Unweaving massive data with sparse coding

Djork-Arné Clevert

Institute of Bioinformatics, Johannes Kepler University Linz

Boston, July 11th, 2014
Objectives

- What is a sparse coding?
- What are some common sparse coding models and what are their fields of application?
- What is the role of assumptions in the data generation process?
- How can domain knowledge be incorporated in order to build an appropriate model?
Sparse coding

Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1?

BRUNO A. OLSHAUSEN,‡ DAVID J. FIELD†

Received 16 July 1996; in revised form 24 December 1996

Given a large set of input patterns, sparse coding attempt to automatically find a small number of representative patterns which reproduce the original input patterns.
Sparse coding and linear generative models

- Some sources 'drive' the model and produce an output distribution which should best match the observed data distribution.
- Observations $x$ are random variables whose distribution depends on model parameters $\theta$.
- Nice advantages of generative models are:
  - they select models using sound model selection techniques such as maximum likelihood or maximum a posterior,
  - signal-to-noise ratios can be computed,
  - they can be compared with each other via the likelihood or posterior,
  - they produce a global model to explain all data.
Some applications

Factor analysis for

- quantitative gene expression analysis
  - Factor Analysis for Robust Microarray Summarization (FARMS)
- discovering copy number variations in array data
  - copy number FARMS (cn.FARMS)
- identifying bicluster in a data matrix
  - Factor Analysis for Bicluster Acquisition (FABIA)
Microarray
Facts and assumptions

- Gene measured by different probes
- Goal: summarize probe intensities to an expression value
- Noise-free probes are correlated
- Replicate probe intensities are Gaussian distributed
The probe level data
FARMS: The idea
Factor analysis

\[ x = \lambda z + \epsilon = \sum_{i=1}^{p} \lambda_i z_i^T + \epsilon \]

- linear generative model
- \( x \in \mathbb{R}^n \) are the observations
- \( z \in \mathbb{R}^p \) are independent \( \mathcal{N}(0,1) \) Gaussian factor
- \( \epsilon \) is independent \( \mathcal{N}(0,\Psi) \) Gaussian noise
- \( x \) is Gaussian with
  \[ P(x) = \int P(z) P(x | z) dx = \mathcal{N}(0,\lambda\lambda^t + \Psi) \]
- where \( \lambda \in \mathbb{R}^{n \times p} \) is the factor loading matrix,
- and the noise covariance matrix \( \Psi \in \mathbb{R}^{p \times p} \) is diagonal
MA-plot MAS5.0 vs. FA

The MA plot shows log fold change as a function of mean log expression level.
Facts and assumptions II

- Gene measured by different probes
- Goal: summarize probe intensities to an expression value
- Noise-free probes are positively correlated
  - Variable probe qualities
  - High quality probes are linear dependent
- Replicate probe intensities are Gaussian distributed

Higher mRNA concentration → larger intensities
FARMS: Bayes framework

Posterior

\[ p(\lambda, \Psi \mid \{x\}) \propto p(\{x\} \mid \lambda, \Psi) \, p(\lambda) \]

Prior knowledge

- Positive \( \lambda \) ensure positive probe correlation
- Most genes show no or small signal (large signals are of interest in a study)

Rectified Gaussian

\[ \lambda_j = \max\{y_j, 0\} \text{ with } y_j \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda) \]
FARMS: EM updates

E-step:

\[ E_{z_i|x_i}(z_i) = \mu_{z_i|x_i} \quad \text{and} \quad E_{z_i|x_i}(z_i^2) = \mu_{z_i|x_i}^2 + \sigma_{z_i|x_i}^2 \]

M-step:

\[
\lambda_j^{\text{Gauss}} = \left( \frac{1}{N} \sum_{i=1}^{N} x_{ij} E_{z_i|x_i}(z_i) + \frac{1}{N} \frac{\mu_{\lambda} \psi_{jj}^{\text{old}}}{\sigma^2_{\lambda}} \right) \left( \frac{1}{N} \sum_{i=1}^{N} E_{z_i|x_i}(z_i^2) + \frac{1}{N} \frac{\psi_{jj}^{\text{old}}}{\sigma^2_{\lambda}} \right)^{-1}
\]

\[
\lambda_j^{\text{new}} = \begin{cases} 
\lambda_j^{\text{Gauss}} & \text{for} \quad \lambda_j^{\text{Gauss}} > 0 \\
0 & \text{for} \quad \lambda_j^{\text{Gauss}} \leq 0 
\end{cases}
\]

\[
\psi_{jj}^{\text{new}} = \left[ \text{diagvect} \left( \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \right) \right]_j - \lambda_j^{\text{new}} \left[ \frac{1}{N} \sum_{i=1}^{N} E_{z_i|x_i}(z_i^2) x_i \right]_j + \frac{1}{N} \frac{\psi_{jj}^{\text{old}}}{\sigma^2_{\lambda}} \lambda_j^{\text{new}} (\mu_{\lambda} - \lambda_j^{\text{new}})
\]
MA-plot FA vs. FARMS

The MA plot shows log fold change as a function of mean log expression level.
FARMS: Filtering by signal variance
**FARMS: z-posterior**

**Variance of $z \mid x$**

Model

$$x = \lambda z + \epsilon$$

and Gaussian $z$-prior $\mathcal{N}(0,1)$ results in the $z$-posterior $p(z \mid x)$:

$$z \mid x \sim \mathcal{N}\left(\mu_{z \mid x}, \sigma^2_{z \mid x}\right)$$

$$\mu_{z \mid x} = (x)^T \Psi^{-1} \lambda \left(1 + \lambda^T \Psi^{-1} \lambda\right)^{-1}$$

$$\sigma^2_{z \mid x} = \left(1 + \lambda^T \Psi^{-1} \lambda\right)^{-1}$$
The variance of $z$ is decomposed into a signal and a noise part:

$$1 = \text{var}(z) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{z_i|x_i} \left( z_i^2 \right) = \frac{1}{N} \sum_{i=1}^{N} \left( \mu_{z_i|x_i}^{2} + \sigma_{z_i|x_i}^{2} \right)$$

$$\frac{1}{N} \sum_{i=1}^{N} \sigma_{z_i|x_i}^{2} = 1 - \frac{1}{N} \sum_{i=1}^{N} \mu_{z_i|x_i}^{2}$$

$$\sigma_{z|x}^{2} = 1 - \frac{1}{N} \sum_{i=1}^{N} \mu_{z_i|x_i}^{2} = \left( 1 + \lambda^T \Psi^{-1} \lambda \right)^{-1}$$

$\sigma_{z|x}^{2}$ is called the "Informative/NonInformative (I/NI) call" and is one minus the signal variance. We see that large $\lambda$ (going with low noise $\Psi$) leads to low variance of $z|x$ which means a precise conditional $z$. 
FARMS: Independent I/NI calls filtering

Independent filtering increases detection power for high-throughput experiments

Richard Bourgon\textsuperscript{a}, Robert Gentleman\textsuperscript{b}, and Wolfgang Huber\textsuperscript{c,1}

\textsuperscript{a}European Bioinformatics Institute, Cambridge CB10 1SD, United Kingdom; \textsuperscript{b}Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080-4990; and \textsuperscript{c}European Molecular Biology Laboratory, 69117 Heidelberg, Germany

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved March 22, 2010 (received for review December 3, 2009)

- For permutation invariant test statistics and for the $t$-test statistic $T$ (only for Gaussian $z$-prior), the I/NI call filter applied to null hypotheses is independent of the statistic.
- This guarantees type I error rate control if first filtering by I/NI calls, then using these statistics, and finally applying correction for multiple testing.
FARMS: I/NI calls distribution

Bimodal distribution
- Enforced by the parameter prior
- Modes clearly separated (insensitive for filtering threshold)
- Works for unbalanced data (few samples contain a signal) in contrast to variance filtering (Bourgon et al. (2010))
- Works for few genes with a signal
A pipeline for gene expression analysis

Probe level data

Background correction
- RMA
- MAS 5.0
- None

Normalisation
- Quantilen
- Cyclic Loess
- Constant
- VSN

PM correction
- PM only
- PM-MM
- IM

FARMS
- Medianpolish
- Tukey Bi-Weight
- LiWong
- AverageDiff

Summarisation

Expression level
### Receiver Operator Characteristics (ROC)

#### Affycomp II / GoldenSpike Benchmark (AUC - area under the curve):

<table>
<thead>
<tr>
<th>INTENSITY</th>
<th>FARMS</th>
<th>RMA</th>
<th>GCRMA</th>
<th>MAS 5.0</th>
<th>MBEI</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOW</td>
<td>0.94</td>
<td>0.51</td>
<td>0.62</td>
<td>0.07</td>
<td>0.21</td>
</tr>
<tr>
<td>MED</td>
<td>0.99</td>
<td>0.91</td>
<td>0.94</td>
<td>0.00</td>
<td>0.43</td>
</tr>
<tr>
<td>HIGH</td>
<td>1.00</td>
<td>0.64</td>
<td>0.59</td>
<td>0.00</td>
<td>0.16</td>
</tr>
<tr>
<td>MEAN</td>
<td>0.95</td>
<td>0.60</td>
<td>0.69</td>
<td>0.05</td>
<td>0.26</td>
</tr>
<tr>
<td>LOW</td>
<td>0.91</td>
<td>0.57</td>
<td>0.45</td>
<td>0.09</td>
<td>-</td>
</tr>
<tr>
<td>MED</td>
<td>1.00</td>
<td>0.91</td>
<td>0.91</td>
<td>0.00</td>
<td>-</td>
</tr>
<tr>
<td>HIGH</td>
<td>0.98</td>
<td>0.96</td>
<td>0.92</td>
<td>0.00</td>
<td>-</td>
</tr>
<tr>
<td>MEAN</td>
<td>0.93</td>
<td>0.65</td>
<td>0.57</td>
<td>0.06</td>
<td>-</td>
</tr>
<tr>
<td>GoldenSpike</td>
<td>0.85</td>
<td>0.76</td>
<td>0.78</td>
<td>0.28</td>
<td>0.39</td>
</tr>
</tbody>
</table>

#### Computational costs for processing 60 arrays

<table>
<thead>
<tr>
<th>COMPUTATIONAL time [s]</th>
<th>FARMS</th>
<th>RMA</th>
<th>MAS 5.0</th>
<th>MBEI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>92</td>
<td>384</td>
<td>851</td>
<td>591</td>
</tr>
</tbody>
</table>
Results I/NI call

- Leads on average to 84 (±1.5)% exclusion rate
  - Applied on 30 real life studies
  - A/P calls excluded only 33 (±1)%
- Validation was carried out on spiked-in data:

Exclusion rate on spiked-in data sets:

<table>
<thead>
<tr>
<th></th>
<th>INFORMATIVE</th>
<th>NON-INFORMATIVE</th>
<th>EXCLUSION RATE</th>
<th>DETECTED SPIKED-INS</th>
<th>DETECTED PSEUDO SPIKED-INS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HGU133A</td>
<td>81</td>
<td>22219</td>
<td>99.63%</td>
<td>42/42</td>
<td>28/28*</td>
</tr>
<tr>
<td>HGU95_V2</td>
<td>56</td>
<td>12570</td>
<td>99.56%</td>
<td>14/14</td>
<td>5/5**</td>
</tr>
<tr>
<td>Hu. Gene 1.0 ST</td>
<td>40</td>
<td>19,753</td>
<td>99.80%</td>
<td>15/15***</td>
<td>-</td>
</tr>
</tbody>
</table>

*McGee et al. 2006; **Wolfinger and Chu 2002; Cope et al. 2004; ***long spiked-in fragments
Copy number variants

- are deletions, duplications, inversions or insertions of chromosomal segments
- are a major source of variation between individual humans
- are an underlying factor in human evolution and in many diseases
- form at a faster rate than other types of mutation
Rare CNV events

Sparse data

CNV data is sparse with an kurtosis larger than 30 \(\rightarrow\) change the model assumption to a Laplacian distributed hidden variable \(z\).

Gauss vs. Laplace

Close up Gauss vs. Laplace
Laplacian cn.FARMS

Data likelihood

\[ p(\{x\} | \lambda, \Psi) = \int p(\{x\} | z, \lambda, \Psi) \ p(z) \ dz \]
Laplacian cn.FARMS

Data likelihood

\[ p(\{x\} | \lambda, \Psi) = \int p(\{x\} | z, \lambda, \Psi) \ p(z) \ dz \]

Problem

- The likelihood is analytically intractable for the non-Gaussian prior
Laplacian cn.FARMS

Data likelihood

\[ p(\{x\} | \lambda, \Psi) = \int p(\{x\} | z, \lambda, \Psi) \ p(z) \ dz \]

Problem

- The **likelihood is analytically intractable** for the non-Gaussian prior

Solution

- Variational EM approach
  - Based on a local Gaussian approximation to the mode
Benchmark data sets

- 30 male and 30 female CEU founders
  - Classification task: distinguish males from females by their copy number on the X chromosome
- Evaluation on:
  - Single-locus / multi-loci classification (window mode)
  - Multi-loci summarization with
    - cn.FARMS
    - Median locus for dChip and CRMA_v2
ROC-Curve (SNP 6.0 arrays)

**TPR / FPR**

True positive rate (TPR) = TP/(TP+FN)
False positive rate (FPR) = FP/(FP+TN)
## Results cn.FARMS

<table>
<thead>
<tr>
<th>Loci</th>
<th>Criteria</th>
<th>cn.FARMS</th>
<th>CRMA_v2</th>
<th>dChip</th>
<th>cn.FARMS</th>
<th>CRMA_v2</th>
<th>dChip</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AUC</td>
<td>0.9852</td>
<td>0.9820</td>
<td>0.9819</td>
<td>0.9838</td>
<td>0.9807</td>
<td>0.9721</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td>8472</td>
<td>9106</td>
<td>9018</td>
<td>56145</td>
<td>68593</td>
<td>77438</td>
</tr>
<tr>
<td></td>
<td>P-VALUE</td>
<td>–</td>
<td>1.8e-65</td>
<td>3.1e-26</td>
<td>–</td>
<td>1e-1160</td>
<td>1e-6049</td>
</tr>
<tr>
<td>2</td>
<td>AUC</td>
<td>0.9983</td>
<td>0.9974</td>
<td>0.9969</td>
<td>0.9983</td>
<td>0.9963</td>
<td>0.9894</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td>1375</td>
<td>1449</td>
<td>1611</td>
<td>9777</td>
<td>11705</td>
<td>18039</td>
</tr>
<tr>
<td></td>
<td>P-VALUE</td>
<td>–</td>
<td>2.7e-4</td>
<td>2.5e-12</td>
<td>–</td>
<td>1e-317</td>
<td>1e-3713</td>
</tr>
<tr>
<td>3</td>
<td>AUC</td>
<td>0.9998</td>
<td>0.9995</td>
<td>0.9992</td>
<td>0.9998</td>
<td>0.9990</td>
<td>0.9953</td>
</tr>
<tr>
<td></td>
<td>FP</td>
<td>240</td>
<td>366</td>
<td>440</td>
<td>1573</td>
<td>3462</td>
<td>6625</td>
</tr>
<tr>
<td></td>
<td>P-VALUE</td>
<td>–</td>
<td>2.6e-38</td>
<td>7.2e-58</td>
<td>–</td>
<td>1e-896</td>
<td>1e-3455</td>
</tr>
</tbody>
</table>

Table: AUC values at the sex classification task for 59 HapMap CEU founders based on the X chromosome copy numbers:
CNV detection benchmark

- “The International HapMap Project” phase 2 data set with Affymetrix SNP 6.0 arrays
  - Goal is to identify true rare CNV regions with a low FDR
  - “True CNV regions” are those regions which were detected and verified by different bio-technologies
    - NimbleGen tiling arrays, Agilent CGH arrays, Illumina Infinium genotyping (Human660W)
  - 2,515 true CNV regions as reference
- CNV calling criteria:
  - I/NI call for cn.FARMS
  - Variance of the raw copy numbers on the samples for dChip and CRMA_v2
CNV detection plot
CNV detection on HapMap (multi-loci 3)

Precision / Recall

Recall = TP/(TP+FN)
Precision = TP/(TP+FP) = 1 - FDR
CNV detection on HapMap (multi-loci 5)

Precision / Recall
Recall = TP/(TP+FN)
Precision = TP/(TP+FP) = 1 - FDR
ICGC copy number data sets

- Glioblastoma multiforme data sets
  - 167 Agilent 415K CGH arrays from Harvard
  - 262 Agilent 244A CGH arrays from Harvard
  - 461 Agilent 244A CGH arrays from MSKCC
  - 533 Affymetrix SNP 6.0 arrays from Broad
  - 432 Illumina HumanHap 550 from Stanford

- CN data for SNP 6.0 and HumanHap 550 were not available
- 167 matched arrays HMS 415K and MSKCC 244A remain
Merged raw data (Chromosome 1)
Prior weight 2.0
Biclustering applications

Definition: Biclustering simultaneously organizes a data matrix into subsets of rows and columns in which the entities of each row subset are similar to each other on the column subset and vice versa.

- Gene expression $\Rightarrow$ columns tissues, rows genes
  - compounds that trigger the same pathway
  - tightly co-expressed gene sets in subgroups of cancer, e.g. patients with bad treatment outcome
- Bioassays $\Rightarrow$ columns compounds, rows bioassay activity
  - compounds that are active on similar targets
- Structural fingerprints $\Rightarrow$ columns compounds, rows chemical fingerprints
  - compounds that share a chemical substructure
Biclustering: The idea
FABIA: The model I

factor $z$

loading matrix $\lambda$

observations $x$

noise $\epsilon$
FABIA: The model II

\[ x = \lambda z + \epsilon = \sum_{i=1}^{p} \lambda_i z_i^T + \epsilon \]

- \( x \) are the observations
- \( \lambda \) is the matrix of factor loadings
- \( z = (z_1, \ldots, z_p)^T \) is the factor matrix
- \( p \) number of biclusters
- \( \lambda_i \in \mathbb{R}^n \) is the sparse prototype vector of the \( i \)-th bicluster
- \( z_i \in \mathbb{R}^l \) is the sparse vector of factors of the \( i \)-th bicluster
- \( \epsilon \in \mathbb{R}^{n \times l} \) is independent additive noise \( \mathcal{N}(0, \Psi) \)-distributed
FABIA: Bayes framework

Loading prior

- Sparseness on the loadings
- Laplace prior
- \( p(\lambda_i) = \left(\frac{1}{\sqrt{2}}\right)^n \prod_{j=1}^{n} e^{-\sqrt{2}|\lambda_{ji}|} \)
FABIA: Bayes framework

**Loading prior**
- Sparseness on the loadings
- Laplace prior
  \[ p(\lambda_i) = \left( \frac{1}{\sqrt{2}} \right)^n \prod_{j=1}^n e^{-\sqrt{2}|\lambda_{ji}|} \]

**Factor prior**
- Sparseness on the factor
- Laplace prior
  \[ p(z) = \left( \frac{1}{\sqrt{2}} \right)^p \prod_{i=1}^p e^{-\sqrt{2}|z_i|} \]
FABIA: Bayes framework

Loading prior
- Sparseness on the loadings
- Laplace prior
  \[ p(\lambda_i) = \left( \frac{1}{\sqrt{2}} \right)^n \prod_{j=1}^{n} e^{-\sqrt{2}|\lambda_{ji}|} \]

Factor prior
- Sparseness on the factor
- Laplace prior
  \[ p(z) = \left( \frac{1}{\sqrt{2}} \right)^p \prod_{i=1}^{p} e^{-\sqrt{2}|z_i|} \]

Problem
Laplace prior on factors leads to intractable likelihood:

\[ p(x | \lambda, \Psi) = \int p(x | z, \lambda, \Psi) \ p(z) \ dz \]

Solution: Prior on factors is replaced by maximum of a Gaussian function family \( \Rightarrow \) variational approach

\[ p(z) \approx \arg \max_{\xi} p(z|\xi) \]
FABIA: Variational EM updates

E-step:

\[ E(\hat{z}_j \mid x_j) = (\lambda^T \Psi^{-1} \lambda + \Xi_j^{-1})^{-1} \lambda^T \Psi^{-1} x_j \quad \text{and} \]
\[ E(\hat{z}_j \hat{z}_j^T \mid x_j) = (\lambda^T \Psi^{-1} \lambda + \Xi_j^{-1})^{-1} + E(\hat{z}_j \mid x_j) E(\hat{z}_j \mid x_j)^T \quad \text{where} \]
\[ \Xi_j = \text{diag}(\text{diagvect}(\sqrt{E(\hat{z}_j \hat{z}_j^T \mid x_j)})) \] is the update for the variational parameter.
FABIA: Variational EM updates

E-step:

\[ E(\tilde{z}_j | x_j) = (\lambda^T \Psi^{-1} \lambda + \Xi_j^{-1})^{-1} \lambda^T \Psi^{-1} x_j \quad \text{and} \]
\[ E(\tilde{z}_j \tilde{z}_j^T | x_j) = (\lambda^T \Psi^{-1} \lambda + \Xi_j^{-1})^{-1} + E(\tilde{z}_j | x_j) E(\tilde{z}_j | x_j)^T \]

where \( \Xi_j = \text{diag} (\text{diagvect}(\sqrt{E(\tilde{z}_j \tilde{z}_j^T | x_j)})) \) is the update for the variational parameter.

M-step:

\[
\lambda^{\text{new}} = \frac{1}{l} \sum_{j=1}^{l} x_j E(\tilde{z}_j | x_j)^T - \frac{\alpha}{l} \Psi \text{sign}(\lambda) \\
\frac{1}{l} \sum_{j=1}^{l} E(\tilde{z}_j \tilde{z}_j^T | x_j)
\]

\[
\Psi^{\text{new}} = \Psi^{\text{EM}} + \text{diag} (\text{diagvect}(\frac{\alpha}{l} \Psi \text{sign}(\lambda) (\lambda^{\text{new}})^T)) , \quad \text{where}
\]

\[
\Psi^{\text{EM}} = \text{diag} (\text{diagvect}(\frac{1}{l} \sum_{j=1}^{l} x_j x_j^T - \lambda^{\text{new}} \frac{1}{l} \sum_{j=1}^{l} E(\tilde{z}_j | x_j) x_j^T)).
\]

\( \alpha \) controls the degree of sparseness (parameter of the Laplacian prior)
Biclustering of copy number variants
Biclustering of copy number variants II
Biclustering of bioassays and compounds

Matrix plot (dim 270,000 x 4,000)

Bioassay data details

- Data source: ChEMBL
- # of assays: ca. 4,000
- # of compounds: ca. 270,000
- Sparseness: ca. 1:2,000
Biclustering of bioassays and compounds

Matrix plot - close up

Bioassay data details
- Data source: ChEMBL
- # of assays: ca. 4,000
- # of compounds: ca. 270,000
- Sparseness: ca. 1:2,000
Biclustering of bioassays and compounds
Compounds of the bioassay bicluster
Biclustering of fingerprints and compounds

Matrix plot (dim $16\times 10^6 \times 1 \times 10^6$)

Bioassay data details
- Data source: ChEMBL
- # of fingerprints: ca. 16,000,000
- # of compounds: ca. 1,000,000
- Sparseness: ca. 1:150,000
Bicluster of fingerprints and compounds I
Bicluster of fingerprints and compounds I
Compounds of the fingerprint bicluster
Compounds of the fingerprint bicluster

- All compounds of this bicluster show kinase bioactivity (urokinase-type plasminogen activator)
Biclustering for recommender systems

Matrix plot (dim many x many)

Recommender

- Data source: Zalando
- # of articles: many
- # of cookies: many
- Sparseness: ca. 1:20,000
Overview 12 random selected bicluster
Top 36 articles of a bicluster
Deep Learning

- emerging machine learning technique
- multiple levels of sparse representations $\Rightarrow$ higher levels representing more abstract concepts
- Google and Facebook now apply deep learning for object recognition, image and information retrieval
- Google recently acquired the deep learning start-up DeepMind for $500M$, winning bidding against Facebook
- Nature and The New York Times covered deep learning with several articles (two front-page articles)
Networks
Sparsity by Linear-Rectified Units
Sparsity by Dropout I

Dropout algorithm

- randomly set units to zero activation
- no derivatives
Sparsity by Dropout II
Model: Rectified Factor Network
Rectified Factor Network: Newton updates

Newton-step:

\[
\begin{align*}
\mu_{h|v} &= W^T (W W^T + \Psi)^{-1} v, \\
\Sigma_{h|v} &= I - W^T (W W^T + \Psi)^{-1} W, \\
E_{h_i|v_i} (h_i) &= \mu_{h_i|v_i} \\
E_{h_i|v_i} (h_i h_i^T) &= \mu_{h_i|v_i} \mu_{h_i|v_i}^T + \Sigma_{h_i|v_i} \\
U &= \frac{1}{n} \sum_{i=1}^{n} v_i E_{h_i|v_i} (h_i) \text{ Hebb rule: input - hidden} \\
S &= \frac{1}{n} \sum_{i=1}^{n} E_{h_i|v_i} (h_i h_i^T) \\
\Delta W &= U S^{-1} - W \text{ Hebb rule: decorrelates factors} \\
\Delta \Psi_{ii} &= [C - U W^T - W U^T + W S W^T]_{ii} - \Psi_{ii}
\end{align*}
\]
**MNIST data set**

- **MNIST 28 x 28 pixel**

- **MNIST**
  - Data source: Yann LeCun
  - # of input pixel: 756
  - # of samples: 70,000
  - Classify digits
Benchmark data

Samples form the various image classification problems. (a): harder variations on the MNIST digit classification problems. (b): artificial binary classification problems.
Motivation

Receptive fields

![Various filters learnt from 1024 hidden units RFN on benchmark data set.](image)

(a) MNIST digits

(b) MNIST digits with random image background

(c) MNIST digits with random noise background

(d) convex and concave shapes

(e) tall and wide rectangular

(f) rectangular images on background images

Various filters learnt from 1024 hidden units RFN on benchmark data set.
## Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>SVM$_{rbf}$</th>
<th>DBN$_i$</th>
<th>DBN$_3$</th>
<th>SAE$_3$</th>
<th>SDAE$_3$</th>
<th>RFN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST basic</td>
<td>1.40±0.23</td>
<td>1.21±0.21</td>
<td>1.24±0.22</td>
<td>1.40±0.23</td>
<td>1.28±0.22</td>
<td>1.27±0.22</td>
</tr>
<tr>
<td>MNIST bg-rand</td>
<td>3.03±0.15</td>
<td>3.94±0.17</td>
<td>3.11±0.15</td>
<td>3.46±0.16</td>
<td>2.84±0.15</td>
<td>2.66±0.14</td>
</tr>
<tr>
<td>MNIST bg-img</td>
<td>14.58±0.31</td>
<td>9.80±0.26</td>
<td>6.73±0.22</td>
<td>11.28±0.28</td>
<td>10.30±0.27</td>
<td>7.94±0.24</td>
</tr>
<tr>
<td>MNIST rect</td>
<td>22.61±0.37</td>
<td>16.15±0.32</td>
<td>16.31±0.32</td>
<td>23.00±0.37</td>
<td>16.68±0.33</td>
<td>16.52±0.32</td>
</tr>
<tr>
<td>MNIST rect-img</td>
<td>2.15±0.13</td>
<td>4.71±0.19</td>
<td>2.60±0.14</td>
<td>2.41±0.13</td>
<td>1.99±0.12</td>
<td>0.63±0.06</td>
</tr>
<tr>
<td>MNIST convex</td>
<td>24.04±0.37</td>
<td>23.69±0.37</td>
<td>22.50±0.37</td>
<td>24.05±0.37</td>
<td>21.59±0.36</td>
<td>20.77±0.36</td>
</tr>
<tr>
<td>MNIST</td>
<td>19.13±0.34</td>
<td>19.92±0.35</td>
<td>18.63±0.34</td>
<td>18.41±0.34</td>
<td>19.06±0.34</td>
<td>16.41±0.32</td>
</tr>
</tbody>
</table>

Test error rate on all considered classification problems is reported together with a 95% confidence interval.
Predicting Drug-Target Interactions

Target prediction

- Data source: ChemBL
- Compounds: 698,425
- Targets: 1,230
- Descriptors: 43,340
- Hidden units: 16,384
- Parameters: 422,232,064
- Computation: Nvidia Tesla K40 with 2,880 CUDA GPU cores

AUC - Area under the ROC curve
Conclusion

- Sparse coding can reliably identify interesting projection in the data
- Sparse coding can be used for biclustering of high-dimensional data
- Sparse coding in drug design can help in selecting compounds with strong on-target effects and thereby helps to impute missing measurements
- Rectified linear units in combination with dropout lead to sparse representations of the data
- Rectified Factor Networks outperform all existing unsupervised deep learning methods and can be used for various problems
Open source software

- FARMS, cn.FARMS and FABIA are publicly available as Bioconductor R packages
- Software homepages:
  - http://www.bioinf.jku.at/software/farms/farms.html
  - http://www.bioinf.jku.at/software/cnfarms/cnfarms.html
  - http://www.bioinf.jku.at/software/fabia/fabia.html