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

Pergamon

PII: S0042-6989(97)00169-7

Vision Res., Vol. 37, No. 23, pp. 3311–3325, 1997

Printed in Great Britain
0042-6989/97 $17.00 + $.00

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

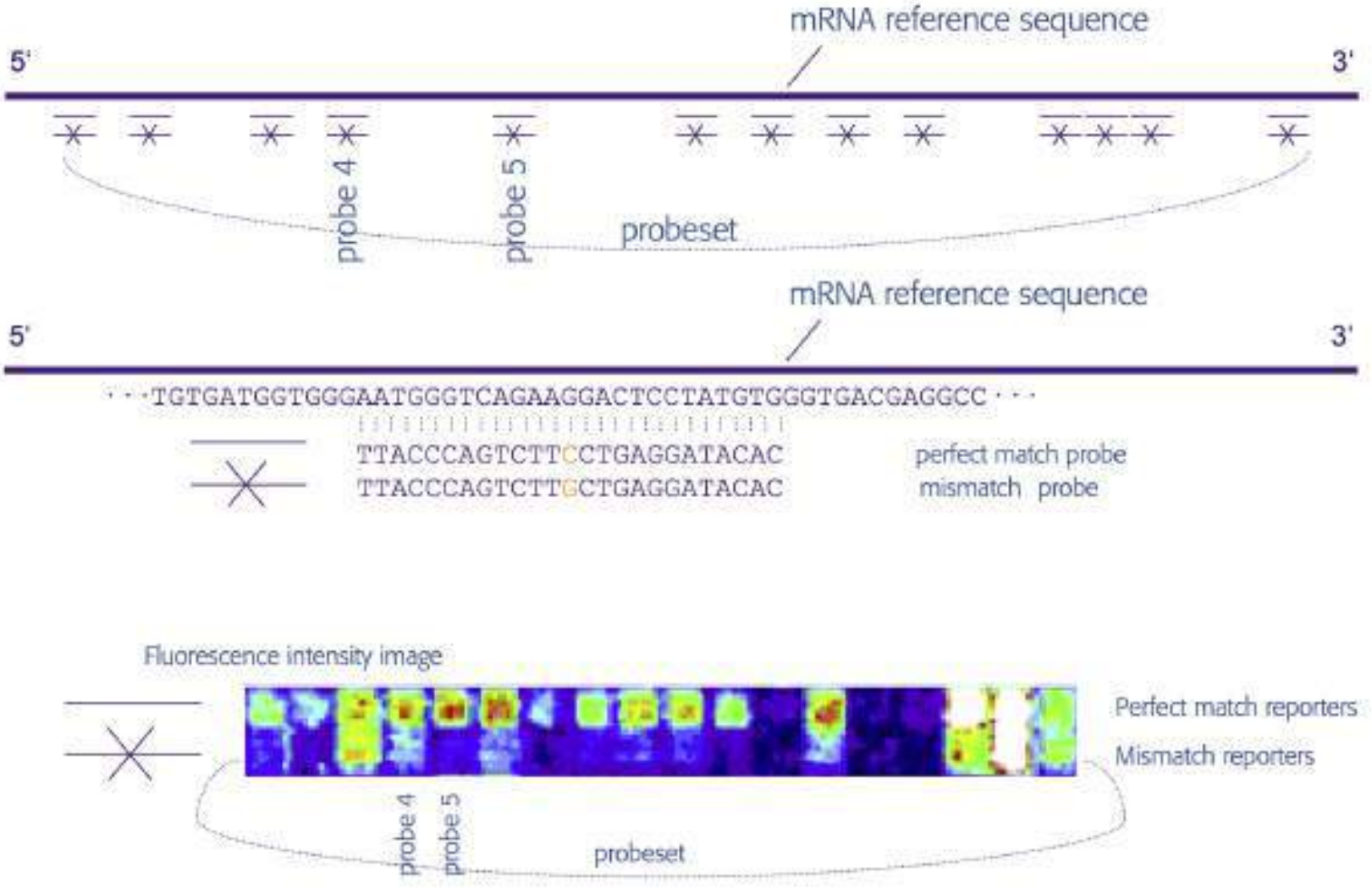
- Some sources 'drive' the model and produce an output distribution which should best match the observed data distribution
- Observations x are **random variables** whose **distribution** depends on model parameters θ
- Nice advantages of generative models are:
  - they select models using sound model 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

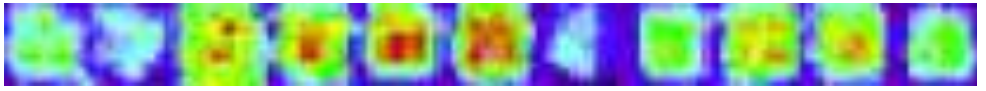
- quantitative gene expression analysis
  - Factor Analysis for Robust Microarray Summarization (FARMS)
- discovering copy number variations in array data
  - copy number FARMS (cn.FARMS)
- identifying bicluster in a data matrix
  - Factor Analysis for Bicluster Acquisition (FABIA)

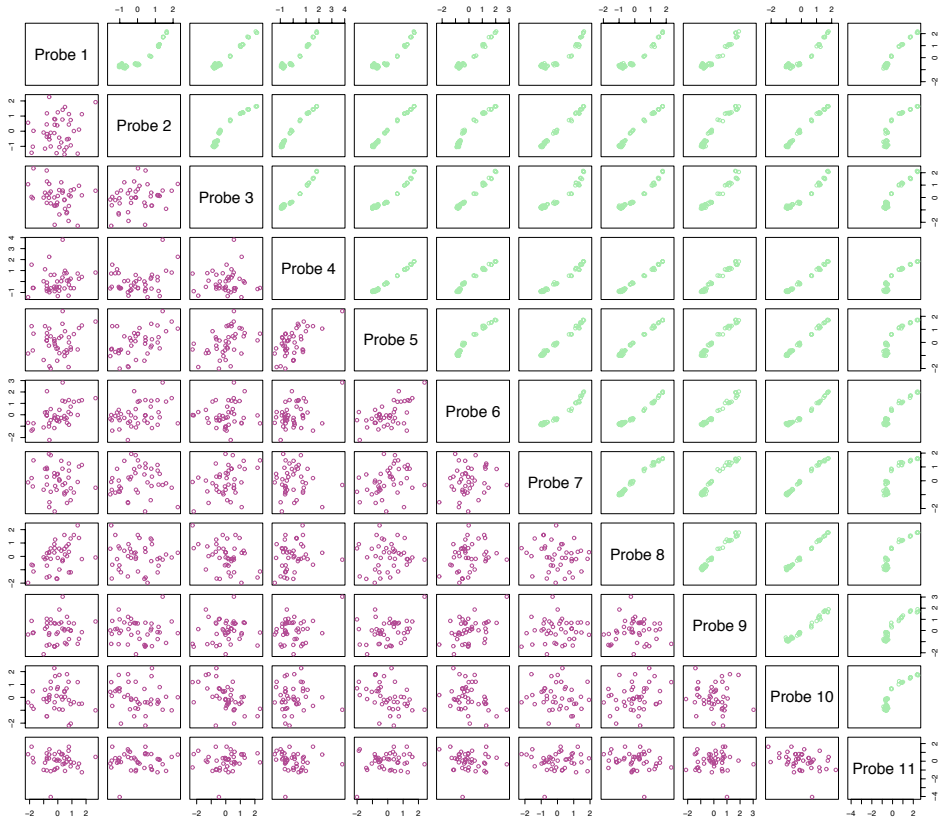
Microarray


mRNA reference sequence
5'
3'
probe 4
probe 5
probeset
mRNA reference sequence
5'
3'
· · ·TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC· · ·
TTACCCAGTCTTCCTGAGGATACAC
TTACCCAGTCTTGCTGAGGATACAC
perfect match probe
mismatch probe
Fluorescence intensity image
Perfect match reporters
Mismatch reporters
probe 4
probe 5
probeset

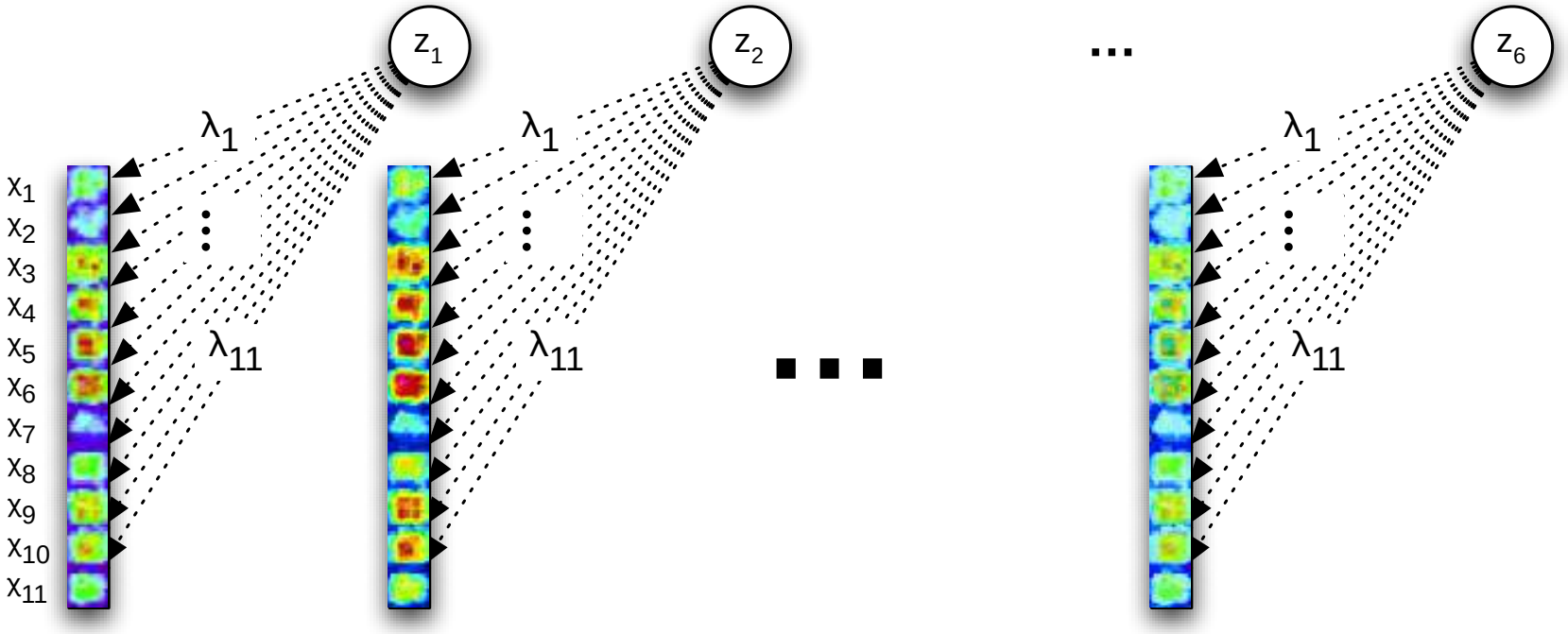
Facts and assumptions

- Gene measured by different probes
- Goal: summarize probe intensities to an expression value
- Noise-free probes are correlated
- Replicate probe intensities are Gaussian distributed

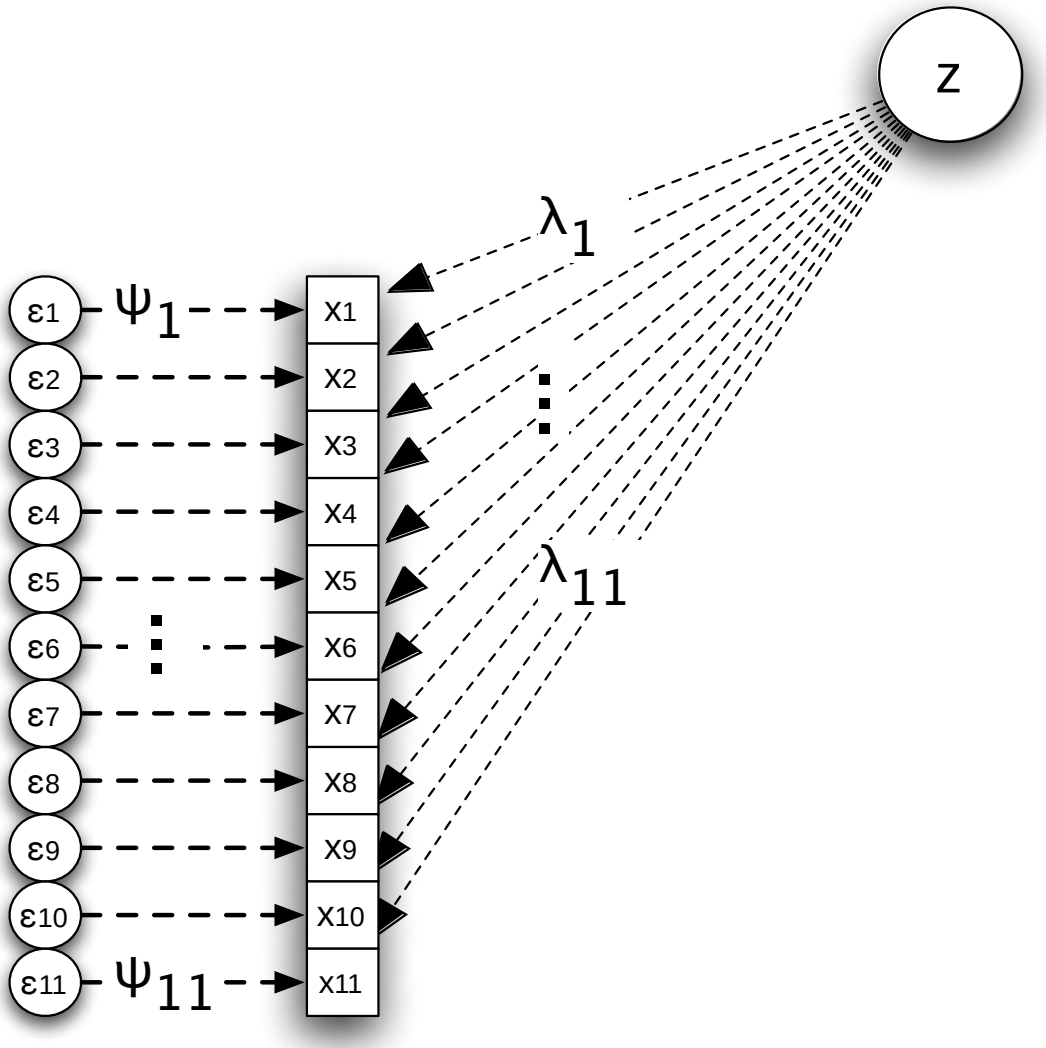




The probe level data



FARMS: The idea



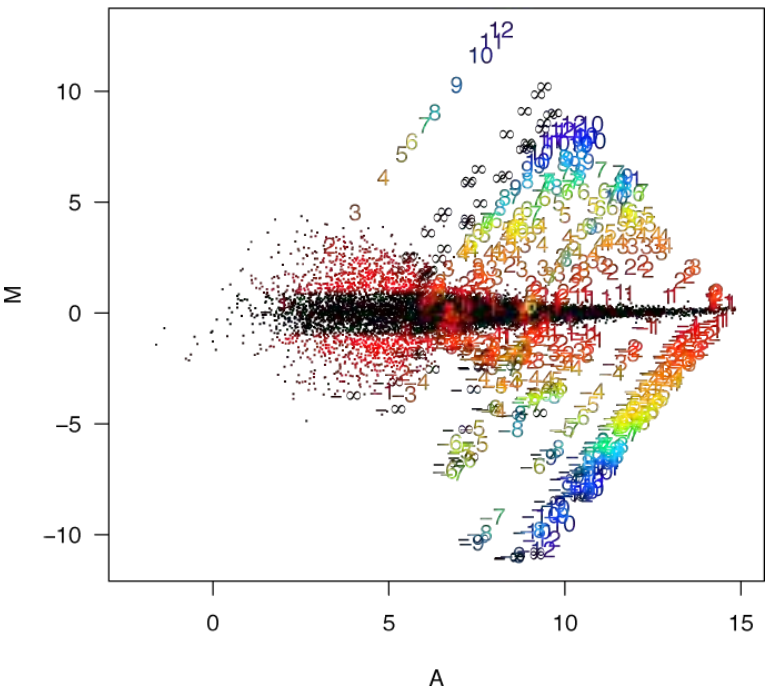
Factor analysis

$$x = \boldsymbol{\lambda} z \; + \; \boldsymbol{\epsilon} \; = \sum_{i=1}^{p} \lambda_i \; z_i^T \; + \; \boldsymbol{\epsilon}$$

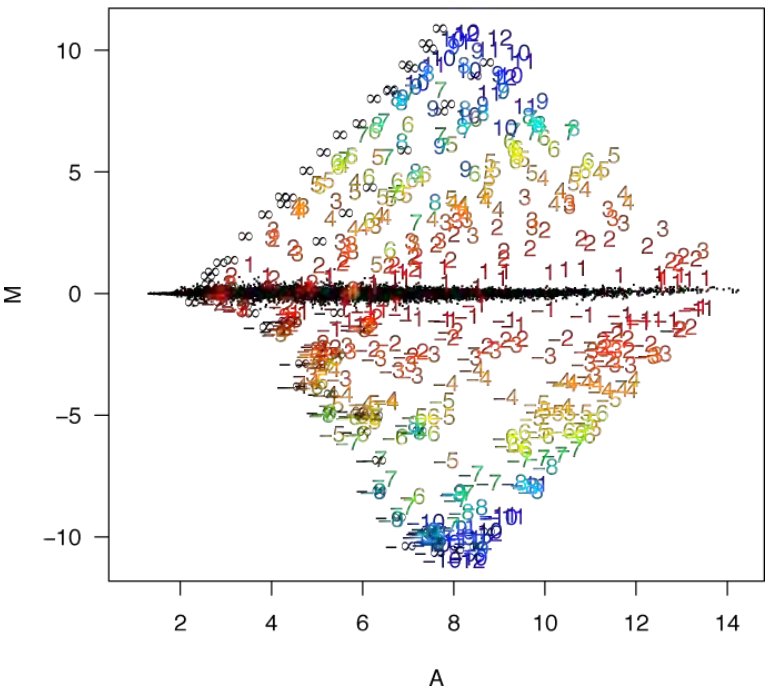
- linear generative model
- $x \in \mathbb{R}^n$ are the observations
- $z \in \mathbb{R}^p$ are independent $\mathcal{N}(0,1)$ Gaussian factor
- $\boldsymbol{\epsilon}$ is independent $\mathcal{N}(0,\Psi)$ Gaussian noise
- x is Gaussian with
  $P(x) = \int P(z)\, P(x \mid z)\, dx \; = \; \mathcal{N}\left(0, \boldsymbol{\lambda}\boldsymbol{\lambda}^t + \Psi\right)$
- where $\boldsymbol{\lambda} \in \mathbb{R}^{n \times p}$ 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

MAS 5.0



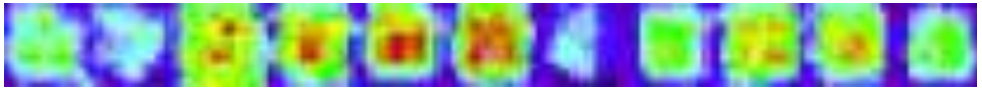
FA

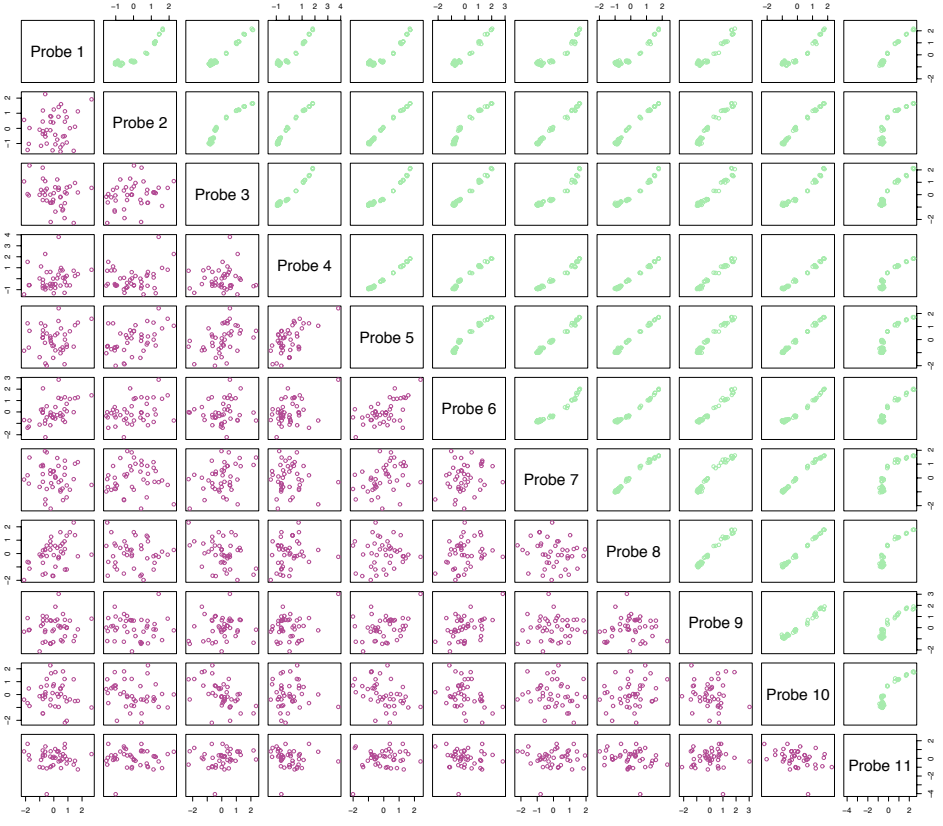


The MA plot shows log fold change as a function of mean log expression level.

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 $\rightarrow$ larger intensities

FARMS: Bayes framework

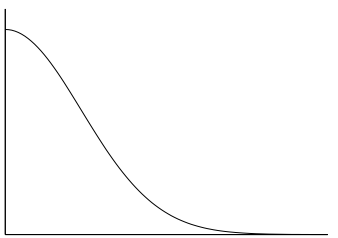
Posterior

$$p(\boldsymbol{\lambda}, \boldsymbol{\Psi} \mid \{x\}) \propto p(\{x\} \mid \boldsymbol{\lambda}, \boldsymbol{\Psi})\ p(\boldsymbol{\lambda})$$

Prior knowledge

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

Rectified Gaussian



$\lambda_j = \max\{y_j, 0\}$ with
$y_j \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda)$

FARMS: EM updates

E-step:

$$\mathrm{E}_{z_i|x_i}(z_i) = \mu_{z_i|x_i} \quad \text{and} \quad \mathrm{E}_{z_i|x_i}(z_i^2) = \mu^2_{z_i|x_i} + \sigma^2_{z_i|x_i}$$

M-step:

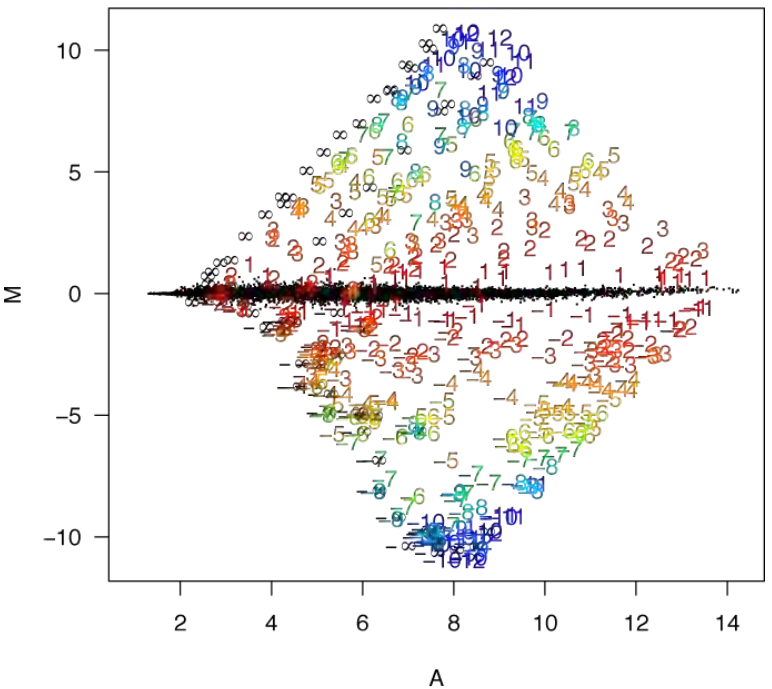
$$\lambda_j^{\text{Gauss}} = \left(\frac{1}{N}\sum_{i=1}^{N} x_{ij}\, \mathrm{E}_{z_i|x_i}(z_i) + \frac{1}{N}\frac{\mu_\lambda\, \Psi_{jj}^{\text{old}}}{\sigma_\lambda^2}\right)\left(\frac{1}{N}\sum_{i=1}^{N} \mathrm{E}_{z_i|x_i}(z_i^2) + \frac{1}{N}\frac{\Psi_{jj}^{\text{old}}}{\sigma_\lambda^2}\right)^{-1}$$

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

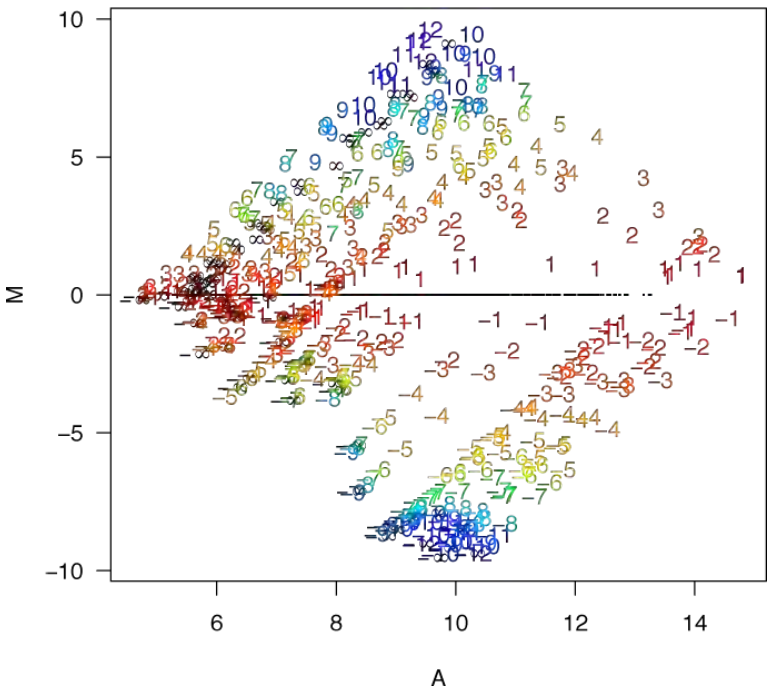
$$\Psi_{jj}^{\text{new}} = \left[\text{diagvect}\left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i^T\right)\right]_j - \lambda_j^{\text{new}}\left[\frac{1}{N}\sum_{i=1}^{N} \mathrm{E}_{z_i|x_i}(z_i)\, x_i\right]_j + \frac{1}{N}\frac{\Psi_{jj}^{\text{old}}}{\sigma_\lambda^2}\lambda_j^{\text{new}}(\mu_\lambda - \lambda_j^{\text{new}})$$

MA-plot FA vs. FARMS

FA

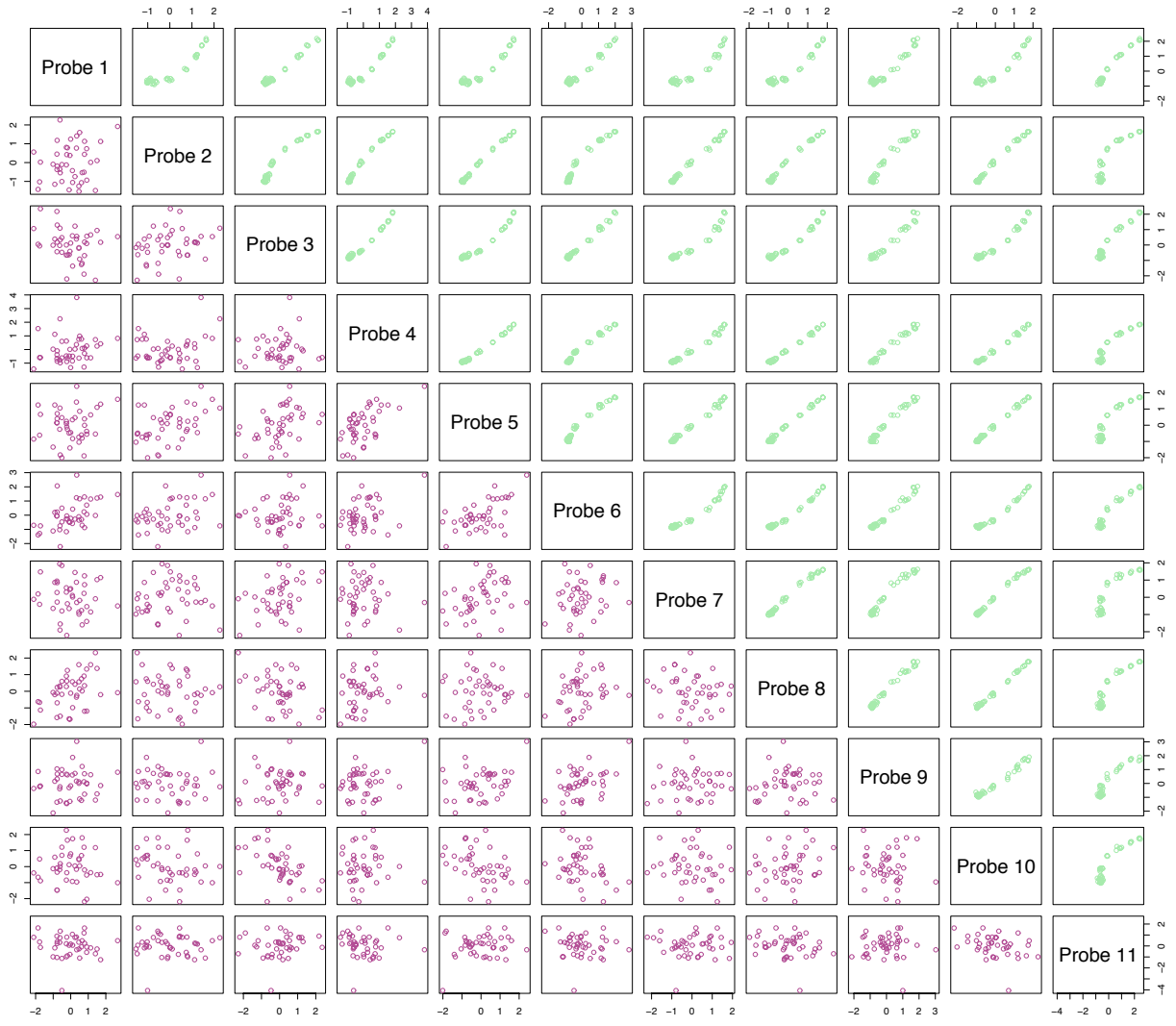


FARMS



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

$$x = \boldsymbol{\lambda}\, z + \boldsymbol{\epsilon}$$

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

$$\begin{aligned} z \mid x &\sim \mathcal{N}\left(\mu_{z|x}\,,\ \sigma^2_{z|x}\right) \\ \mu_{z|x} &= (x)^T\ \boldsymbol{\Psi}^{-1}\boldsymbol{\lambda}\ \left(1 + \boldsymbol{\lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\lambda}\right)^{-1} \\ \sigma^2_{z|x} &= \left(1 + \boldsymbol{\lambda}^T\boldsymbol{\Psi}^{-1}\boldsymbol{\lambda}\right)^{-1} \end{aligned}$$

FARMS: The I/NI call

The variance of z is decomposed into a signal and a noise part:

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

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

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

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

FARMS: Independent I/NI calls filtering

Independent filtering increases detection power for high-throughput experiments

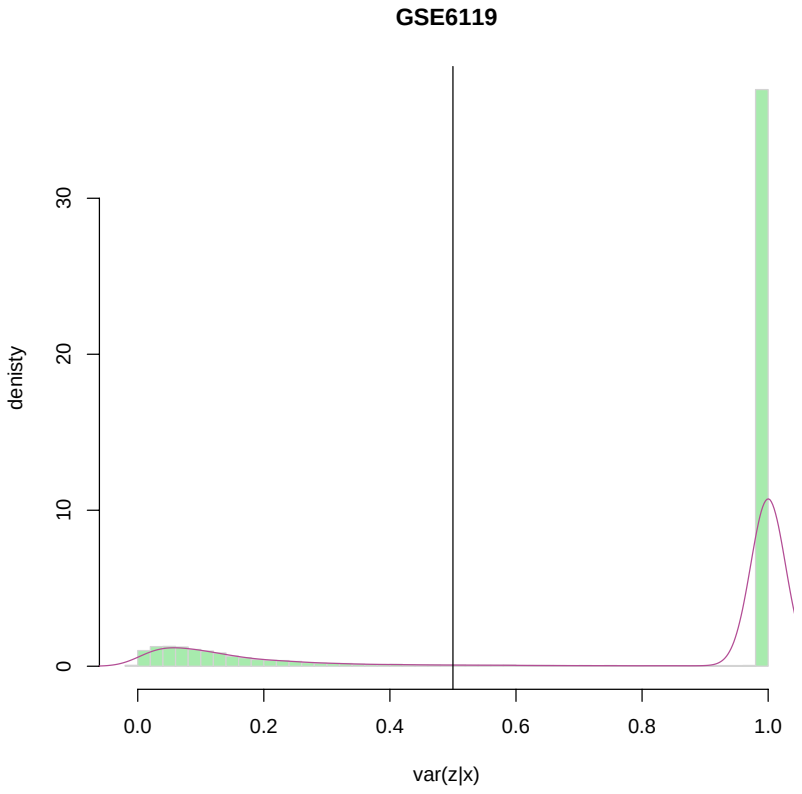
Richard Bourgon[a], Robert Gentleman[b], and Wolfgang Huber[c,1]

[a]European Bioinformatics Institute, Cambridge CB10 1SD, United Kingdom; [b]Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080-4990; and [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.
- http://www.bioinf.jku.at/software/cnfarms/proof_ini.pdf

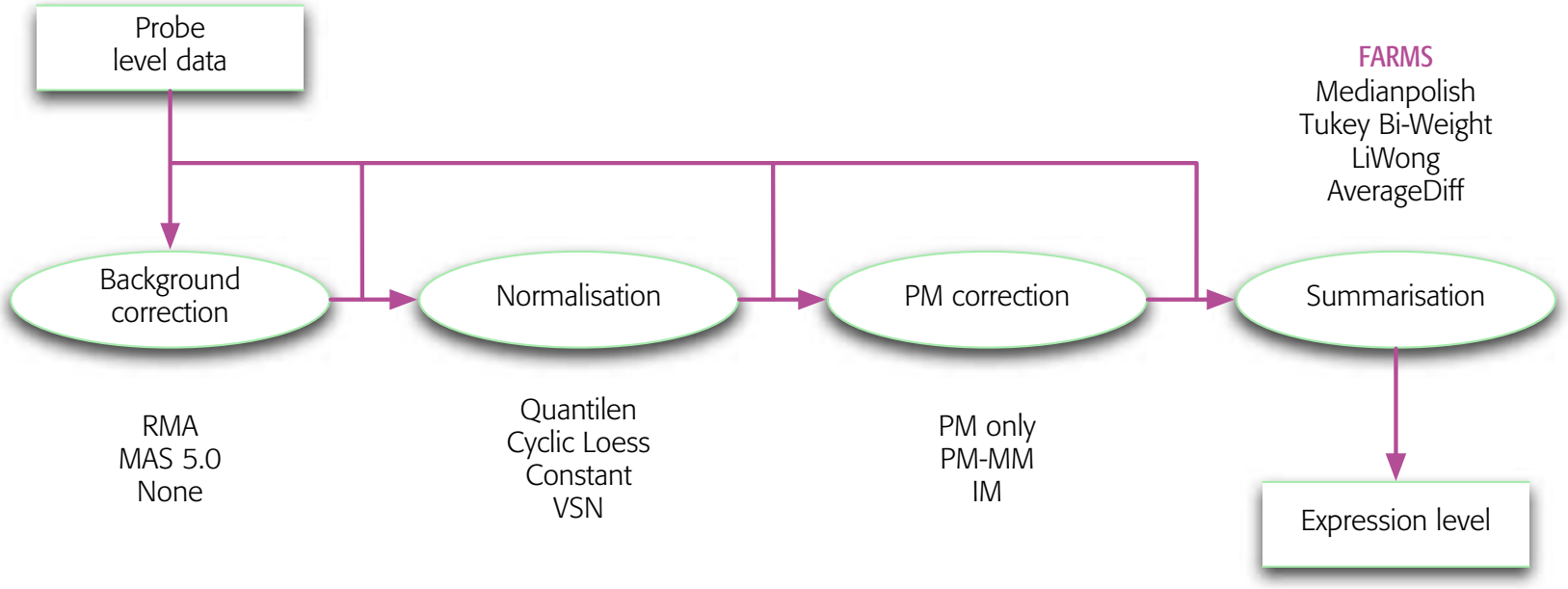
FARMS: I/NI calls distribution



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



Receiver Operator Characteristics (ROC)

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

	Intensity	FARMS	RMA	GCRMA	MAS 5.0	MBEI
HGU133	Low	**0.94**	0.51	0.62	0.07	0.21
	Med	**0.99**	0.91	0.94	0.00	0.43
	High	**1.00**	0.64	0.59	0.00	0.16
	Mean	**0.95**	0.60	0.69	0.05	0.26
HGU95	Low	**0.91**	0.57	0.45	0.09	-
	Med	**1.00**	0.91	0.91	0.00	-
	High	**0.98**	0.96	0.92	0.00	-
	Mean	**0.93**	0.65	0.57	0.06	-
GoldenSpike		**0.85**	0.76	0.78	0.28	0.39

Computational costs for processing 60 arrays

	FARMS	RMA	MAS 5.0	MBEI
Computational time [s]	**92**	384	851	591

Results I/NI call

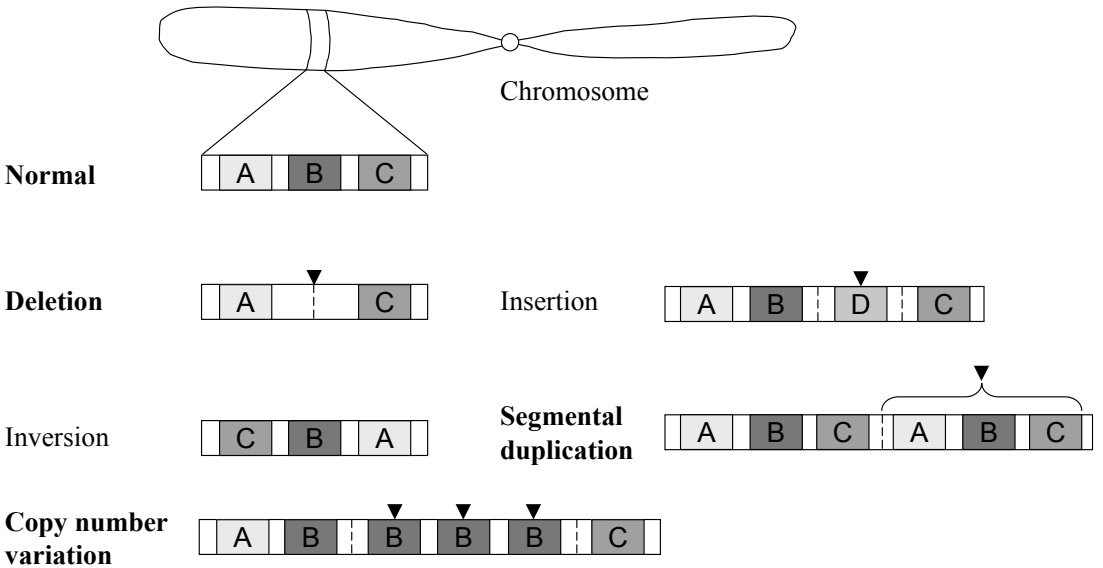
- Leads on average to 84 ($\pm$1.5)% exclusion rate
  - Applied on 30 real life studies
  - A/P calls excluded only 33 ($\pm$1)%
- Validation was carried out on spiked-in data:

Exclusion rate on spiked-in data sets:

	Informative	Non-informative	Exclusion rate	Detected Spiked-ins	Detected Pseudo Spiked-ins
HGU133A	81	22219	99.63%	42/42	28/28*
HGU95_V2	56	12570	99.56%	14/14	5/5**
Hu. Gene 1.0 ST	40	19,753	99.80%	15/15***	-

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

Copy number variants



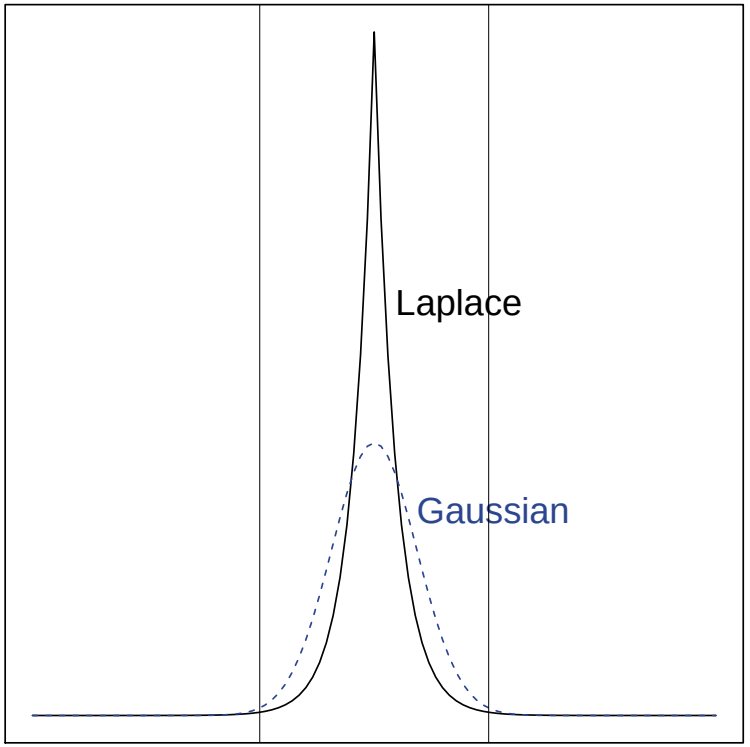
- are deletions, duplications, inversions or inserations of chromosomal segments
- are a major source of variation between individual humans
- are an underlying factor in human evolution and in many diseases
- form at a faster rate than other types of mutation

Rare CNV events

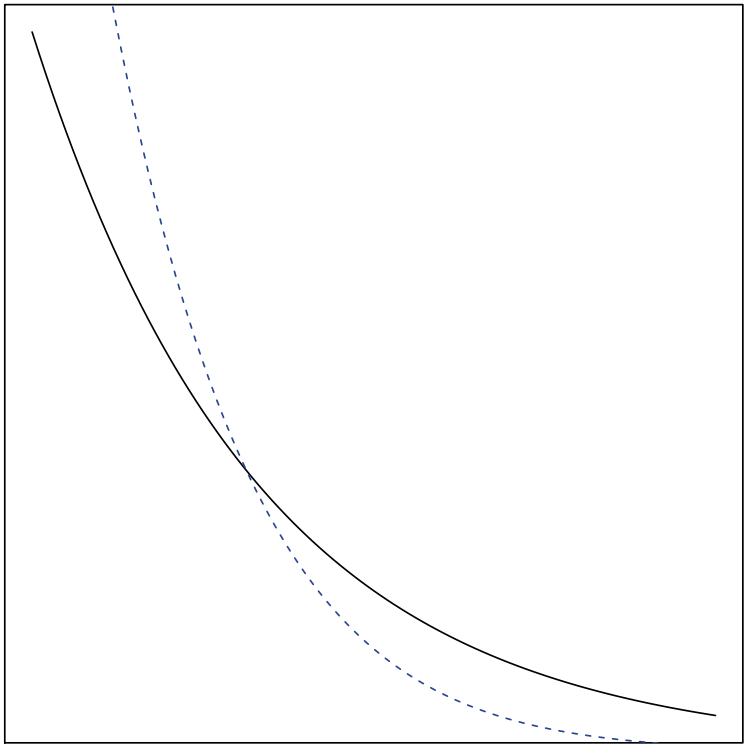
Sparse data

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

$$p(\{x\} \mid \boldsymbol{\lambda}, \boldsymbol{\Psi}) = \int p(\{x\} \mid z, \boldsymbol{\lambda}, \boldsymbol{\Psi}) \; p(z) \; dz$$

Laplacian cn.FARMS

Data likelihood

$$p(\{x\} \mid \boldsymbol{\lambda}, \boldsymbol{\Psi}) = \int p(\{x\} \mid z, \boldsymbol{\lambda}, \boldsymbol{\Psi}) \; p(z) \; dz$$

Problem

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

Laplacian cn.FARMS

Data likelihood

$$p(\{x\} \mid \boldsymbol{\lambda}, \boldsymbol{\Psi}) = \int p(\{x\} \mid z, \boldsymbol{\lambda}, \boldsymbol{\Psi}) \; p(z) \; dz$$

Problem

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

Solution

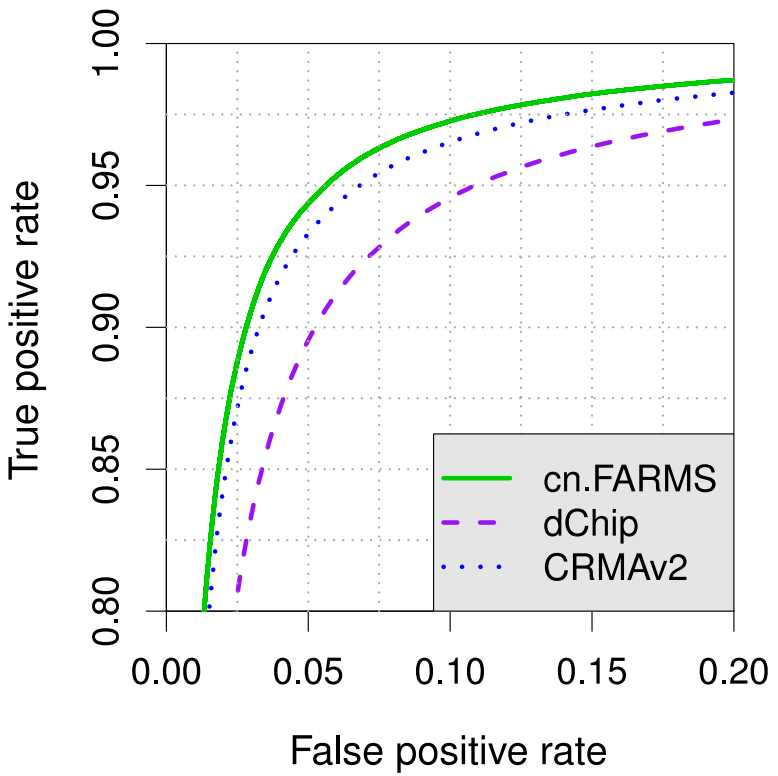
- Variational EM approach
- Based on a local Gaussian approximation to the mode

Benchmark data sets

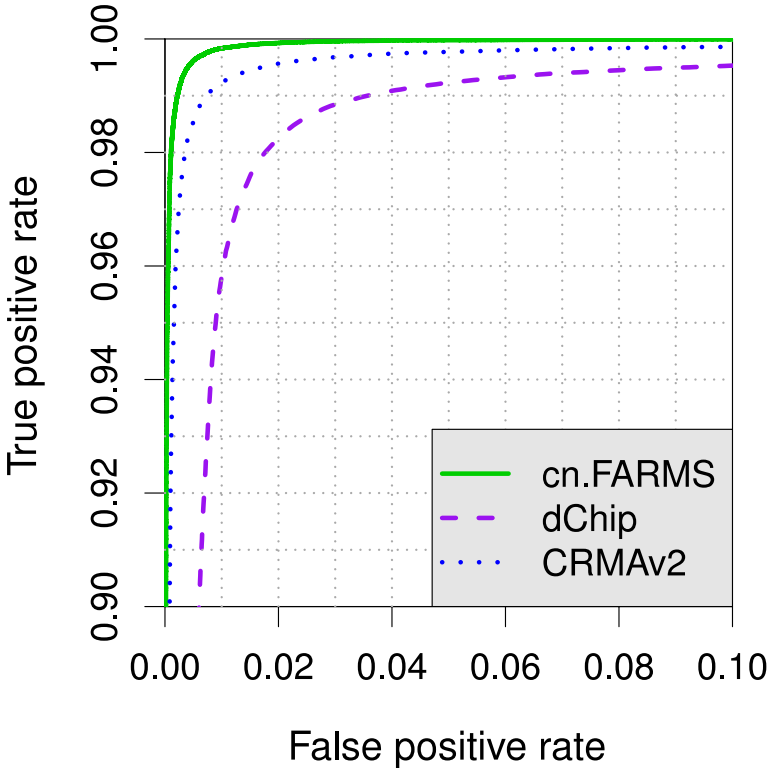
- 30 male and 30 female CEU founders
  - Classification task: distinguish males from females by their copy number on the X chromosome
- Evaluation on:
  - Single-locus / multi-loci classification (window mode)
  - Multi-loci summarization with
    - cn.FARMS
    - Median locus for dChip and CRMA_v2

ROC-Curve (SNP 6.0 arrays)

single-locus



multi-loci, 3 markers



TPR / FPR

True positive rate (TPR) = TP/(TP+FN)

False positive rate (FPR) = FP/(FP+TN)

Results cn.FARMS

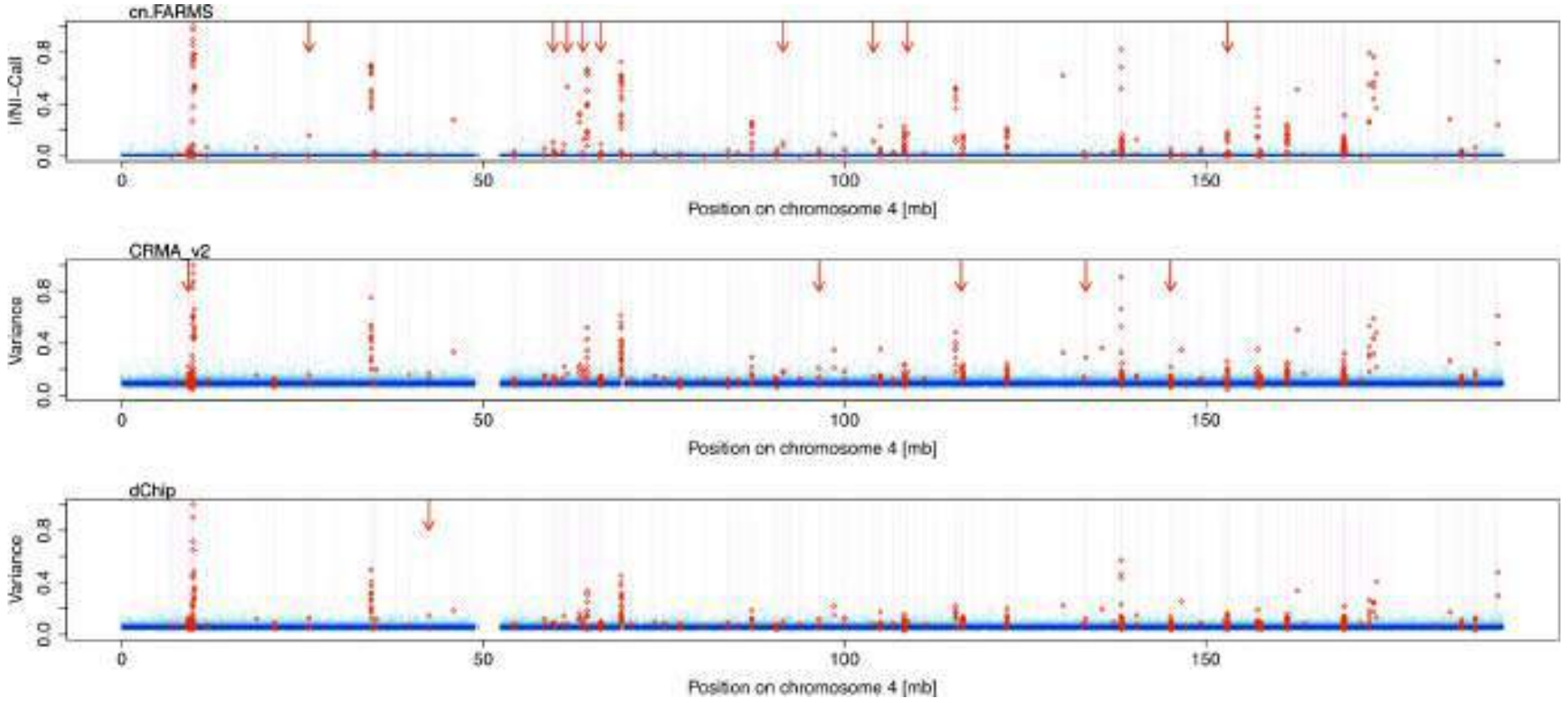
		Affymetrix Mapping250K_NSP			Affymetrix SNP 6.0		
Loci	Criteria	**cn.FARMS**	CRMA_v2	dChip	**cn.FARMS**	CRMA_v2	dChip
1	AUC	**0.9852**	0.9820	0.9819	0.9838	0.9807	0.9721
	FP	**8472**	9106	9018	56145	68593	77438
	P-value	–	1.8e-65	3.1e-26	–	1e-1160	1e-6049
2	AUC	**0.9983**	0.9974	0.9969	0.9983	0.9963	0.9894
	FP	**1375**	1449	1611	9777	11705	18039
	P-value	–	2.7e-4	2.5e-12	–	1e-317	1e-3713
3	AUC	**0.9998**	0.9995	0.9992	0.9998	0.9990	0.9953
	FP	**240**	366	440	1573	3462	6625
	P-value	–	2.6e-38	7.2e-58	–	1e-896	1e-3455

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

CNV detection benchmark

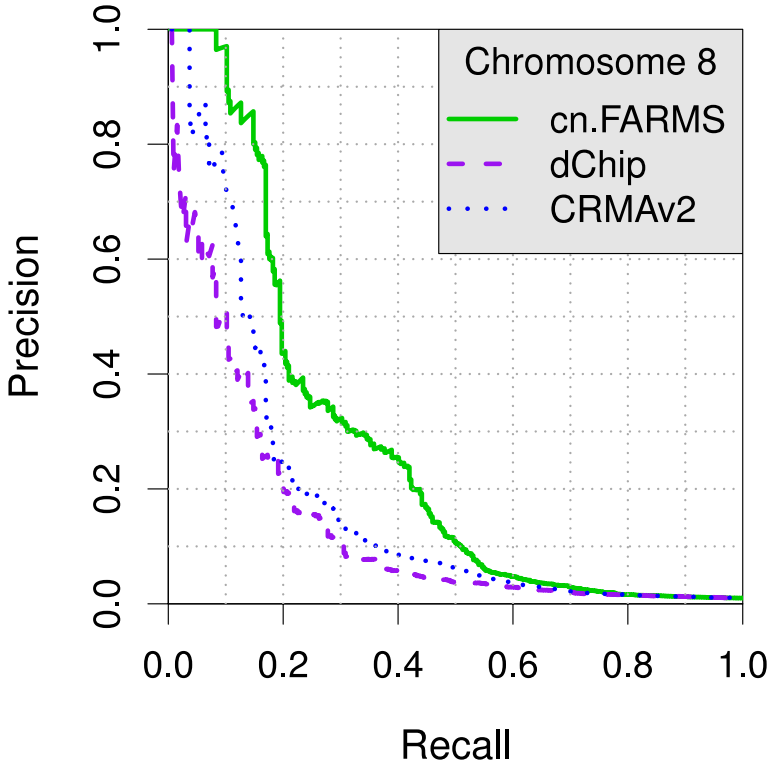
- "The International HapMap Project" phase 2 data set with Affymetrix SNP 6.0 arrays
  - Goal is to identify true rare CNV regions with a low FDR
  - "True CNV regions" are those regions which were detected and verified by different bio-technologies
    - NimbleGen tiling arrays, Agilent CGH arrays, 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

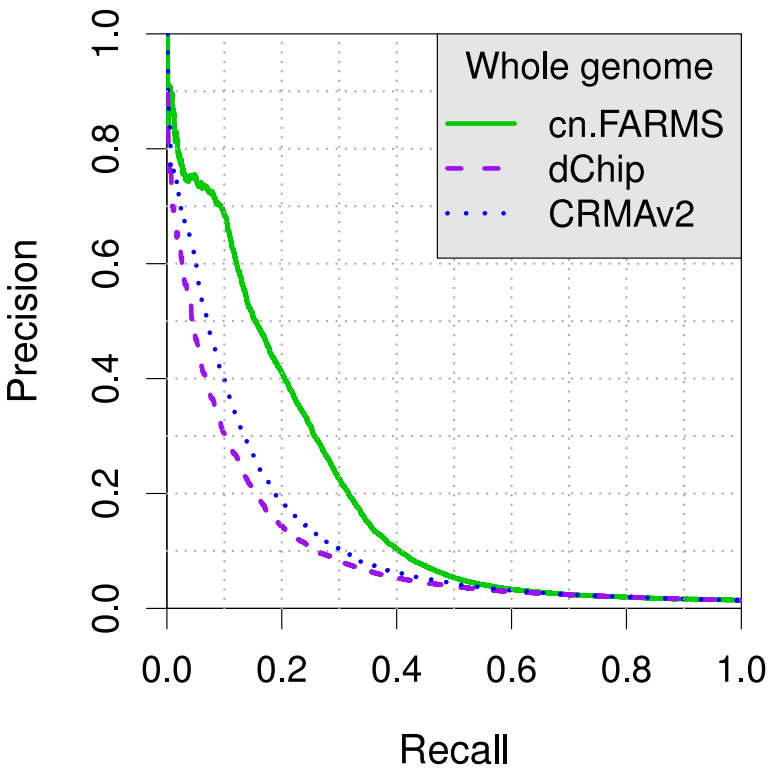


CNV detection on HapMap (multi-loci 3)

Chromosome 8



Whole genome



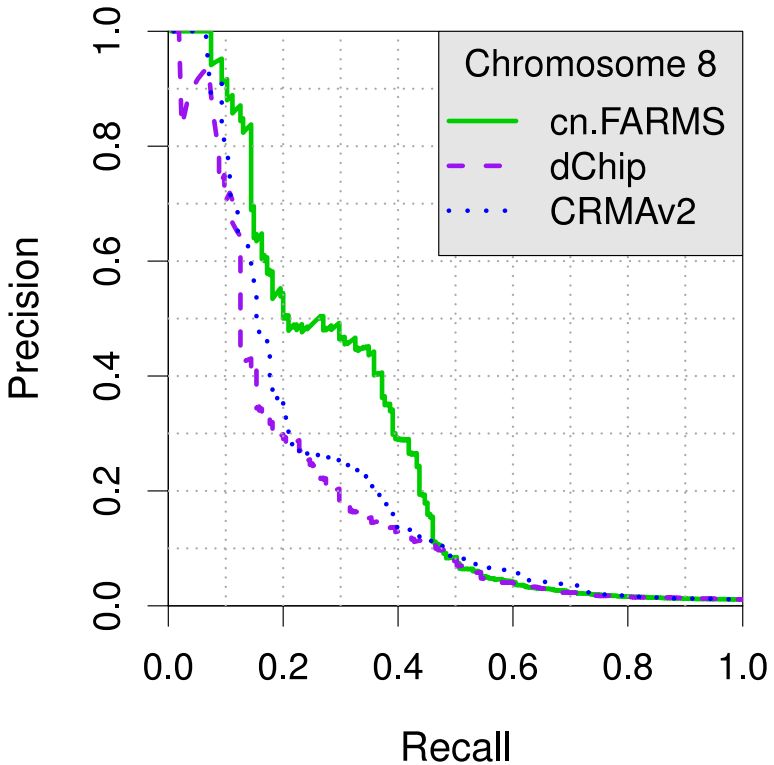
Precision / Recall

Recall = TP/(TP+FN)

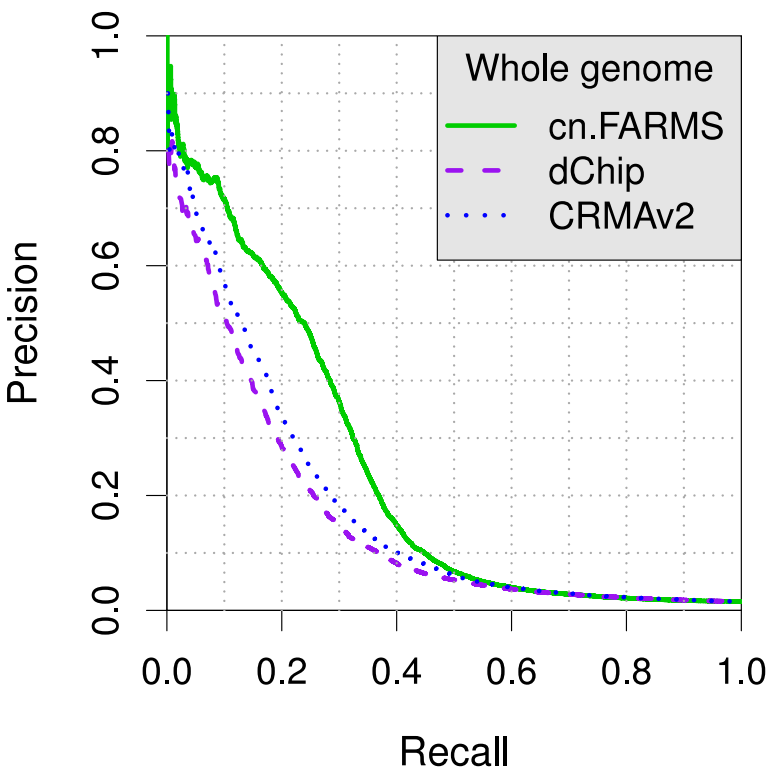
Precision = TP/(TP+FP) = 1 - FDR

CNV detection on HapMap (multi-loci 5)

Chromosome 8



Whole genome



Precision / Recall

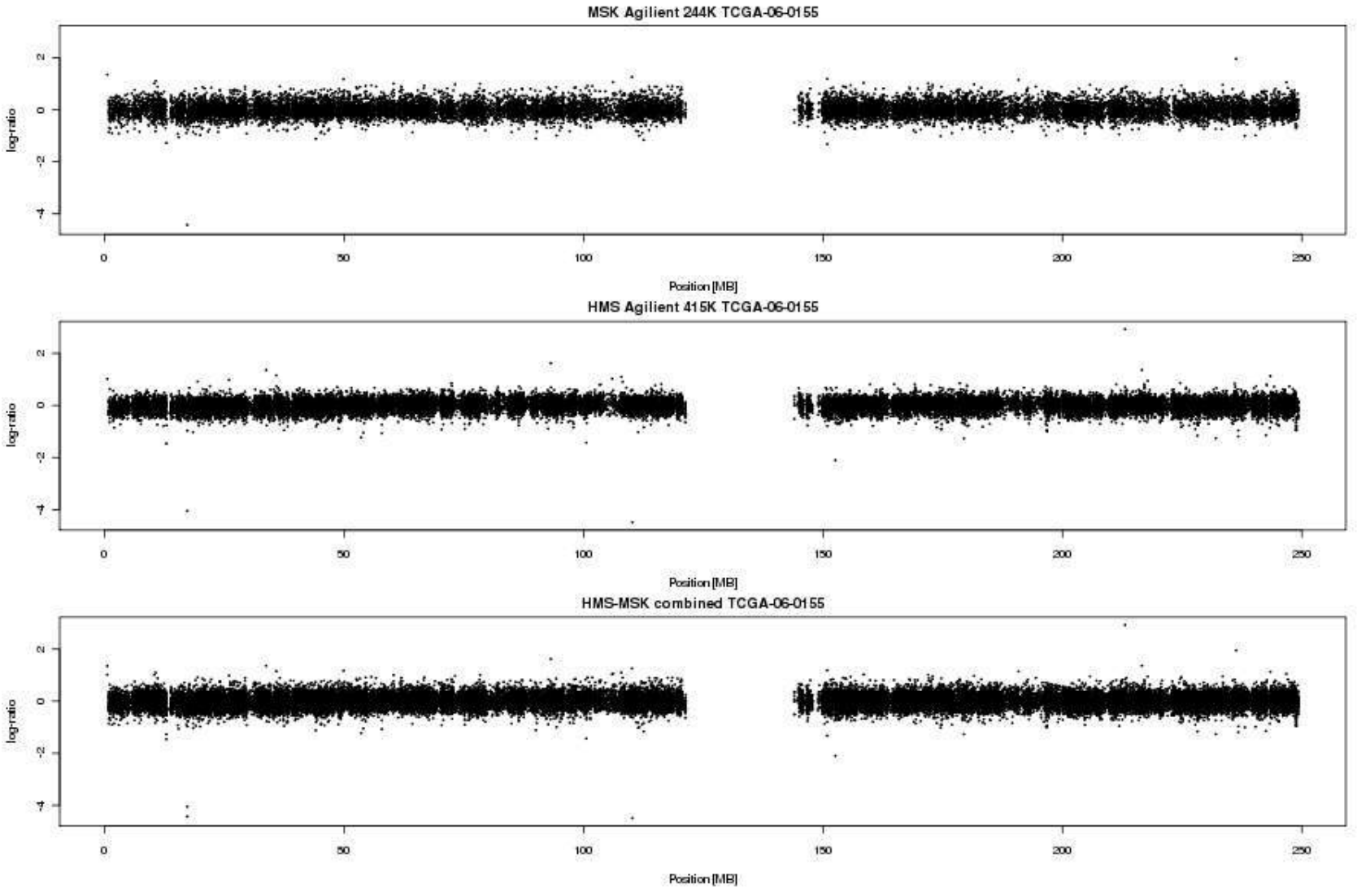
Recall = TP/(TP+FN)

Precision = TP/(TP+FP) = 1 - FDR

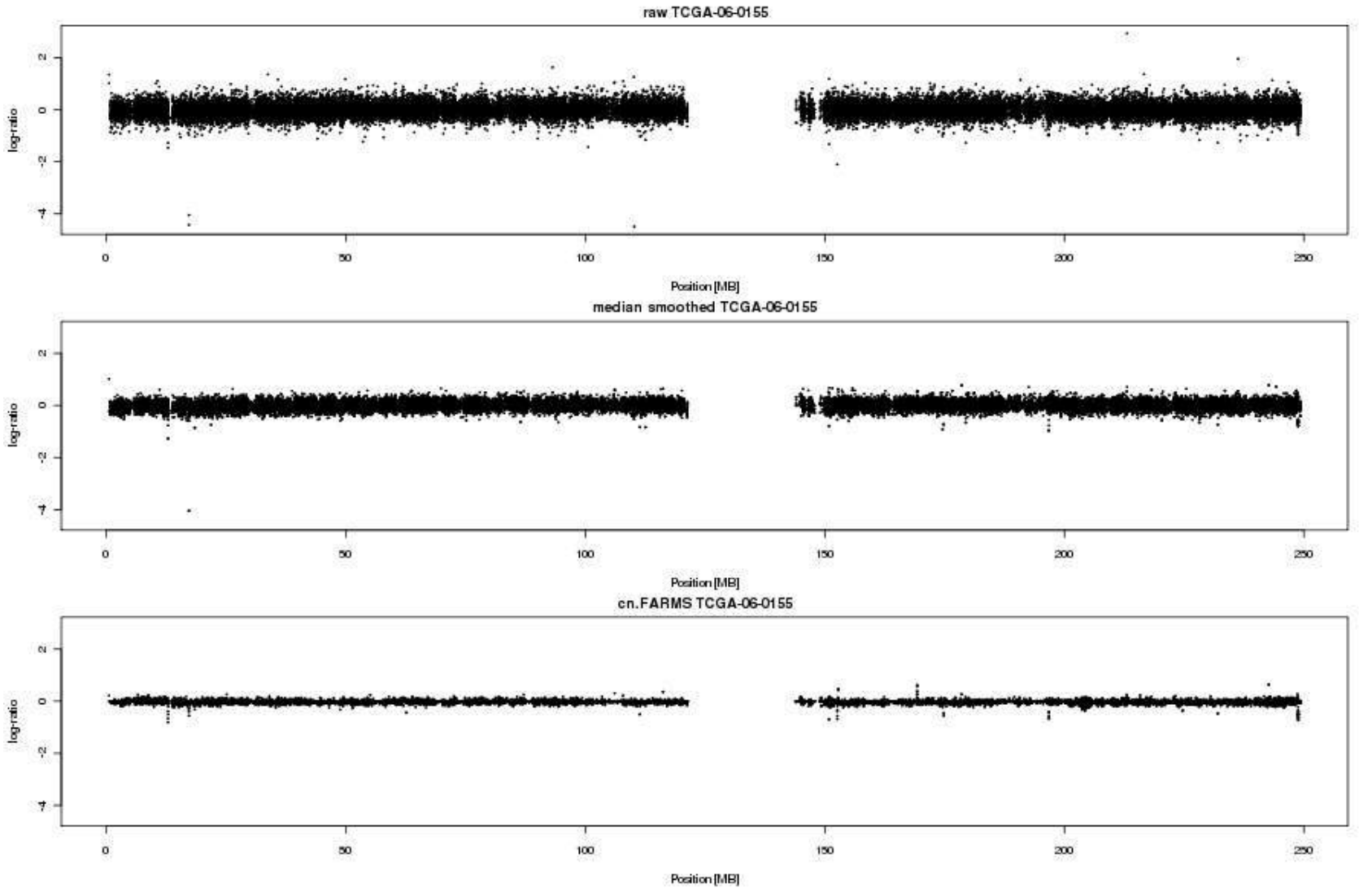
ICGC copy number data sets

- 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

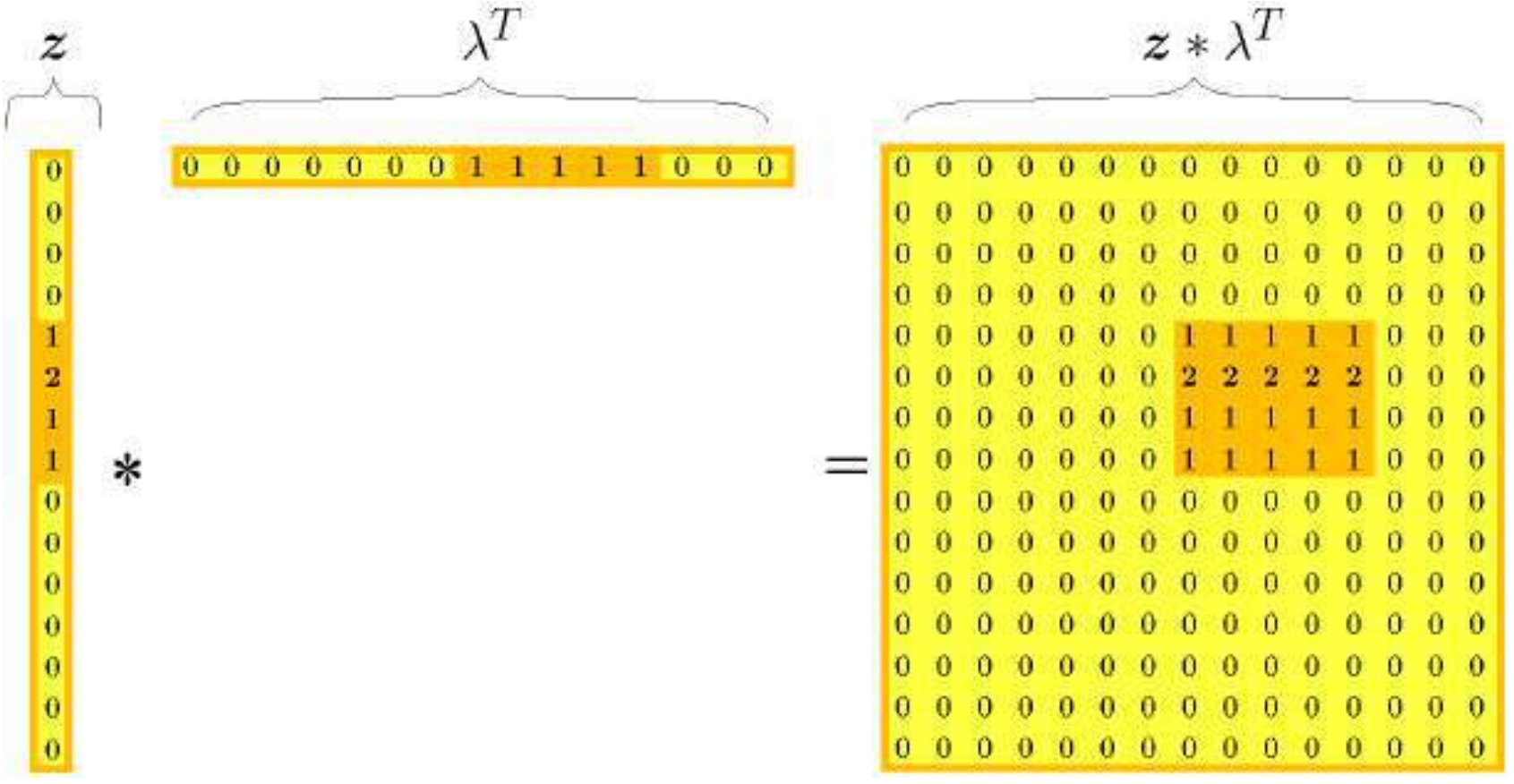


Biclustering applications

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



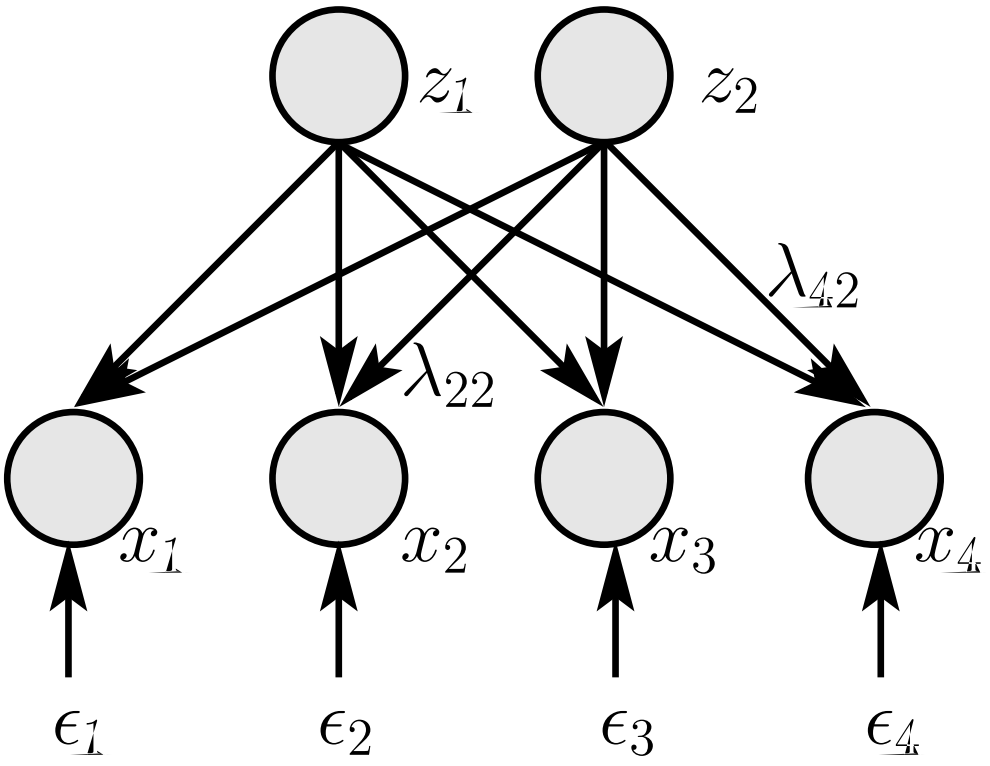
FABIA: The model I

factor z

loading matrix Λ

observations x

noise ϵ



FABIA: The model II

$$x = \boldsymbol{\lambda} z \; + \; \boldsymbol{\epsilon} \; = \sum_{i=1}^{p} \lambda_i \; z_i^T \; + \; \boldsymbol{\epsilon}$$

- x are the observations
- $\boldsymbol{\lambda}$ 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 $\mathcal{N}(0, \boldsymbol{\Psi})$-distributed

FABIA: Bayes framework

Loading prior

- Sparseness on the loadings
- Laplace prior
- $p(\boldsymbol{\lambda}_i) = \left(\frac{1}{\sqrt{2}}\right)^n \prod_{j=1}^{n} e^{-\sqrt{2}|\lambda_{ji}|}$

FABIA: Bayes framework

Loading prior

- Sparseness on the loadings
- Laplace prior
- $p(\boldsymbol{\lambda}_i) = \left(\frac{1}{\sqrt{2}}\right)^n \prod_{j=1}^{n} e^{-\sqrt{2}|\lambda_{ji}|}$

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
- $p(\boldsymbol{\lambda}_i) = \left(\frac{1}{\sqrt{2}}\right)^n \prod_{j=1}^{n} e^{-\sqrt{2}|\lambda_{ji}|}$

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:

$$p(x \mid \boldsymbol{\lambda}, \boldsymbol{\Psi}) = \int p(x \mid z, \boldsymbol{\lambda}, \boldsymbol{\Psi}) \; p(z) \; dz$$

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:

$\mathrm{E}(\tilde{z}_j \mid x_j) = (\boldsymbol{\lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda} + \Xi_j^{-1})^{-1} \boldsymbol{\lambda}^T \boldsymbol{\Psi}^{-1} x_j$ and

$\mathrm{E}(\tilde{z}_j \tilde{z}_j^T \mid x_j) = (\boldsymbol{\lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda} + \Xi_j^{-1})^{-1} + \mathrm{E}(\tilde{z}_j \mid x_j) \mathrm{E}(\tilde{z}_j \mid x_j)^T$ where

$\Xi_j = \mathrm{diag}(\mathrm{diagvect}\left(\sqrt{\mathrm{E}(\tilde{z}_j \tilde{z}_j^T \mid x_j)}\right))$ is the update for the variational parameter.

FABIA: Variational EM updates

E-step:

$\mathrm{E}(\tilde{z}_j \mid x_j) = (\boldsymbol{\lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda} + \Xi_j^{-1})^{-1} \boldsymbol{\lambda}^T \boldsymbol{\Psi}^{-1} x_j$ and

$\mathrm{E}(\tilde{z}_j \tilde{z}_j^T \mid x_j) = (\boldsymbol{\lambda}^T \boldsymbol{\Psi}^{-1} \boldsymbol{\lambda} + \Xi_j^{-1})^{-1} + \mathrm{E}(\tilde{z}_j \mid x_j)\, \mathrm{E}(\tilde{z}_j \mid x_j)^T$ where

$\Xi_j = \mathrm{diag}(\mathrm{diagvect}\left(\sqrt{\mathrm{E}(\tilde{z}_j \tilde{z}_j^T \mid x_j)}\right))$ is the update for the variational parameter.

M-step:

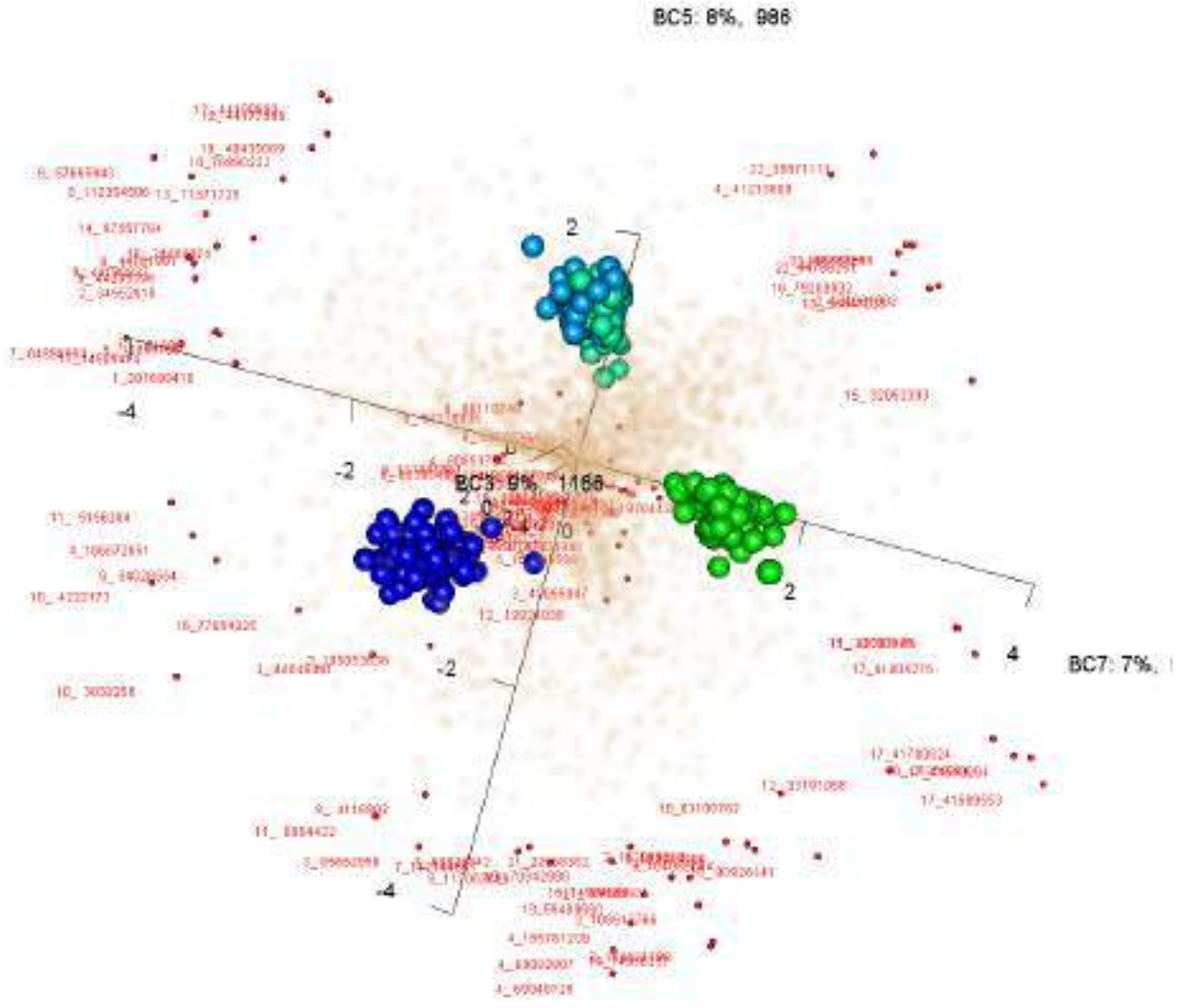
$$\boldsymbol{\lambda}^{\mathrm{new}} = \frac{\frac{1}{l} \sum_{j=1}^{l} x_j\, \mathrm{E}(\tilde{z}_j \mid x_j)^T - \frac{\alpha}{l} \boldsymbol{\Psi}\, \mathrm{sign}(\boldsymbol{\lambda})}{\frac{1}{l} \sum_{j=1}^{l} \mathrm{E}(\tilde{z}_j \tilde{z}_j^T \mid x_j)}$$

$$\boldsymbol{\Psi}^{\mathrm{new}} = \boldsymbol{\Psi}^{\mathrm{EM}} + \mathrm{diag}(\mathrm{diagvect}\left(\frac{\alpha}{l} \boldsymbol{\Psi}\, \mathrm{sign}(\boldsymbol{\lambda})(\boldsymbol{\lambda}^{\mathrm{new}})^T\right)), \quad \text{where}$$

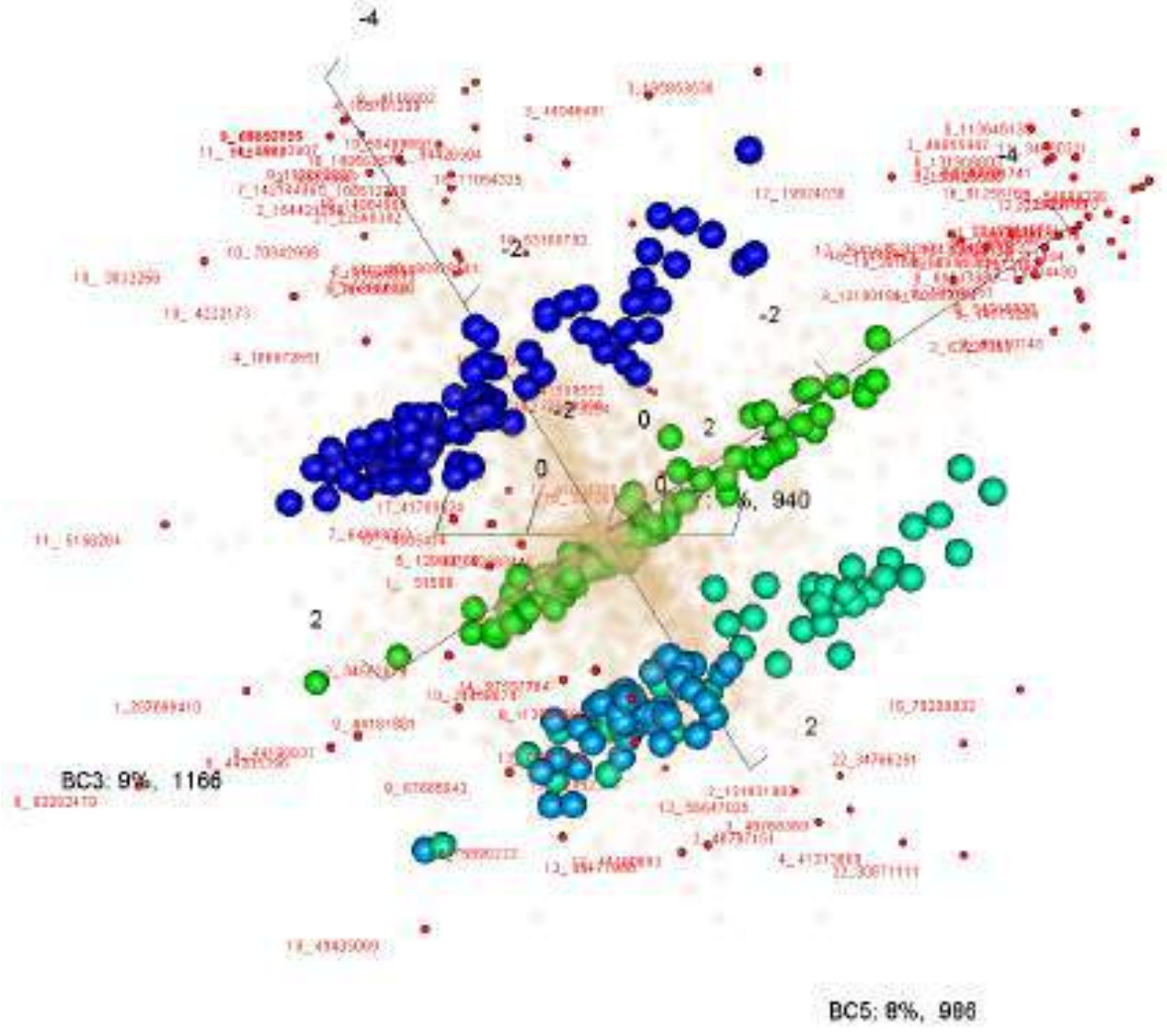
$$\boldsymbol{\Psi}^{\mathrm{EM}} = \mathrm{diag}(\mathrm{diagvect}\left(\frac{1}{l} \sum_{j=1}^{l} x_j x_j^T - \boldsymbol{\lambda}^{\mathrm{new}} \frac{1}{l} \sum_{j=1}^{l} \mathrm{E}(\tilde{z}_j \mid x_j)\, x_j^T\right)).$$

α controls the degree of sparseness (parameter of the Laplacian prior)

Biclustering of copy number variants I



Biclustering of copy number variants II



Biclustering of bioassays and compounds

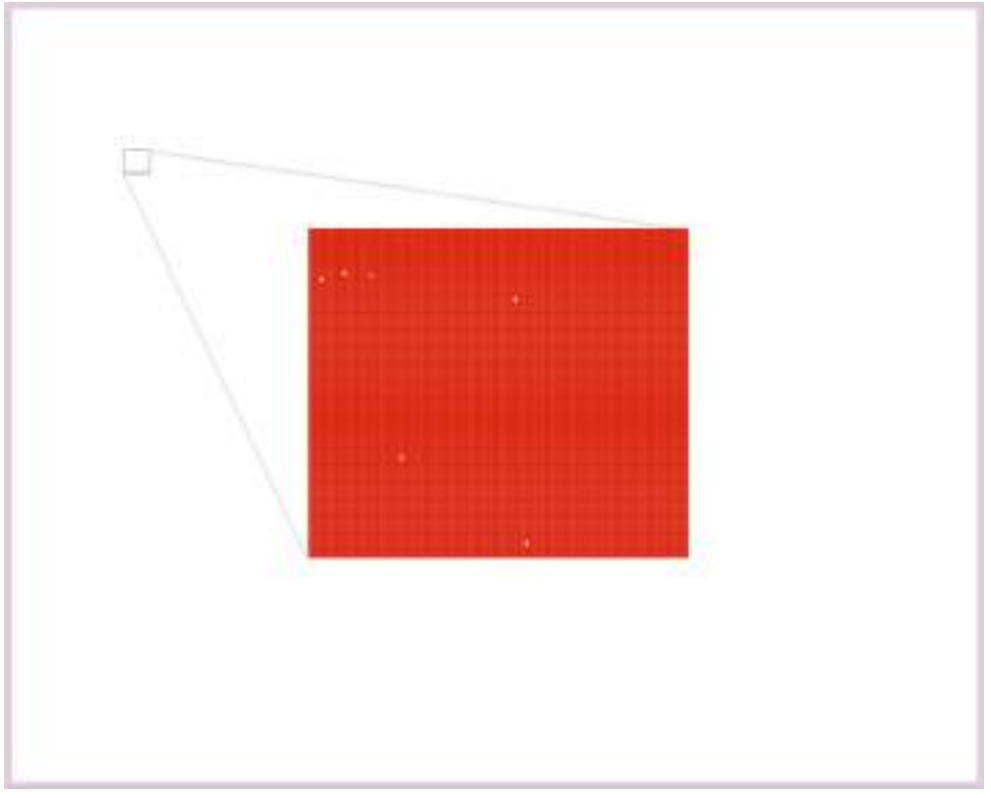
Matrix plot (dim 270,000 x 4,000)

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

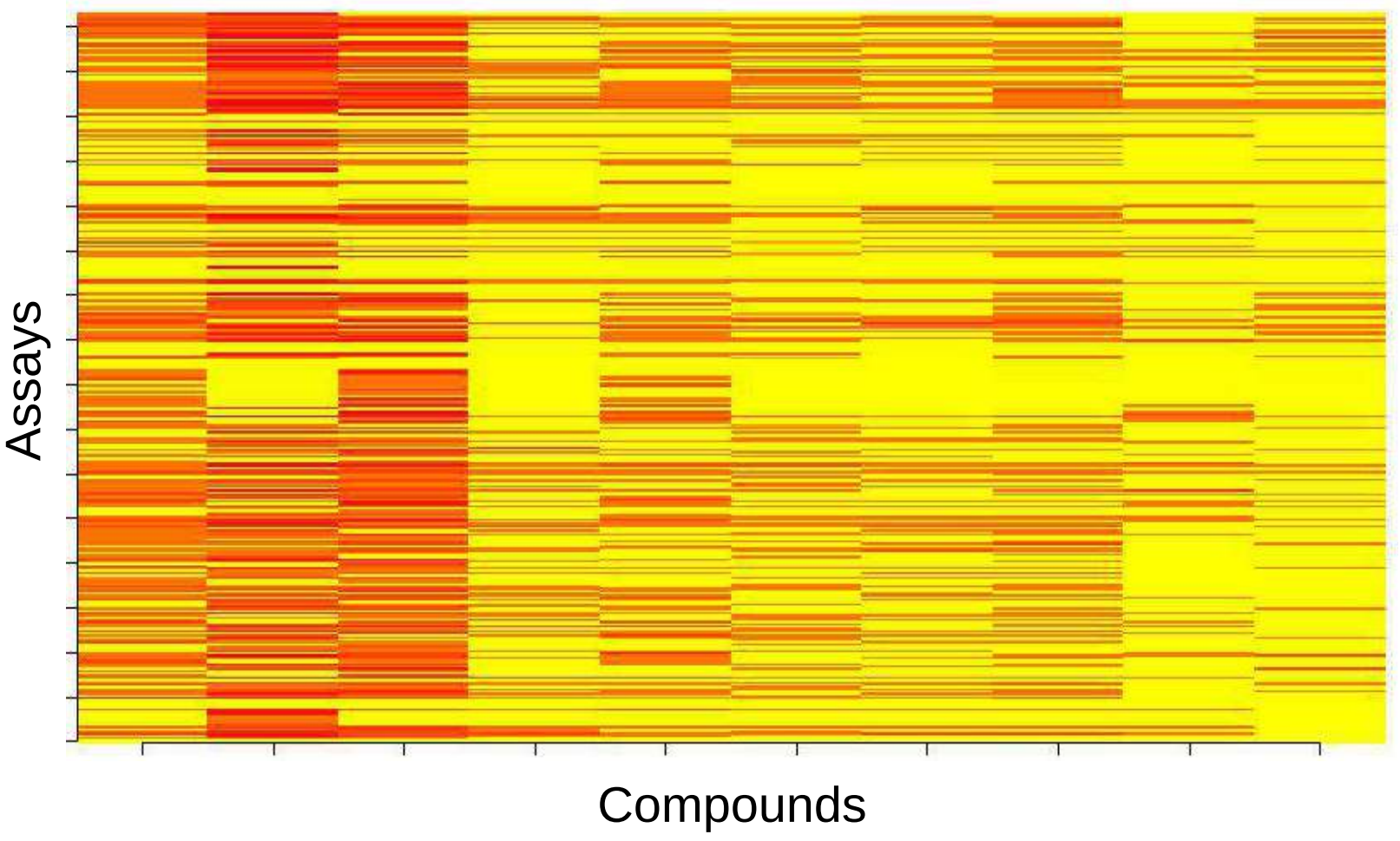
Matrix plot - close up



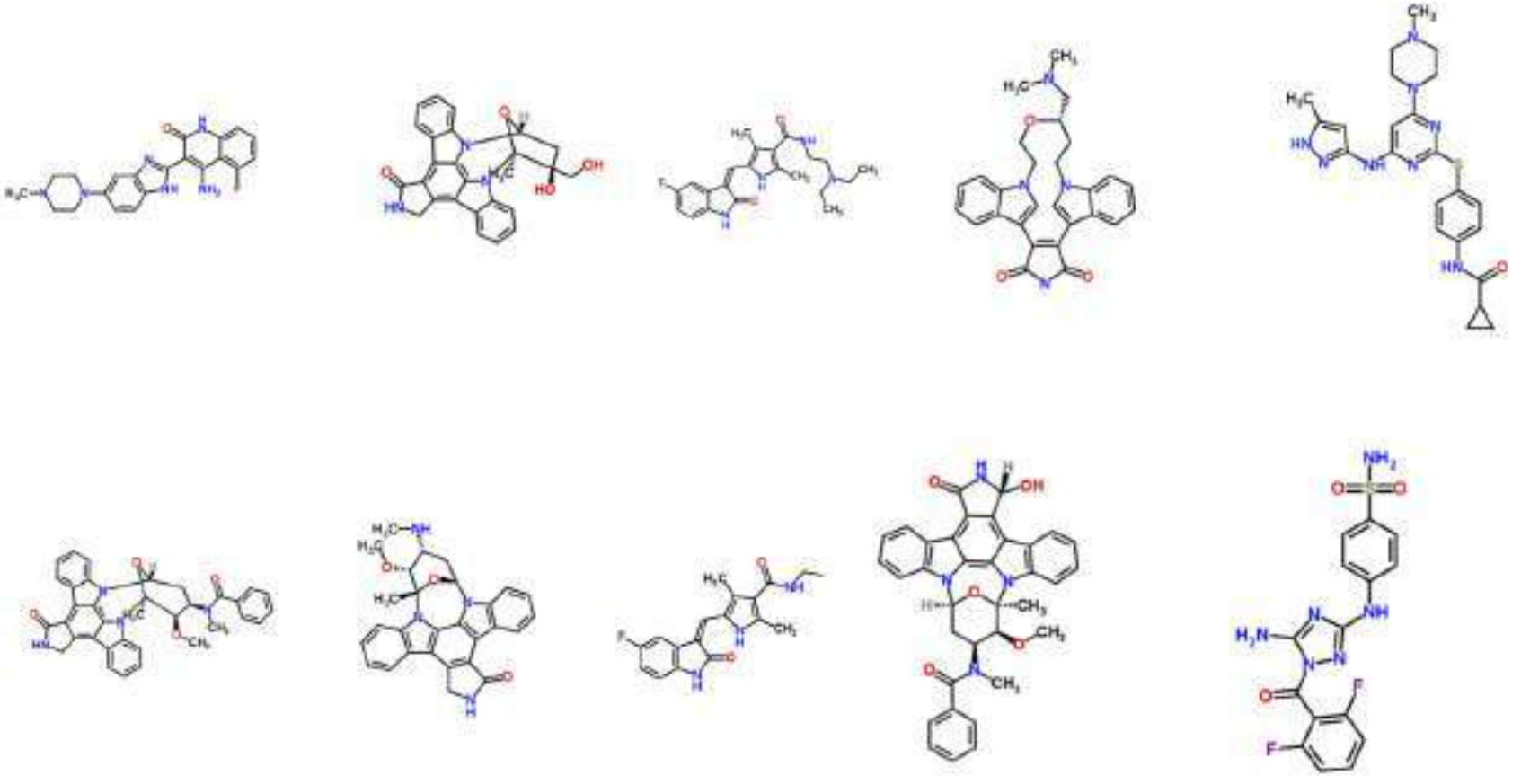
Bioassay data details

- 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



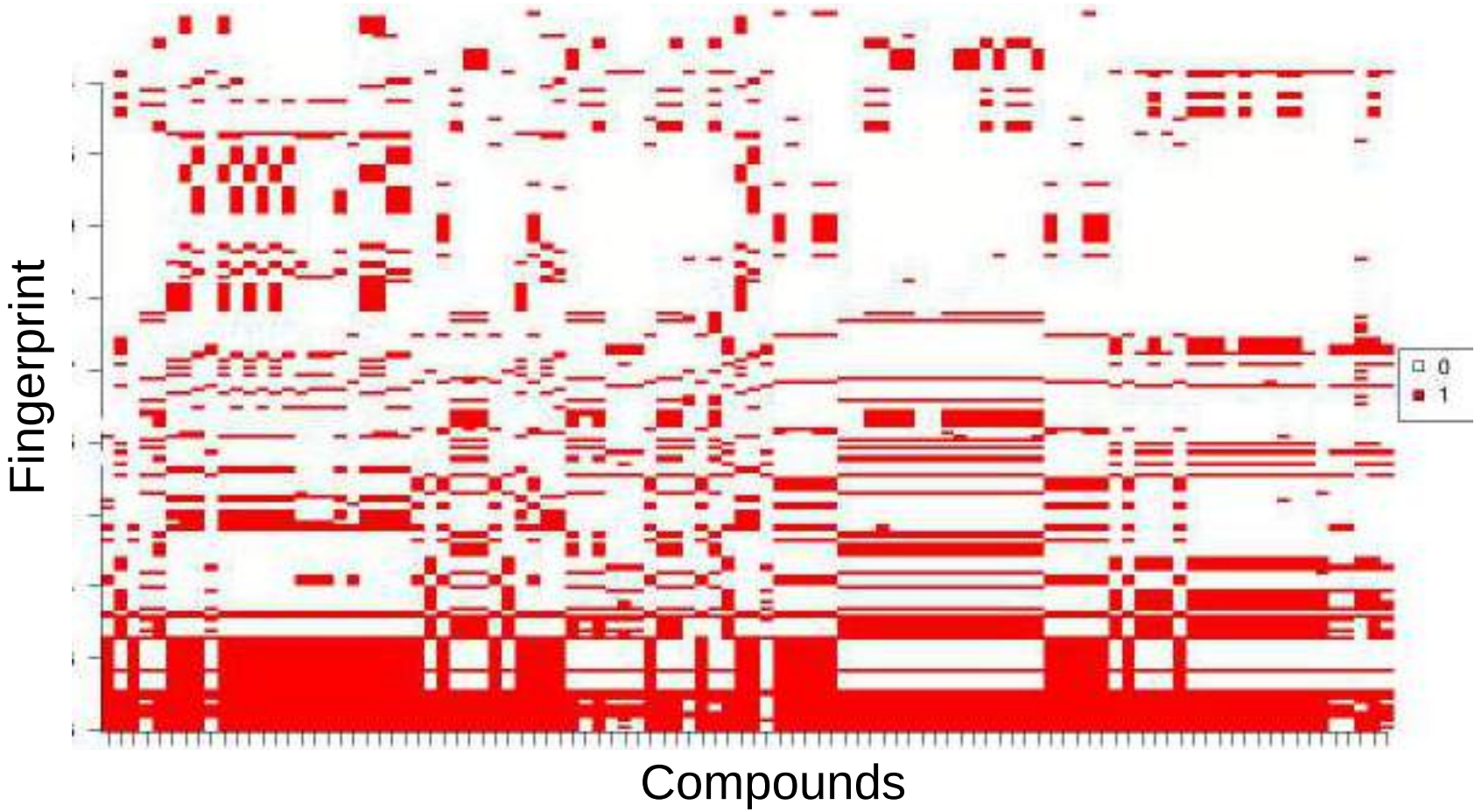
Biclustering of fingerprints and compounds

Matrix plot (dim 16e+6 x 1e+6)

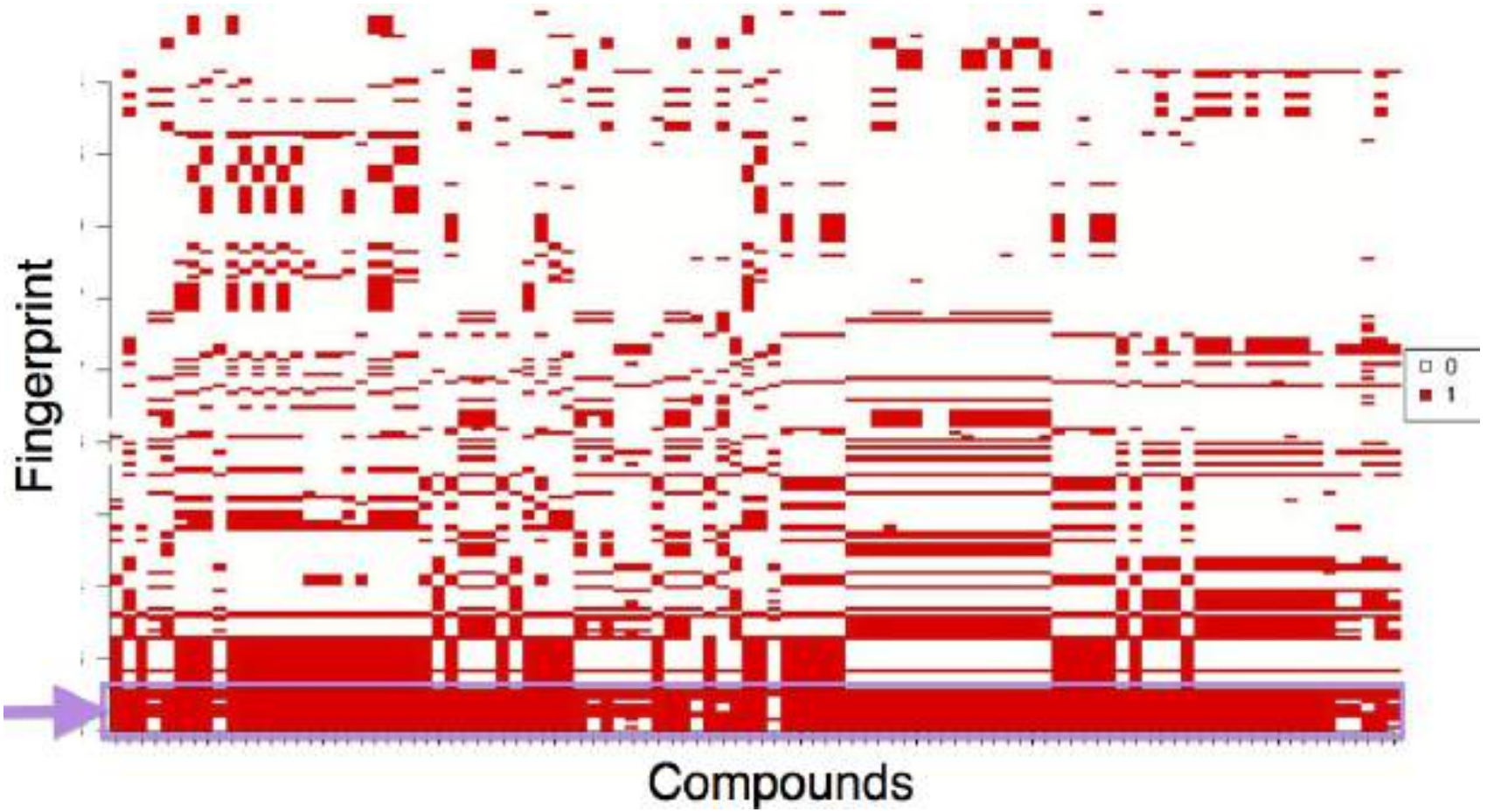
Bioassay data details

- Data source: ChEMBL
- # of fingerprints: ca. 16,000,000
- # of compounds: ca. 1,000,000
- Sparseness: ca. 1:150,000

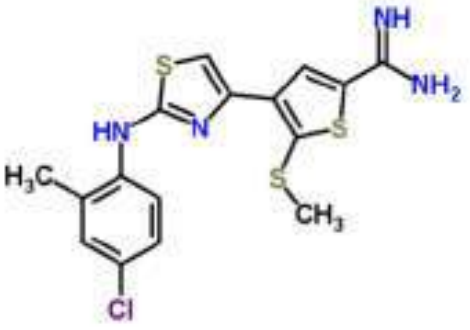
Bicluster of fingerprints and compounds I

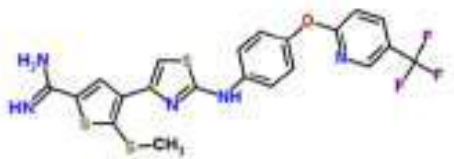


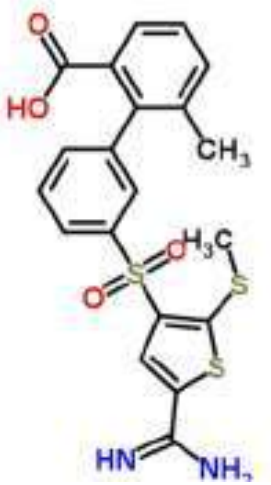
Bicluster of fingerprints and compounds I



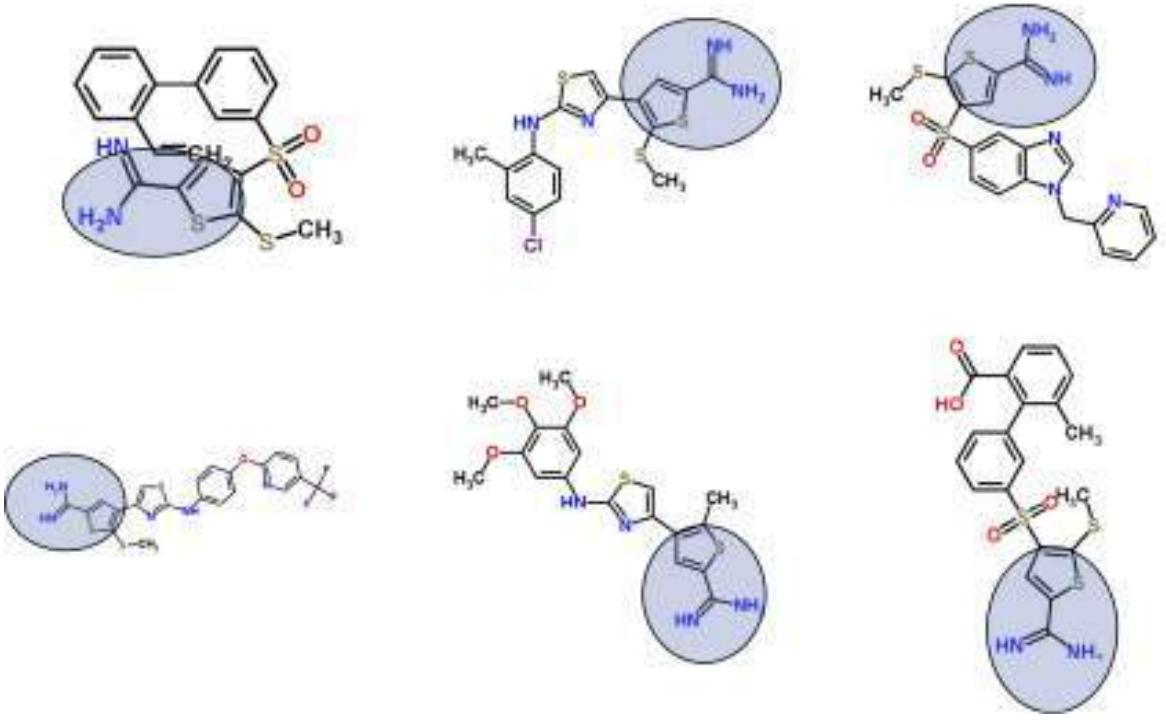
Compounds of the fingerprint bicluster







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

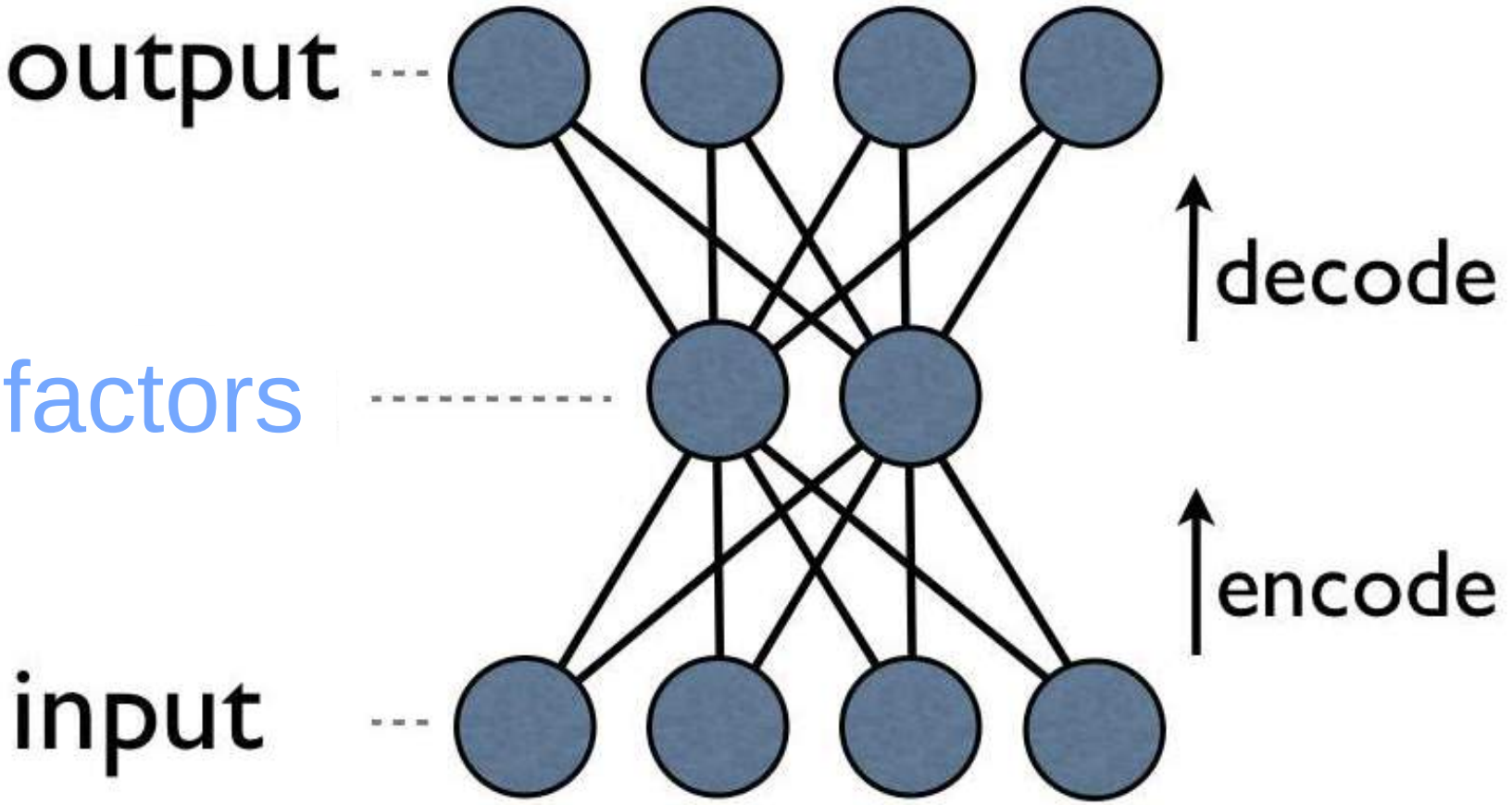
Overview 12 random selected bicluster

Top 36 articles of a bicluster

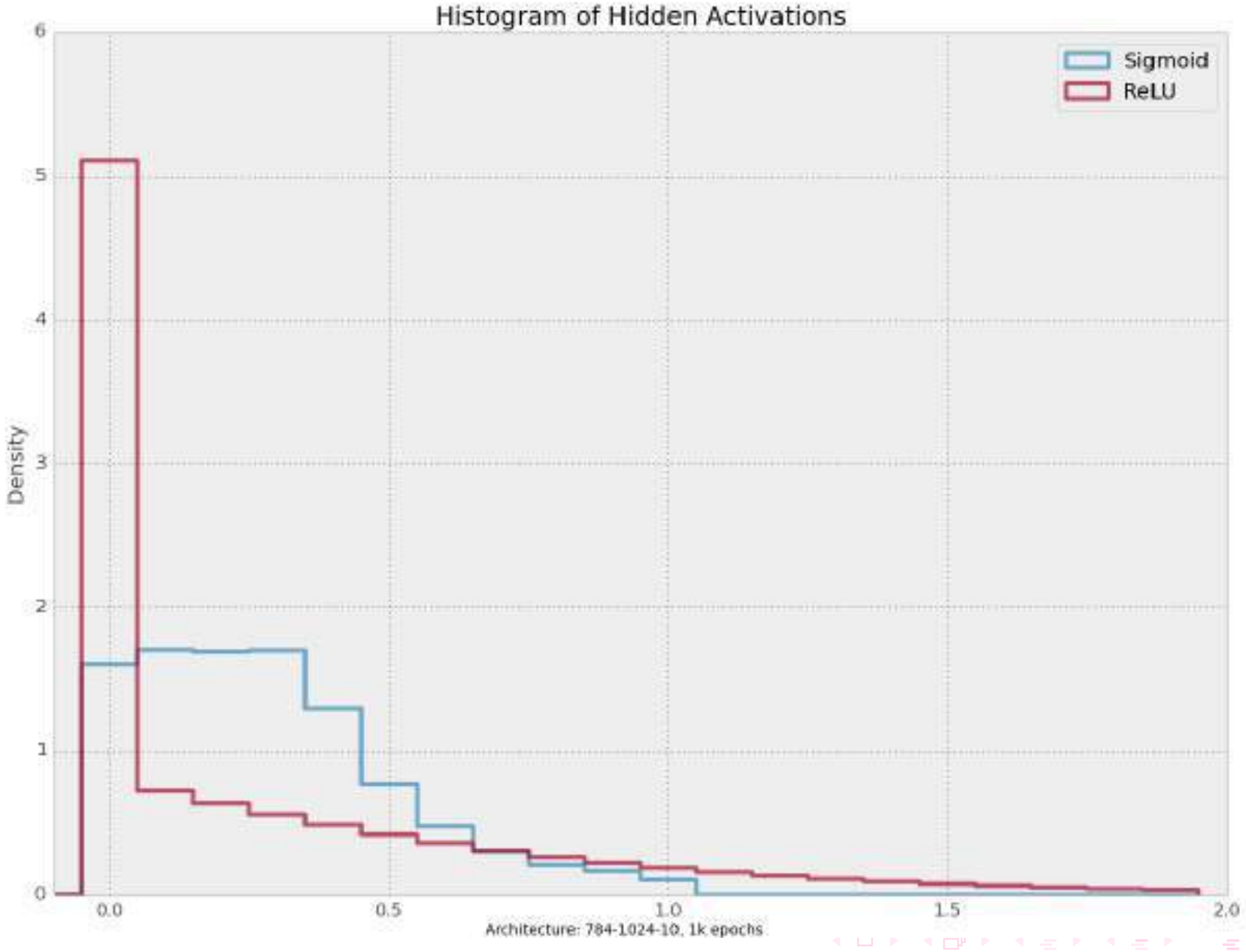
Deep Learning

- 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 serveral articles (two front-page articles)

Networks

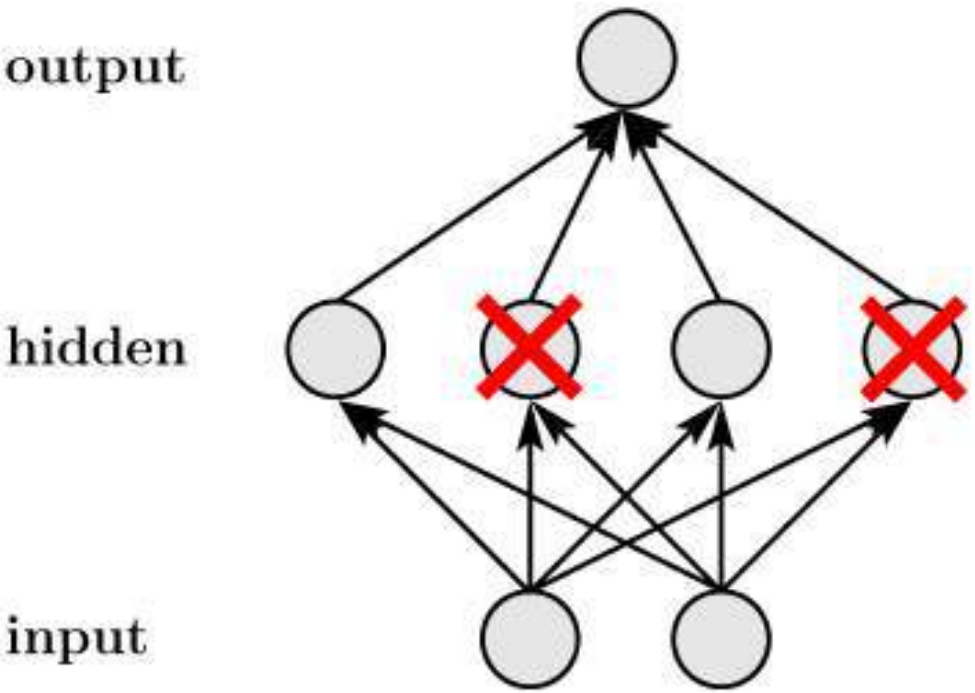


Sparsity by Linear-Rectified Units



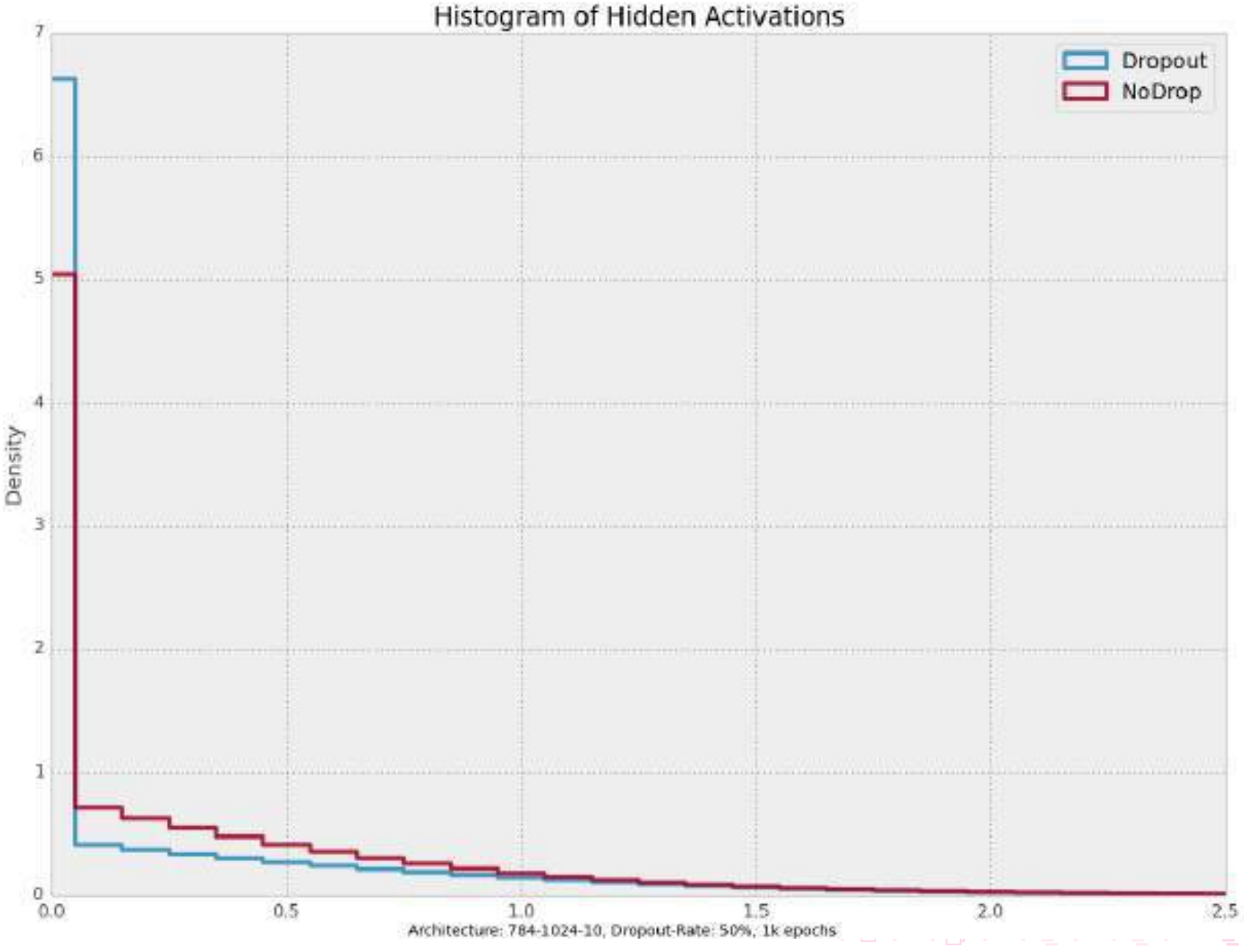
Sparsity by Dropout I

Dropout algorithm

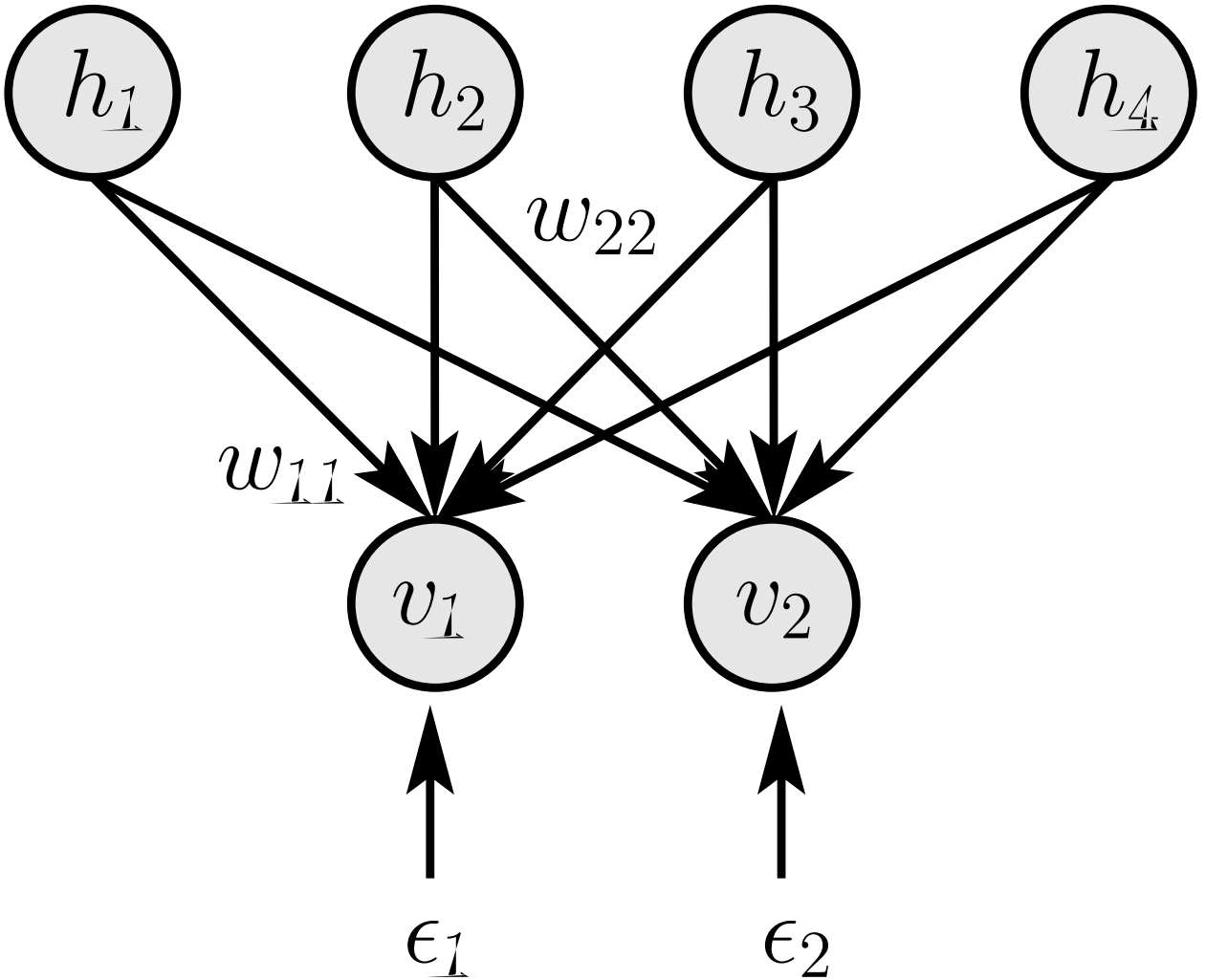


- randomly set units to zero acitvation
- no derivatives

Sparsity by Dropout II



Model: Rectified Factor Network



Rectified Factor Network: Newton updates

Newton-step:

$$
\begin{aligned}
\boldsymbol{\mu}_{h|v} &= W^T \left(W\, W^T + \boldsymbol{\Psi}\right)^{-1} v\,, \\
\boldsymbol{\Sigma}_{h|v} &= I - W^T \left(W\, W^T + \boldsymbol{\Psi}\right)^{-1} W\,, \\
\mathbb{E}_{h_i|v_i}(h_i) &= \boldsymbol{\mu}_{h_i|v_i} \\
\mathbb{E}_{h_i|v_i}\left(h_i\; h_i^T\right) &= \boldsymbol{\mu}_{h_i|v_i}\; \boldsymbol{\mu}_{h_i|v_i}^T + \boldsymbol{\Sigma}_{h_i|v_i} \\
U &= \frac{1}{n} \sum_{i=1}^{n} v_i \mathbb{E}_{h_i|v_i}^T(h_i) \quad \text{Hebb rule: input - hidden} \\
S &= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{h_i|v_i}\left(h_i\; h_i^T\right)
\end{aligned}
$$

Gradients:

$$
\begin{aligned}
\Delta W &= U\, S^{-1} - W \quad \text{Hebb rule: decorrelates factors} \\
\Delta \Psi_{ii} &= \left[C - U\, W^T - W\, U^T + W\, S\, W^T\right]_{ii} - \Psi_{ii}
\end{aligned}
$$

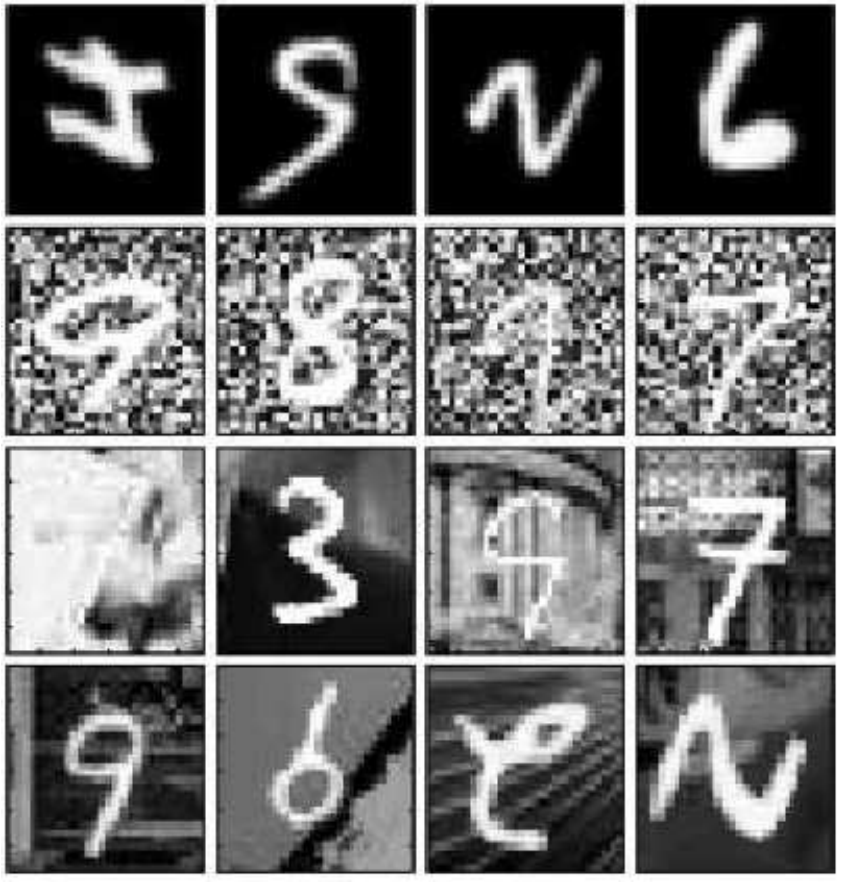
MNIST data set

MNIST 28 x 28 pixel

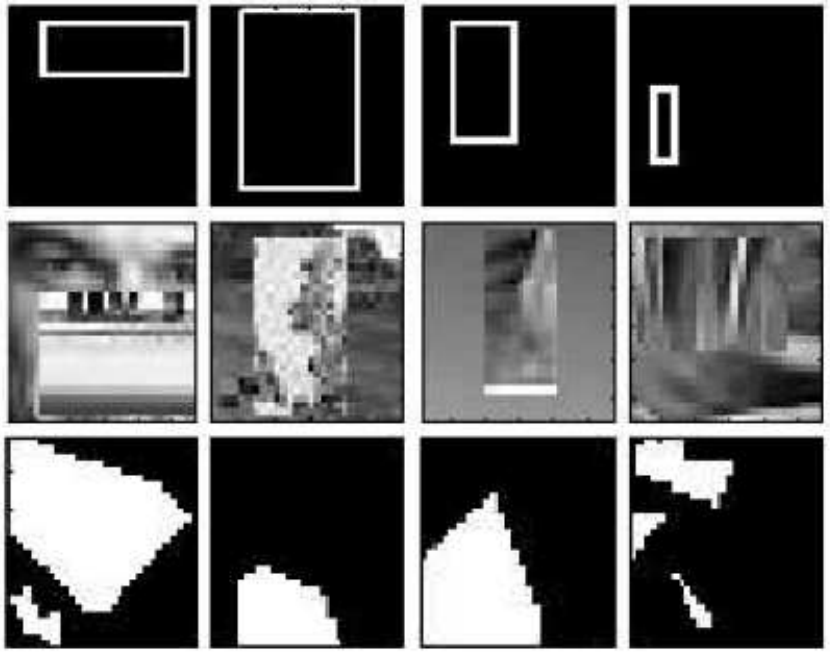
MNIST

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

Benchmark data



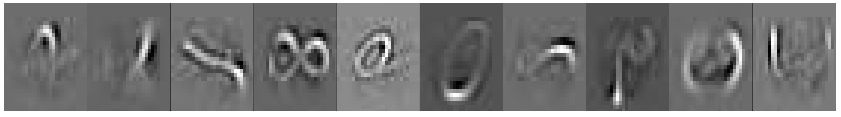
(a) *rot*, *bg-rand*, *bg-img*, *bg-img-rot*



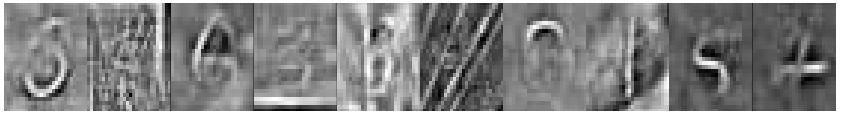
(b) *rect*, *rect-img*, *convex*

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

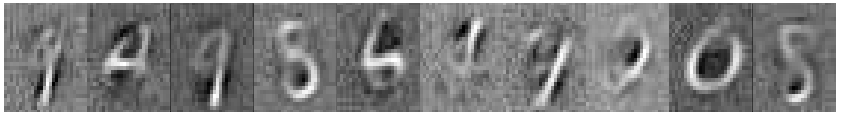
Receptive fields



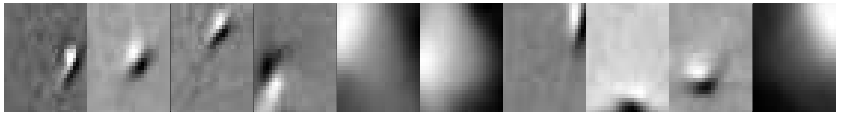
(a) MNIST digits



(b) MNIST digits with random image background

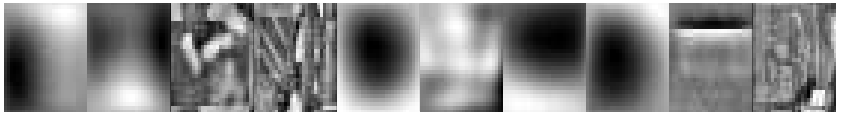


(c) MNIST digits with random noise background



(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.

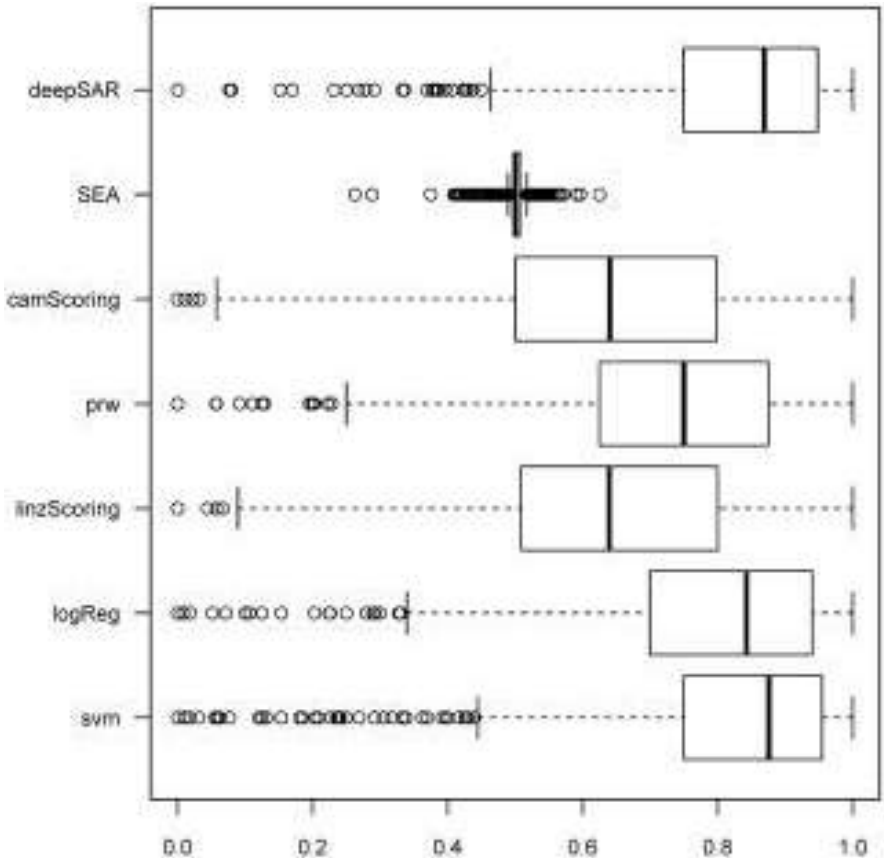
Results

Dataset		SVM_{rbf}	DBN_1	DBN_3	SAE_3	$SDAE_3$	RFN	
MNIST	50k-10k-10k	**1.40**$\pm$**0.23**	**1.21**$\pm$**0.21**	**1.24**$\pm$**0.22**	**1.40**$\pm$**0.23**	**1.28**$\pm$**0.22**	**1.27**$\pm$**0.22**	(1)
basic	10k-2k-50k	3.03$\pm$0.15	3.94$\pm$0.17	3.11$\pm$0.15	3.46$\pm$0.16	**2.84**$\pm$**0.15**	**2.66**$\pm$**0.14**	(1)
bg-rand	10k-2k-50k	14.58$\pm$0.31	9.80$\pm$0.26	**6.73**$\pm$**0.22**	11.28$\pm$0.28	10.30$\pm$0.27	7.94$\pm$0.24	(3)
bg-img	10k-2k-50k	22.61$\pm$0.37	**16.15**$\pm$**0.32**	**16.31**$\pm$**0.32**	23.00$\pm$0.37	**16.68**$\pm$**0.33**	**16.52**$\pm$**0.32**	(1)
rect	1k-0.2k-50k	2.15$\pm$0.13	4.71$\pm$0.19	2.60$\pm$0.14	2.41$\pm$0.13	1.99$\pm$0.12	**0.63**$\pm$**0.06**	(1)
rect-img	10k-2k-50k	24.04$\pm$0.37	23.69$\pm$0.37	22.50$\pm$0.37	24.05$\pm$0.37	21.59$\pm$0.36	**20.77**$\pm$**0.36**	(1)
convex	10k-2k-50k	19.13$\pm$0.34	19.92$\pm$0.35	18.63$\pm$0.34	18.41$\pm$0.34	19.06$\pm$0.34	**16.41**$\pm$**0.32**	(1)

Test error rate on all considered classification problems is reported together with a 95% confidence interval.

Predicting Drug-Target Interactions

Target prediction



AUC - Area under the ROC curve

- 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

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 methds 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