The Abdus Salam
International Centre
for Theoretical Physics

2361-17

School on Large Scale Problems in Machine Learning and Workshop on Common Concepts in Machine Learning and Statistical Physics

20 - 31 August 2012

MACHINE LEARNING IN SYSTEMS BIOLOGY: Factor Modeling 'Identifiability and Sparsity - Learning Models of Genomic Data'

Ole WINTHER

Technical University of Denmark DTU and University of Copenhagen KU
Denmark

Strada Costiera, 11 - 34151 - Trieste - Italy • Tel. +39 0402240111 • Fax. +39 040224163 • sci_info@ictp.it • www.ictp.it
ICTP is governed by UNESCO, IAEA, and Italy, and it is a UNESCO Category 1 Institute

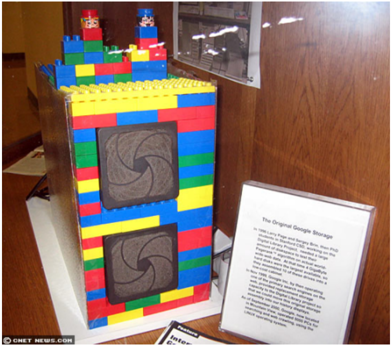
Factor Modeling

Identifiability and Sparsity - Learning models of genomic data

Ole Winther

Technical University of Denmark (DTU)

August 21, 2012



Google vision: develop the "perfect search engine," defined by co-founder Larry Page as something that, "understands exactly what you mean and gives you back exactly what you want."

The random surfer

$$\mathbf{p}^{(t)} = \mathbf{T}\mathbf{p}^{(t-1)}$$

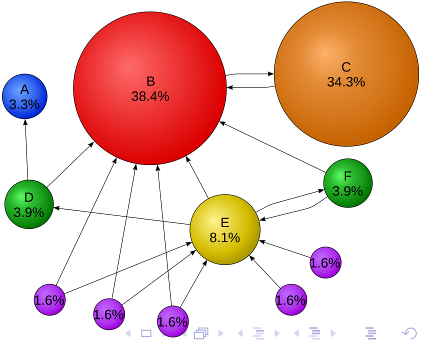
Motivation – multivariate data

The Anatomy of a Large-Scale Hypertextual Web Search Engine

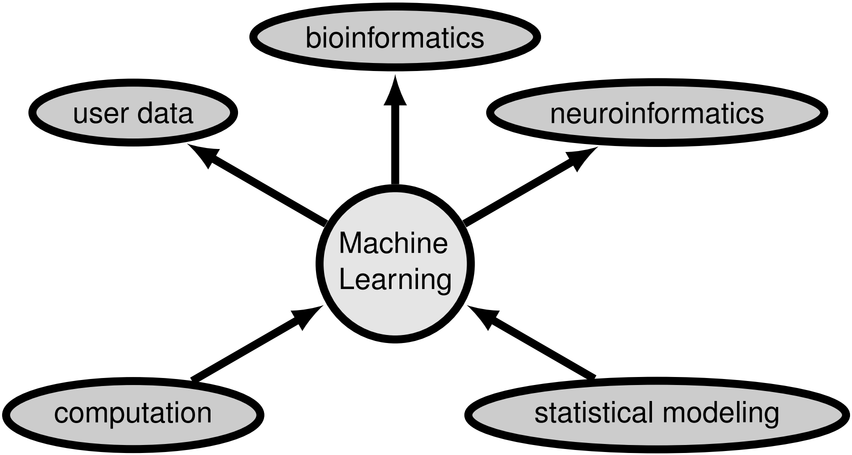
Sergey Brin and Lawrence Page

Computer Science Department,
Stanford University, Stanford, CA 94305, USA
sergey@cs.stanford.edu and page@cs.stanford.edu

$$\mathbf{p}^{(t)} = \mathbf{T}\mathbf{p}^{(t-1)}$$

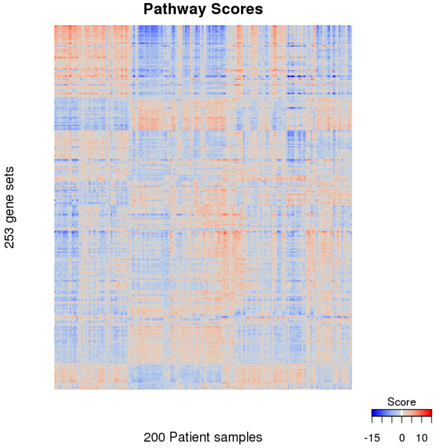


Motivation – multivariate data



Statistical machine learning merges statistics, modeling and computational sciences

- Learning can be
- Supervised $p(y|x)$
  - classification
  - regression
- Unsupervised $p(x)$
  - clustering
  - factor analysis



- Data is often (but not always) represented as a matrix of d features and N samples:

$$\text{size}(\mathbf{X}) = [d \;\; N]$$

- In stats $d = p$, $N = n$ and data matrix transposed $\mathbf{X} \rightarrow \mathbf{X}^T$
- Collaborative filtering:

$$\mathbf{X} = \text{item-user matrix}$$

- Gene expression:

$$\mathbf{X} = \text{gene-tissue matrix}$$

- Text analysis:

$$\mathbf{X} = \text{term-document matrix}$$

- Neuro-informatics: $\mathbf{X}$ = sensor-time series

Continuous latent variable models

18,000 movies; 480,000 users

x	1	1	x	...	x
x	x	x	5	...	x
x	x	3	x	...	x
x	4	3	x	...	2
...	x	x	x	...	x
x	5	x	1	...	x
x	x	3	3	...	x
x	1	x	x	...	2

- $\mathbf{v}_n$: "taste" vector of viewer n, $\mathrm{length}(\mathbf{v}_n) = K$.
- $\mathbf{u}_m$: "profile" vector movie m.
- Rating model:

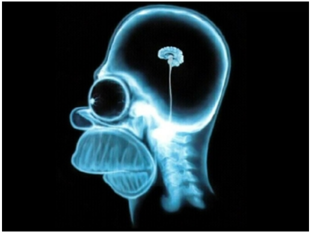
$$r_{mn} = \mathbf{u}_m \cdot \mathbf{v}_n + \epsilon_{mn}$$

- Learn $\mathbf{U}$ and $\mathbf{V}$ from rating matrix. Computation!

- Functional magnetic resonance imaging (fMRI) data
- Decomposing data into independent sources

$$\mathbf{X} = \sum_k \mathbf{u}_k \mathbf{v}_k^T + \epsilon$$

- $\mathbf{u}_k$ is the brain-image of the kth process
- $\mathbf{v}_k$ is the time-series of the kth process
- Play video



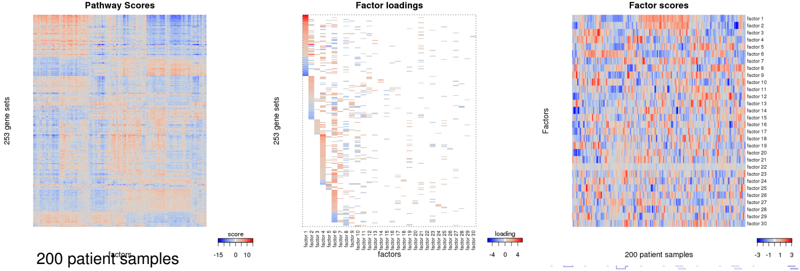
Bag of words representation – term-document matrix

Terms	Documents													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0

$\mathbf{u}_k$ is the kth latent topic and $\mathbf{v}_k$ is the usage of that topic across documents.

- Gene expression profiling – simultaneous measurement of 50k genes (mRNA levels).
- Use library of gene sets representing response to genetic and chemical perturbations.
- Covariation (redundancy) – use factor model

$$\mathbf{X} = \mathbf{W}\mathbf{Z} + \mathbf{E}$$

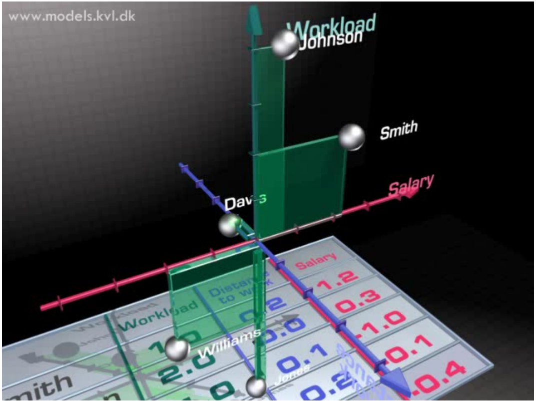


Roadmap / (learning objectives)

- Principal component analysis (PCA)
- Independent component analysis (ICA) - identifiability
- Factor modeling - Bayesian formulation with sparsity
- Insights from physics - fundamental limitations for learning covariance structure
- Sparse linear identifiable modeling (SLIM) - learning models of genomic data
- + exercises and breaks!

- Reading material:
- PCA: Bishop (Pattern Recogniyion and Machine Learning) 12-12.2.1, 12.1.2 and 12.2.1
- ICA: Bishop 12.4-12.4.1
- Factor analysis: 12.2.4 and in case story below
- Covariance learning: Hoyle and Rattray, 2003+2004; Alexei Onatski, 2007
- Henao and Winther, 2011; Shimizu et. al., 2006; Carvalho et. al., 2008. `http://cogsys.imm.dtu.dk/slim`

- Principal Component Analysis (PCA) is the number one multivariate data analysis method.
- Play video from `models.life.ku.dk`

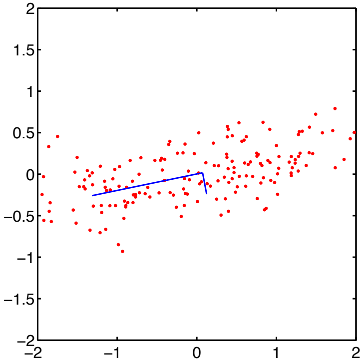


- Principal components (PCs): orthogonal directions with most variance.
- Empirical co-variance (centered) data:

$$\mathbf{S} = \frac{1}{N}\mathbf{X}\mathbf{X}^T$$

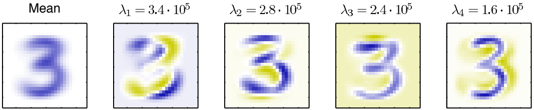
- `size(`**S**`) = [ d d ]`
- PCs: eigen-vectors of **S**

$$\mathbf{S}\mathbf{u}_i = \lambda_i \mathbf{u}_i$$



Plot axis $\sqrt{\lambda_i}\mathbf{u}_i$

- Database of N, $d = 28 \times 28 = 784$ pixel values
- Mean and first four PCs

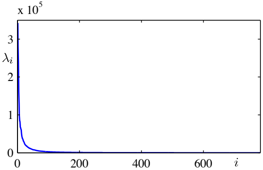


- Reconstruction

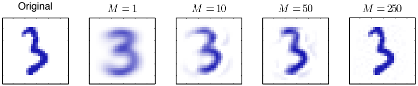
$$\tilde{\mathbf{x}}_n = \bar{\mathbf{x}} + \sum_{i=1}^{M} \left[(\mathbf{x}_n - \bar{\mathbf{x}})^T \mathbf{u}_i \right] \mathbf{u}_i$$

- Projections and reconstruction can be computed efficiently with singular value decomposition (SVD), see exercise.

Where is the signal?



(a)



Probabilistic PCA

- Tipping and Bishop, 1999 proposed:

$$\begin{aligned} p(\mathbf{z}) &= \mathrm{Norm}(\mathbf{z}; \mathbf{0}, \mathbf{I}) \\ p(\epsilon; \sigma^2) &= \mathrm{Norm}(\epsilon; \mathbf{0}, \sigma^2 \mathbf{I}) \end{aligned}$$

- $\Rightarrow$ **x** Gaussian with mean and covariance

$$\begin{aligned} \overline{\mathbf{x}} &= \mathbf{W}\overline{\mathbf{z}} + \overline{\epsilon} = \mathbf{0} \\ \overline{\mathbf{x}\mathbf{x}^T} &= \mathbf{W}\overline{\mathbf{z}\mathbf{z}^T}\mathbf{W}^T + \overline{\epsilon\epsilon^T} = \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I} \\ p(\mathbf{x}; \mathbf{W}, \sigma^2) &= \mathrm{Norm}(\mathbf{x}; \mathbf{0}, \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I}) \end{aligned}$$

- Log likelihood for $\mathbf{W}$ and σ^2 is joint distribution of all data:

$$\begin{aligned} \log L(\theta;\mathbf{X}) &= \sum_n \log p(\mathbf{x}_n|\mathbf{W},\sigma^2) \\ &= -\frac{N}{2}\left\{\log\det 2\pi\Sigma + \mathrm{Tr}\left[\Sigma^{-1}\mathbf{S}\right]\right\} \end{aligned}$$

- Model covariance: $\Sigma = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$
- Empirical covariance: $\mathbf{S} = \frac{1}{N}\mathbf{X}\mathbf{X}^T$
- Maximum likelihood: $\mathbf{W}_{\mathrm{ML}}$ is spanned by first M PCs
- The remaining variance is fitted by $\sigma^2\mathbf{I}$,

$$\sigma^2_{\mathrm{ML}} = \sum_{i=M+1}^{d} \lambda_i/(d-M) \ .$$

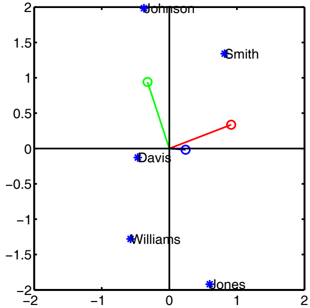
- Example of structured covariance estimation.

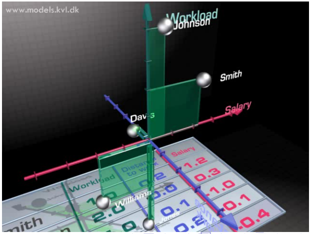
- We now have a probabilistic model for the data

$$p(\mathbf{x}; \mu_{\mathrm{ML}}, \Sigma_{\mathrm{ML}}) = \mathrm{Norm}(\mathbf{x}; \mu_{\mathrm{ML}}, \Sigma_{\mathrm{ML}})$$

- Projected distribution $M = 2$: $\mathbf{U}_M = [\mathbf{u}_1\ \mathbf{u}_2]$:

$$p(\mathbf{U}_M^T \mathbf{x}; \mu_{\mathrm{ML}}, \Sigma_{\mathrm{ML}}) = \mathrm{Norm}\left(\mathbf{U}_M^T\, \mathbf{x}; \mathbf{U}_M^T \mu_{\mathrm{ML}}, \mathbf{U}_M^T \Sigma_{\mathrm{ML}} \mathbf{U}_M\right)$$



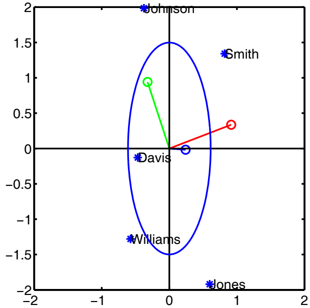


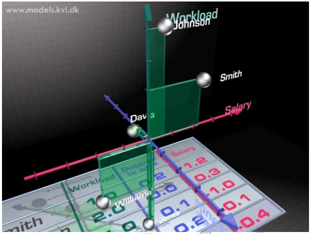
- We now have a probabilistic model for the data

$$p(\mathbf{x}; \mu_{\mathrm{ML}}, \Sigma_{\mathrm{ML}}) = \mathcal{N}(\mathbf{x}; \mu_{\mathrm{ML}}, \Sigma_{\mathrm{ML}})$$

- Projected distribution $M = 2$: $\mathbf{U}_M = [\mathbf{u}_1 \mathbf{u}_2]$:

$$p(\mathbf{U}_M^T \mathbf{x}; \mu_{\mathrm{ML}}, \Sigma_{\mathrm{ML}}) = \mathcal{N}\left(\mathbf{U}_M^T\, \mathbf{x}; \mathbf{U}_M^T \mu_{\mathrm{ML}}, \mathbf{U}_M^T \Sigma_{\mathrm{ML}} \mathbf{U}_M\right)$$





- Let us try to solve the cocktail party problem:

$$\text{Recordings} = \text{Mixing} \times \text{Speakers}$$

or

$$\mathbf{x} = \mathbf{W}\mathbf{z}$$

- Use PCA to estimate **W** (and **z**).
- Ignore complications of room acoustics.

- Stop sign! Non-uniqueness of solution!
- Likelihood only depends upon $\mathbf{W}$ through $\Sigma = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}$
- Rotate $\mathbf{W}$:

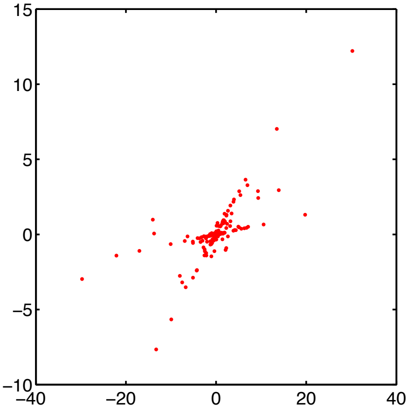
$$\mathbf{W} \leftarrow \widetilde{\mathbf{W}}\mathbf{U}$$

- leave covariance unchanged

$$\mathbf{W}\mathbf{W}^T = \widetilde{\mathbf{W}}\mathbf{U}\mathbf{U}^T\widetilde{\mathbf{W}} = \widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^T \ .$$

Independent component analysis (ICA)

- Prior knowledge to the rescue!
- Real signals are not Gaussian
- Example $\mathbf{x} = \mathbf{w}_1 z_1 + \mathbf{w}_2 z_2$
- with z_1 and z_2 independent and heavy tailed.
- Include this prior information in our modeling!



- Bell and Sejnowski Algorithm aka InfoMax
- Assumption square mixing and no noise

$$\mathbf{x} = \mathbf{W}\mathbf{z} \qquad \mathbf{W} : \quad d \times d$$

- Likelihood - one sample

$$p(\mathbf{x}|\mathbf{W}) = \int d\mathbf{z}\, P(\mathbf{x}|\mathbf{W}, \mathbf{z}) P(\mathbf{z}) = \int d\mathbf{z}\, \delta(\mathbf{x} - \mathbf{W}\mathbf{z}) P(\mathbf{z})$$

- Make change of variables $\mathbf{y} = \mathbf{W}\mathbf{z}$ and $d\mathbf{y} = |\mathbf{W}| d\mathbf{z}$:

$$\begin{aligned} p(\mathbf{x}|\mathbf{W}) &= \frac{1}{|\mathbf{W}|} \int d\mathbf{y} \delta(\mathbf{x} - \mathbf{y}) P(\mathbf{W}^{-1}\mathbf{y}) \\ &= \frac{1}{|\mathbf{W}|} P(\mathbf{W}^{-1}\mathbf{x}) \end{aligned}$$

- Maximize log likelihood: $\sum_n \log P(\mathbf{x}_n|\mathbf{W})$.

- If a statistical model $p(\mathbf{x};\theta)$ has the property that

$$p(\mathbf{x};\theta) = p(\mathbf{x};\theta') \quad \Rightarrow \quad \theta = \theta' \quad \text{for all } \ \theta, \theta' \in \Theta.$$

- then the model is said to be identifiability.
- The pPCA model is not identifiable since $\mathbf{W}$ and $\mathbf{WU}$ give same model.
- Many variants of ICA can be proven to be identifiable, Kagan et. al., 1973 and Comon, 1994.
- up to arbitrary permutation $\mathbf{P}$ and sign $\mathbf{Sign}$:

$$\mathbf{z} \rightarrow \mathbf{Sign}\,\mathbf{P}\,\mathbf{z} \quad \mathbf{W} \rightarrow -\mathbf{W}\,\mathbf{P}^{-1}\,\mathbf{Sign}$$

- PCA not strictly a statistical model, but PC projections identifiable up to sign.

- Matlab exercises on these topics available from

  http://www.imm.dtu.dk/Forskning/ISP/Undervisning/02901_2012.aspx

- (For offline use)
- Student Exercise 1
  - PCA and ICA on cocktail party problem - identifiability
- Student Exercise 2
  - Singular value decomposition (SVD) for PCA

Factor modeling

- A bit of history
  - Invented by Spearman 1904. A *single* underlying *g*-factor can explain most of variation in cognitive tests.
  - Raymond Cattell expanded on Spearman's idea of a two-factor theory of intelligence and developed 16 Personality Factors.
  - Widely used in any field working with multivariate data: Psychology, Economy, Bioinformatics,...
- Vanilla Bayes - Gaussian factors and Gaussian weights.
- Non-identifiable and identifiable models
- Sparsity and model selection

- We will only consider iid observations with Gaussian noise:

$$\mathbf{x}_n = \mathbf{W}\mathbf{z}_n + \epsilon_n$$

$$\mathbf{X}|\mathbf{W},\mathbf{Z},\Psi \sim \prod_{n=1}^{N} \mathrm{Norm}(\mathbf{x}_n|\mathbf{W}\mathbf{z}_n,\Psi) \ ,$$

where $\mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_N]$ and $\mathbf{Z} = [\mathbf{z}_1 \ldots \mathbf{z}_N]$.

- Priors:

$$\psi_i^{-1}|s_s,s_r \sim \mathrm{Gamma}(\psi_i^{-1},s_s,s_r)$$

$$\mathbf{z}_n \sim \mathrm{Norm}(\mathbf{z}_n|\mathbf{0},\mathbf{I}) \ , \quad \mathbf{w}_j \sim \mathrm{Norm}(\mathbf{w}_j|\mathbf{0},\mathbf{D})$$

- $\psi_i = \psi$ is a (non-identifiable) Bayesian version of pPCA.

- Access significance of PCA findings in:
  - Frequentist sense - how different PCs for another dataset?
  - Bayesian sense - posterior over $\mathbf{W}$ and Ψ - how much does subspace vary?
- Bayesian answer depends upon the prior!
- Depending upon whom you ask this is big weakness, or
- an advantage as our assumptions are explicit.

- Closer look at the distribution for the (inverse) noise variance

$$\psi_i^{-1} | s_s, s_r \ \sim \ \mathrm{Gamma}(\psi_i^{-1}, s_s, s_r)$$

- s_s shape and s_r rate:

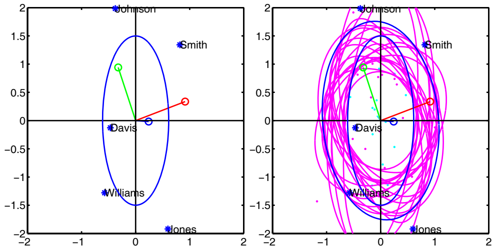
$$< \psi_i >= \frac{s_r}{s_s - 1} \ , \quad < \psi_i^2 > - < \psi_i >^2 = \frac{s_r^2}{(s_s - 1)^2 (s_s - 2)}$$

- Shape s_s needs to be $s_s > 1$ and $s_s > 2$ for mean and variance to be defined.
- Rate s_r scales mean and standard deviation of variance

- MCMC: draw samples, $\mathbf{W}^{(r)}, \Psi^{(r)}$ from posterior $p(\mathbf{W}, \Psi|\mathbf{X})$
- Plot contours of covariance samples and expected value

$$\langle \Sigma \rangle_{\mathbf{W}|\mathbf{X}} \approx \frac{1}{R} \sum_{r}^{R} \left[\mathbf{W}^{(r)} (\mathbf{W}^{(r)})^T + \Psi^{(r)} \right]$$

- Prior $s_s = 2$ and $s_r = 1$: $\mathrm{var}(\psi_i) = \frac{s_r^2}{(s_s-1)^2(s_s-2)} = \infty$.

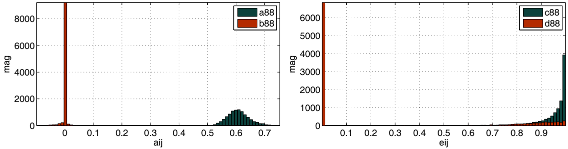


- Bayesian sparsity - slab and spike!
- Two component mixture of a continuous component and a point-mass at zero:

$$W_{ij}|\eta_{ij} \sim (1-\eta_{ij})\delta(W_{ij}) + \eta_{ij}\mathcal{N}(W_{ij}|0,\tau_{ij})$$

- Parsimoneous two-level model, Carvalho et. al., JASA, 2008.

$$\eta_{ij}|\nu_j \sim (1-\nu_j)\delta(\eta_{ij}) + \nu_j \mathrm{Beta}(\eta_{ij}|\alpha_p\alpha_m, \alpha_p(1-\alpha_m))$$



- Matlab exercises on these topics available from

  `http://www.imm.dtu.dk/Forskning/ISP/Undervisning/02901_2012.aspx`

- (For offline use)
- Student Exercise 3:
  - Bayesian factor analysis using MCMC inference on some simple datasets.
  - Inference summaries
  - Identifiable quantities and multivariate data.
- Student Exercise 4:
  - Model interpretation - slab and spike
  - Sparsity
  - Model selection - how many factors?
- Break!

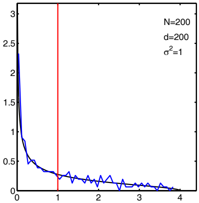
- Factor model $\mathbf{x} = \mathbf{Wz} + \epsilon$, true covariance

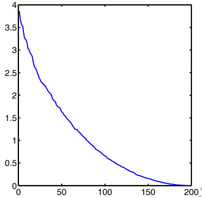
$$\mathbf{C} = \overline{\mathbf{x}\mathbf{x}^T} = \overline{(\mathbf{Wz} + \epsilon)(\mathbf{Wz} + \epsilon)^T} = \mathbf{W}\mathbf{W}^T + \Psi$$

- Empirical covariance:

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T$$

- Marcenko-Pastur - eigenvalue spectrum of $\mathbf{S}$: $p(\lambda)$
- for $d \to \infty$ and $\alpha \equiv d/N$ finite, example $\mathbf{C} = \sigma^2 I$





- How much data do we need to learn direction **w**?
- Hoyle and Rattray, 2003+2004; Alexei Onatski, 2007; (Halkjær and Winther, 1997).
- Bulk spectrum - contribution from $\mathbf{C} = \sigma^2 I$

$$p(\lambda) = (1-\alpha)\Theta(1-\alpha)\delta(\lambda) + \alpha \frac{\sqrt{(\lambda-\lambda_-)(\lambda_+-\lambda)}}{2\pi\lambda\sigma^2}$$

with $\lambda_\pm = \sigma^2(1 \pm \sqrt{\alpha})^2/\alpha$

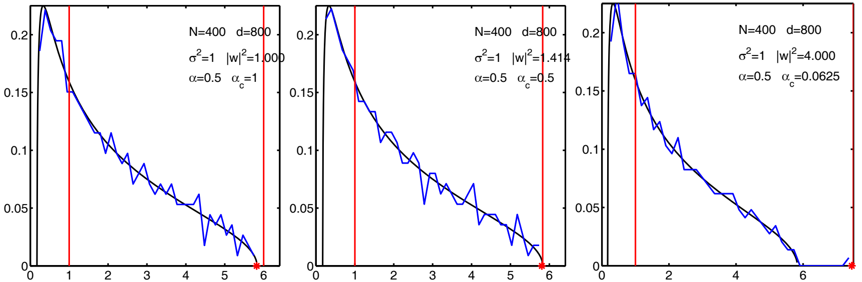
- Single direction $\mathbf{w}_m \mathbf{w}_m^T$ only visible in spectrum when $\alpha \geq \alpha_c$ as $\delta(\lambda - \lambda_u)$ with

$$\lambda_u = (\sigma^2 + |\mathbf{w}_m|^2)\left(1 + \frac{\sigma^2}{\alpha|\mathbf{w}_m|^2}\right)$$

- Learning transition at $\lambda_u(\alpha_c) = \lambda_+(\alpha_c)$: $\alpha_c = \left(\frac{\sigma^2}{|\mathbf{w}_m|^2}\right)^2$

Learning factors - symmetry breaking

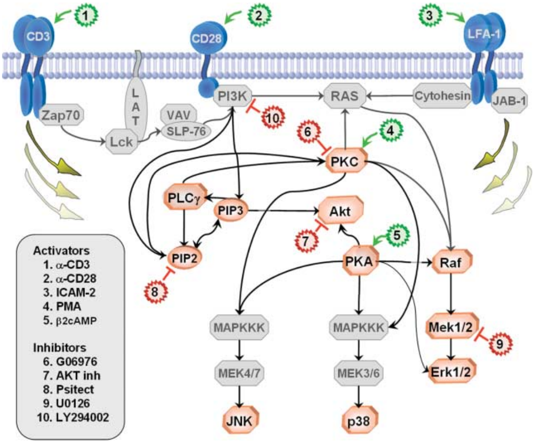
$\alpha < \alpha_c$ $\qquad$ $\alpha = \alpha_c$ $\qquad$ $\alpha > \alpha_c$



- The number of factors should adapt to data.
- Knowles and Ghahramani, 2004+2011 model the sparsity pattern in **W** with an Indian Buffet Process.
- Model has an infinite number of factors but only a finite number is active.
- Number of active factors adapt approximately to the number of orthogonal directions with

$$|\mathbf{w}_m^2| \geq \frac{d}{N}\sigma^2$$

Protein signalling network textbook – Sachs et. al. Science, 2005.



- Sparse linear identifiable modeling (SLIM)
  - Use connection between factor model and Bayes network
  - to learn structure of both and do model comparison
  - Henao and Winther, JMLR, 2011. `http://cogsys.imm.dtu.dk/slim`

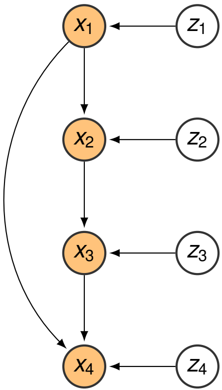
- Factor model (FM)

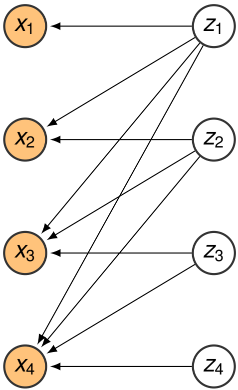
$$\mathbf{x} = \mathcal{A}\mathbf{z} + \epsilon$$

- Linear Bayes network (LBN)

$$\mathbf{x} = \mathbf{C}\mathbf{x} + \mathbf{z}$$

- $\mathbf{C} = \mathbf{P}^{-1}\mathbf{B}\mathbf{P}$ with $\mathbf{P}$ permutation matrix
- $\mathbf{B}$ upper triangular $\Leftrightarrow$ $\mathbf{C}$ defines a directed acyclic graph (DAG)
- Sparse $\mathcal{A}$ and $\mathbf{B}$ (parsimonious)
- Identifiable - $\mathbf{z}$ must be non-Gaussian – no rotation ambiguity
- Learn both FM and LBN and perform quantitative test likelihood model comparison
- Models complementary and aid in scientific discovery
- Sparsity and non-Gaussianty often justifiable assumptions


x_1
x_2
x_3
x_4
z_1
z_2
z_3
z_4


x_1
x_2
x_3
x_4
z_1
z_2
z_3
z_4

- Factor model

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \epsilon$$

- Bayesian network (BN)

$$\mathbf{x} = \mathbf{C}\mathbf{x} + \mathbf{z}$$

- Rewriting BN as factor model:

$$\mathbf{x} = (\mathbf{I} - \mathbf{C})^{-1}\mathbf{z}$$

- Inspired by LiNGAM, Shimizu et. al., JMLR, 2006.
- Novelty here: explicit model of sparsity, stochastic order search and quantitative model comparison.
- Stochastic ˘ search $\mathbf{P}^{-1}\mathbf{W}\mathbf{P}_f$ should be approximately triangular.

- Model comparison, parameter shorthand θ
  - Marginal likelihood

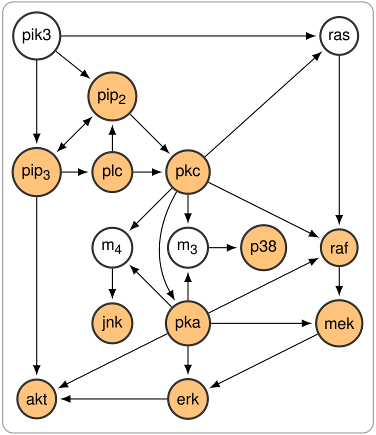
$$p(\mathbf{X}|\mathcal{M}) = \int p(\mathbf{X}|\theta, \mathcal{M}) p(\theta|\mathcal{M}) d\theta$$

  - Test likelihood

$$p(\mathbf{X}^{\text{test}}|\mathbf{X}) = \int p(\mathbf{X}^{\text{test}}|\theta, \mathcal{M}) p(\theta|\mathbf{X}, \mathcal{M}) d\theta$$

- Test likelihood relatively easy to compute!
- Extensions, see `http://cogsys.imm.dtu.dk/slim`,
  - Non-linear DAGS
  - Latent variables - raises new identifiability problems
  - time-series data - temporal smoothness with Gaussian process factors

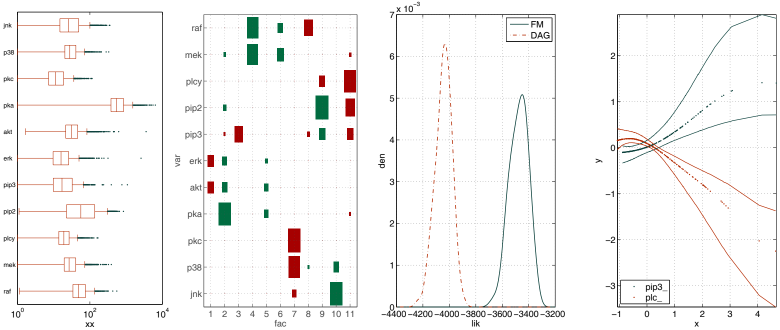
Single cell flow cytometry data



- Single cell flow cytometry measurements of 11 phosphorylated proteins and phospholipids.
- Data was generated from a series of stimulatory cues and inhibitory interventions.
- Observational data: 1755 general stimulatory conditions,
- Experimental data $\sim$ 80% not used in our approach.
- Not "small *n* large *p*"!

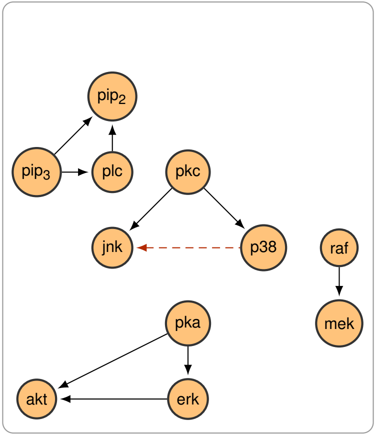
Single cell flow cytometry data

$$\log\langle\mathcal{L}_{\mathrm{FM}}\rangle = -3.46e3$$

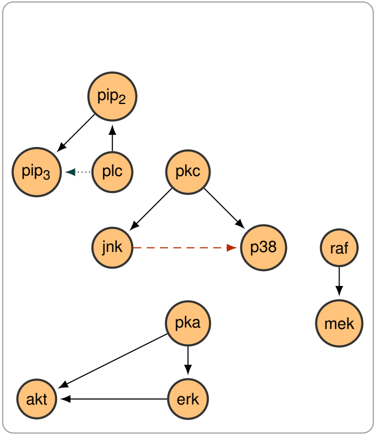


Single cell flow cytometry data

$$\log\langle\mathcal{L}_{\text{DAG}}\rangle = -4.30e3$$

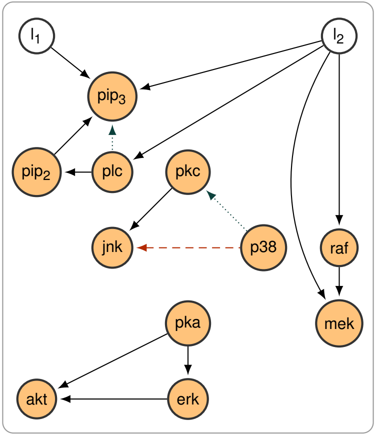


$$\log\langle\mathcal{L}_{\mathrm{DAG}}\rangle = -4.10e3$$



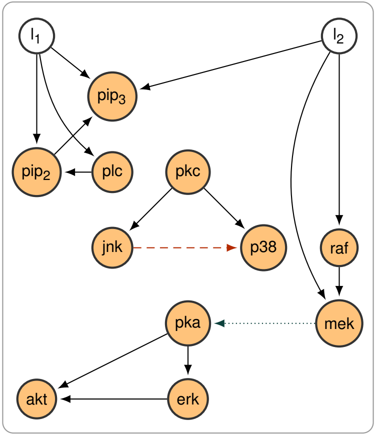
Single cell flow cytometry data

$$\log\langle\mathcal{L}_{\mathrm{DAG}}\rangle = -3.4e3$$



Single cell flow cytometry data

$$\log\langle\mathcal{L}_{\mathrm{DAG}}\rangle = -3.70e3$$



- Factor models - from PCA to
    - identifiable models (ICA) and
    - sparsity (model selection)
- We can learn learn model structure when $N \gg d$
- Markov chain Monte Carlo - used as standard inference tool.
- Thank you!