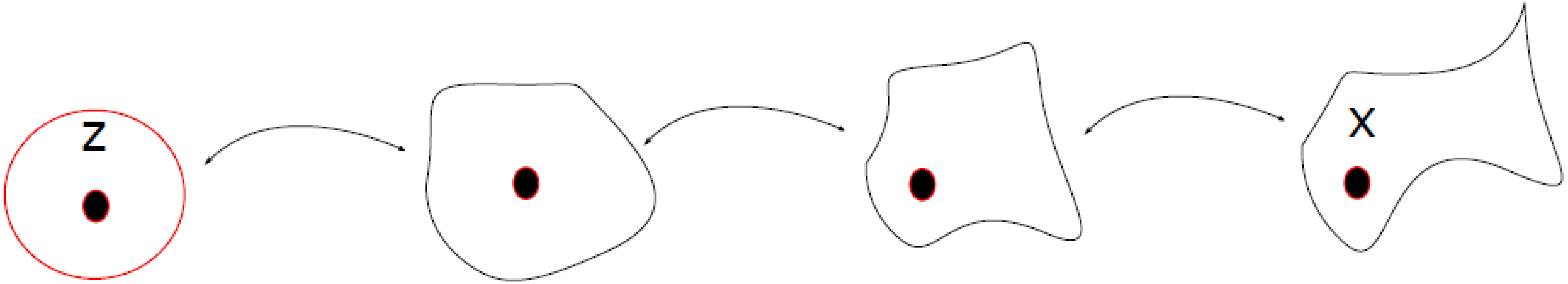
Back to Matrix

Dong Wang

2019/10/27

Basic principles of function mapping



$$\ln p(x) = \ln p(z(x)) + \ln |\mathrm{det}(\frac{\mathrm{d}z}{\mathrm{d}x})|$$

$$H(x) = H(z) + \int p(z) ln(|det(\frac{dz}{dx})|) dz$$

$$H(x) \leq \sum_i H(x_i)$$

Encoding and decoding scheme

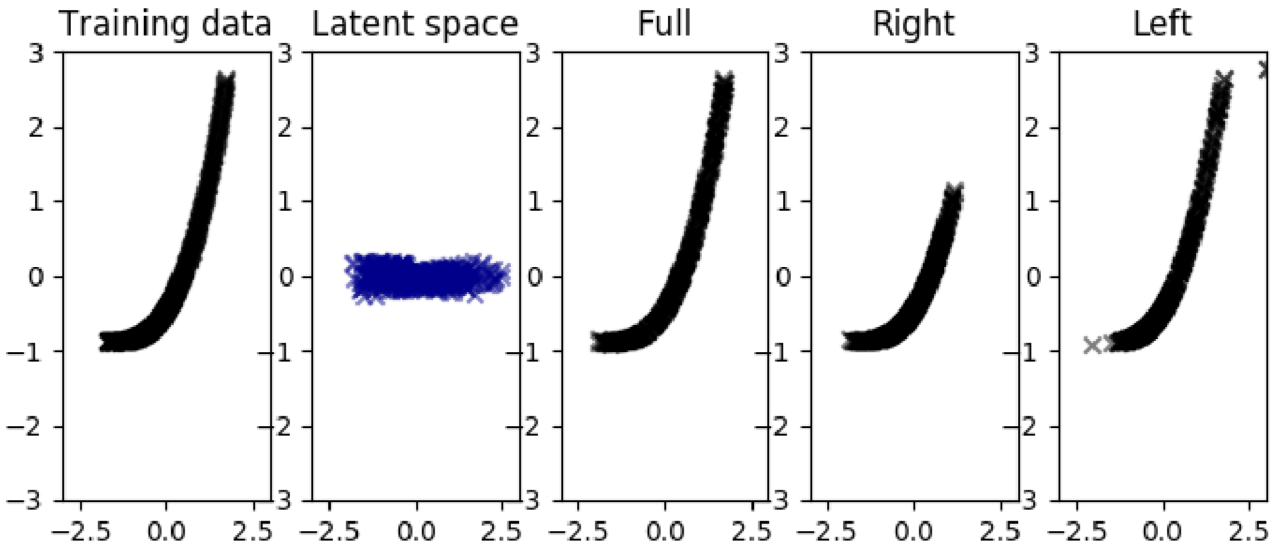
- Perfect reconstruction, therefore happy to match distribution p(z)
- VAE: accumulated posterior matching p(h)
- Directed Genreative Auto-encoder, matching p(h)

Two main problems of flows

- High dimensionality
- No supervision, no intuition

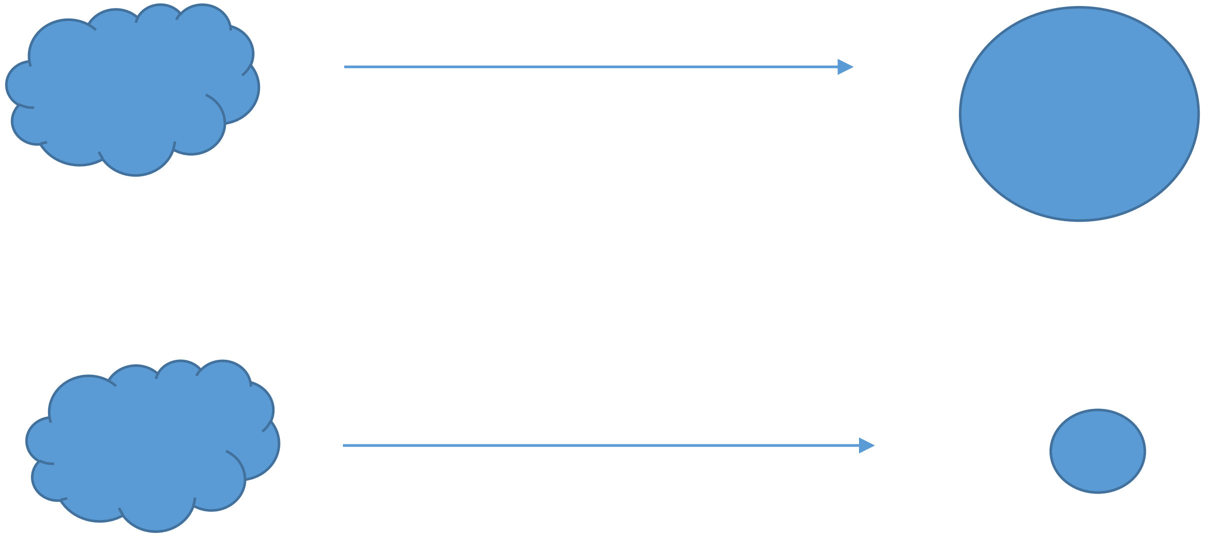
Initial attemp for subspace flow

- Control the variance of the output
- Failed as no contraint



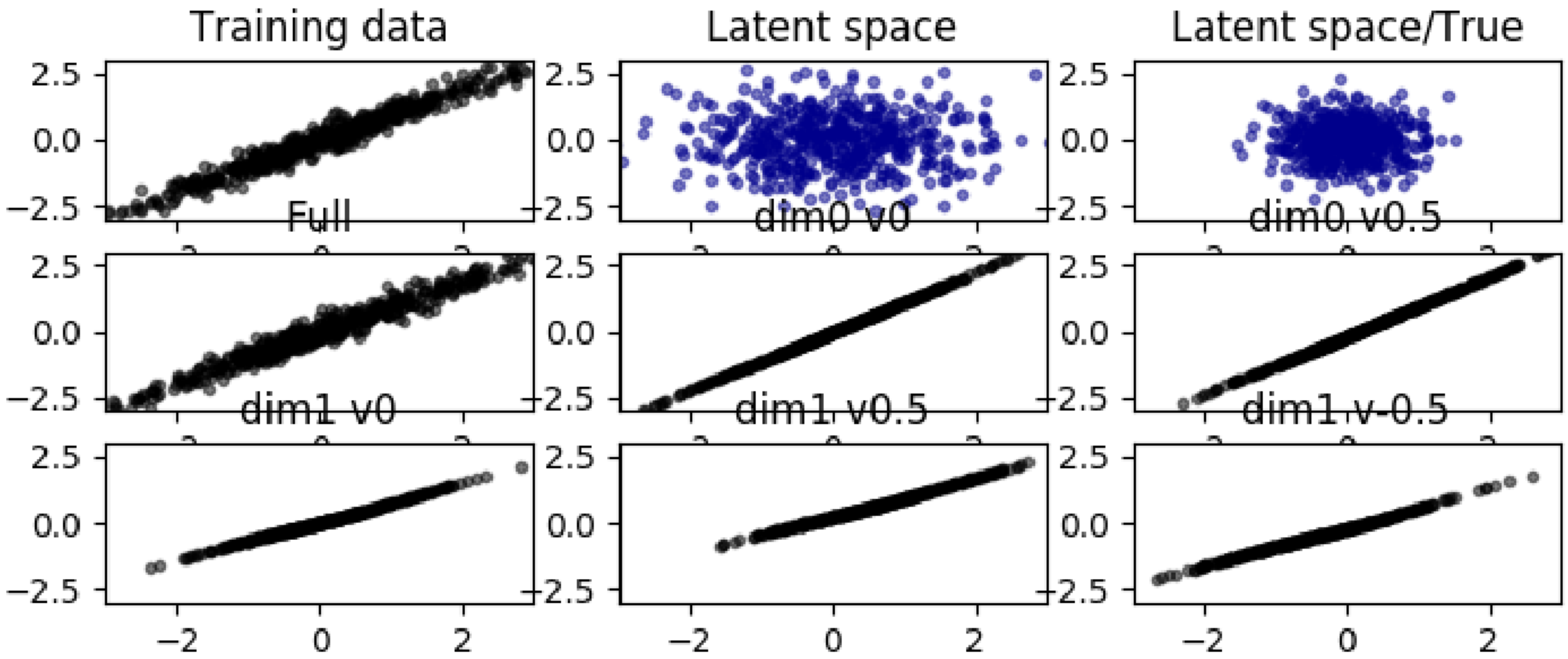
VP and NVP flow

- VP: For feature extraction
- NVP: For sampling and generation



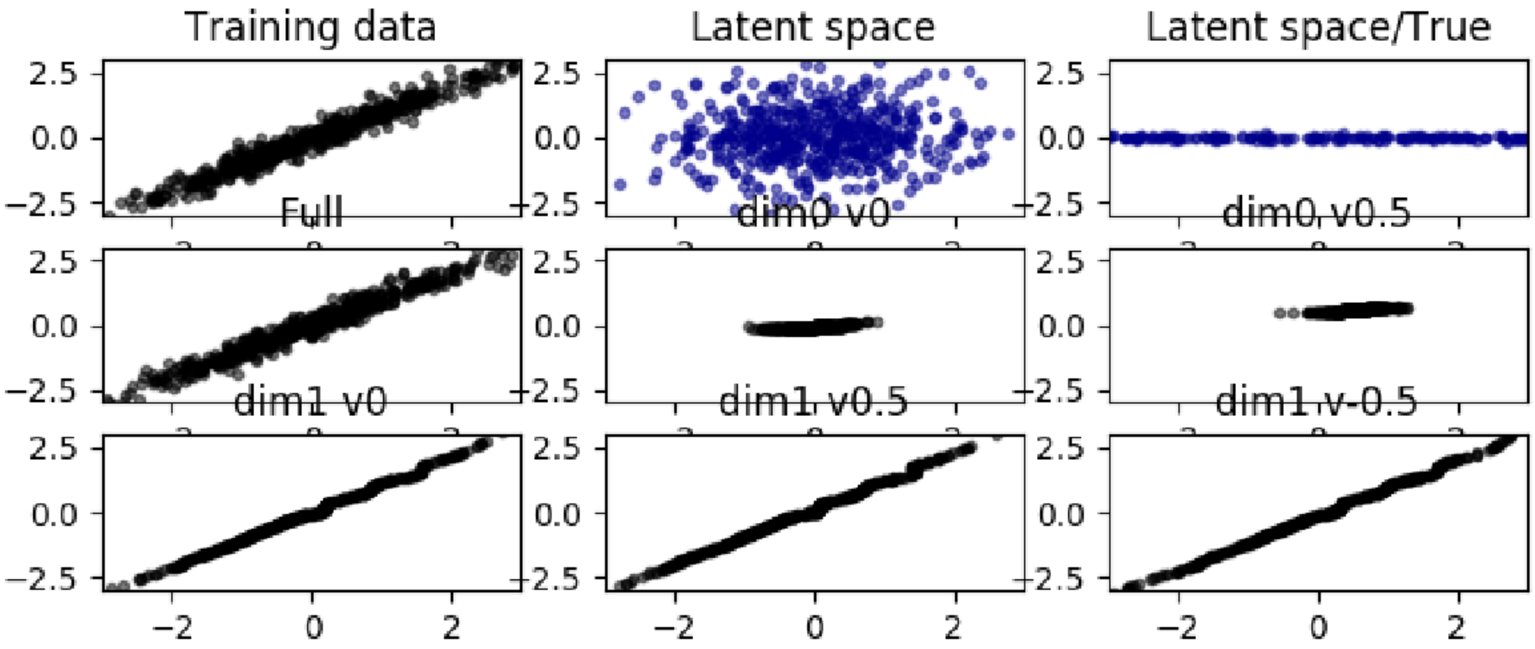
Subspace flow

- VP with unconstrained evolution



Subspace flow

- VP with constraind evolution



Unsupervised learning by Barlow

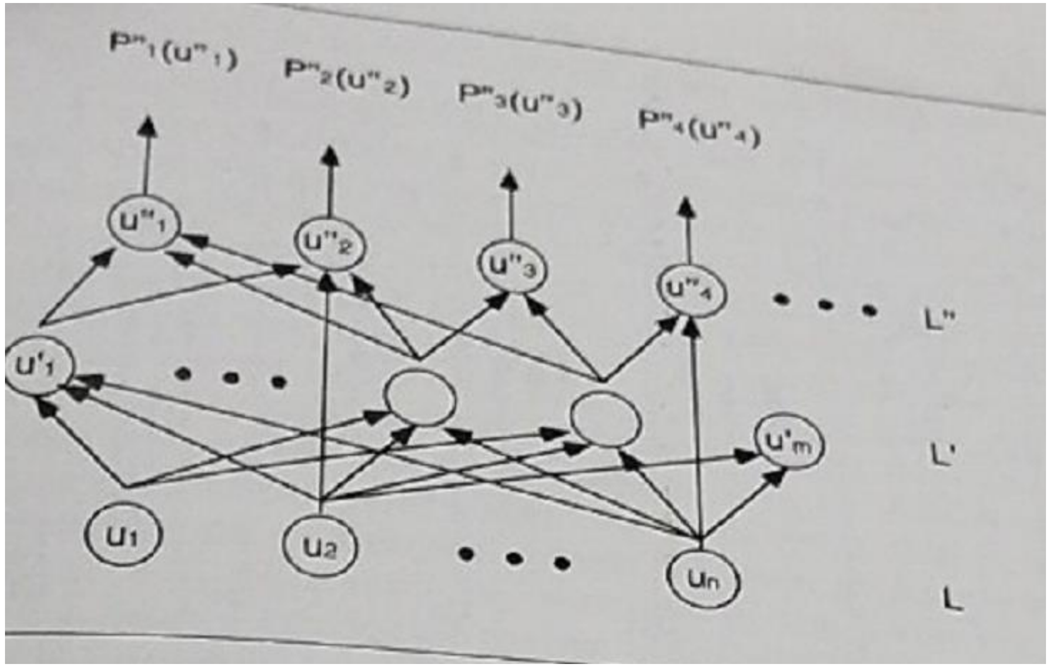
- Information is containd in redundancy; knowledge is represented as mean, variance, and covariance.
- Patterns are supervising (Laws of thought, economy of thought)
- Minimum entropy

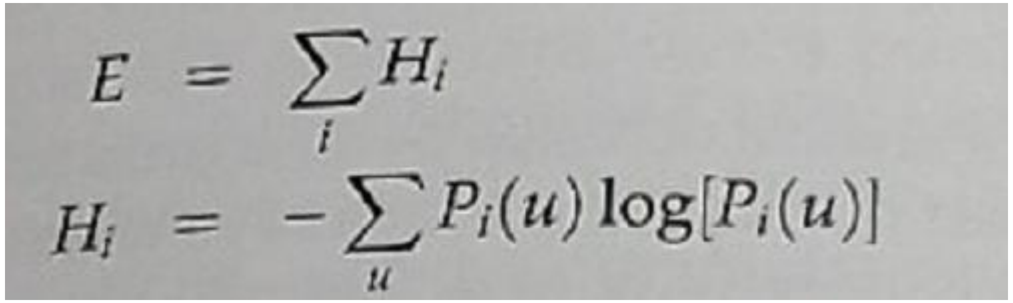
H. Barlow. (1989) Unsupervised Learning. Neural Computation, 1,295-311.
Boole, G., 1854, An investigation of the Laws of Thought
Barlow, H.B., and Foldiak, P. 1989, Adaptation and decorrelation in the cortex. The computing Neuron.

Minimum entropy with VP

- Minimizing entropy of the output
- Preserving the entropy

$$H(x) \leq \sum_i H(x_i)$$





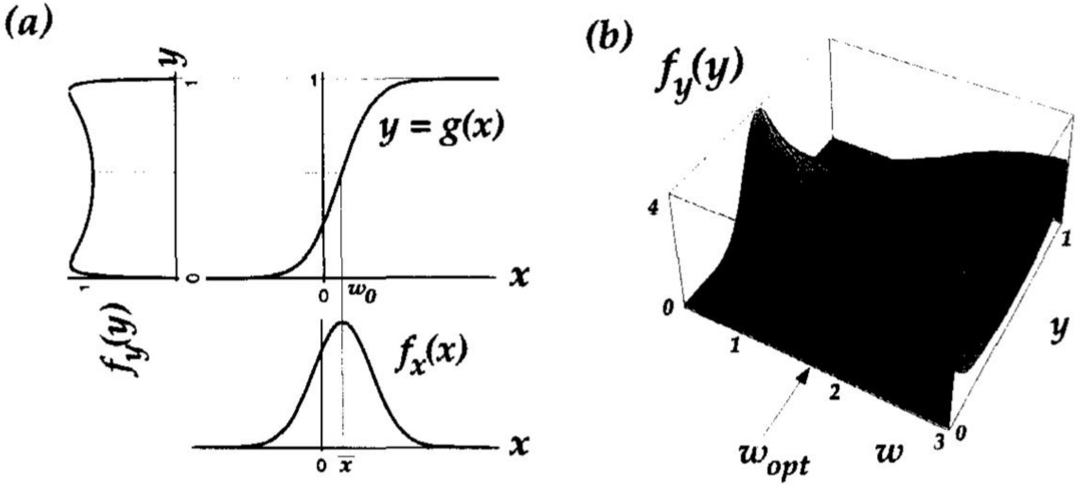
$$E = \sum_i H_i$$

$$H_i = -\sum_u P_i(u) \log[P_i(u)]$$

A. N. Redlich. (1993) Supervised Factorial Learning. Neural Computation, 5, 750-766.

Maximum mutual information with NVP

- When inputs are to be passed through a sigmoid function, maximum information transmission can be achieved when the sloping part of the sigmoid is optimally lined up with the highdensity parts of the inputs. As we show, this can be achieved in an adaptive manner, using a stochastic gradient ascent rule.
- Note that if two rows/columns in Jacobin, then two outputs seem correlated. This leads to unstable solution.



$$H(x) \leq H(z) + \int p(z) ln(|det(\frac{dx}{dz})|)dz$$

Anthony J. Bell and Terrence I. Sejnowski , “An Information-Maximization Approach to Blind Separation and Blind Deconvolution”, Neural computation, 1995.

Other ways

- Anti-Hebbian synapsese (Barlow, 1989)
- Sparse coding
- Decorrelation

Unsupervised nonlinear ICA

- **(1)High-order statistic for non-Gaussian; (2)no-linear for nolinear decorellation**
- Necessary condition for factorization with high-order statistics

$$\vec{y} = \vec{F}(\vec{x})$$

$$H(\vec{y}) \le H(\vec{x}) + \int P(\vec{x})\, ln\,(det\,(\frac{\partial\vec{F}}{\partial\vec{x}}))\, d\vec{x}$$

$$C_{ij} = 0, \quad if(i \neq j)$$

$$C_{ijk} = 0, \quad if(i \neq j \vee i \neq k)$$

$$C_{ijkl} = 0, \quad if(\{i \neq j \vee i \neq k \vee i \neq l\} \wedge \neg L)$$

$$C_{iijj} - C_{ii}C_{jj} = 0, \quad if(i \neq j)\,.$$

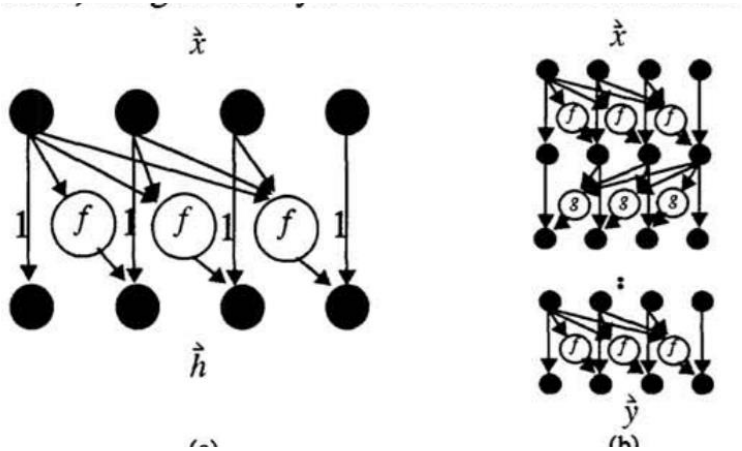
$$L = \{(i = j \wedge k = l \wedge j \neq k) \vee (i = k \wedge j = l \wedge i \neq j) \vee (i = l \wedge j = k \wedge i \neq j)\}$$

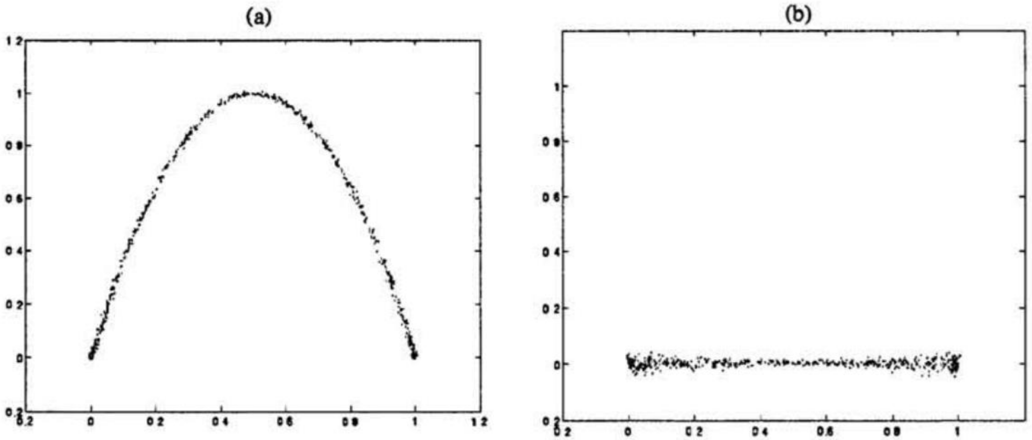
$$E = \alpha\sum_{i<j} C_{ij}^2 + \beta\sum_{i<j\le k} C_{ijk}^2 + \gamma\sum_{i<j\le k\le l} C_{ijkl}^2 + \delta\sum_{i<j}(C_{iijj} - C_{ii}C_{jj})^2$$

Gustavo Deco , Wilfried Brauer, Higher Order Statistical Decorrelation without Information Loss, NIPS 94

Unsupervised nonlinear ICA

- Volume preserving transform





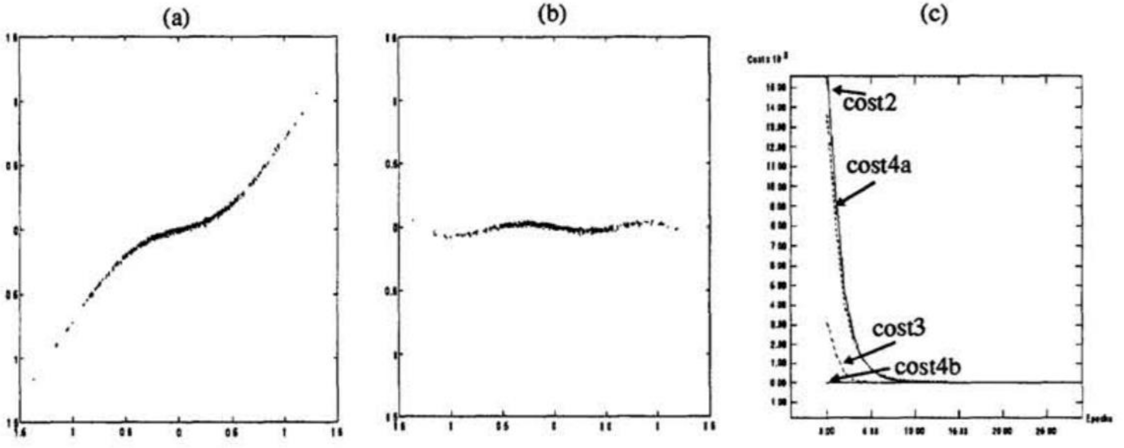


Figure 4: Input and Output space distribution after training with a two-layer polynomial volume-conserving network of order for the noisy curve of eq. (4.2). (a) input space; (b) output space (c) Development of the four summands of the cost function (eq. 2.18) during learning: (cost 2) first summand (second order correlation tensor); (cost 3) second summand (third correlation order tensor); (cost 4a) third summand (fourth order correlation tensor); (cost4b) fourth summand (fourth order correlation tensor).

Gustavo Deco , Wilfried Brauer, Higher Order Statistical Decorrelation without Information Loss, NIPS 94

Some interesting things to do

- Generation model
  - Model data directly, as deep Gaussian SID
  - Data augmentation, data fixing (noise inpainting), data expansion (bandwidth)
  - Speech conversion
  - Multimodality models
  - Channel mapping
- Feature learning
  - Nonlinear LDA, ICA
  - Anti-attack

Most prominent

- Turn over the big data approach, by finding the ‘true’ distribution.
- Turn over the blind-training approach, by finding a sensory space in the mind, where things can be explained.
- Establish the information-based speech generation framework