Two-way Regularized Matrix Decomposition

Jianhua Huang

Texas A&M University

SVD and regularization

Examples of two-way regularized SVD

Scale-invariance in two formulations of regularized SVD

Singular value decomposition

- SVD: $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$
- $\mathbf{X}$ $(n \times p)$
- $\mathbf{U}$ $(n \times m)$, $\mathbf{D}$ $(m \times m)$, $\mathbf{V}$ $(p \times m)$, $m = \min(n, p)$
- truncated SVD: $\mathbf{X} = \mathbf{U}_k\mathbf{D}_k\mathbf{V}_k^T$, $k \ll m$
  $k = 1$: $\mathbf{X} = d\mathbf{u}\mathbf{v}^T$
- Eckart-Young theorem
  $\min \|\mathbf{X} - \widehat{\mathbf{X}}\|^2$ subject to rank constraint to $\widehat{\mathbf{X}}$

One-way regularized SVD

- $(\mathbf{u}_1, \mathbf{v}_1) = arg \min_{\mathbf{u},\mathbf{v}} ||\mathbf{X} - \mathbf{u}\mathbf{v}^T||_F^2 + \lambda \mathcal{P}(\mathbf{v})$
- functional PCA
  using roughness penalty

$$\mathbf{v}^T \boldsymbol{\Omega} \mathbf{v} = \sum_{i=2}^{n-1} \{v_{i-1} - 2v_i + v_{i+1}\}^2$$

- sparse PCA
  using sparsity-inducing penalty

$$|\mathbf{v}| = \sum_{i=1}^{n} |v_i|$$

Two-way structured data

- two-way functional data:
  - row and column domains are structured
  - mortality rate as a function of time and age
- functional-sparse structured data, e.g., fMRI data:
  - row from temporal space, change continuously with time - smooth
  - column from spatial space, active region only a small proportion - sparse
- checkerboard structure data:
  biclustering problem

Regularized SVD

- Standard SVD

$$(\mathbf{u}_1, \mathbf{v}_1) = arg \min_{\mathbf{u},\mathbf{v}} ||\mathbf{X} - \mathbf{u}\mathbf{v}^T||_F^2$$

- Regularized SVD!

$$(\mathbf{u}_1, \mathbf{v}_1) = arg \min_{\mathbf{u},\mathbf{v}} ||\mathbf{X} - \mathbf{u}\mathbf{v}^T||_F^2 + \mathcal{P}(\mathbf{u}, \mathbf{v})$$

- squared-error loss can be replaced
- How do we choose $\mathcal{P}(\mathbf{u}, \mathbf{v})$?
- Other formulations use constrained optimization: Allen, Witten, etc.

SVD and regularization

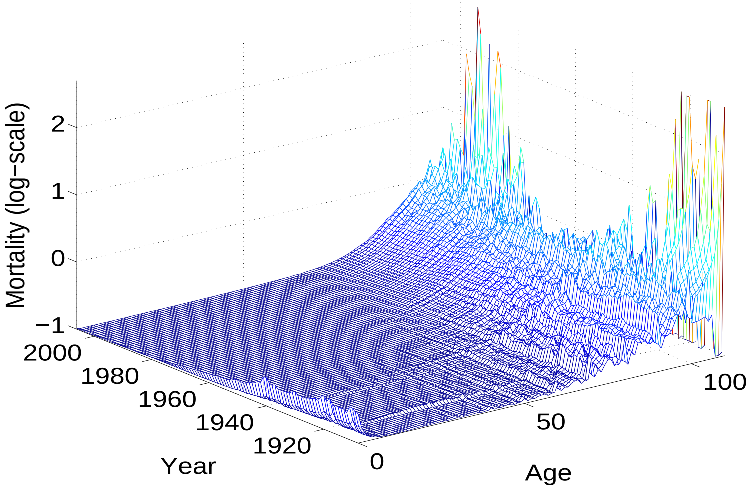
Examples of two-way regularized SVD

Scale-invariance in two formulations of regularized SVD

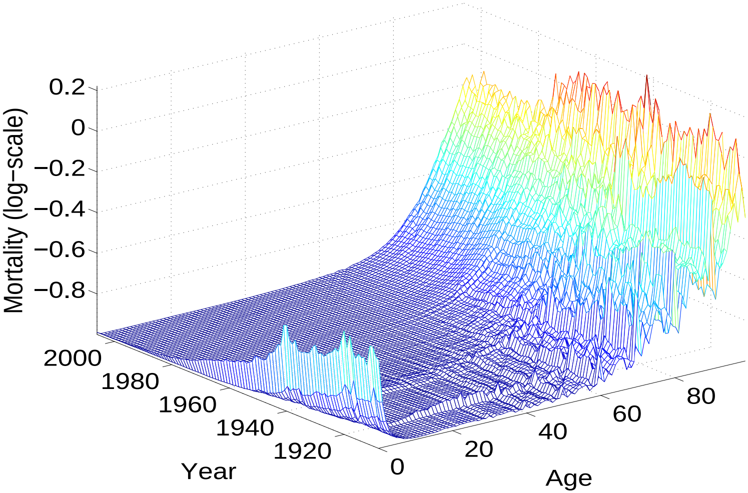
Spanish mortality rate

- available in the Human Mortality Database
- each row: a year between 1908 and 2007
- each column: an age group from 0 to 110
- each cell: the mortality rate for a particular age group during that year
- two-way functional structured
- $\log(x + 1/2)$

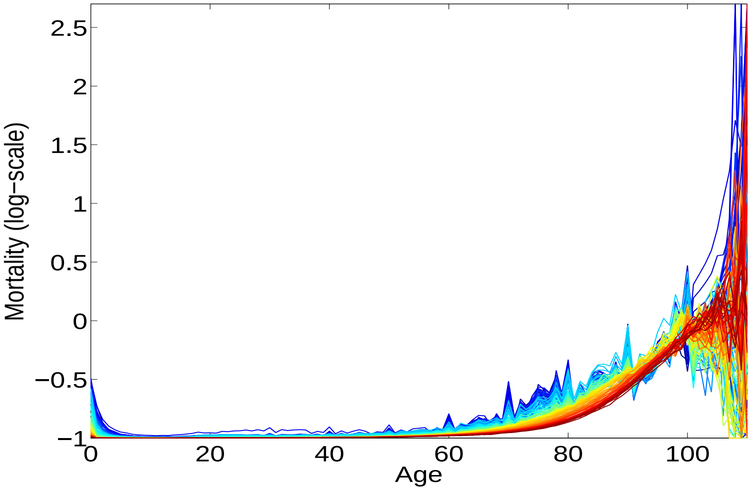
3-d view of the data



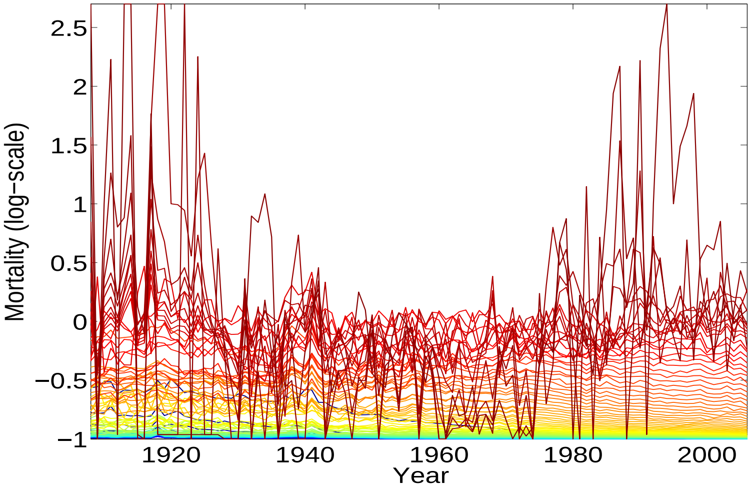
3-d view of the data (zoomed)



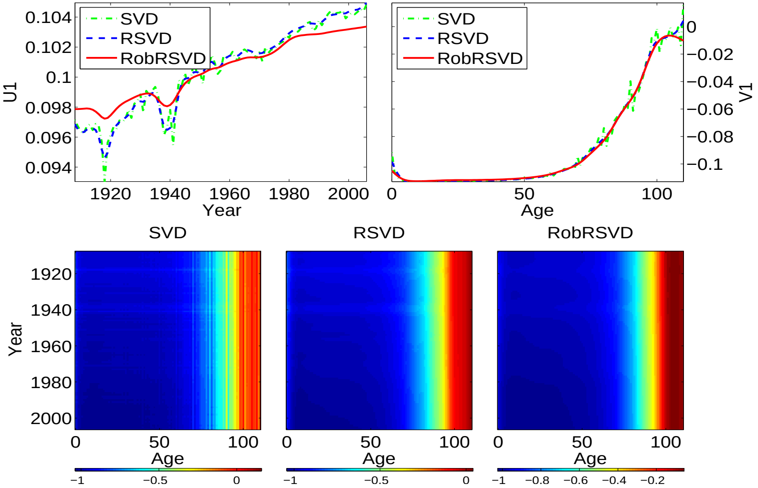
Age plot of the data



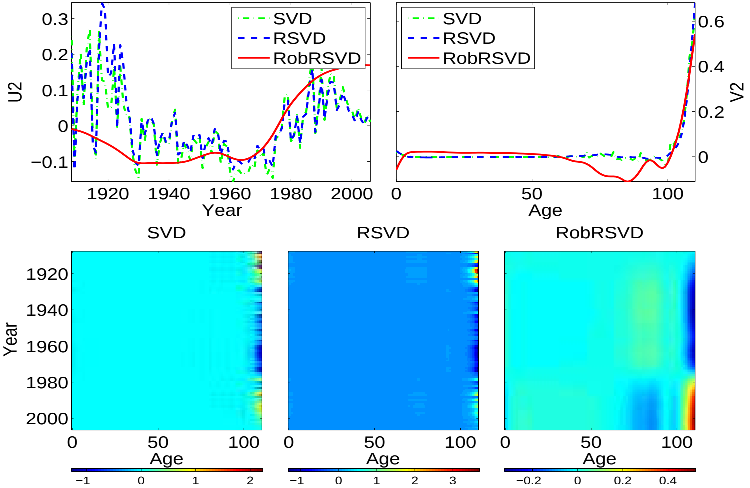
Year plot of the data



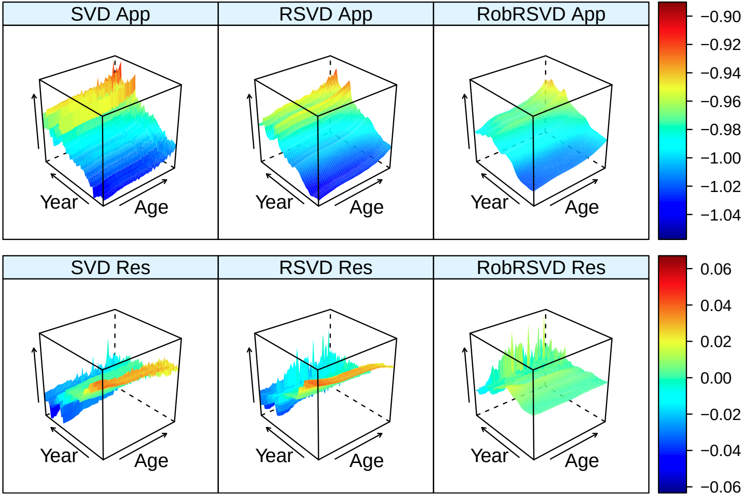
First component of SVD



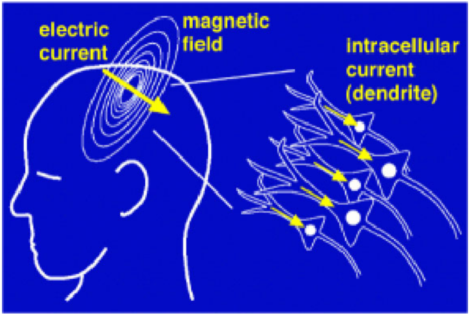
Second component of SVD



Fitted and residual plot of the rank-2 model



Inverse problem of MEG source reconstruction



Imaging methods

- $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$
- $\mathbf{Y} \in \mathbf{R}^{n \times s}$: measured MEG data ($n$ sensors s time points).
- $\mathbf{B} \in \mathbf{R}^{p \times s}$: the potential source time courses in the cortical area (p source components, $p \gg n$).
- $\mathbf{X} \in \mathbf{R}^{n \times p}$: forward operator
  can be derived using a head model
- $\mathbf{E} \in \mathbf{R}^{n \times s}$: noise
- Goal: solving for **B**—ill-posed

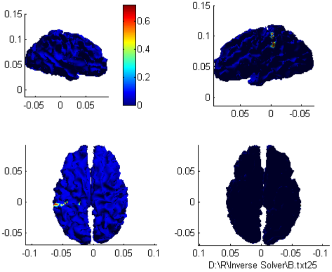
Two-way regularization

- $\mathbf{B} = \mathbf{A}\mathbf{G}^T \quad p \times s$
- $\mathbf{G} \in \mathbf{R}^{s \times q}$ contains the temporal features
- $\mathbf{A} \in \mathbf{R}^{p \times q}$ captures the spatial signals
- $q \leq s$

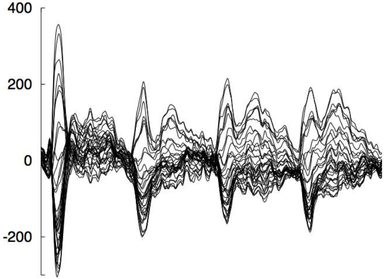
Penalized least squares problem

$$\min_{\mathbf{a},\mathbf{G}} \left\{ \|\mathbf{Y} - \mathbf{X}\mathbf{A}\mathbf{G}^T\|_F^2 + \mathcal{P}(\mathbf{A}, \mathbf{G}) \right\}$$

Desired properties

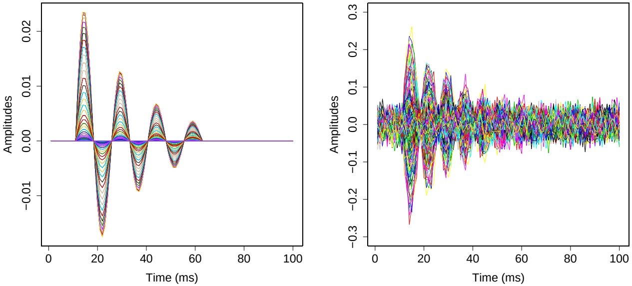


(a) Spatial focality



(b) Temporal smoothness

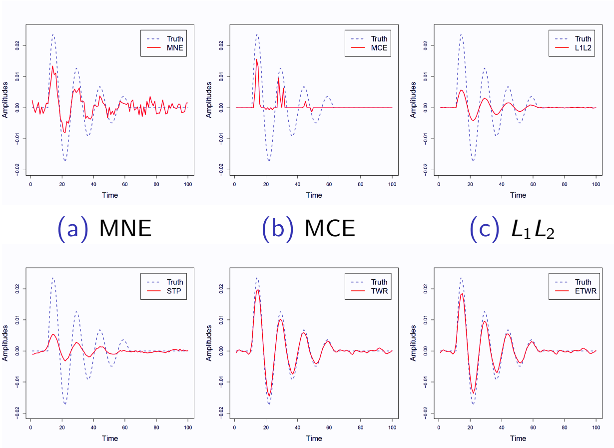
Synthetic example



(a) Simulated source time courses

(b) Simulated sensor signals

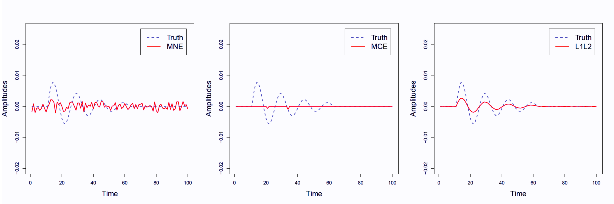
Figure: (a) simulated source time course using a sine-exponential function; (b) synthetical sensor time courses (SNR=6dB).



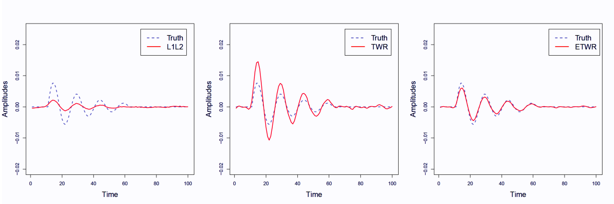
(a) MNE (b) MCE (c) $L_1 L_2$

(d) STP (e) TWR (f) ETWR

Figure: Reconstructed time courses by different methods at the center of the active area.



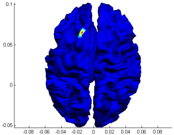
(a) MNE (b) MCE (c) L_1L_2



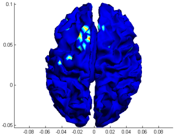
(d) STP (e) TWR (f) ETWR

Figure: Reconstructed time courses by different methods at an arbitrary location near the edge of the active area (SNR=6dB).

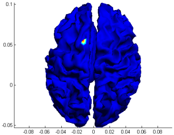
Synthetic example



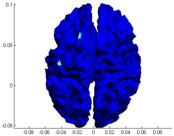
(a) Truth



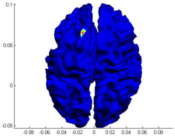
(b) MNE



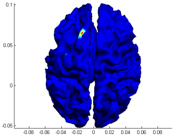
(c) MCE



(d) L_1L_2



(e) TWR



(f) ETWR

Figure: Overviews of brain mapping by different methods at 14 ms (SNR=6dB).

SVD and regularization

Examples of two-way regularized SVD

Scale-invariance in two formulations of regularized SVD

Two formulations of regularized SVD

$\min_{\mathbf{u},\mathbf{v}} ||\mathbf{X} - \mathbf{u}\mathbf{v}^T||_F^2 + \mathcal{P}(\mathbf{u}, \mathbf{v})$

- Huang, Shen and Buja (2009, JASA)

$$\mathcal{P}_1(\mathbf{u}, \mathbf{v}) = \lambda_u \mathcal{P}_\mathbf{u}(\mathbf{u}) \cdot \mathbf{v}\mathbf{v}^T + \lambda_v \mathcal{P}_\mathbf{v}(\mathbf{v}) \cdot \mathbf{u}\mathbf{u}^T + \lambda_\mathbf{u}\lambda_\mathbf{v}\mathcal{P}_\mathbf{u}(\mathbf{u})\mathcal{P}_\mathbf{v}(\mathbf{v})$$

  - scale invariant

$$\mathcal{P}_1(c \cdot \mathbf{u}, \mathbf{v}/c) = \mathcal{P}_1(\mathbf{u}, \mathbf{v}), \ \forall c \neq 0$$

- Hong and Lian (2013, JMVA)

$$\mathcal{P}_2(\mathbf{u}, \mathbf{v}) = \lambda_u \mathcal{P}_\mathbf{u}(\mathbf{u}) + \lambda_v \mathcal{P}_\mathbf{v}(\mathbf{v})$$

  - not scale-invariant

"Advantages of ignoring scale invariance"

$\mathcal{P}_2(\mathbf{u}, \mathbf{v}) = \lambda_u \mathcal{P}_{\mathbf{u}}(\mathbf{u}) + \lambda_v \mathcal{P}_{\mathbf{v}}(\mathbf{v})$

- adjust the tuning parameters for varying scale
- scale-shift between **u** and **v**, only need one effective tuning parameter
- set $\lambda_{\mathbf{v}} = 1$, only $\lambda_{\mathbf{u}}$ to be tuned
- reduce computation for tuning parameter selection

Do we lose anything?

Smooth-smooth problem

- Huang, Shen and Buja (2009):

$$-2\mathbf{u}^T\mathbf{X}\mathbf{v} + \mathbf{u}^T(\mathbf{I} + \lambda_{\mathbf{u}}\boldsymbol{\Omega})\mathbf{u} \cdot \mathbf{v}^T(\mathbf{I} + \lambda_{\mathbf{v}}\boldsymbol{\Omega})\mathbf{v}$$

- Hong and Lian (2013):

$$-2\mathbf{u}^T\mathbf{X}\mathbf{v} + \mathbf{u}^T\mathbf{u} \cdot \mathbf{v}^T\mathbf{v} + \lambda_{\mathbf{u}}\mathbf{u}^T\boldsymbol{\Omega}\mathbf{u} + \lambda_{\mathbf{v}}\mathbf{v}^T\boldsymbol{\Omega}\mathbf{v}$$

Stationary equations

- Huang, Shen and Buja (2009):

$$\mathbf{u} = \frac{1}{\sqrt{\mathbf{v}^T(\mathbf{I} + \lambda_{\mathbf{v}}\mathbf{\Omega}_{\mathbf{v}})\mathbf{v}}} \cdot (\mathbf{I} + \lambda_{\mathbf{u}}\mathbf{\Omega}_{\mathbf{u}})^{-1} \frac{\mathbf{X}\mathbf{v}}{\sqrt{\mathbf{v}^T(\mathbf{I} + \lambda_{\mathbf{v}}\mathbf{\Omega}_{\mathbf{v}})\mathbf{v}}}$$

$$\mathbf{v} = \frac{1}{\sqrt{\mathbf{u}^T(\mathbf{I} + \lambda_{\mathbf{u}}\mathbf{\Omega}_{\mathbf{u}})\mathbf{u}}} \cdot (\mathbf{I} + \lambda_{\mathbf{v}}\mathbf{\Omega}_{\mathbf{v}})^{-1} \frac{\mathbf{X}^T\mathbf{u}}{\sqrt{\mathbf{u}^T(\mathbf{I} + \lambda_{\mathbf{u}}\mathbf{\Omega}_{\mathbf{u}})\mathbf{u}}}$$

- Hong and Lian (2013):

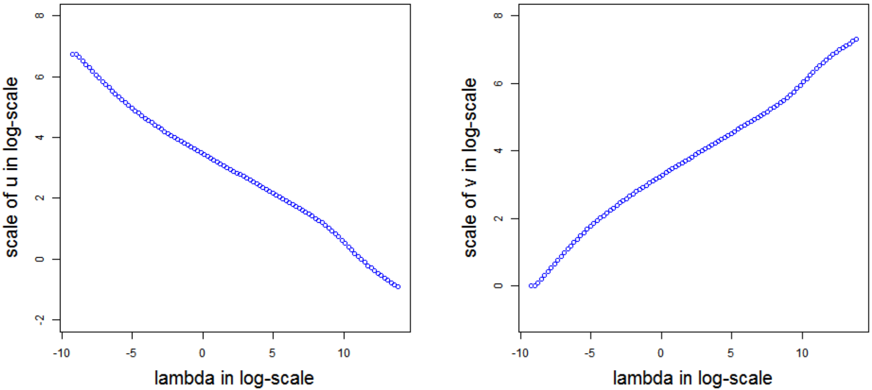
$$\mathbf{u} = \frac{1}{\sqrt{\mathbf{v}^T\mathbf{v}}} \cdot (\mathbf{I} + \frac{\lambda_{\mathbf{u}}}{\mathbf{v}^T\mathbf{v}}\mathbf{\Omega}_{\mathbf{u}})^{-1} \frac{\mathbf{X}\mathbf{v}}{\sqrt{\mathbf{v}^T\mathbf{v}}}$$

$$\mathbf{v} = \frac{1}{\sqrt{\mathbf{u}^T\mathbf{u}}} \cdot (\mathbf{I} + \frac{\lambda_{\mathbf{v}}}{\mathbf{u}^T\mathbf{u}}\mathbf{\Omega}_{\mathbf{v}})^{-1} \frac{\mathbf{X}^T\mathbf{u}}{\sqrt{\mathbf{u}^T\mathbf{u}}}$$

Confounding of scale and penalty parameter

- actual penalty parameters:
  - $\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}}$ (Huang, Shen and Buja 2009)
  - $\frac{\lambda_{\mathbf{u}}}{\mathbf{v}^T\mathbf{v}}, \frac{\lambda_{\mathbf{v}}}{\mathbf{u}^T\mathbf{u}}$ (Hong and Lian 2013)
- penalty parameters $(\lambda_{\mathbf{u}}, \lambda_{\mathbf{v}})$ and scales $(\mathbf{u}^T\mathbf{u}, \mathbf{v}^T\mathbf{v})$ are confounded in Hong and Lian (2013)
- no confounding in Huang, Shen and Buja (2009)

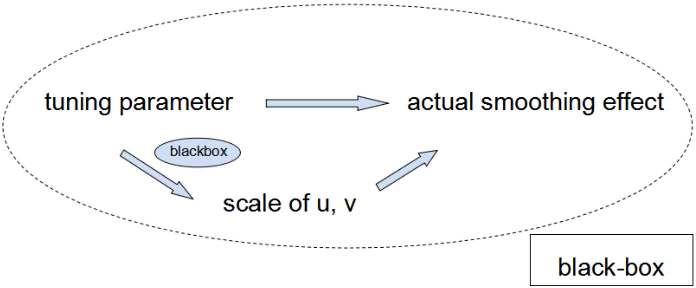
Scale at convergence as a function of penalty parameter



1st consequence: difficulty in defining optimal tuning

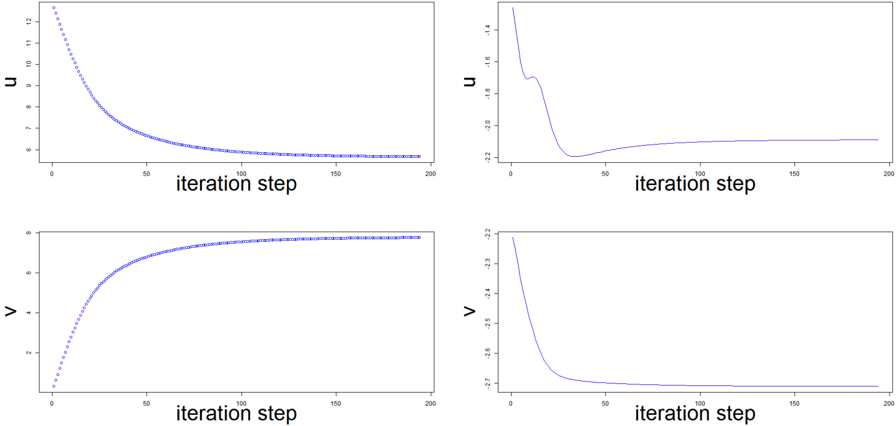
- Huang, Shen and Buja (2009):

- Hong and Lian (2013):



Scale and roughness as function of # of iterations

(log) path of scales | (log) path of roughnesses

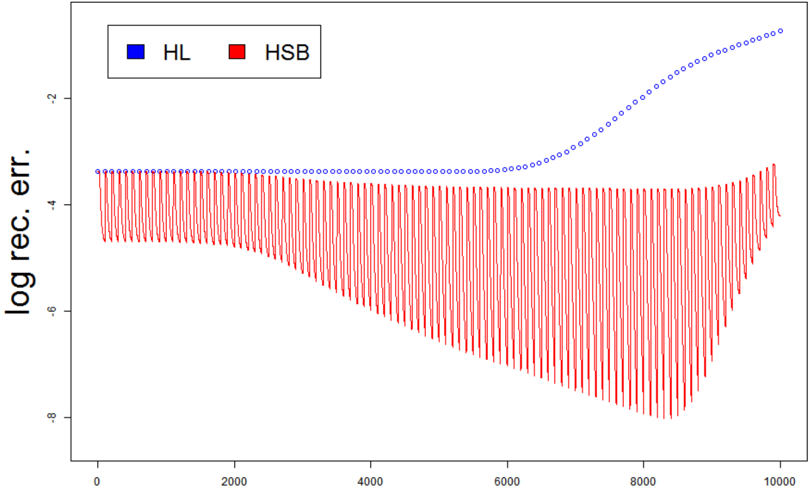


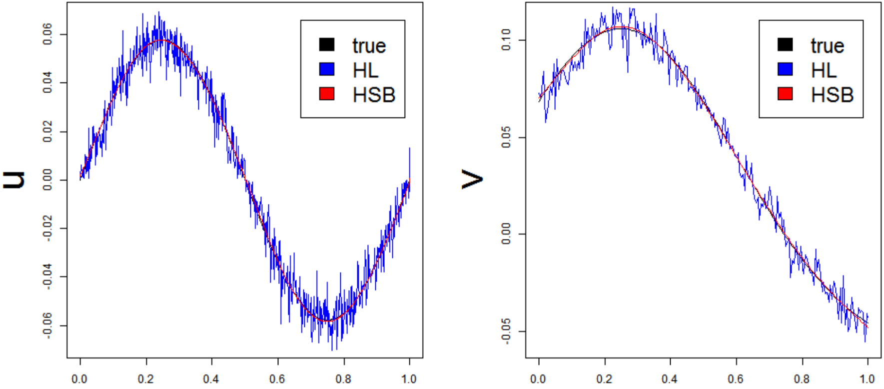
2nd consequence: redundant Iterations

- signal is being processed at appropriate level of smoothness, only when scale is adjusted to the right level
- most of iteration steps used to adjust the scale, not smoothness
- according to simulation, scale-adjustment uses 75% of steps
- result in much more steps to convergence than Huang, Shen and Buja (2009)
- # of iterations (HL: 100 para., HSB: 100 $\times$ 100 para.)

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
HL	16.00	112.00	175.50	550.80	939.00	1853.00
HSB	5.00	7.00	10.00	9.54	12.00	14.00

3rd consequence: bad recovery of signals




u
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
0.0
0.2
0.4
0.6
0.8
1.0
true
HL
HSB
v
0.10
0.05
0.00
-0.05
0.0
0.2
0.4
0.6
0.8
1.0
true
HL
HSB

Sparse-smooth problem: stationary equations

- stationary equations:

$$\mathbf{u} = \frac{1}{\sqrt{\mathbf{v}^T\mathbf{v}}} \cdot \mathbf{sparse}\left(\frac{\mathbf{Xv}}{\sqrt{\mathbf{v}^T\mathbf{v}}}; \frac{\lambda_{\mathbf{u}}}{\sqrt{\mathbf{v}^T\mathbf{v}}}\right)$$

$$\mathbf{v} = \frac{1}{\sqrt{\mathbf{u}^T\mathbf{u}}} \cdot \left(\mathbf{I} + \frac{\lambda_{\mathbf{v}}}{\mathbf{u}^T\mathbf{u}}\boldsymbol{\Omega}\right)^{-1} \cdot \frac{\mathbf{Xv}}{\sqrt{\mathbf{v}^T\mathbf{v}}}$$

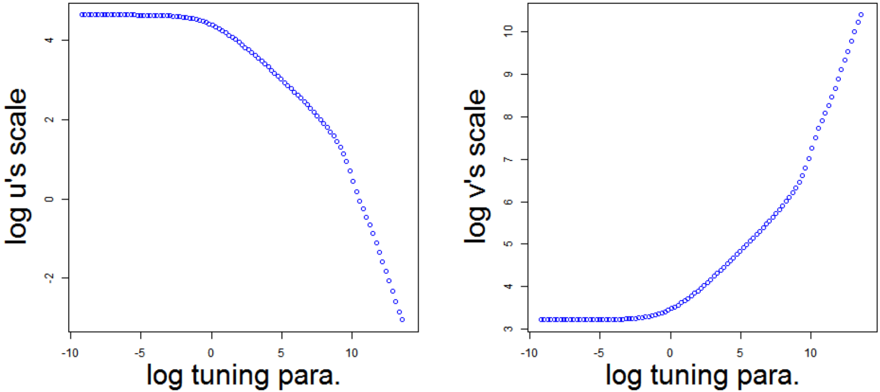
- $\mathbf{sparse}(\mathbf{y}; \lambda)$ is solution of

$$\min_{\mathbf{x}} ||\mathbf{y} - \mathbf{x}||_2^2 + \lambda||\mathbf{x}||_1$$

- still confounding of scale and penalty parameter

1st consequence: difficulty in defining the optimal tuning

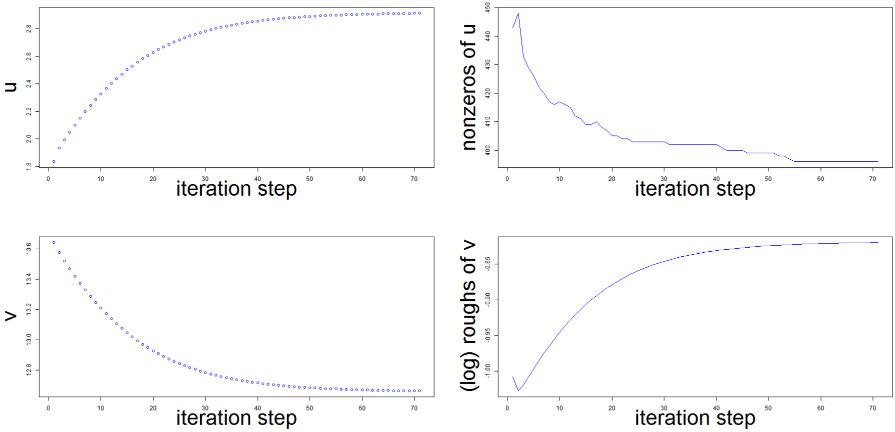
scales at converging for given λ



2nd consequence: redundant Iterations

(log) path of scales

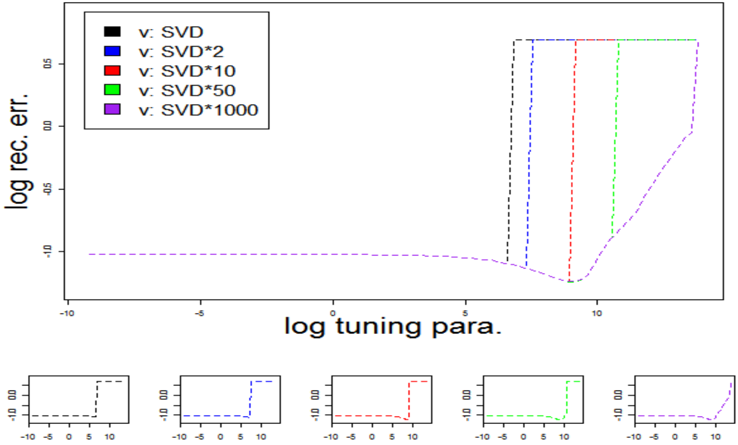
path of sparsities/roughnesses



3rd consequence: "threshold-to-zero"

- $\mathbf{sparse}(\mathbf{y}; \lambda) = \textit{arg}\min_{\mathbf{x}} ||\mathbf{y} - \mathbf{x}||_2^2 + \lambda||\mathbf{x}||_1$
- $\mathbf{sparse}(\mathbf{y}; \lambda) = \mathbf{0}$, if λ is too large
- $\mathbf{u} = \frac{1}{\sqrt{\mathbf{v}^T\mathbf{v}}} \cdot \mathbf{sparse}(\frac{\mathbf{X}\mathbf{v}}{\sqrt{\mathbf{v}^T\mathbf{v}}}; \frac{\lambda_{\mathbf{u}}}{\sqrt{\mathbf{v}^T\mathbf{v}}})$
- if starting with wrong scale before convergence, threshold **u** all into zero

Solution-path given initialization with different scales



Sparse-sparse problem: stationary equations

- stationary equations:

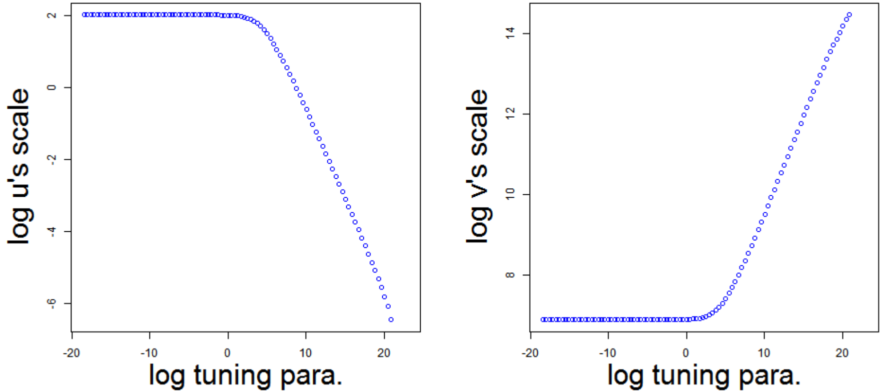
$$\mathbf{u} = \frac{1}{\sqrt{\mathbf{v}^T\mathbf{v}}} \cdot \textbf{sparse}\left(\frac{\mathbf{X}\mathbf{v}}{\sqrt{\mathbf{v}^T\mathbf{v}}}; \frac{\lambda_{\mathbf{u}}}{\sqrt{\mathbf{v}^T\mathbf{v}}}\right)$$

$$\mathbf{v} = \frac{1}{\sqrt{\mathbf{u}^T\mathbf{u}}} \cdot \textbf{sparse}\left(\frac{\mathbf{X}^T\mathbf{u}}{\sqrt{\mathbf{u}^T\mathbf{u}}}; \frac{\lambda_{\mathbf{v}}}{\sqrt{\mathbf{u}^T\mathbf{u}}}\right)$$

- still confounding of scale and penalty parameter

1st consequence: difficulty in defining optimal tuning

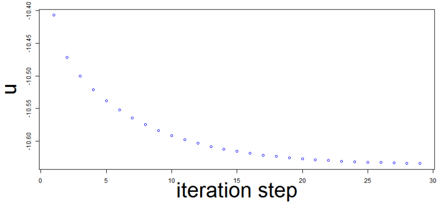
scales at convergence given different λ

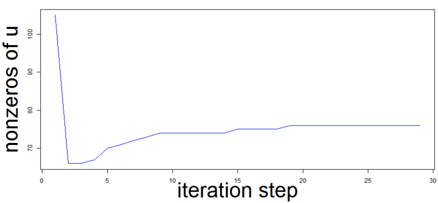


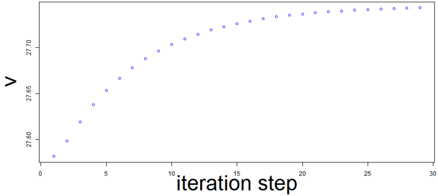
2nd consequence: redundant Iterations

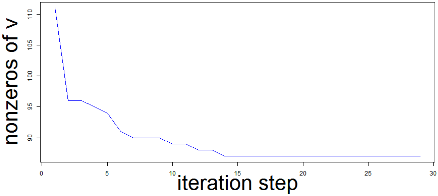
(log) path of scales

path of sparsities





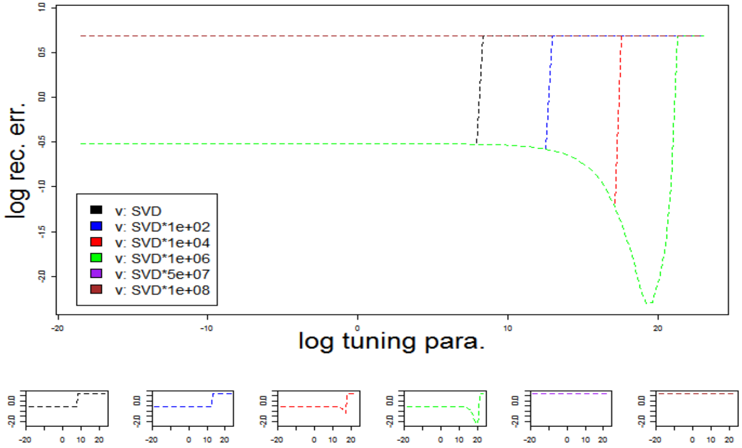




3rd consequence: two-sided "threshold-to-zero"

- $\mathbf{u} = \frac{1}{\sqrt{\mathbf{v}^T\mathbf{v}}} \cdot \textbf{sparse}(\frac{\mathbf{X}\mathbf{v}}{\sqrt{\mathbf{v}^T\mathbf{v}}}; \frac{\lambda_\mathbf{u}}{\sqrt{\mathbf{v}^T\mathbf{v}}})$
- $\mathbf{v} = \frac{1}{\sqrt{\mathbf{u}^T\mathbf{u}}} \cdot \textbf{sparse}(\frac{\mathbf{X}^T\mathbf{u}}{\sqrt{\mathbf{u}^T\mathbf{u}}}; \frac{\lambda_\mathbf{v}}{\sqrt{\mathbf{u}^T\mathbf{u}}})$
- "two-sided":
  - if initial **v** too small, **u** is thresholded to zero
  - if initial **v** too large, **v** is thresholded to zero
- sensitivity to initialization

Solution-path given initialization with different scales



Summary

- Matrix decomposition has wide application.
- Scale-invariance is important in the design of two-way regularization penalty.
- Consequence of ignoring scale-invariance:
  - confunding of scale and penalty parameter
  - # of iterations of the algorithm
  - non-flexibility of using single penalty parameter
  - threshold-all-to-zero problem

Acknowledgement

- Collaborators:
  - Andreas Buja (U Penn)
  - Xin Gao (KAUST)
  - Jianhua Hu (MD Anderson)
  - Seokho Lee (Hankuk University of Foreign Studies, Korea)
  - Mehdi Maadooliat (Marquette University)
  - Haipeng Shen (UNC)
  - Siva Tian (U Houston)
  - Senmao Liu, Lan Zhou (Taxas A&M)
  - Lingsong Zhang (Purdue)
- Grants
  - National Science Foundation
  - King Abdullah University of Science and Technology (KAUST), Saudi Arabia