Regularized Sparse Kernel Slow Feature Analysis

Wendelin Böhmer Steffen Grünewälder
Hannes Nickisch Klaus Obermayer

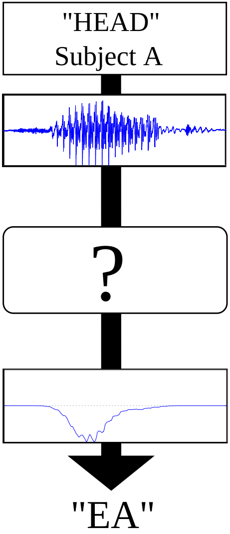
contact: WENDELIN@CS.TU-BERLIN.DE

Neural Information Processing Group
Technische Universität Berlin

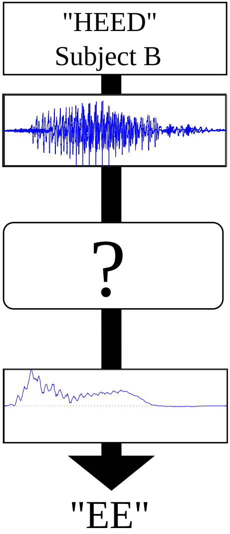
2011/09/08

Bernstein Focus: Neurotechnology Berlin

Linear solutions to non-linear problems

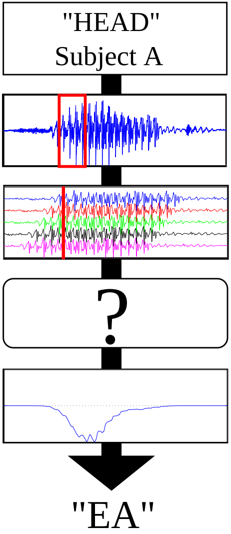


- Classification or regression w.r.t. latent variables Θ
- Example: vowel classification[1]



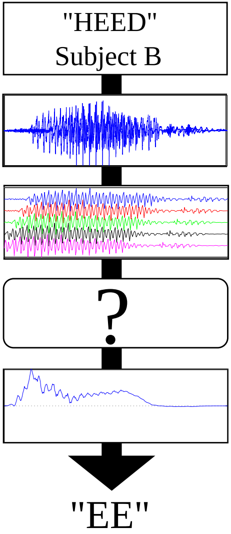
[1] North Texas vowel database (Assmann et al., 2008)

Linear solutions to non-linear problems



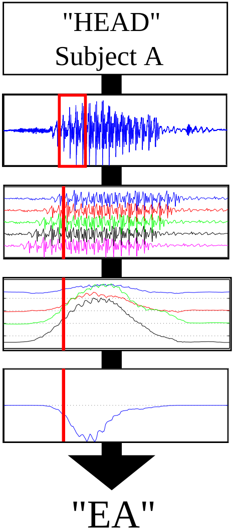
- Classification or regression w.r.t. latent variables Θ
  - Θ non-linearly embedded
  - Solution non-linear in Θ

- Example: vowel classification[1]

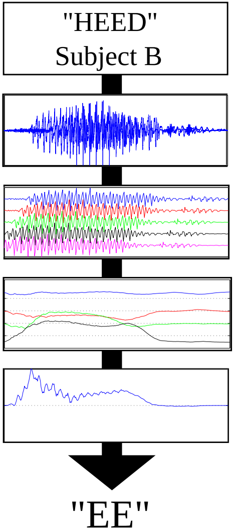


[1] North Texas vowel database (Assmann et al., 2008)

Linear solutions to non-linear problems



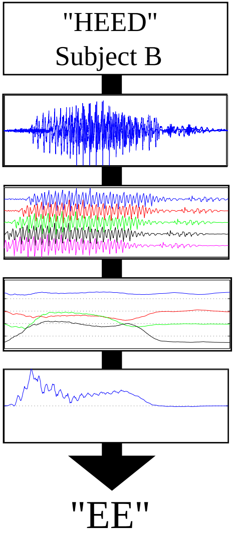
- Classification or regression w.r.t. latent variables Θ
  - Θ non-linearly embedded
  - Solution non-linear in Θ
- Mapping into feature space Φ
  - Φ is non-linear in data
  - Φ is functional basis in Θ
  - Φ is low dimensional
- Example: vowel classification[1]



[1] North Texas vowel database (Assmann et al., 2008)

Unsupervised non-linear feature extraction

- How to choose a feature space Φ?
  - Construct non-linear features from data
  - Here we investigate unlabelled data
- Unsupervised non-linear feature extraction
  - Non-linear PCA[2] depends on function space
  - Non-linear SFA[3] features in the limit independent of function space



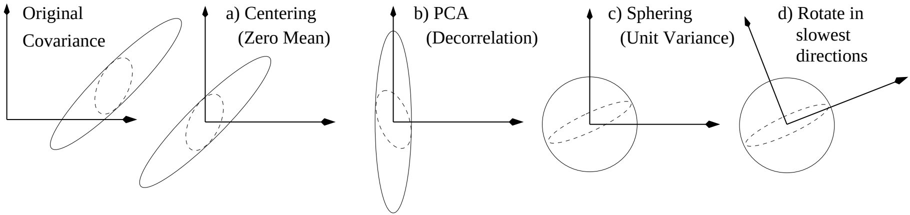
[2] Kernel Principal Component Analysis (Schölkopf et al., 1998)
[3] Slow Feature Analysis (Wiskott and Sejnowski, 2002; Wiskott, 2003)

RSK-**SFA** - slow feature analysis

- Filter temporally coherent signals in time series $\{x_t\}_{t=1}^n$

$$\min_{\phi \in \mathcal{F}^p} \mathbb{E}_t\left[\left\|\dot{\phi}(x_t)\right\|_2^2\right] \quad \text{(slowness)}$$

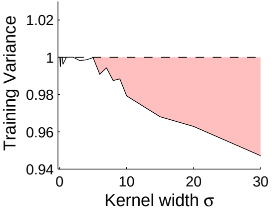
$$\begin{aligned} \text{s.t.} \quad \mathbb{E}_t\left[\phi(x_t)\right] &= \mathbf{0} && \text{(zero mean)} \\ \mathbb{E}_t\left[\phi(x_t)\phi(x_t)^\top\right] &= \mathbf{I} && \text{(unit variance \& decorrelation)} \end{aligned}$$



- Non-linear SFA features converge[4] to Fourier basis in Θ

[4] In the limit of an infinite time series and unrestricted function class (Wiskott, 2003)

RSK-SFA - kernel slow feature analysis



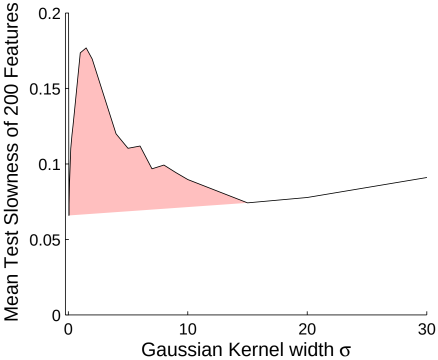
- Has up to now only been studied provisionally[5]
- Employs *reproducing kernel Hilbert spaces* $\mathcal{H}$

$$\phi_i(y) = \sum_{t=1}^{n} A_{ti}\, \kappa(y, x_t) \;-\; c_i$$

$$\min_{\mathbf{A}\in\mathbb{R}^{n\times p}} \quad \frac{1}{n-1}\,\mathbf{tr}\left(\mathbf{A}^\top \dot{\mathbf{K}}\dot{\mathbf{K}}^\top \mathbf{A}\right)$$

$$\text{s.t.} \qquad \frac{1}{n}\mathbf{A}^\top\mathbf{K}\mathbf{1} \;=\; \mathbf{0}$$

$$\frac{1}{n}\mathbf{A}^\top\mathbf{K}\mathbf{K}^\top\mathbf{A} \;=\; \mathbf{I}$$



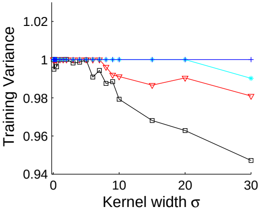
- Complexity $O(n^3)$
- K-SFA exhibits **over-fitting** and **numerical instabilities**[6]

[5] Bray and Martinez (2002)

[6] Shown analytically for the related Kernel CCA (Fukumizu et al., 2007)

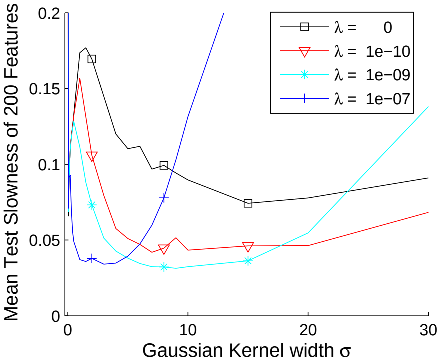
RSK-SFA - penalizing complex functions

- Power of functions class grows with n
- Regularization of function complexity
- Penalize Hilbert norm $\|\phi_i(\cdot)\|_{\mathcal{H}}$



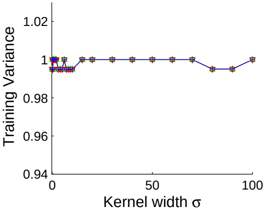
$$\min_{\phi \in \mathcal{H}^p} \mathbb{E}_t \left[\|\dot{\phi}(x_t)\|_2^2 \right] + \lambda \sum_{i=1}^{p} \|\phi_i(\cdot)\|_{\mathcal{H}}^2$$

$$\equiv \min_{\mathbf{A} \in \mathbb{R}^{n \times p}} \mathbf{tr} \left(\mathbf{A}^\top \left(\tfrac{1}{n-1} \dot{\mathbf{K}} \dot{\mathbf{K}}^\top + \lambda \mathbf{K} \right) \mathbf{A} \right)$$



- Little computational overhead
- λ must be fitted to kernel
- λ can become extremely small

R**S**K-SFA - preventing complex functions

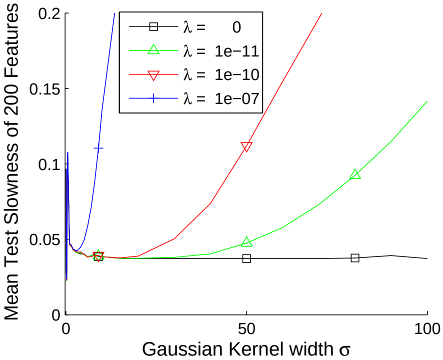


- Use subset of training data to express functions
- Restricts solution to a subspace of $\mathcal{H}$
- Implicit regularization of function complexity

$$\phi_i(y) = \sum_{j=1}^{m} A_{ji}\, \kappa(y, z_j) \;-\; c_i$$

$$\{z_j\}_{j=1}^{m} \subset \{x_t\}_{t=1}^{n}, \qquad m << n$$

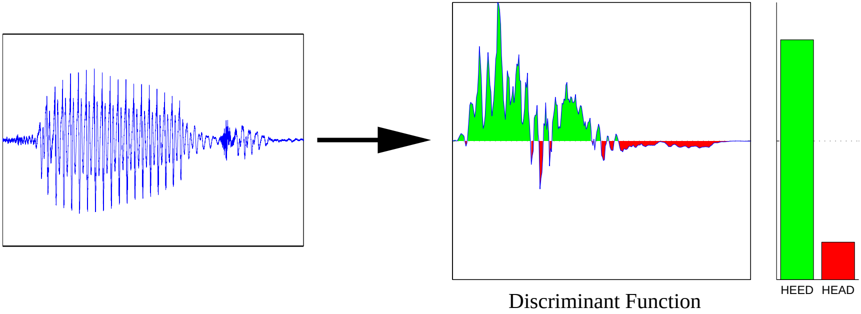
$$K_{jt} := \kappa(z_j, x_t), \quad \mathbf{K} \in \mathbb{R}^{m \times n}$$



- Reduces complexity to $O(m^2 n)$
- Efficient over many kernels
- Sensitive to subset selection[7]

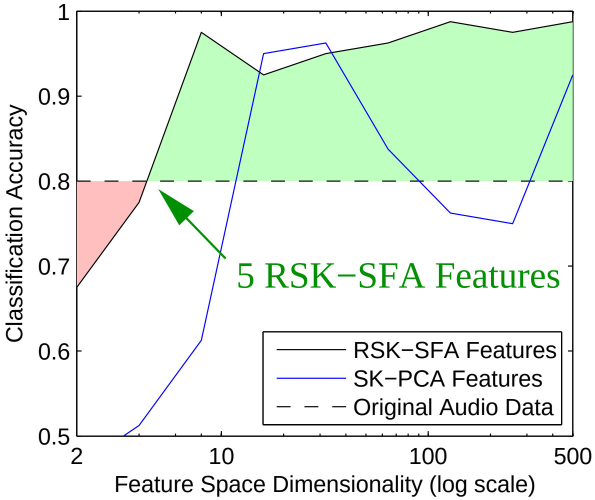
[7] See selection algorithms in Smola and Schölkopf (2000); Csató and Opper (2002)

Feature validation: vowel classification (1)



1. Delayed embedding ("windowing")
2. RSK-SFA/PCA feature extraction
3. Quadratic Discriminant Analysis (QDA)
4. Compare area above and below zero

Feature validation: vowel classification (2)



Take home message

Context	Linear classification/regression w.r.t. latent variables Θ
Data	Complex time series data with a reasonable kernel
Problem	No idea how to construct a proper feature space
Suggestion	Try RSK-SFA to approximate Fourier basis in Θ

Thank you for your attention!

P.F. Assmann, T.M. Nearey, and S. Bharadwaj. Analysis and classification of a vowel database. *Canadian Acoustics*, 36(3):148–149, 2008.

A. Bray and D. Martinez. Kernel-based extraction of Slow features: Complex cells learn disparity and translation invariance from natural images. *Neural Information Processing Systems*, 15:253–260, 2002.

L. Csató and M. Opper. Sparse on-line gaussian processes. *Neural Computation*, 14(3):641 – 668, 2002.

K. Fukumizu, F.R. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8: 361–383, 2007.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

A.J. Smola and B. Schölkopf. Sparse greedy matrix approximation for machine learning. In *Proceedings to the 17th International Conference Machine Learning*, pages 911–918, 2000.

L. Wiskott. Slow feature analysis: A theoretical analysis of optimal free responses. *Neural Computation*, 15(9):2147–2177, 2003.

L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.