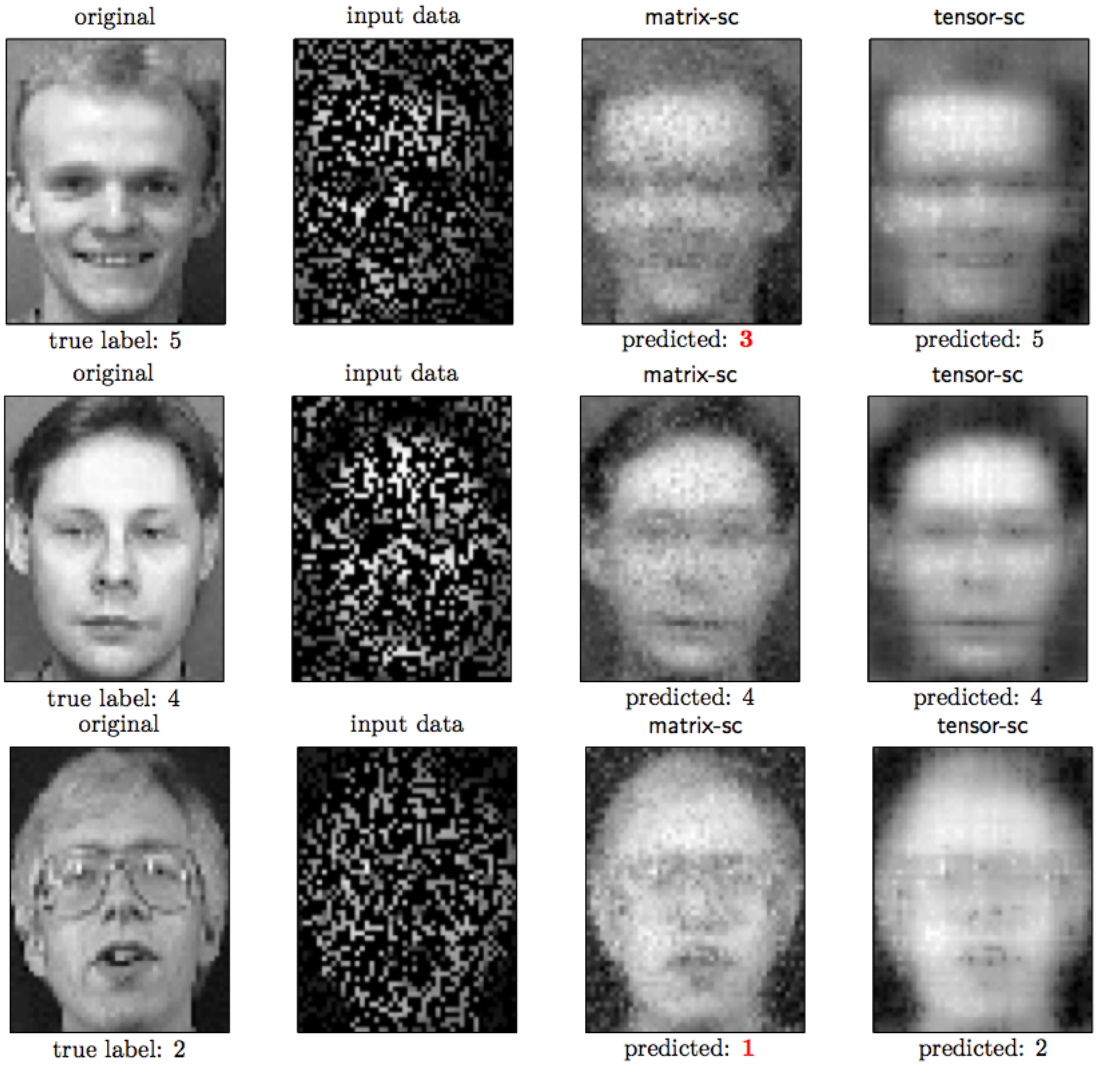


Figure: soft-completion (5 classes, 1-vs-1 encoding)

tensor-sc

$\hat{y}$ \ y	1	2	3	4	5
1	**19**	3	0	0	3
2	0	**29**	0	0	0
3	0	0	**20**	0	0
4	0	0	1	**23**	0
5	0	0	0	0	**27**

matrix-sc

$\hat{y}$ \ y	1	2	3	4	5
1	17	3	1	0	4
2	2	27	0	0	0
3	0	0	**20**	0	0
4	0	0	3	21	0
5	1	0	2	0	24

imp+ls-svm

$\hat{y}$ \ y	1	2	3	4	5
1	14	5	0	1	5
2	0	**29**	0	0	0
3	0	0	19	0	1
4	0	0	2	22	0
5	0	0	0	0	**27**

Table: confusion matrices

System Identification with Missing Data

System Identification with Missing Inputs and Outputs [Liu et al., 2013]

$$\min_{u,y} \lambda_1 \left\| \mathcal{S}_u(u) - u_{\text{meas}} \right\|^2 + \lambda_2 \left\| \mathcal{S}_y(y) - y_{\text{meas}} \right\|^2 + \left\| F(u,y) \right\|_1$$

where:

$$F : (u,y) \mapsto \begin{bmatrix} \mathcal{H}(u) \\ \hline \mathcal{H}(y) \end{bmatrix}, \; \mathcal{H}(x) = \begin{bmatrix} x_1 & x_2 & \cdots & x_T \\ x_2 & x_3 & \cdots & x_{T+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_I & x_{I+1} & \cdots & x_{T+I-1} \end{bmatrix}$$

Z. Liu, A. Hansson, L. Vandenberghe, Nuclear norm system identification with missing inputs and outputs, *System and Control Letters* 62, 605-612, 2013

System Identification with Missing Data

Shared Dictionary Across Multiple Systems

- M systems with common dynamics: $y_t^{(m)} = \sum_{d=1}^{D} \sum_i w_d^{(m)} \psi_i^{(d)} u_{t-i}$
- $\psi^{(d)}$ impulse response function of a stable ARMA(P, Q), $P \geq Q$

System Identification with Missing Data

Shared Dictionary Across Multiple Systems

if $\mathcal{Z} \in \mathbb{R}^{2I \times T \times M}$ satisfies $z_{jtm} = \left(F^{(m)}\right)_{jt}$, $F^{(m)}(u, y) = \begin{bmatrix} \mathcal{H}(y^{(m)}) \\ \hline \mathcal{H}(u) \end{bmatrix}$

then

- $\text{rank}(M_2(\mathcal{Z})) = \text{rank}(\mathcal{H}(u)) + DP \ll T$
- $\text{rank}(M_3(\mathcal{Z})) = D \ll M$

System Identification with Missing Data

Shared Dictionary Across Multiple Systems

if $\mathcal{Z} \in \mathbb{R}^{2I \times T \times M}$ satisfies $z_{jtm} = \left(F^{(m)}\right)_{jt}$, $F^{(m)}(u,y) = \begin{bmatrix} \mathcal{H}(y^{(m)}) \\ \hline \mathcal{H}(u) \end{bmatrix}$

then

- $\text{rank}(M_2(\mathcal{Z})) = \text{rank}(\mathcal{H}(u)) + DP \ll T$
- $\text{rank}(M_3(\mathcal{Z})) = D \ll M$

Multitask System Identification as Completion of a Structured Tensor

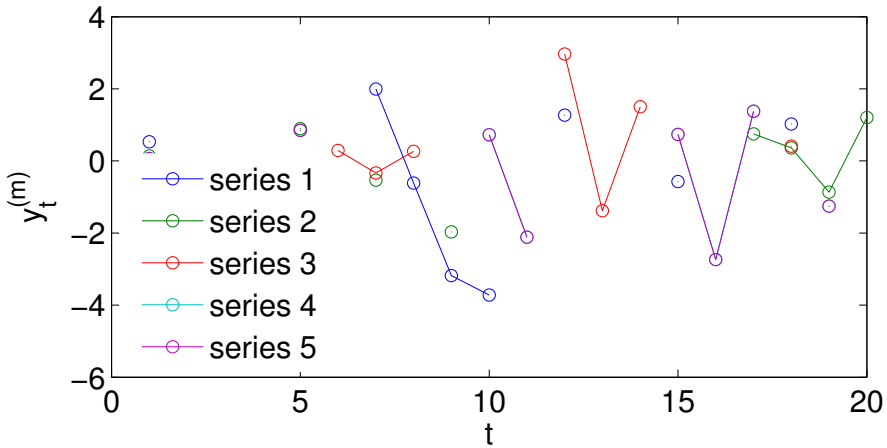
$$\min_{u,y,\mathcal{Z}} \quad f_{\mathscr{D}_N,\lambda}(u,y) + \sum_{j=2}^{3} \|M_j(\mathcal{Z})\|_1$$

$$\text{subject to} \quad z_{jtm} = \left(F^{(m)}(u,y)\right)_{jt} \text{ for all } j,t,m$$

System Identification with Missing Data

Toy Example

- 20 systems
- $T = 60, D = 3,\ P = 3,\ Q = 2$
- 20% missing inputs
- 40% missing outputs



Multitask System Identification as Completion of a Structured Tensor

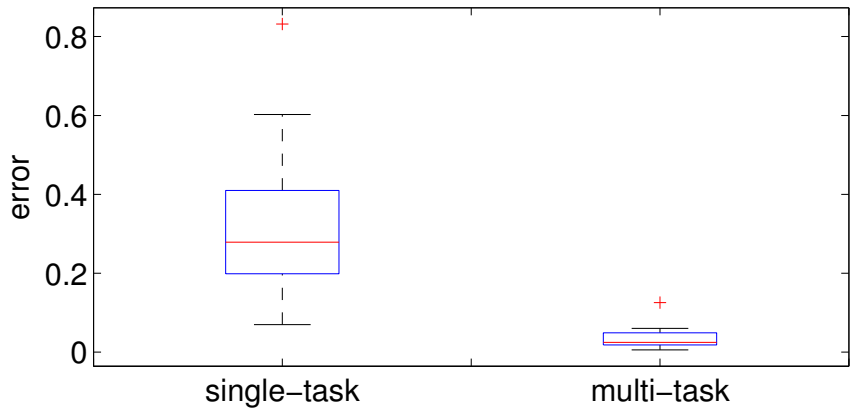
$$\min_{u,y,\mathcal{Z}} \quad f_{\mathscr{D}_N,\lambda}(u,y) + \sum_{j=2}^{3} \|M_j(\mathcal{Z})\|_1$$

$$\text{subject to} \quad z_{jtm} = \left(F^{(m)}(u,y)\right)_{jt} \text{ for all } j,t,m$$

System Identification with Missing Data

Toy Example

- 20 systems
- $T = 60, D = 3,\ P = 3,\ Q = 2$
- 20% missing inputs
- 40% missing outputs
- error=impulse responses distance



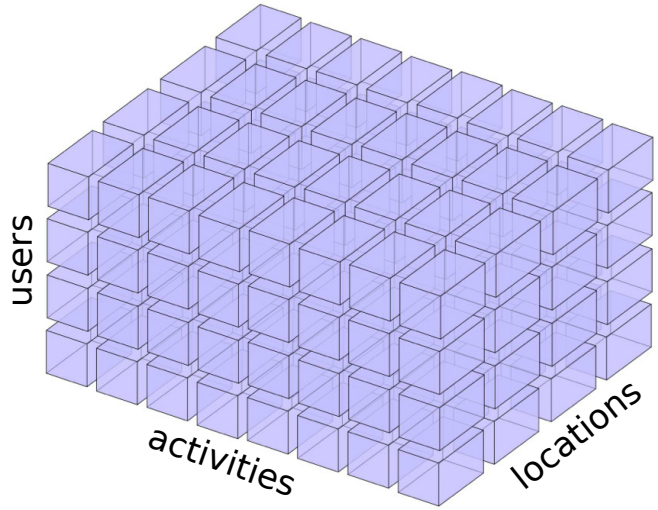
Multitask System Identification as Completion of a Structured Tensor

$$\min_{u,y,\mathcal{Z}} \quad f_{\mathscr{D}_N,\lambda}(u,y) + \sum_{j=2}^{3} \|M_j(\mathcal{Z})\|_1$$

$$\text{subject to} \quad z_{jtm} = \left(F^{(m)}(u,y)\right)_{jt} \quad \text{for all } j,t,m$$

Overcoming Limits

- what is the kernel function induced by completion?
- how to transfer knowledge to new entries?
- how to incorporate attributes (user profiles, object features,...)?



Co-rating Problem

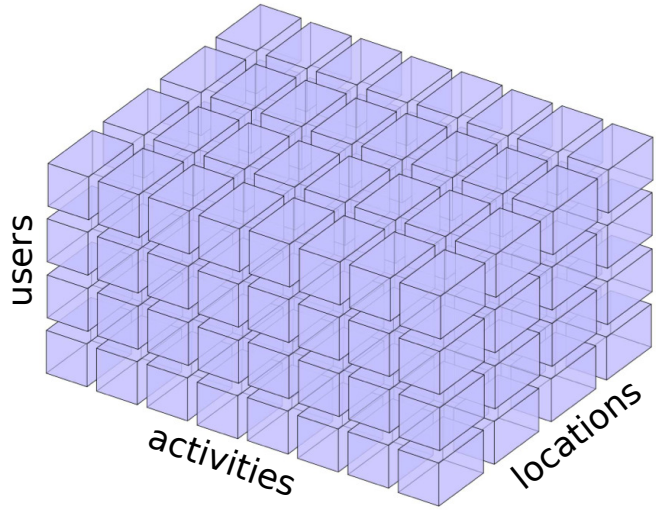
Estimate a model:

$$f : \text{user}_1 \times \text{user}_2 \times \cdots \times \text{user}_N \times \text{movie} \rightarrow \mathcal{Y}$$

based upon:
- ratings of the individual users
- users and/or movies features (or similarities)

Overcoming Limits

- what is the kernel function induced by completion?
- how to transfer knowledge to new entries?
- how to incorporate attributes (user profiles, object features,...)?



Higher Order Recommender Systems / Analysis of Survey Data

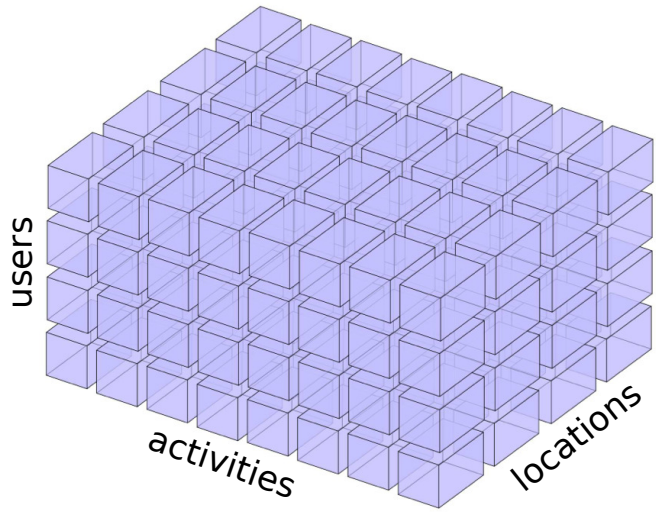
Estimate a model:

$$f : \text{user} \times \text{activity} \times \text{location} \times \cdots \rightarrow \mathcal{Y}$$

based upon:
- score/preference/frequency on selected tuples
- features (or similarities) per mode

Overcoming Limits

- what is the kernel function induced by completion?
- how to transfer knowledge to new entries?
- how to incorporate attributes (user profiles, object features,...)?



Proposed Approach

- from finite dimensional tensors to tensor product functions
- tools for tensors (revisited)
- general inference problem and representer theorem

(Partial) Tensor Product RKHS $\boldsymbol{\mathcal{H}}_\beta$

- $\left(\mathcal{H}_q, \langle\cdot,\cdot\rangle_q, k^{(q)}\right)$ RKHS of functions defined on $\mathcal{X}_q$, $q \in \mathbb{N}_Q$

(Partial) Rank-1 tensor product function ($\beta \subseteq \mathbb{N}_Q$)

$$\otimes_{q\in\beta} f^{(q)} : x^{(\beta)} = \left(x^{(q)} \ : \ q \in \beta\right) \mapsto \prod_{q\in\beta} f^{(q)}\left(x^{(q)}\right), \quad f^{(q)} \in \mathcal{H}_q$$

- finite-rank function: $\boldsymbol{f}^{(\beta)} = \sum_\nu \alpha_\nu \prod_{q\in\beta} f_\nu^{(q)}$

Reproducing Property

$$\boldsymbol{f}(x) = \langle \boldsymbol{f}^{(\beta)}, \boldsymbol{k}_x^{(\beta)} \rangle = \sum_\nu \alpha_\nu \prod_{q\in\beta} \langle f_\nu^{(q)}, k_x^{(q)} \rangle_q = \sum_\nu \alpha_\nu \prod_{q\in\beta} f_\nu^{(q)}\left(x^{(q)}\right)$$

(Partial) Tensor Product RKHS $\boldsymbol{\mathcal{H}}_\beta$

- $\left(\mathcal{H}_q, \langle\cdot,\cdot\rangle_q, k^{(q)}\right)$ RKHS of functions defined on $\mathcal{X}_q$, $q \in \mathbb{N}_Q$

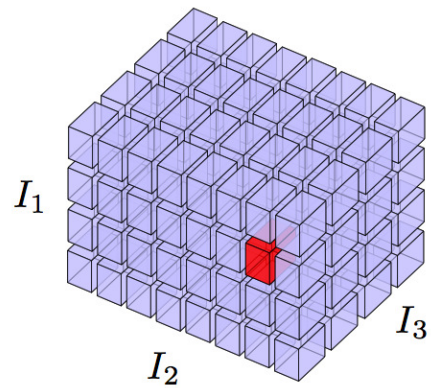
Special Case: Finite Dimensional Tensors

- $\left(\mathcal{H}_q, \langle\cdot,\cdot\rangle_q, k^{(q)}\right)$ RKHS of functions on $\mathbb{I}_q = \{1, 2, \ldots, I_q\}$
- $\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_Q \ni x = (i_1, i_2, \ldots, i_Q)$
- evaluation functional and reproducing property:

$$\left(f^{(1)} \otimes f^{(2)} \otimes \cdots \otimes f^{(Q)}\right)_{i_1 i_2 \ldots i_Q} =$$

$$\left\langle f^{(1)} \otimes f^{(2)} \otimes \cdots \otimes f^{(Q)}, \boldsymbol{k}_{(i_1 i_2 \ldots i_Q)} \right\rangle =$$

$$f^{(1)}_{i_1} f^{(2)}_{i_2} \cdots f^{(Q)}_{i_Q}$$



Functional Unfolding

Functional Unfolding [W. Hackbusch. Tensor Spaces and Tensor Calculus, 2012]

$$M_\beta : \quad \otimes_{q=1}^{Q} f^{(q)} \quad \mapsto \quad \boldsymbol{f}^{(\beta)} \otimes \boldsymbol{f}^{(\beta^c)} : \quad \begin{array}{ccc} \boldsymbol{\mathcal{H}}_{\beta^c} & \to & \boldsymbol{\mathcal{H}}_{\beta} \\ \boldsymbol{z} & \mapsto & \langle \boldsymbol{f}^{(\beta^c)}, \boldsymbol{z} \rangle_{\beta^c} \boldsymbol{f}^{(\beta)} \end{array}$$

Functional Unfolding

Functional Unfolding [W. Hackbusch. Tensor Spaces and Tensor Calculus, 2012]

$$M_\beta : \quad \otimes_{q=1}^{Q} f^{(q)} \quad \mapsto \quad \boldsymbol{f}^{(\beta)} \otimes \boldsymbol{f}^{(\beta^c)} : \quad \begin{array}{ccc} \boldsymbol{\mathcal{H}}_{\beta^c} & \to & \boldsymbol{\mathcal{H}}_{\beta} \\ \boldsymbol{z} & \mapsto & \langle \boldsymbol{f}^{(\beta^c)}, \boldsymbol{z} \rangle_{\beta^c} \boldsymbol{f}^{(\beta)} \end{array}$$

Special Case: Matricization of a $I_1 \times I_2 \times I_3$ Tensors

$$M_{\{1\}}(\boldsymbol{f}) = \begin{bmatrix} \boldsymbol{f}_{111} & \boldsymbol{f}_{112} & \cdots & \boldsymbol{f}_{11I_3} & \boldsymbol{f}_{121} & \boldsymbol{f}_{122} & \cdots & \boldsymbol{f}_{1I_2I_3} \\ \boldsymbol{f}_{211} & \boldsymbol{f}_{212} & \cdots & \boldsymbol{f}_{21I_3} & \boldsymbol{f}_{221} & \boldsymbol{f}_{222} & \cdots & \boldsymbol{f}_{2I_2I_3} \\ \vdots & & \ddots & & & \cdots & \vdots & \\ \boldsymbol{f}_{I_111} & \boldsymbol{f}_{I_112} & \cdots & \boldsymbol{f}_{I_11I_3} & \boldsymbol{f}_{I_121} & \boldsymbol{f}_{I_122} & \cdots & \boldsymbol{f}_{I_1I_2I_3} \end{bmatrix}$$

Multilinear Spectral Regularization

$$\text{if } M_\beta(\boldsymbol{f}) \text{ is compact} \quad \longrightarrow \quad M_\beta(\boldsymbol{f}) = \textstyle\sum_r \sigma_r^{(\beta)} \boldsymbol{u}_r \otimes \boldsymbol{v}_r$$

$\mathcal{B}$-multlilinear rank

For $\mathcal{B}$ partition of $\mathbb{N}_Q$: $\quad \mathrm{mlrank}_{\mathcal{B}}(\boldsymbol{f}) := (\mathrm{rank}(M_\beta(\boldsymbol{f})) \ : \ \beta \in \mathcal{B})$

Multilinear Spectral Penalties (MLSPs)

- sum of *Hilbert-Frobenius norms* (a.k.a. Shatten-2 norms):

$$\Omega_{\mathcal{B}}(\boldsymbol{f}) = \sum_{\beta \in \mathcal{B}} \|M_\beta(\boldsymbol{f})\|_2^2, \text{ where } \|A\|_2^2 := \langle A, A \rangle_{\mathsf{HF}} = \sum_r \sigma_r^2(A)$$

Multilinear Spectral Regularization

$$\text{if } M_\beta(\boldsymbol{f}) \text{ is compact} \quad \longrightarrow \quad M_\beta(\boldsymbol{f}) = \textstyle\sum_r \sigma_r^{(\beta)} \boldsymbol{u}_r \otimes \boldsymbol{v}_r$$

$\mathcal{B}$-multlilinear rank

For $\mathcal{B}$ partition of $\mathbb{N}_Q$: $\quad \mathrm{mlrank}_{\mathcal{B}}(\boldsymbol{f}) := (\mathrm{rank}(M_\beta(\boldsymbol{f})) \ : \ \beta \in \mathcal{B})$

Multilinear Spectral Penalties (MLSPs)

- sum of *nuclear norms* (a.k.a. trace norms or Shatten-1 norms):

$$\Omega_{\mathcal{B}}(\boldsymbol{f}) = \sum_{\beta \in \mathcal{B}} \|M_\beta(\boldsymbol{f})\|_1, \text{ where } \|A\|_1 := \sum_r \sigma_r(A)$$

Multilinear Spectral Regularization

$$\text{if } M_\beta(\boldsymbol{f}) \text{ is compact} \quad \longrightarrow \quad M_\beta(\boldsymbol{f}) = \sum_r \sigma_r^{(\beta)} \boldsymbol{u}_r \otimes \boldsymbol{v}_r$$

$\mathcal{B}$-multlilinear rank

For $\mathcal{B}$ partition of $\mathbb{N}_Q$: $\quad \operatorname{mlrank}_{\mathcal{B}}(\boldsymbol{f}) := (\operatorname{rank}(M_\beta(\boldsymbol{f})) \ : \ \beta \in \mathcal{B})$

Multilinear Spectral Penalties (MLSPs)

- mlrank-constrained variation of the previous MLSPs:

$$\Omega_{\mathcal{B}}(\boldsymbol{f}) = \begin{cases} \sum_{q \in \mathbb{N}_Q} \|M_{\beta_q}(\boldsymbol{f})\|_p^p, & \text{if } \operatorname{rank}(M_{\beta_q}(\boldsymbol{f})) \leq R_q \ \forall q \in \mathbb{N}_Q \\ \infty, & \text{otherwise .} \end{cases}$$

Main Inference Problem

$$\mathcal{D}_N = \left\{ \left(x_n^{(1)}, \ldots, x_n^{(Q)}, y_n \right) \in (\mathcal{X}_1 \times \cdots \times \mathcal{X}_Q) \times \mathcal{Y} : n \in \mathbb{N}_N \right\}$$

Penalized Empirical Risk Minimization Problem

$$\inf_{\boldsymbol{f} \in \boldsymbol{H}_{\mathcal{B}}} \sum_{n \in \mathbb{N}_N} l\left(y_n, \langle \boldsymbol{k}_{x_n}, \boldsymbol{f} \rangle\right) + \lambda \, \Omega_{\mathcal{B}}(\boldsymbol{f})$$

Special Case: Regularization Networks [T. Evgeniou et al., 2000]

- $\|M_\beta(\boldsymbol{f})\|_2^2 = \|\boldsymbol{f}\|^2 \longrightarrow \Omega_{\mathcal{B}}(\boldsymbol{f}) = \sum_{\beta \in \mathcal{B}} \|M_\beta(\boldsymbol{f})\|_2^2 = |\mathcal{B}| \|\boldsymbol{f}\|^2$
- $l(a, b) = (a - b)^2$

$$\min_{\boldsymbol{f} \in \mathcal{H}} \sum_{n \in \mathbb{N}_N} \left(y_n - \langle \boldsymbol{k}_{x_n}, \boldsymbol{f} \rangle\right)^2 + \lambda \, |\mathcal{B}| \|\boldsymbol{f}\|^2$$

Main Inference Problem

$$\mathcal{D}_N = \left\{ \left(x_n^{(1)}, \ldots, x_n^{(Q)}, y_n \right) \in (\mathcal{X}_1 \times \cdots \times \mathcal{X}_Q) \times \mathcal{Y} : n \in \mathbb{N}_N \right\}$$

Penalized Empirical Risk Minimization Problem

$$\inf_{\boldsymbol{f} \in \boldsymbol{H}_{\mathcal{B}}} \sum_{n \in \mathbb{N}_N} l\left(y_n, \langle \boldsymbol{k}_{x_n}, \boldsymbol{f} \rangle\right) + \lambda\, \Omega_{\mathcal{B}}(\boldsymbol{f})$$

Special Case: Tensor Completion [Gandy et al., 2011; Signoretto et al., 2011]

- $\left(\mathcal{H}_q, \langle \cdot, \cdot \rangle_q, k^{(q)}\right)$ RKHS of functions on $\mathbb{I}_q = \{1, 2, \ldots, I_q\}$
- $\Omega_{\mathcal{B}}(\boldsymbol{f}) = \sum_{q \in \mathbb{N}_Q} \|M_{\{n\}}(\boldsymbol{f})\|_1$

$$\min_{\boldsymbol{f} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_Q}} \sum_{n \in \mathbb{N}_N} l\left(y_n, \boldsymbol{f}_{i_1^{(n)} i_2^{(n)} \ldots i_Q^{(n)}}\right) + \lambda \sum_{q \in \mathbb{N}_Q} \|M_{\{n\}}(\boldsymbol{f})\|_1$$

Main Inference Problem

$$\mathcal{D}_N = \left\{ \left(x_n^{(1)}, \ldots, x_n^{(Q)}, y_n \right) \in (\mathcal{X}_1 \times \cdots \times \mathcal{X}_Q) \times \mathcal{Y} : n \in \mathbb{N}_N \right\}$$

Penalized Empirical Risk Minimization Problem

$$\inf_{\boldsymbol{f} \in \boldsymbol{H}_{\mathcal{B}}} \sum_{n \in \mathbb{N}_N} l\left(y_n, \langle \boldsymbol{k}_{x_n}, \boldsymbol{f} \rangle\right) + \lambda \, \Omega_{\mathcal{B}}(\boldsymbol{f})$$

Special Case: Operator Estimation

- for $Q = 2$ (second order case) closely related to operator estimation.

J. Abernethy, F. Bach, T. Evgeniou and J. Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization, *The Journal of Machine Learning Research*, 10:803–826, 2009

Representer Theorem for arbitrary MLSPs

Let $\mathcal{B} = \{1, 2, 3\}$ and consider the problem:

$$\inf_{\boldsymbol{f} \in \boldsymbol{H}_{\mathcal{B}}} \mathcal{P}_\lambda(\boldsymbol{f}), \quad \mathcal{P}_\lambda(\boldsymbol{f}) := \sum_{n \in \mathbb{N}_N} l\left(y_n, \langle \boldsymbol{k}_{x_n}, \boldsymbol{f} \rangle\right) + \lambda\, \Omega_{\mathcal{B}}(\boldsymbol{f}) \ . \tag{1}$$

- If the set of solutions to (1) is non-empty, then a solution is :

$$\widehat{\boldsymbol{f}} = \sum_{i_1 \in \mathbb{I}_1} \sum_{i_2 \in \mathbb{I}_2} \sum_{i_3 \in \mathbb{I}_3} \widehat{\boldsymbol{\alpha}}_{i_1 i_2 i_3}\, u_{i_1}^{(1)} \otimes u_{i_2}^{(2)} \otimes u_{i_3}^{(3)} \quad (*)$$

$(*)$ $\left\{u_i^{(q)}\right\}$ orthobasis for $\operatorname{span}\left\{k^{(q)}\left(x^{(q)}, \cdot\right) \ : (x, y) \in \mathcal{D}_N\right\}, \ q \in \{1, 2, 3\}$

Representer Theorem for arbitrary MLSPs (cont'd)

- $\widehat{\boldsymbol{\alpha}}$ can be found solving the finite dimensional problem:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^{\mathbb{I}_1 \times \mathbb{I}_2 \times \mathbb{I}_3}} \left\{ \sum_{n \in \mathbb{N}_N} l\left(y_n, \boldsymbol{\alpha} \times_1 \boldsymbol{F}_{n\cdot}^{(1)} \times_2 \boldsymbol{F}_{n\cdot}^{(1)} \times_3 \boldsymbol{F}_{n\cdot}^{(3)}\right) + \lambda\, \Omega_{\mathcal{B}}\left(\boldsymbol{\alpha}\right) \right\}$$

  where $\boldsymbol{F}^{(q)} \in \mathbb{R}^{\mathbb{N}_N \times \mathbb{I}_q}$ satisfies:

$$\boldsymbol{K}^{(q)} = \boldsymbol{F}^{(q)} \boldsymbol{F}^{(q)\top} \text{ and } \boldsymbol{K}_{ij}^{(q)} := k^{(q)}\left(x_i^{(q)}, x_j^{(q)}\right)$$

- $\widehat{\boldsymbol{g}}$ can be equivalently stated in terms of the kernel-based expansion:

$$\sum_{n_1 \in \mathbb{N}_N} \sum_{n_2 \in \mathbb{N}_N} \sum_{n_2 \in \mathbb{N}_N} \boldsymbol{\gamma}_{n_1 n_2 n_3} \boldsymbol{k}^{(1)}\left(x_{n_1}^{(1)}, .^{(1)}\right) \otimes \boldsymbol{k}^{(2)}\left(x_{n_2}^{(2)}, .^{(2)}\right) \otimes \boldsymbol{k}^{(3)}\left(x_{n_3}^{(3)}, .^{(3)}\right)$$

Computational Approach

$$\Omega_{\mathcal{B}}(\boldsymbol{f}) = \begin{cases} \sum_{q\in\mathbb{N}_3} \|M_q(\boldsymbol{f})\|_1, & \text{if } \operatorname{rank}(M_q(\boldsymbol{f})) \leq R_q \forall q \in \mathbb{N}_3 \\ \infty, & \text{otherwise .} \end{cases}$$

Finite Dimensional Optimization Problem

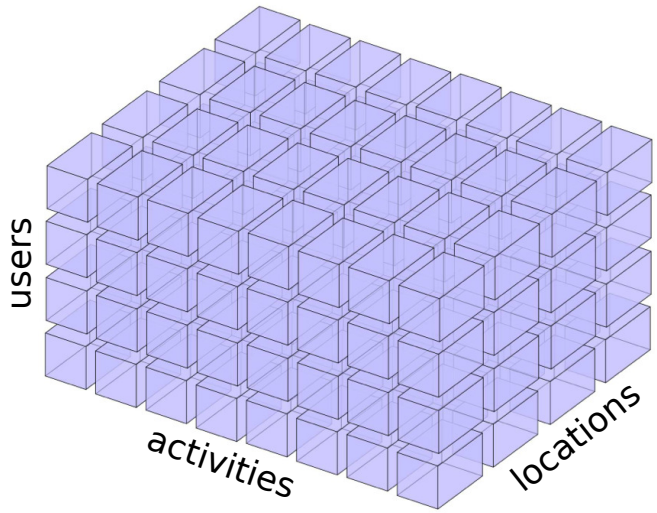
$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \tfrac{1}{2}\sum_{n=1}^{N}\left(y_n - \boldsymbol{\alpha} \times_1 \boldsymbol{F}^{(1)}_{n\cdot} \times_2 \cdots \times_3 \boldsymbol{F}^{(3)}_{n\cdot}\right)^2 + \tfrac{\lambda}{3}\sum_{q=1}^{3}\|M_q(\boldsymbol{\alpha})\|_1 \\ \text{s. t.} \quad & \operatorname{rank}(M_q(\boldsymbol{\alpha})) \leq R_q, \ \ q \in \mathbb{N}_3 \end{aligned}$$

- Tucker parametrization: $\boldsymbol{\alpha} = \boldsymbol{\beta} \times_1 \boldsymbol{U}^{(1)} \times_2 \boldsymbol{U}^{(2)} \times_3 \boldsymbol{U}^{(3)}$
- Turned into a smooth problem in $\left\{\boldsymbol{\beta}, \boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}\right\}$
- Solved by block-coordinate descent (nonlinear Gauss-Seidel)

Toy Problem

Experimental Set-up

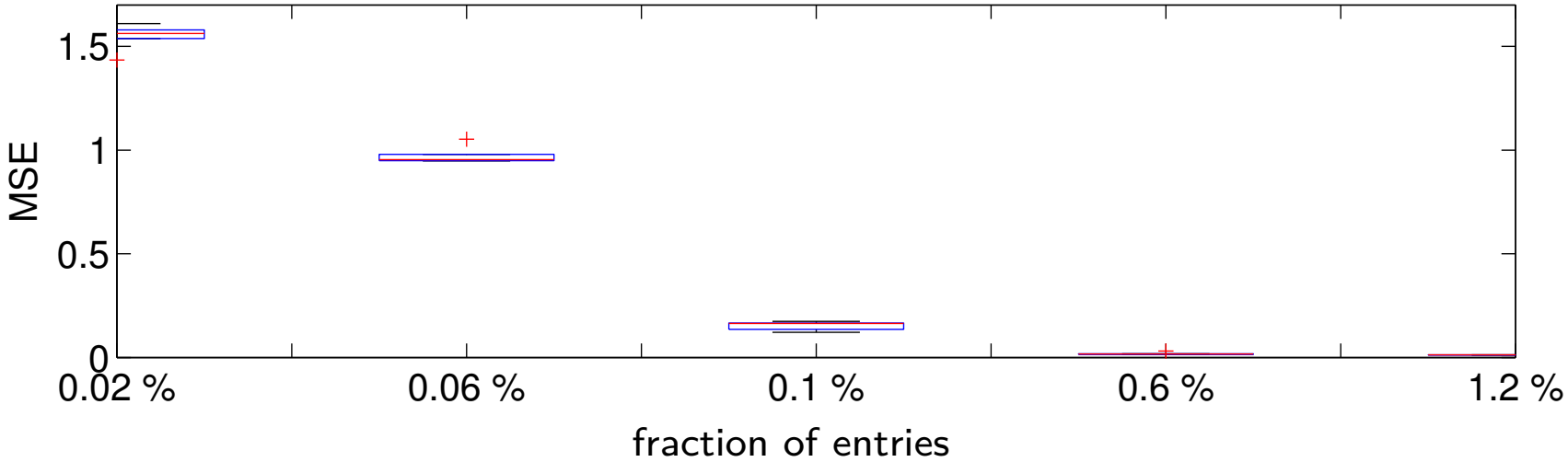
- $50 \times 50 \times 50$ (125k entries)
- 200 dimensional per-mode features
- generating function with mlrank$= (2, 2, 2)$



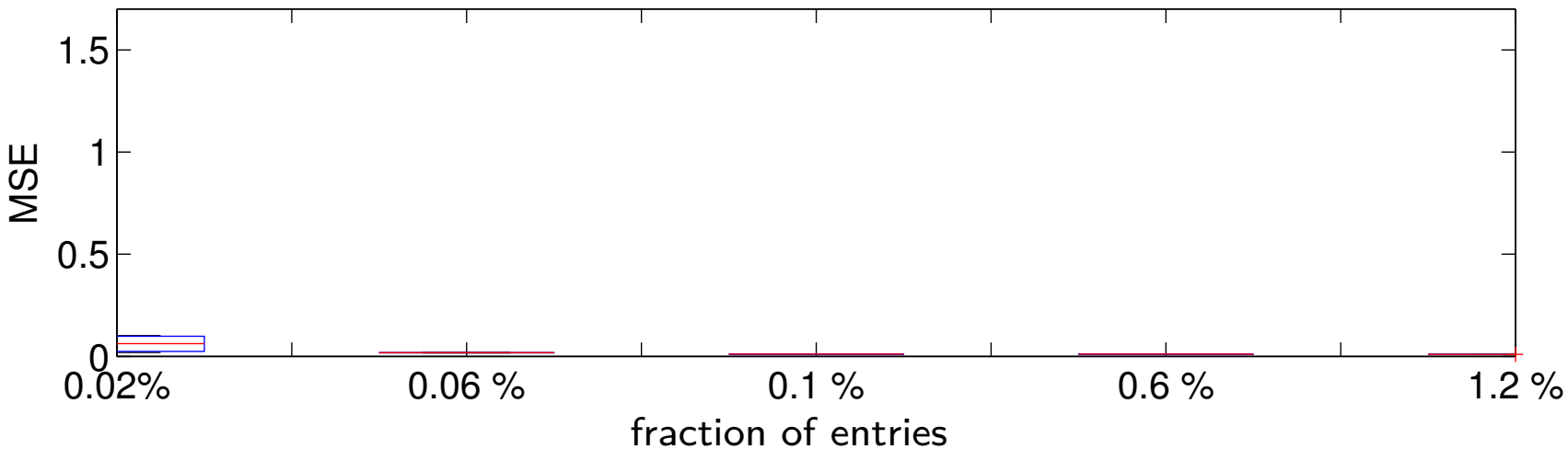
- compute models based upon increasing fraction of entries
- compared against LS-SVM (standard kernel-based approach based on quadratic regularization)

Toy Problem

LS-SVM (RBF kernel)



multilinear spectral regularization



Health Satisfaction Index (HSI)

Survey on German health care system; 7293 patients - 22835 observations

Data per patient

sex

age

schooling

...

doctor visits

hospitalizations

privately insured

...

health satisfaction in t_1

health satisfaction in t_2

...

Health Satisfaction Index (HSI)

Survey on German health care system; 7293 patients - 22835 observations

Data per patient

sex age schooling ...	anagraphical
# doctor visits # hospitalizations privately insured ...	health condition
health satisfaction in t_1 health satisfaction in t_2 ...	time dependent

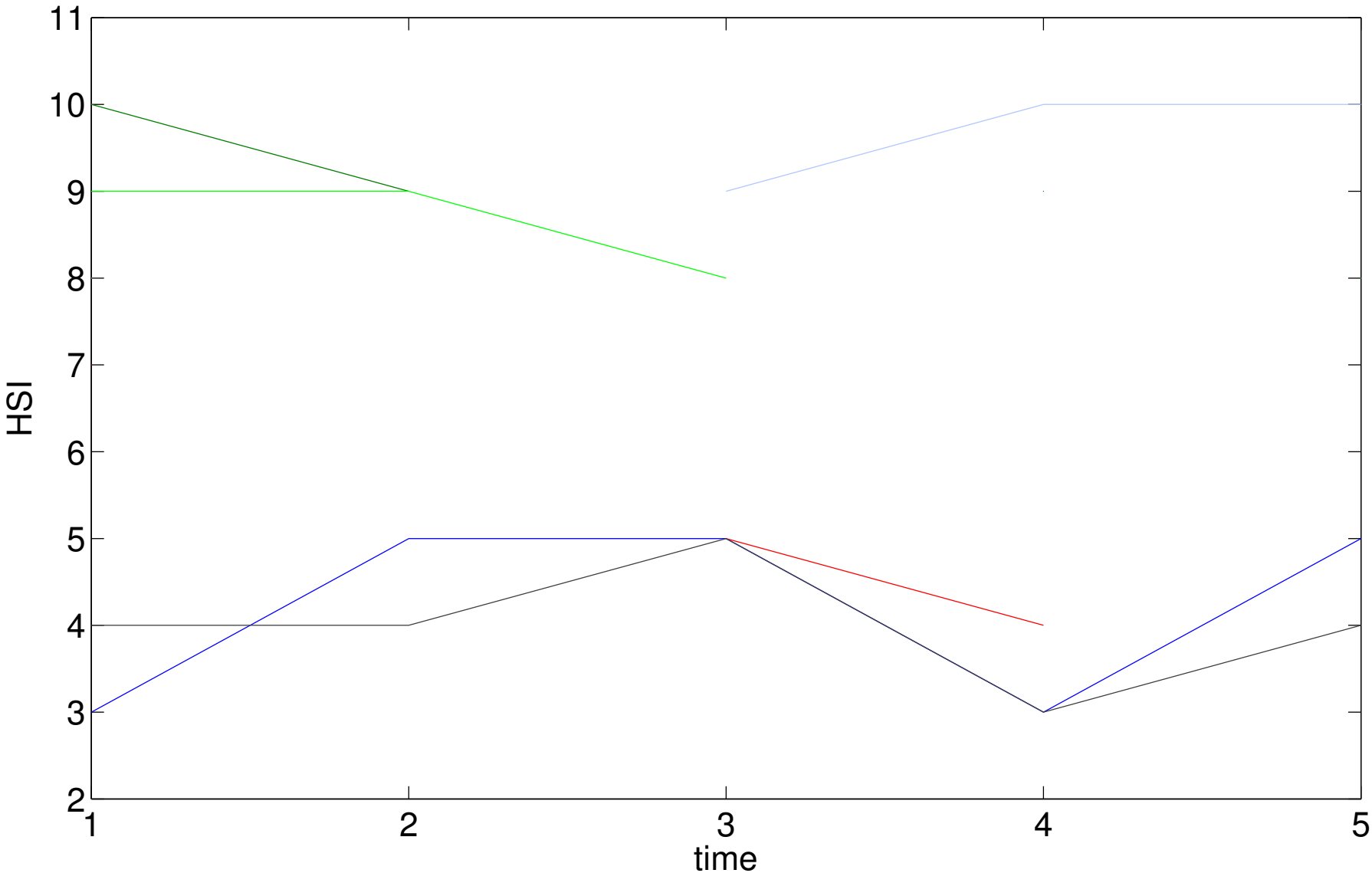
Health Satisfaction Index (HSI)

Survey on German health care system; 7293 patients - 22835 observations

Data per patient

sex	patient features
age	
schooling	
...	
# doctor visits	
# hospitalizations	
privately insured	
...	
health satisfaction in t_1	time dependent
health satisfaction in t_2	
...	

Health Satisfaction Index (HSI)



Health Satisfaction Index (HSI)

Survey on German health care system; 7293 patients - 22835 observations

Estimation of prospect HSI (evolution in time)

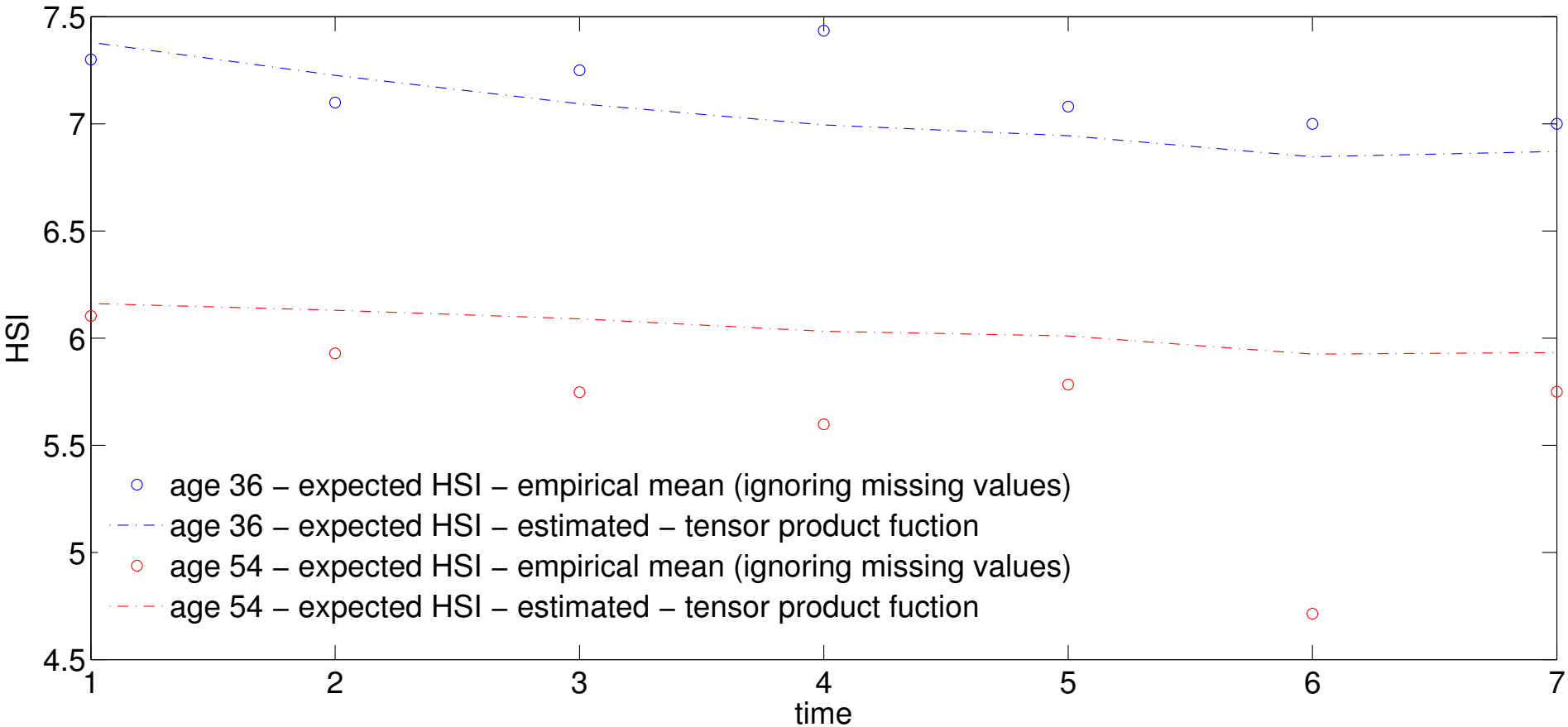
- time mode: spline kernel $k(t_1, t_2) = 1 + \min\{t_1, t_2\}$
- anagraphical/health condition (patient features): linear kernel
- third order outperforms second order approach (25% test - MSE):

III order	II order
4.30(0.01)	4.44(0.01)

Health Satisfaction Index (HSI)

Survey on German health care system; 7293 patients - 22835 observations

Model Based Evaluation (age comparison)



Conclusions

To Wrap Up

- Learning with tensors
- Transductive and inductive problems with structured data
- Inference problem in RKHSs
- Representer theorem and computational approach

Open Problems/Future Directions

- Novel application(s)?
- Efficient algorithms for learning kernel-based models via multilinear spectral penalties

Selected References

- M. Signoretto, Q. Tran Dinh, L. De Lathauwer, and J. A. K. Suykens. Learning with Tensors: a Framework Based on Convex Optimization and Spectral Regularization. *Machine Learning*, 2013.
- M. Signoretto, R. Van de Plas, B. De Moor, and J. A. K. Suykens. Tensor versus matrix completion: a comparison with application to spectral data. *IEEE Signal Processing Letters, 18(7):403–407, 2011.*
- M. Signoretto, E. Olivetti, L. De Lathauwer and J. A. K. Suykens. Classification of multichannel signals with cumulant-based kernels. *IEEE Transactions on Signal Processing*, 60(5), 2304–2314, 2012.
- M. Signoretto, L. De Lathauwer and J. Suykens Nuclear Norms for Tensors and Their Use for Convex Multilinear Estimation. Internal Report 10-186, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2011
- B. Hunyadi, M. Signoretto, S. Debener, S. Van Huffel, M. De Vos, Classification of structured EEG tensors using nuclear norm regularization: improving P300 classification , *Workshop on Pattern Recognition in NeuroImaging* (PRNI) 2013
- B. Hunyadi, M. Signoretto, W. Van Paesschen, J. Suykens, S. Van Huffel, and M. De Vos, Incorporating structural information from the multichannel EEG improves patient-specific seizure detection, *Clinical Neurophysiology*, 2012