Bayesian training of RBF networks for nonlinear system modeling

Erhard Rank

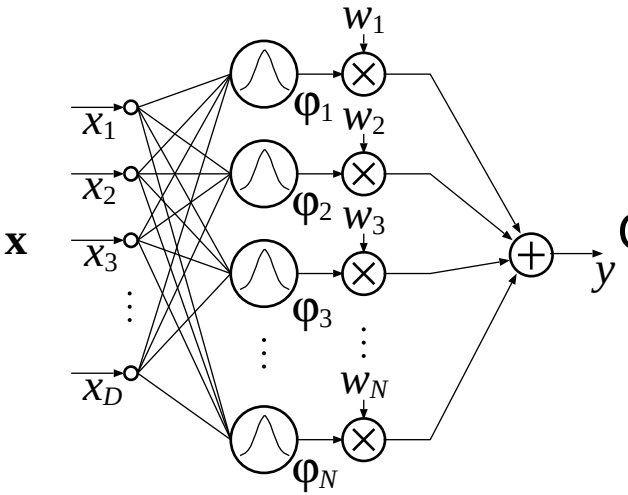
Institute of Communications and
Radio-Frequency Engineering,
Vienna University of Technology

Bayes seminar, TU Graz, 12. 5. 2003

Contents

- Radial basis function (RBF) networks
- Data model
- Bayesian training of Gaussian RBF networks
- Relation to
  - Regularized RBF networks
  - Relevance vector machine
- Oscillator model for nonlinear system modeling
  - Lorenz system
  - Speech with mixed excitation
- Summary

RBF network



$$y(\mathbf{x}) = \sum_{n=1}^{N} w_n \, \varphi_n(\mathbf{x})$$

Gaussian BF:

$$\varphi_n(\mathbf{x}) = e^{-\frac{\|\mathbf{x}-\mathbf{c}_n\|}{2\sigma}}$$

$\mathbf{c}_n$ . . . network centers

RBF networks (cont.)

- Fixed centers (on hyper-grid)
- Fixed BF (Gaussian, same variance)
- Regularization
  - Regularization of matrix inversion during network learning
  - Generalized Regularization Networks (GRN, Poggio&Girosi 89)
  - Bayesian regularization (MacKay 92)
  - Relevance Vector Machine (Tipping 01)

Data model

Data model:

$$t_i = \sum_{n=1}^{N_c} w_n \, \varphi_n(\mathbf{x}_i) + \epsilon_i, \quad p(\epsilon) = \mathcal{N}(0, \sigma^2)$$

- Training data: input $\mathbf{x}_i$ (dimension D), and output t_i, $i = 1 \ldots P$
- Network parameters
  - N_c and $\varphi_n()$ (fixed)
  - $\mathbf{w} = [w_1, w_2, \ldots w_{N_c}]^T$ (unknown)
- Noise variance σ^2 (unknown)

Data model (cont)

- $\mathbf{w}$ and σ^2 are random variables
- $\varphi_n()$ are Gaussian
- Prefer small weights (regularization/weight decay)

$$p(\mathbf{w}|\alpha) = \left(\frac{\alpha}{2\pi}\right)^{-\frac{N}{2}} \exp\left(-\frac{\alpha}{2}\|\mathbf{w}\|^2\right)$$

- Prior pdf for hyper-parameters σ^2 and α

$$p(\alpha) = \begin{cases} \propto \frac{1}{\alpha} & \alpha > 0, \\ 0 & \alpha \leq 0. \end{cases}$$

Bayesian learning:

Find parameters that maximize $p(\mathbf{w}, \alpha, \sigma^2|\mathbf{t})$

Bayesian training

$$p(\mathbf{w}, \alpha, \sigma^2|\mathbf{t}) = p(\mathbf{w}|\mathbf{t}, \alpha, \sigma^2)\, p(\alpha, \sigma^2|\mathbf{t})$$

$$p(\mathbf{w}|\mathbf{t}, \alpha, \sigma^2) = (2\pi)^{-\frac{N}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}}\, e^{\left(-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})\right)}$$

$$\boldsymbol{\Sigma} = (\frac{1}{\sigma^2}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \alpha\boldsymbol{I})^{-1},$$

$$\boldsymbol{\mu} = \frac{1}{\sigma^2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{t} \qquad (1)$$

$$\hat{\mathbf{w}}_{\text{bay}} = \boldsymbol{\mu}$$

Bayesian training (cont)

$$p(\alpha, \sigma^2|\mathbf{t}) \propto p(\mathbf{t}|\alpha, \sigma^2)\, p(\alpha)\, p(\sigma^2)$$

Maximizing leads to update equations for the hyper-parameters

$$\begin{aligned} \alpha^{\text{new}} &= \frac{\gamma}{\boldsymbol{\mu}^T \boldsymbol{\mu}}, \\ \left(\frac{1}{\sigma^2}\right)^{\text{new}} &= \frac{\|\mathbf{t} - \boldsymbol{\Phi}\boldsymbol{\mu}\|^2}{P - \gamma}, \\ \gamma &= N_c - \alpha\, \text{Trace}(\boldsymbol{\Sigma}). \end{aligned} \qquad (2)$$

Eqs. 1 and 2 are iterated until hyper-parameters converge.

Relation to regularized RBF networks

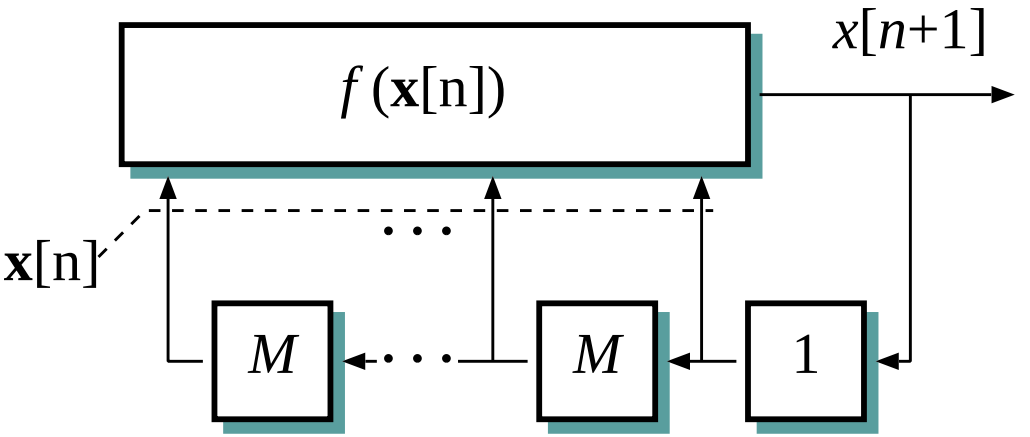
Besides the weights $\hat{\mathbf{w}}_{\text{bay}}$ the Bayesian training algorithm provides

- Regularization factor $\lambda_{\text{bay}} = \alpha\,\sigma^2$
  compare eq. 1 to regularization:
  - Regularized pseudo-inverse
    $$\hat{\mathbf{w}}_{\text{reg}} = (\mathbf{\Phi}^T\mathbf{\Phi} + \lambda \boldsymbol{I})^{-1}\mathbf{\Phi}^T\,\mathbf{t}$$
  - GRN
    $$\hat{\mathbf{w}}_{\text{reg}} = (\mathbf{\Phi}^T\mathbf{\Phi} + \lambda \mathbf{\Phi}_0)^{-1}\mathbf{\Phi}^T\,\mathbf{t}$$
- Estimate of noise level σ^2
- Needs no cross-validation (all training data can be used)

Relation to relevance vector machine

- Same priors ($\mathbf{w}, \alpha, \sigma^2$)
- Same optimization
- Only one α (variance of *all* weigths)
- No pruning

Oscillator model

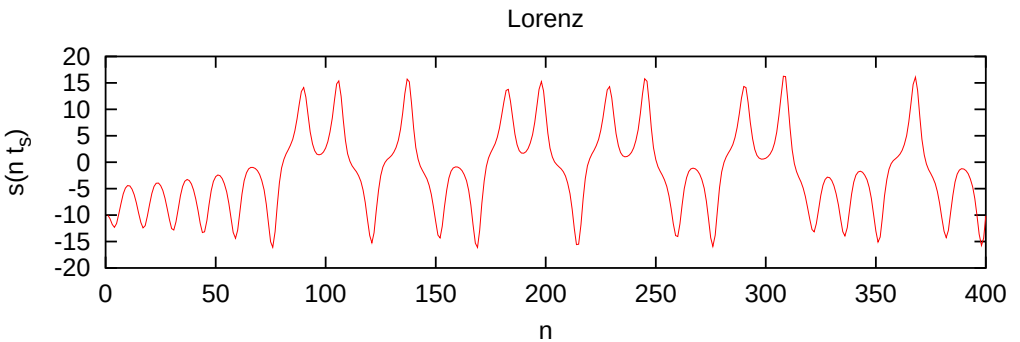


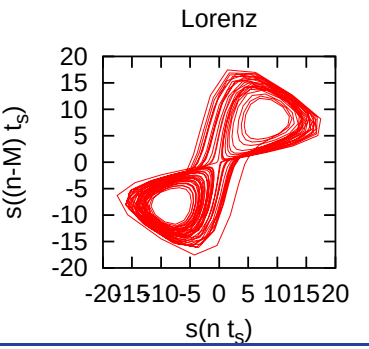
- Modeling oscillatory systems
- $f(\mathbf{x}[n])$ . . . RBF network
- Rules to optimize *embedding parameters* M and D (dimension of $\mathbf{x}[n]$)

Lorenz system

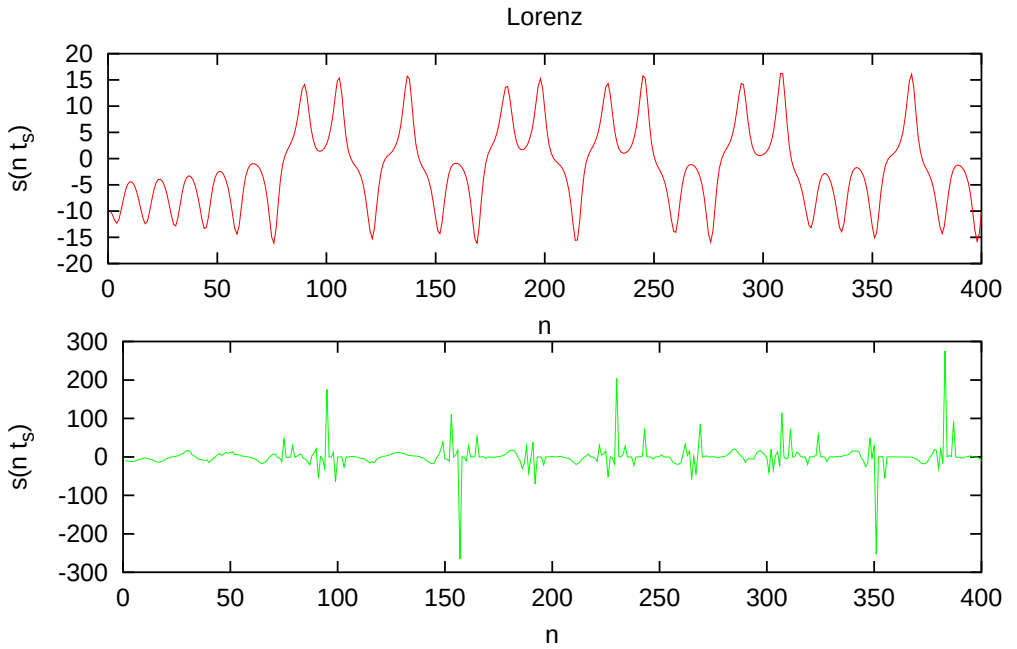
$$\begin{aligned}
\frac{dx(t)}{dt} &= -\sigma x(t) + \sigma y(t), \\
\frac{dy(t)}{dt} &= -x(t)z(t) + rx(t) - y(t), \\
\frac{dz(t)}{dt} &= x(t)y(t) - bz(t).
\end{aligned}$$

Parameters: $\sigma = 10$, $b = 8/3$ and $r = 28$





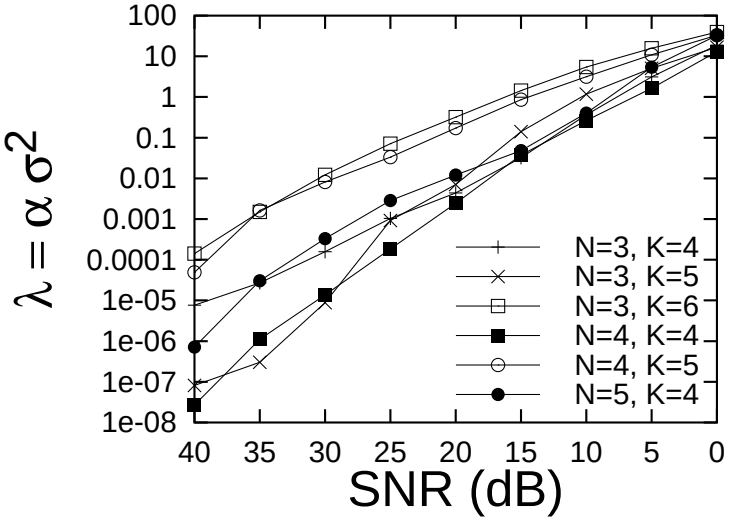
Lorenz system (cont)

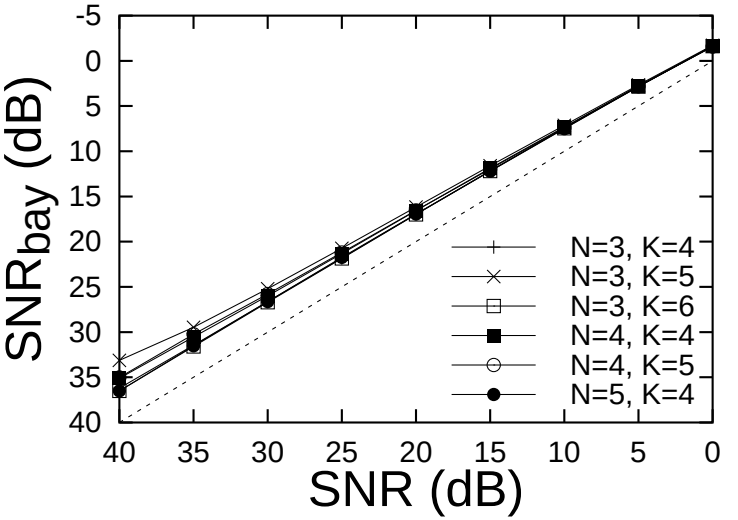


- Need regularization (Haykin&Principe 98)
- Does Bayesian training work (compared to cross-validation)?

Lorenz system (cont)

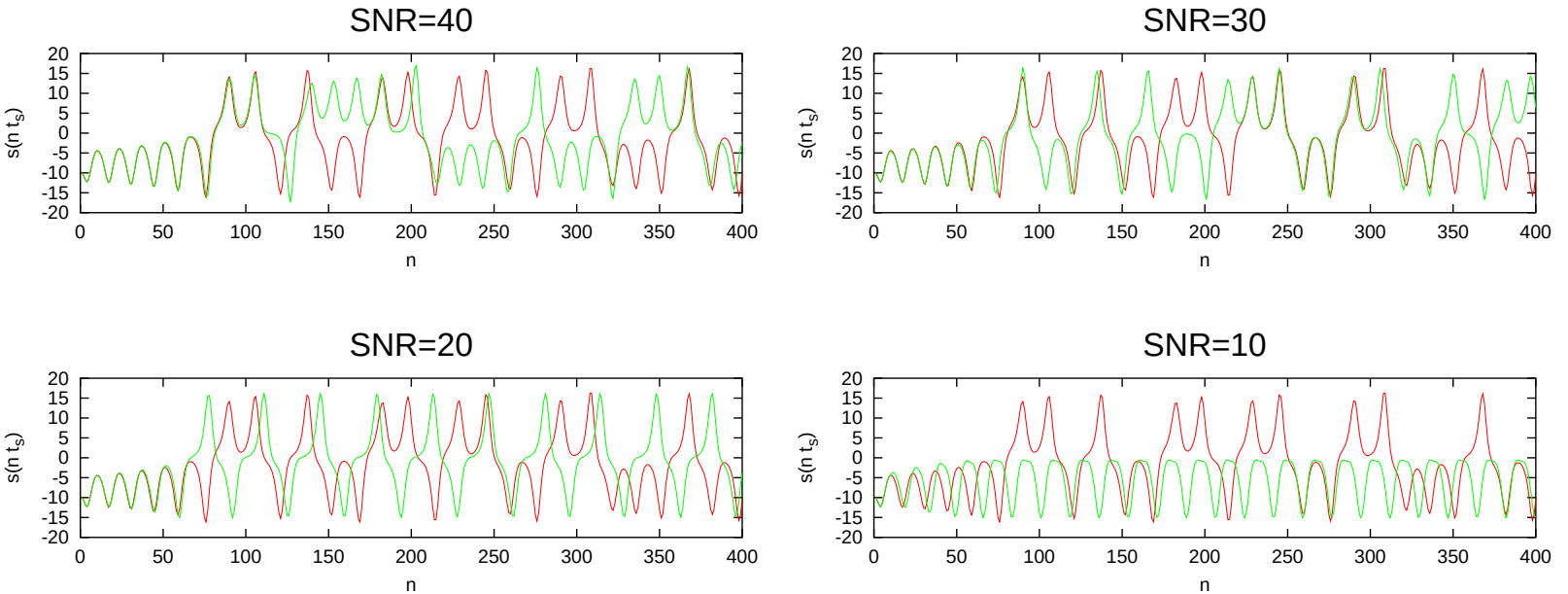
Oscillator model with RBF trained for varying SNR, with different embedding dimension N and number of network centers per dimension K



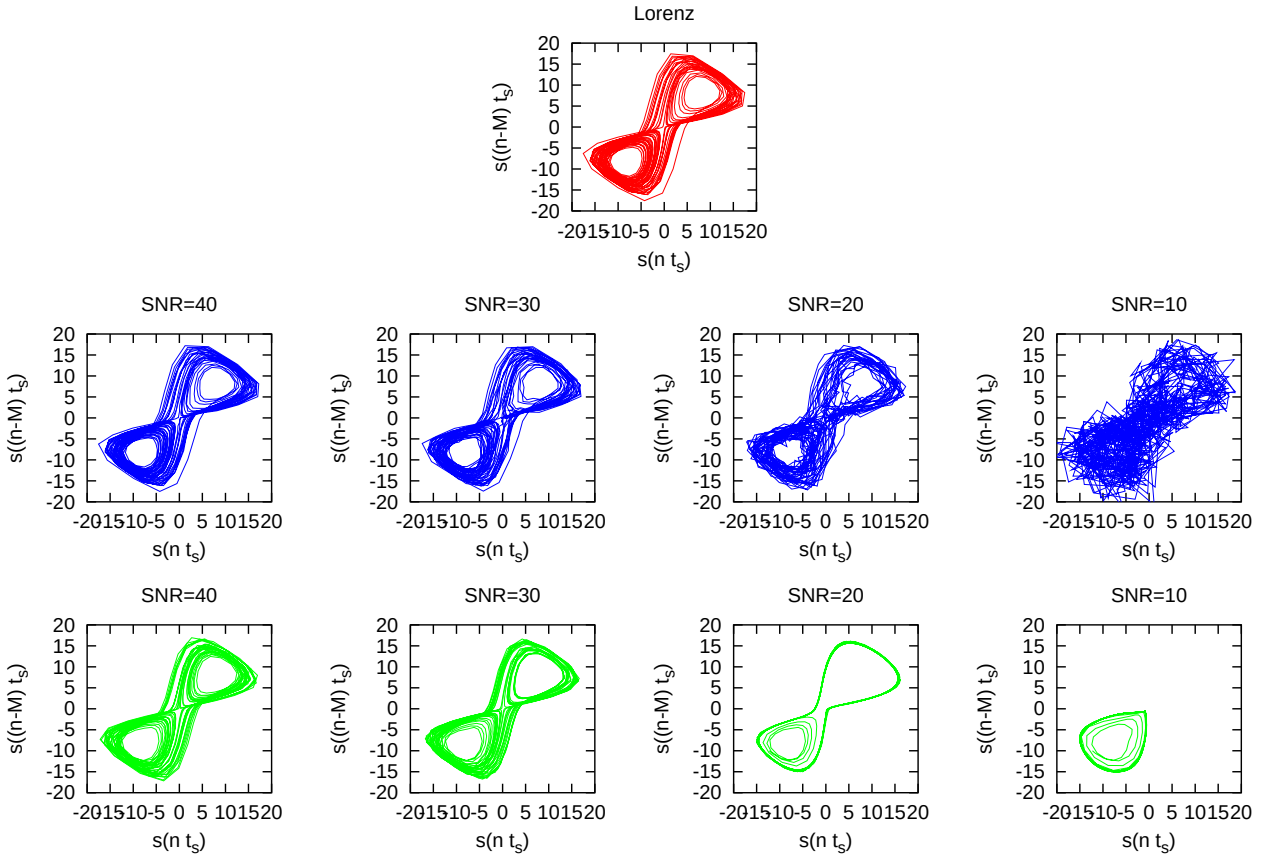


Lorenz system (cont)

Reconstructed signals, $N = 3$, $K = 4$ (64 centers):



Lorenz system (cont)



Lorenz system (cont)

Invariant measures, $d_E = 3$ (since $d_c < 3$)

SNR (dB)	λ_1	λ_2	λ_3	d_{KY}
40	$+2.13$	-1.89	-6.39	2.04
35	$+2.14$	-2.05	-6.16	2.01
30	$+2.29$	-2.13	-6.32	2.02
25	$+2.30$	-2.28	-5.98	2.0
original	$+2.80$	-0.27	-6.32	2.4
analytic	$+0.906$	0	-14.575	2.06

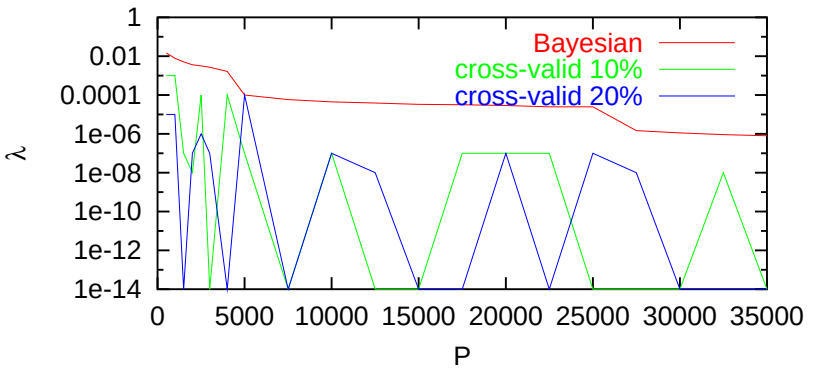
C. f.: un-regularized RBF (Haykin&Principe)

$d_E = 5$, $\boldsymbol{\lambda} = [13.2, 5.9, -3.1, -18.0, -47.1]$

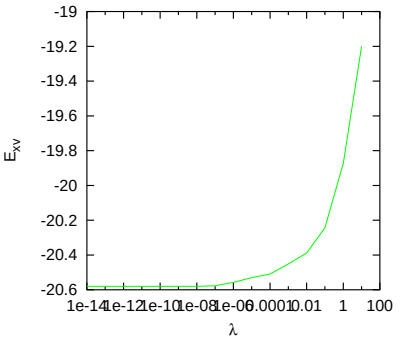
Lorenz system (cont)

Comparison to cross-validation (SNR = 25 dB)

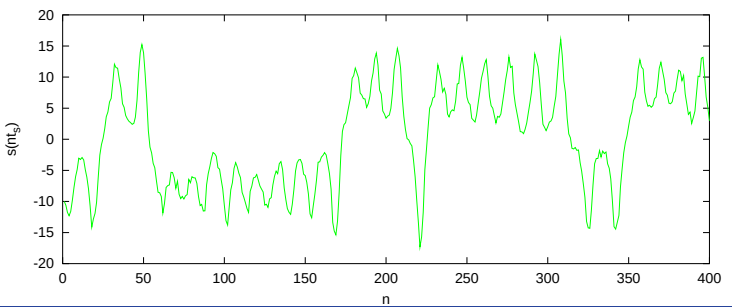
λ (training length P)

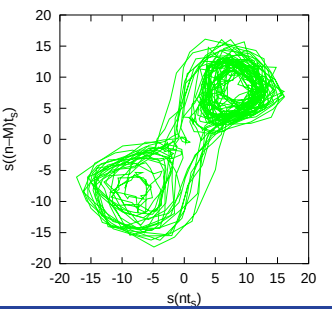


x-valid. error (λ)



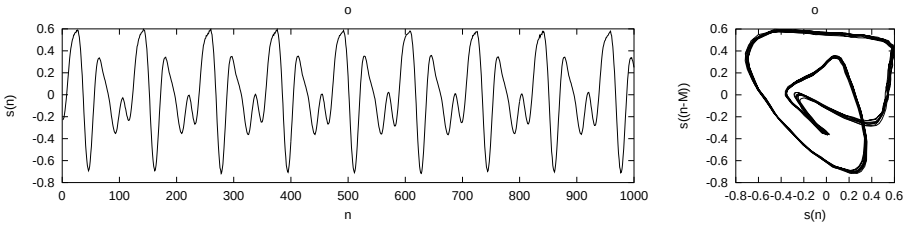
Reconstructed signal ($P = 15000 \rightarrow \lambda = 10^{-14}$)



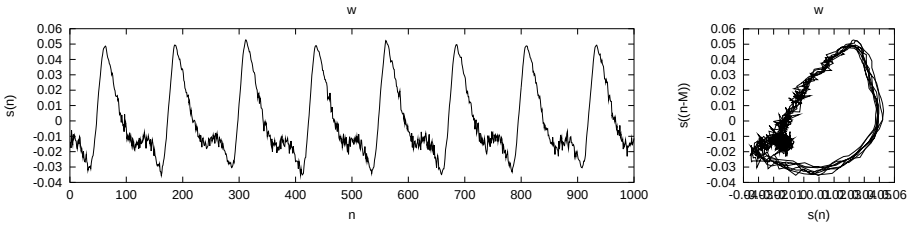


Speech

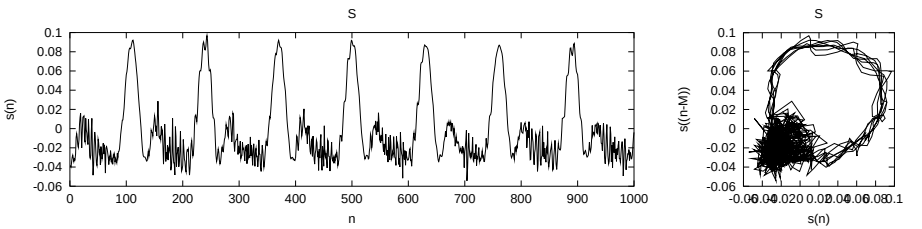
Voiced:



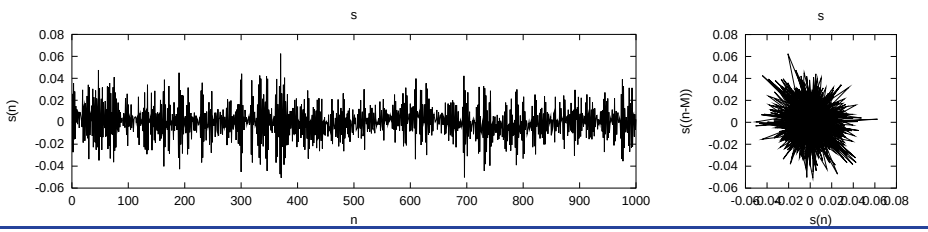
Mixed (1):



Mixed (2):



Unvoiced:



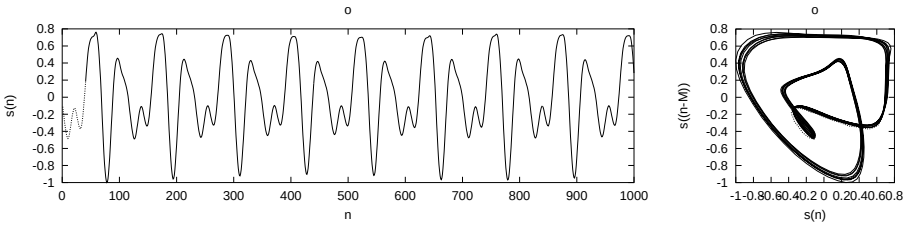
Speech (cont)

voiced	/aː/	/ɛ/	/oː/
SNR_{bay} (dB)	25.4	29.8	36.1
λ_{bay}	$2.2\,10^{-4}$	$3.3\,10^{-1}$	$1.5\,10^{-3}$
mixed	**/ʋ/**	**/ʒ/**	**/z/**
SNR_{bay} (dB)	19.1	12.7	4.9
λ_{bay}	$9.9\,10^{-2}$	$2.4\,10^{-1}$	2.9
unvoiced	**/s/**	**/f/**	
SNR_{bay} (dB)	5.5	0.39	
λ_{bay}	2.0	88	

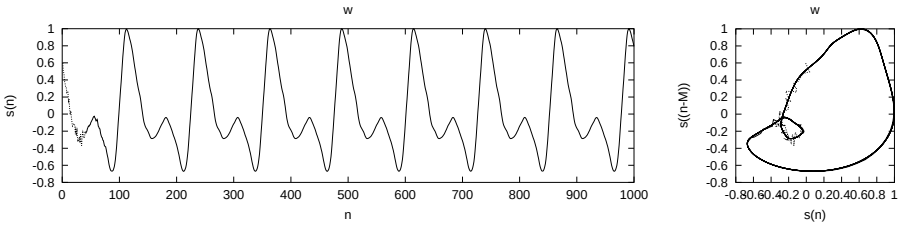
Amount of regularization correlates with unvoiced excitation

Speech (cont)

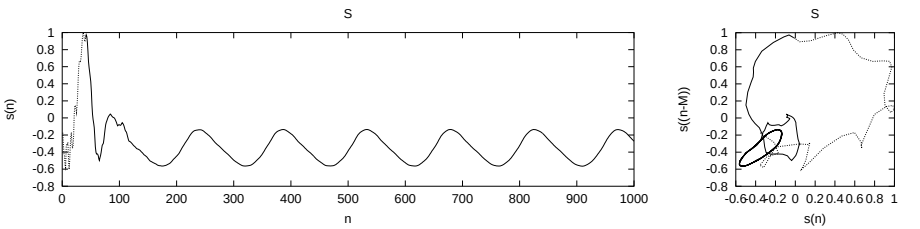
Voiced:



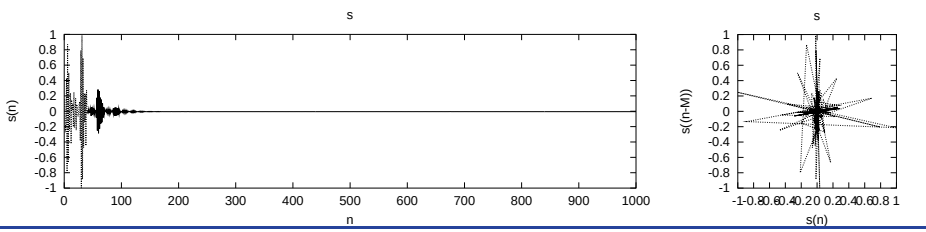
Mixed (1):



Mixed (2):



Unvoiced:



Speech (cont)

Audio examples:

- /o/ original/modeled
- /v/ original/modeled
- /ʒ/ original/modeled
- /s/ original/modeled

More examples on
http://www.nt.tuwien.ac.at/ erank/work.

Summary

- Bayesian training of RBF networks
  - Iteratively assigns amount of regularization
  - Estimates noise level
  - No cross-validation data
- Models chaotic behavior of Lorenz system
- Models voiced part of speech signals

References

Tomaso Poggio and Federico Girosi. A theory of networks for approximation and learning. A.I. Memo 1140, Massachusetts Institute of Technology, 1989.
ftp://publications.ai.mit.edu/ai-publications/1000-1499/AIM-1140.ps.Z

David J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.
http://spsc.inw.tugraz.at/courses/asp/ss03/docs/mackay1992_1.pdf

Michael E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.
http://spsc.inw.tugraz.at/courses/asp/ss03/docs/tipping2001.pdf

Simon Haykin and José Príncipe. Making sense of a complex world. *IEEE Signal Processing Magazine*, 15(3):66–81, May 1998.