Learning Models for Continuous Nonnegative Data

Olly Downs
Hopfield Group, Princeton University

http://olly.downs.net

Overview

- Probabilistic Data Modeling and Maximum Entropy Distributions
- Maximum Entropy for nonnegative data
- The Nonnegative Boltzmann Machine (Downs, MacKay & Lee, 2000)
- 'Advanced' Mean Field Theory for the NNBM
- The Mean Field NNBM with Hidden Units
- Testing and Application
    - Orientation Tuning in Visual Cortex
    - The Nonnegative Matrix Factorisation and Models for Human Faces
- Summary

Probabilistic Models for Symmetric Networks

(Hopfield, Ackley et. al.)

- The symmetric neural network is a highly simplified model of associative memory.
- It is 'taught' in a Hebbian fashion, forming the synaptic weights of the network. (We consider binary neurons here).

$$\Rightarrow T_{ij} \sim \text{correlation}(V_i, V_j)$$
$$\Rightarrow \mathbf{T} = \left\langle \mathbf{V}\mathbf{V}^T \right\rangle_{\text{memories}}$$

- The 'neurons' obey the dynamical equations (V_j=sign(u_j) here)

$$\dot{u}_i = -u_i + \sum_j T_{ij} V_j + I_i$$

- It has been shown that these dynamics minimise the *Lyapunov Function*

$$L = -\frac{1}{2}\sum_{i,j} T_{ij} V_i V_j + \sum_i I_i V_i$$

- The Lyapunov Function defines an 'energy' for all 2^N possible binary configurations, under the prevailing T_{ij}

Learning Models for Continuous Nonnegative Data
3

Enabling the *future* through innovation.
Microsoft Research

Probabilistic Models for Symmetric Networks

- We can therefore define a probability distribution over all the states

$$P(\mathbf{x}) = \frac{1}{Z(\mathbf{T})}\exp(-L(\mathbf{x})) = \frac{1}{Z(\mathbf{T})}\exp\left(\frac{1}{2}\mathbf{x}^T\mathbf{T}\mathbf{x}\right)$$
$$\text{with } Z = \sum_{\substack{\text{all binary} \\ \text{configs of } \underline{x}}} \exp(-L(\mathbf{x}))$$

- This is a *Boltzmann Distribution*
- *Note:* When we have deterministically settled into the state which minimises L we have found the state with the greatest probability under this Boltzmann Distribution
- Moving away from the deterministic neural dynamics to the stochastic rule
  - Compute $\alpha_i = \sum_j T_{ij} V_j$ (the change in energy on flipping u_i)
  - Then set

$$u_i = -u_i \text{ with probability} \left(1 + \exp\left(\frac{\alpha_i}{\kappa}\right)\right)^{-1} \text{ where } \kappa = \text{"Temperature"}$$

Learning Models for Continuous Nonnegative Data
4

Enabling the *future* through innovation.
Microsoft Research

Deterministic Networks to Boltzmann Machines

- The stochastic dynamics visits different configurations of the system with frequency given by the Boltzmann distribution
- The Probability model gives us a means of *learning* T_{ij}
- The patterns $\{\mathbf{V}\}$ we ideally wish to store define a probability distribution over all configurations of the network $Q(\mathbf{x})$
- We want to match $P(\mathbf{x})$ to the probability distribution of the patterns.
- We can derive a learning rule by performing gradient descent w.r.t. $\boldsymbol{T}$ on the Kullback-Leibler distance

$$D(Q \,\|\, P) = -\sum_{m \text{ patterns}} Q\left(\mathbf{V}^{(m)}\right) \log\left(\frac{P\left(\mathbf{V}^{(m)}\right)}{Q\left(\mathbf{V}^{(m)}\right)}\right)$$

Boltzmann Machines

(Ackley et. al.)

- Thus we have a learning rule for T

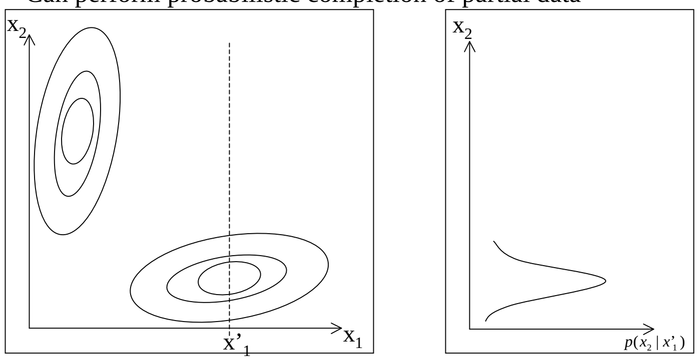
$$\boxed{\Delta\mathbf{T} = \frac{1}{2}\left[\left\langle \mathbf{x}\mathbf{x}^T \right\rangle_{\text{patterns}} - \left\langle \mathbf{x}\mathbf{x}^T \right\rangle_{\substack{\text{all binary}\\ \text{configs}}}\right]}$$

$$\boxed{\left\langle \mathbf{x}\mathbf{x}^T \right\rangle_{patterns} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{V}^{(m)}\mathbf{V}^{(m)^T}}$$

$$\boxed{\left\langle \mathbf{x}\mathbf{x}^T \right\rangle_{\substack{\text{all binary}\\ \text{configs}}} = \sum_{i=1}^{\substack{\#\text{ binary}\\ \text{configs}}} \mathbf{x}^{(i)}\mathbf{x}^{(i)^T} P\left(\mathbf{x}^{(i)}\right) = \left\langle \mathbf{x}\mathbf{x}^T \right\rangle_{free}}$$

Why make probabilistic models of data?

- How do we use our learned generative model?
  - Sampling from the model generates new prototype data
  - Can perform probabilistic completion of partial data



- We can also use the model for recognition, by considering the 'log-likelihood' $\ln p\left(\mathbf{x} \mid \mathbf{V}^{(i)}\right) >$ Threshold

Enabling the *future* through innovation. Research

Maximum Entropy Data Modeling

- Consider the abstract problem of selecting a *natural* probability model to capture different classes of data.
- Consider the 'least-committal' model, the maximum entropy model, and constrain it to match certain statistics of the data.
- Simple Lagrange multiplier problem

$$\max_{p(\mathbf{x})}\left[\int -p(\mathbf{x})\log p(\mathbf{x})d\mathbf{x} - \lambda\Phi - \nu\Gamma - \varpi\Psi - \ldots\right]$$

$$\Phi = \int p(\mathbf{x})d\mathbf{x} - 1$$

$$\Gamma = \int \mathbf{x}p(\mathbf{x})d\mathbf{x} - \boldsymbol{\mu}, \qquad \boldsymbol{\mu} \text{ is the mean of the data}$$

$$\Psi = \int \mathbf{x}\mathbf{x}^T p(\mathbf{x})d\mathbf{x} - \mathbf{C}, \ \mathbf{C} \text{ is the data cross-correlation matrix}$$

- We simply attempt to match 1st and 2nd order statistics

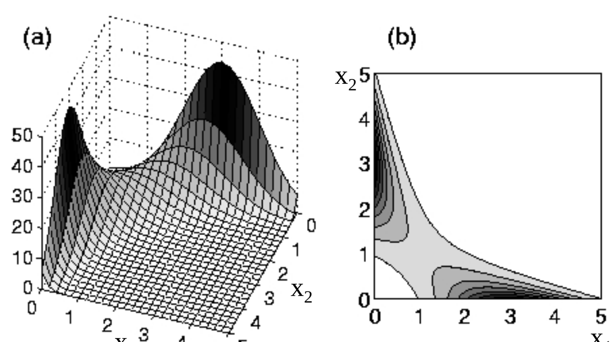
Enabling the *future* through innovation. Research

Maximum Entropy Data Modeling Contd.

- For continuous, unbounded data $p(\boldsymbol{x})$ is Gaussian.
- For binary data the maximum-entropy distribution is the *Boltzmann Distribution* with quadratic energy.
  - This is indeed the probability distribution learned and explored by the *Boltzmann Machine*.
  - ***The Boltzmann Machine learns the maximum entropy probability distribution of binary data based upon 1st and 2nd order statistics.***
- For continuous data *bounded to the nonnegative orthant* the distribution is the *Nonnegative Boltzmann Distribution*

$$p(\mathbf{x}) = \begin{cases} \frac{1}{Z} e^{-\beta \mathbf{x}^T \mathbf{A}\mathbf{x} + \mathbf{b}^T \mathbf{x}}, \forall\, x_i \geq 0 \\ 0, \quad \text{otherwise} \end{cases}$$

The Nonnegative Boltzmann Distribution



Surface plot

Contour plot

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \ \mathbf{b} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$$

The Nonnegative Boltzmann Machine

(Downs et. al.)

- So how do we learn the NNBD to best represent some set of nonnegative data {**V**}?
- Search for *maximum likelihood* parameters, given the data, under the assumption that $p(\{\mathbf{V}\}|\,\text{Parameters}) = NNBD$
- Bayes

$$\max_{Parameters}\left[p(Parameters\,|\,\{\mathbf{V}\})\right] = \max_{Parameters}\left[\frac{p(\{\mathbf{V}\}|\,Parameters)\,p(Parameters)}{p(\{\mathbf{V}\})}\right]$$
$$\Rightarrow \max_{Parameters}\ \ln p(\{\mathbf{V}\}|\,Parameters)$$
$$\Rightarrow \max_{\mathbf{A},\mathbf{b}}\left[-\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} - \ln Z\right], \mathbf{x} \in \{\mathbf{V}\}$$

- We perform Gradient Ascent to maximise this log-likelihood function

Learning Models for Continuous Nonnegative Data
11

Enabling the *future* through innovation.
Microsoft Research

The NNBM Contd.

- Learning rules, (cf. Boltzmann Machine)

$$\Delta\mathbf{A} = \eta_A \frac{\partial L}{\partial \mathbf{A}} = \eta_A\left[\left\langle \mathbf{V}^{(\alpha)}\mathbf{V}^{(\alpha)^T}\right\rangle_\alpha - \left\langle \mathbf{x}\mathbf{x}^T\right\rangle_{free}\right]$$
$$\Delta\mathbf{b} = \eta_b \frac{\partial L}{\partial \mathbf{b}} = \eta_b\left[\left\langle \mathbf{x}\right\rangle_{free} - \left\langle \mathbf{V}^{(\alpha)}\right\rangle_\alpha\right]$$

- Learning is made hard by $\langle \_ \rangle_{free}$ which is a multidimensional Gaussian integral over the nonnegative orthant.
- Ways ahead
  - Integrate
  - Mean-Field/Variational Approximation
  - Markov Chain Monte-Carlo methods

Learning Models for Continuous Nonnegative Data
12

Enabling the *future* through innovation.
Microsoft Research

Practical Implementation

- Why not integrate?
  - Consider

$$\mathbf{A} = \begin{pmatrix} a & b & \dots & 0 & 0 \\ b & a & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & a & b \\ 0 & 0 & \cdots & b & a \end{pmatrix}, \text{ with } b > a$$

  - This defines a rectified Gaussian distribution with $2^{N/2}$ modes.
  - Hard to integrate efficiently if we must first find the modes in a high-dimensional space.
- Use of Mean-Field Approximation
  - Allows analytic approximate solutions for **A,b**
  - Fundamental assumption - *factorisability* of the model density

$$q(\mathbf{x}) = \prod_i \frac{1}{\tau_i} \exp\left(-\frac{x_i}{\tau_i}\right)$$

  - Factorised fit density is *unimodal*

Learning Models for Continuous Nonnegative Data
13

Enabling the *future* through innovation. Microsoft Research

Practical Implementation

- Markov Chain Monte-Carlo Sampling
  - Problem is to generate a set of 'data' representative of the topography of the probability density.
  - Develop dynamics satisfying *detailed balance*, s.t. evolving random points asymptotically produces samples characterising the probability density

$$\text{Detailed balance; } T(\mathbf{x}_1 \mid \mathbf{x}_0) P(\mathbf{x}_0) = T(\mathbf{x}_0 \mid \mathbf{x}_1) P(\mathbf{x}_1)$$

  - For NNBD
    - Gaussianity of conditional densities
    - Quadratic $L(\mathbf{x})$

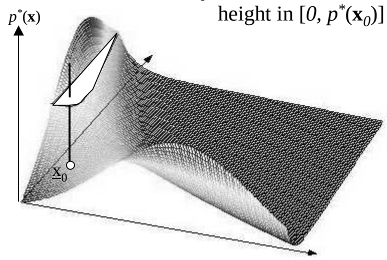
Learning Models for Continuous Nonnegative Data
14

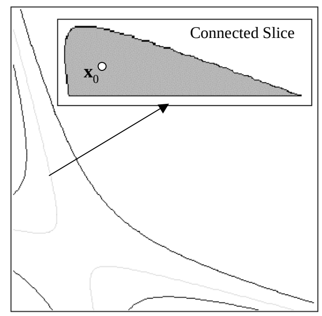
Enabling the *future* through innovation. Microsoft Research

MCMC for the NNBM

Reflective Slice Sampling (Neal 1997)

From starting point, $\mathbf{x}_0$ pick slice at random height in $[0, p^*(\mathbf{x}_0)]$

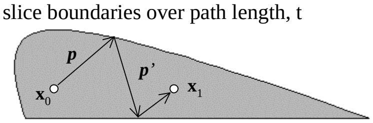




$\mathbf{g}$ = gradient of contour boundary

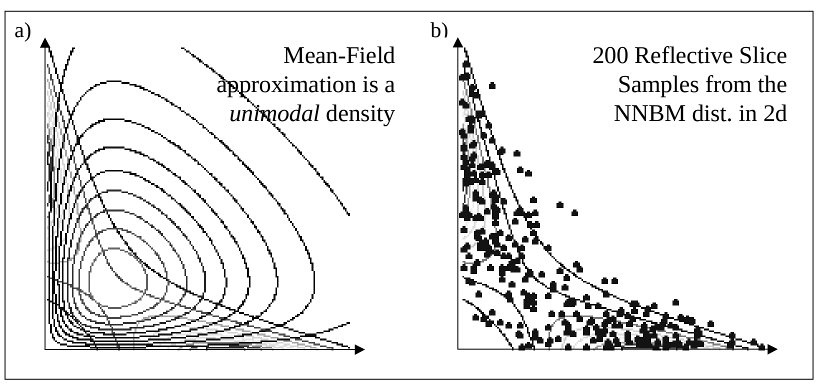
$$\mathbf{p}' = \mathbf{p} - 2\mathbf{g}\frac{\mathbf{p}\cdot\mathbf{g}}{|\mathbf{g}|^2}$$

Choose random momentum, $\boldsymbol{p}$ bouncing randomly from slice boundaries over path length, t



Learning Models for Continuous Nonnegative Data
15

MFT and Monte-Carlo Comparison



Learning Models for Continuous Nonnegative Data
16

Advanced Mean Field Theory for the NNBM

- The nature of the ML learning rule, suggests parameters *are* expressible as $f\left(\langle \mathbf{V}\rangle, \langle \mathbf{V}\mathbf{V}^T\rangle\right)$
- Confident some *useful* deterministic theory exists
- High Temperature Expansions (Georges & Yedidia 1995)
  - Consider the 'Free Energy'

$$-\beta G(\beta,\mathbf{m}) = \ln \prod_{k=1}^{N}\int_0^\infty dx_k \exp\left(-\beta\sum_{i,j} x_i A_{ij} x_j - \sum_i \lambda_i(\beta)(x_i - m_i)\right)$$

  - Compare this with Z the normalisation of the NNBD

$$\ln Z = -\beta G(\beta,\mathbf{m}) + \text{Constant}$$
$$\text{with } b_i = -\lambda_i(\beta)$$
$$\text{and } m_i = \left\langle V_i^{(\alpha)}\right\rangle_\alpha$$

Learning Models for Continuous Nonnegative Data
17

Enabling the *future* through innovation. Microsoft Research

Advanced Mean Field Theory for the NNBM

- The term $\sum_i \lambda_i(\beta)(x_i - m_i)$ enforces the Lagrange constraint that the 'mean-field' of the system, matches the 'mean-field' of the patterns (cf. 2nd part of NNBM learning rule) $\left\langle \mathbf{V}^{(\alpha)}\right\rangle_\alpha$
- In high-temperature limit (β=0) the local interactions in the system vanish
- We solve for the Lagrange multipliers by explicitly forcing the constraint

$$m_i = \langle x_i\rangle_{free} = \langle x_i\rangle_{free,\beta=0} = \frac{\prod_k \int_0^\infty dx_k x_i \exp\left(-\sum_j \lambda_j(\beta)(x_j - m_j)\right)}{\prod_k \int_0^\infty dx_k \exp\left(-\sum_j \lambda_j(\beta)(x_j - m_j)\right)} = \frac{1}{\lambda_i(0)}$$

- We consider Taylor expansion of -βG about β=0

$$-\beta G(\beta,\mathbf{m}) \approx -\beta G(0,\mathbf{m}) + \beta \left.\frac{\partial(-\beta G)}{\partial \beta}\right|_{\beta=0} + \frac{\beta^2}{2}\left.\frac{\partial(-\beta G)}{\partial \beta}\right|_{\beta=0} + \ldots$$

Learning Models for Continuous Nonnegative Data
18

Enabling the *future* through innovation. Microsoft Research

Advanced Mean Field Theory for the NNBM

- The first derivative, reproduces the naive MFT

$$\left.\frac{\partial(-\beta G)}{\partial\beta}\right|_{\beta=0} = \frac{\prod_k \int_0^\infty dx_k \left(-\sum_{i,j} A_{ij}x_i x_j - \sum_i (x_i - m_i)\frac{\partial\lambda}{\partial\beta}\right)\exp\left(-\sum_l \lambda_l(\beta)(x_i - m_i)\right)}{\prod_k \int_0^\infty dx_k \exp\left(-\sum_l \lambda_l(\beta)(x_i - m_i)\right)}$$
$$= \sum_{i,j}(1+\delta_{ij})A_{ij}m_i m_j$$

- The second derivative is of the same order as the TAP-Onsager term for the Ising model/Boltzmann Machine (Kappen & Rodriguez, 1997)

$$\left.\frac{\partial^2(-\beta G)}{\partial\beta^2}\right|_{\beta=0} = \left\langle\left(\sum_{i,j} A_{ij}x_i x_j\right)^2\right\rangle_{\beta=0} - \left(\sum_{i,j}(1+\delta_{ij})A_{ij}m_i m_j\right)^2$$
$$= \sum_{i,j,k,l}\alpha_{ijkl}A_{ij}A_{kl}m_i m_j m_k m_l$$

Learning Models for Continuous Nonnegative Data
19

Enabling the *future* through innovation.
Microsoft Research

Advanced Mean Field Theory for the NNBM

- We are interested in approximating

$$\langle x_i x_j\rangle_{free} = -\frac{\partial \ln Z}{\partial A_{ij}} = -\frac{\partial(-\beta G)}{\partial A_{ij}}$$

- From the high-temperature expansion, we obtain

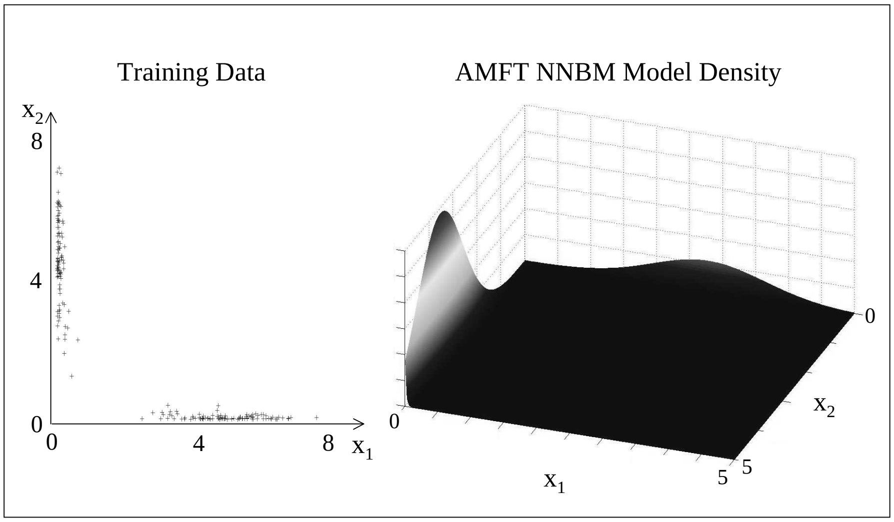
$$\langle x_i x_j\rangle_{free} \approx (1+\delta_{ij})m_i m_j - \beta\sum_{k,l}\alpha_{ijkl}A_{kl}m_k m_l m_i m_j$$

- We can substitute this into the learning rule, and either
  - solve directly for **A** setting ΔA=0
  - Iterate the learning rule

Learning Models for Continuous Nonnegative Data
20

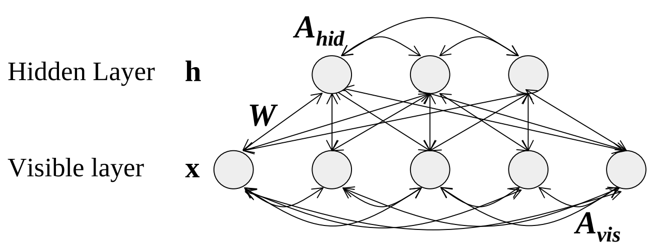
Enabling the *future* through innovation.
Microsoft Research

Advanced Mean Field Theory for the NNBM



The Mean Field NNBM with Hidden Units

- With the Stochastic learning rule, we have to sample for both the clamped and free learning phases
- Using the Mean Field theory, we can totally avoid sampling, after some simple assumptions



$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{vis} & \mathbf{W} \\ \mathbf{W}^{-1} & \mathbf{A}_{hid} \end{pmatrix}$$

The Mean Field NNBM with Hidden Units

- Recall the learning rule

$$\Delta\mathbf{A} = \eta_A \frac{\partial L}{\partial \mathbf{A}} = \eta_A \left[\left\langle \mathbf{V}^{(\alpha)} \mathbf{V}^{(\alpha)^T} \right\rangle_\alpha - \left\langle \mathbf{x}\mathbf{x}^T \right\rangle_{free} \right]$$

- With hidden units we have

$$\Delta\mathbf{A} = \left\langle \begin{matrix} \mathbf{V}^{(\alpha)}\mathbf{V}^{(\alpha)^T} & \mathbf{V}^{(\alpha)}\mathbf{h}^T \\ \mathbf{h}\mathbf{V}^{(\alpha)^T} & \mathbf{h}\mathbf{h}^T \end{matrix} \right\rangle_{\alpha \text{ patterns}, \mathbf{h} \text{ free}} - \left\langle \begin{matrix} \mathbf{x}\mathbf{x}^T & \mathbf{x}\mathbf{h}^T \\ \mathbf{h}\mathbf{x}^T & \mathbf{h}\mathbf{h}^T \end{matrix} \right\rangle_{\text{free}}$$

- For the mean-field case we assume
  - The hidden units form a *distributed* representation of the data - hence their mean-fields are equal and constant in the 'free' phase
  - In the 'clamped' phase the hidden units are restrained by the clamping of the visible ones, and explore an ensemble adequately characterised by 'backpropagation' of the patterns

$$\left\langle \mathbf{h}\mathbf{h}^T \right\rangle_{\text{clamped}} \approx \left\langle \left(\mathbf{W}^{-1}\mathbf{V}^{(\alpha)}\right)\left(\mathbf{W}^{-1}\mathbf{V}^{(\alpha)}\right)^T \right\rangle_{\alpha \text{ patterns}}, \left\langle \mathbf{h}\mathbf{V}^{(\alpha)} \right\rangle_{\alpha \text{ patterns}} \approx \left\langle \left(\mathbf{W}^{-1}\mathbf{V}^{(\alpha)}\right)\mathbf{V}^{(\alpha)^T} \right\rangle_{\alpha \text{ patterns}}$$

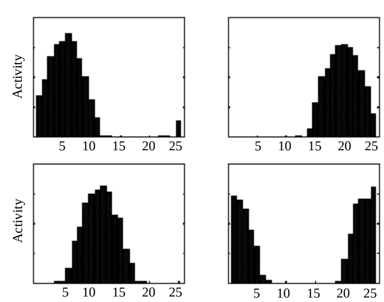
Enabling the ***future*** through innovation.
Microsoft Research

Orientation Tuning in Visual Cortex

- Model of Ben-Yishai et al. (1995) has parameters which map directly onto a NNBD

$$A_{ij} = \delta_{ij} + \frac{1}{N} - \frac{4}{N}\cos\left(\theta_i - \theta_j\right), b_i = 1$$

- Trained on stable activation patterns of their model, NNBM captures parameter set for the data, and can generate valid activity patterns

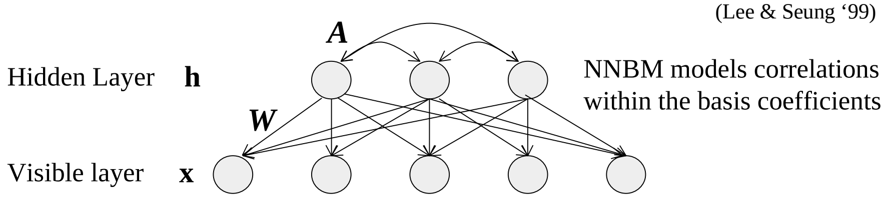


Enabling the ***future*** through innovation.
Microsoft Research

Hierarchical Nonnegative Models

- We build a hierarchical nonnegative computing structure by Combining NNBM with *Nonnegative Matrix Factorisation*

(Lee & Seung '99)



- NMF learns a *nonnegative* factorisation of the dataset such that

$$\mathbf{Wh} = \mathbf{x}$$

  - Nonnegative coefficients, nonnegative basis
  - Basis of 'qualitative features' of the data (cf. PCA/SVD, VQ)
- Use NNBM on the set of hidden layer coefficients, to learn the valid coactivations of basis features (columns of **W**)

Enabling the *future* through innovation.
Research

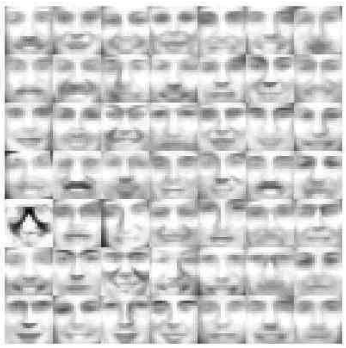
Hierarchical Nonnegative Models

- NMF Basis for database of face images

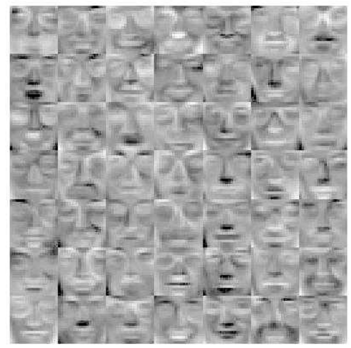
Enabling the *future* through innovation.
Research

Hierarchical Nonnegative Models

- Compare NNBM performance, with (unrectified) Gaussian in learning correlations between NMF coefficients
- Sampling from the NNBM model generates more plausible prototypes, as compared with samples from the unimodal Gaussian fit to the data.



a) Samples from NNBM distribution, trained on human faces.



b) Samples from Gaussian Fit to the data.

Summary

- View of Boltzmann Machines as learning a probabilistic model for binary data
- Bounding to the nonnegative orthant makes drastic changes to the maximum entropy model for the data - NNBD
- The NNBM describes the framework for maximum-likelihood learning of the NNBD for a given set of data
- Efficient (but still slow) MCMC can be implemented to sample the NNBD, Naive Mean-Field theory fails
- The high-temperature expansion allows corrections of arbitrarily high order and may lead to deterministic NNBM learning
- Under simple assumptions the AMFT can be applied to computation with hidden units
- Work continues in transforming the NNBM into a useable and interpretable model for continuous nonnegative data

Future Work

- High-dimensional application using advanced mean-field theory - inferring super-resolution images
- Interpretable nonnegative feature extraction under the NNBD as probabilistic model (cf. Gaussian model for PCA)
- Application with Hidden Units
- Linear Response Theory (with Bert Kappen)
- Mixed-constraint latent variable models (with David Barber)
  - Additive combination of parts generative models for continuous data
  - Gaussian visible units + NNBD over hidden units (application to DNA Array data?)
- Learning models via continuous-time dynamics

Learning Models for Continuous Nonnegative Data
29

Enabling the *future* through innovation. Research

References

- Hinton & Sejnowski - *Optimal perceptual learning*, IEEE Conference on Vision and Pattern Recognition, Washington D.C. (1983)
- Ackley, Hinton & Sejnowski - *A learning algorithm for Boltzmann machines,* Cognitive Science **9**:147-169 (1985)
- Lee & Seung - *Learning the parts of objects by nonnegative matrix factorization,* Nature, **401,** 788-791 (1999)
- Downs, MacKay & Lee - *The Nonnegative Boltzmann Machine*, Advances in Neural Information Processing Systems, **12** (2000) *(to appear)*
- Georges & Yedidia - *How to expand around mean field theory using high-temperature expansion,* Journal of Physics A, **24**, 2173-2192 (1991)

Learning Models for Continuous Nonnegative Data
30

Enabling the *future* through innovation. Research