Introduction to Bayesian networks and graphical models

Statistical Modeling and Machine Learning in Computational Systems Biology June 22-26, 2009, Tampere, Finland

Ole Winther

Technical University of Denmark (DTU) & University of Copenhagen (KU)

June 24, 2009

All lectures

1. Introduction to graphical models and Bayesian networks
2. Estimating the size of the transcriptome
3. Using biological prior information in motif discovery
4. Learning linear Bayes networks with sparse Bayesian models

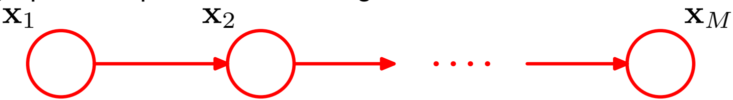
Common theme:

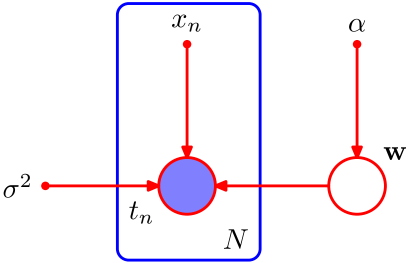
- Complex Bayesian model building possible and advantageous
- Model checking – prediction, marginal- and test-likelihood

Lecture 1

- Introduction to graphical models and Bayesian networks
- Machine learning
- Example application – collaborative filtering $1M\$$-prize
- Summary and reading

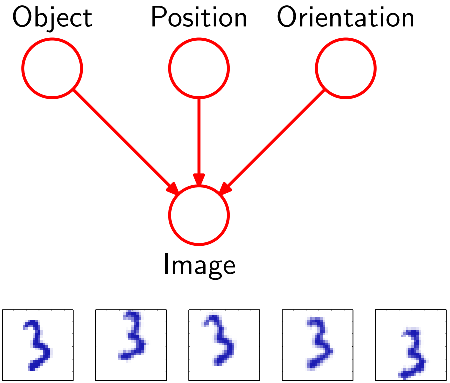
- Graphical representation of conditional probabilities and independence
- All standard probabilistic statistical models can be given a graphical representation – e.g. Markov





- or regression

Generative models



- Variables may be latent and unobserved
- Bayesian networks – directed acyclic graphs (DAGs)
- Also undirected graphs – Markov random fields.

Understanding conditional probabilities

- Smokers are more likely to have lung cancer than random person:

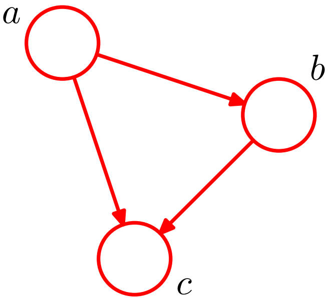
$$P(\text{Lung cancer}|\text{Smoking}) > P(\text{Lung cancer})$$

- Bayes theorem relate joint to conditionals

$$\begin{aligned} P(X,Y) &= P(X|Y)P(Y) = P(Y|X)P(X) \\ P(Y|X) &= \frac{P(X|Y)P(Y)}{P(X)} \\ P(X) &= \sum_Y P(X,Y) = \sum_Y P(Y|X)P(Y) \end{aligned}$$

- We can use Bayes theorem to calculate $P(\text{Lung cancer}|\text{Smoking})$.

Structured probabilistic models – directed acyclic graphs (DAGs)



Graph reveals conditional independence (in example non).

$$P(a, b, c) = P(c|a, b)P(b|a)P(a)$$

The structure can be exploited to make effective inference

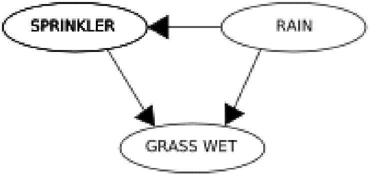
- predictions

$$P(\text{“financial crisis 2010”}|\text{“economy 2009”})$$

- learning model parameters
- learning network structure

Example Sprinkler

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99

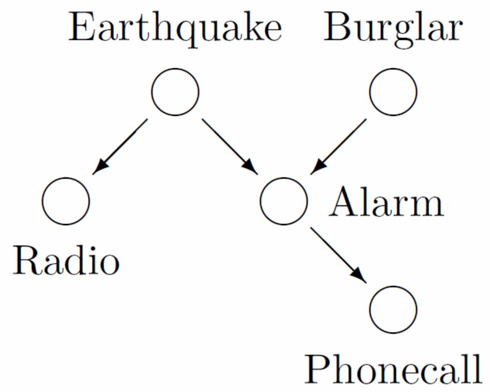


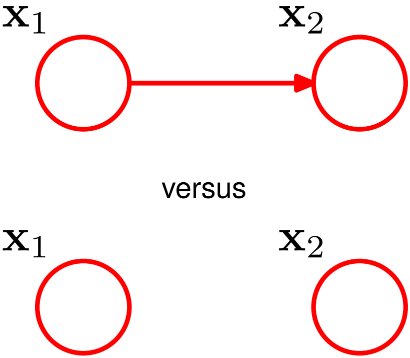
RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

$$P(GW, S, R) = P(GW|S, R)P(S|R)P(R)$$

Burglar alarm – explaining away





Aims – test for

1. Independence versus dependence
2. Directionality, who are the parents of a node.

- $\mathcal{H}_0$ null hypothesis independence
- $\mathcal{H}_1$ dependence: no factorization
- This is a classical frequentist statistical test situation

$$\Lambda = \frac{L(\widehat{\boldsymbol{\theta}}_1; \mathbf{X}, \mathcal{H}_1)}{L(\widehat{\boldsymbol{\theta}}_0; \mathbf{X}, \mathcal{H}_0)} \qquad \chi^2\text{-distributed with } \ |\boldsymbol{\theta}_1| - |\boldsymbol{\theta}_0| \ \text{ d.f.}$$

- Many dimensions: $\mathcal{O}(d!2^{d(d-1)/2})$ possible structures
- Bayesian approach: specify “probability of everything”

- Marginal likelihood - independent model
- $\mathcal{H}_0$ independence: Likelihood: $\boldsymbol{\theta}_0 = \{\boldsymbol{\theta}_0(1), \boldsymbol{\theta}_0(2)\}$

$$p(\mathbf{x}_1, \mathbf{x}_2 | \boldsymbol{\theta}_0, \mathcal{H}_0) = p(\mathbf{x}_1 | \boldsymbol{\theta}_0(1), \mathcal{H}_0)\, p(\mathbf{x}_2 | \boldsymbol{\theta}_0(2), \mathcal{H}_0)$$

- Specify priors - for example independent

$$p(\boldsymbol{\theta}_0 | \mathcal{H}_0) = p(\boldsymbol{\theta}_0(1) | \mathcal{H}_0)\, p(\boldsymbol{\theta}_0(2) | \mathcal{H}_0)$$

- Model likelihood (marginal likelihood)

$$p(\mathcal{D}|\mathcal{H}_0) = \int p(\mathcal{D}|\boldsymbol{\theta}_0, \mathcal{H}_0)\, p(\boldsymbol{\theta}_0|\mathcal{H}_0)\, d\boldsymbol{\theta}_0 = p(\mathbf{X}_1|\mathcal{H}_0)\, p(\mathbf{X}_2|\mathcal{H}_0)$$

with data $\mathcal{D} = \{\mathbf{X}_1, \mathbf{X}_2\}$ and $\mathbf{X}_d = \{\mathbf{x}_{id}\}_{i=1,\ldots,n}$.

- Dependent model
- $\mathcal{H}_1$ dependence: No factorization in likelihood nor prior

$$p(\mathcal{D}|\mathcal{H}_1) = \int p(\mathcal{D}|\boldsymbol{\theta}_1, \mathcal{H}_1)\, p(\boldsymbol{\theta}_1|\mathcal{H}_1)\; d\boldsymbol{\theta}_1 \ .$$

- Bayes factor

$$\frac{p(\mathcal{D}|\mathcal{H}_1)}{p(\mathcal{D}|\mathcal{H}_0)}$$

  replace log likelihood ratio test.
- Sampling distribution considerations possible, but not widely used (Gelman, Carlin, Stern & Rubin).

Example - discrete data (MacKay 2003)

	$x_2=0$	$x_2=1$	
$x_1=0$	760	5	765
$x_1=1$	190	45	235
	950	50	

- Likelihood

$$p(\mathcal{D}|\boldsymbol{\theta}) = \theta_{00}^{n_{00}}\,\theta_{01}^{n_{01}}\,\theta_{10}^{n_{10}}\,\theta_{11}^{n_{11}}$$

- Independence $\mathcal{H}_0$:

$$\theta_{kl} = \theta_k(1)\;\theta_l(2)$$

Counts are the sufficient statistics $n_k = \sum_{i=1}^{n} x_{ik}$:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{n_k}$$

Enter a very convenient prior - the Dirichlet

$$p(\boldsymbol{\theta};\boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \delta(\sum_{k'} \theta_{k'} - 1)$$

Normalizer:

$$Z(\boldsymbol{\alpha}) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)} \; .$$

Discrete data

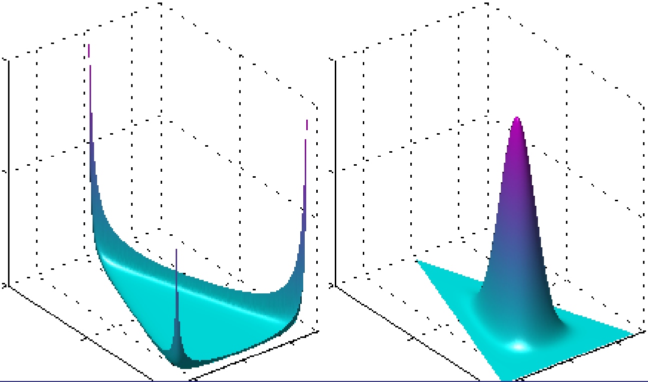
$$p(\boldsymbol{\theta};\boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha})}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}\delta(\sum_{k'}\theta_{k'}-1)$$

$K = 3$,

$\alpha_k = \alpha$,

left $\alpha < 1$ and

right $\alpha > 1$.



Multinomial likelihood

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{n_k}$$

Dirichlet prior

$$p(\boldsymbol{\theta};\boldsymbol{\alpha}) = \frac{1}{Z(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \delta(\sum_{k'} \theta_{k'} - 1)$$

Dirichlet posterior

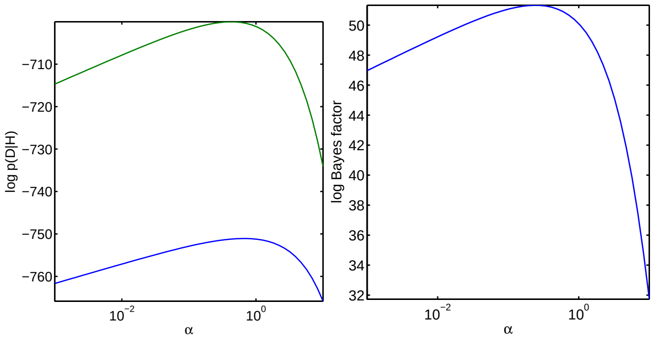
$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta};\boldsymbol{\alpha})}{p(\mathcal{D};\boldsymbol{\alpha})}$$

Polya marginal likelihood

$$p(\mathcal{D};\boldsymbol{\alpha}) = \frac{Z(\boldsymbol{\alpha} + \mathbf{n})}{Z(\boldsymbol{\alpha})}$$

$\mathcal{H}_1$ $(K = 4)$ versus $\mathcal{H}_0$ $(2 \times [K = 2])$

	x_2=0	x_2=1	
x_1=0	760	5	765
x_1=1	190	45	235
	950	50	



Learning of parenthood

- We are now ready for the harder task of making inference about parenthood.
- What does this actually mean?
- Likelihood equivalence

$$p(\mathbf{x}_1, \mathbf{x}_2) = p(\mathbf{x}_1|\mathbf{x}_2)p(\mathbf{x}_2) = p(\mathbf{x}_2|\mathbf{x}_1)p(\mathbf{x}_1)$$

- So from the observational data alone we cannot say anything about parenthood.
- Heckerman, Geiger and Chickering, 1995: choose prior such that marginal likelihood equivalent.

- We can still test different hypotheses about parenthood, but strong assumptions needed!
- Consider example and $p(x_1|x_2)p(x_2)$ – we have 3 binomials

$$p(x_1|x_2 = 0), p(x_1|x_2 = 1) \;\text{ and }\; p(x_2)$$

- We assume independence between prior distributions

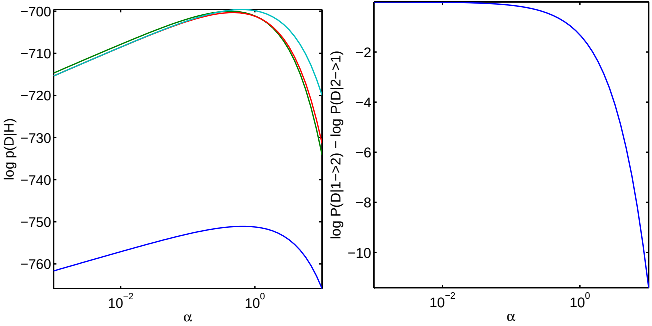
$$p(\theta|\mathcal{H}_{2\to1}) = p(\theta_{\cdot|0}|\mathcal{H}_{2\to1})p(\theta_{\cdot|1}|\mathcal{H}_{2\to1})p(\theta(2)|\mathcal{H}_{2\to1})$$

- We call this model $\mathcal{H}_{2\to1}$ but all that we are really testing is how well the data agrees with this specific parameter independence assumption.

Learning of parenthood

Comparing $\mathcal{H}_0, \mathcal{H}_1, \mathcal{H}_{1\to2}$ and $\mathcal{H}_{2\to1}$

	$x_2=0$	$x_2=1$	
$x_1=0$	760	5	765
$x_1=1$	190	45	235
	950	50	



Learning of parenthood

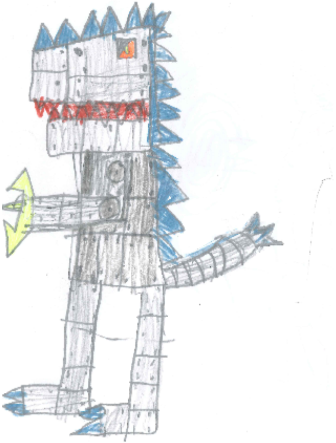
- Can we make causal inference from data?
- Distinguish between observational and experimental data
- Judea Pearl and others:

  no go for learning from (observational) data.
- Some Bayesians:

  We can still test different hypotheses about parenthood,
  but we have to make assumptions explicit.
- If you want to avoid trouble - use directionality instead of causality.

Machine learning

- Predictive and often statistical – grand goal is to achieve human like generalization.
- From wikipedia: “Applications for machine learning include natural language processing, syntactic pattern recognition, search engines, medical diagnosis, bioinformatics, brain-machine interfaces and cheminformatics,. . . ”
- The “Google paradigm”. . .



- . . . more data is different


EXPERT OPINION
Contact Editor: **Brian Brannon**, bbrannon@computer.org

The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

IEEE Intelligent Systems, 2009.

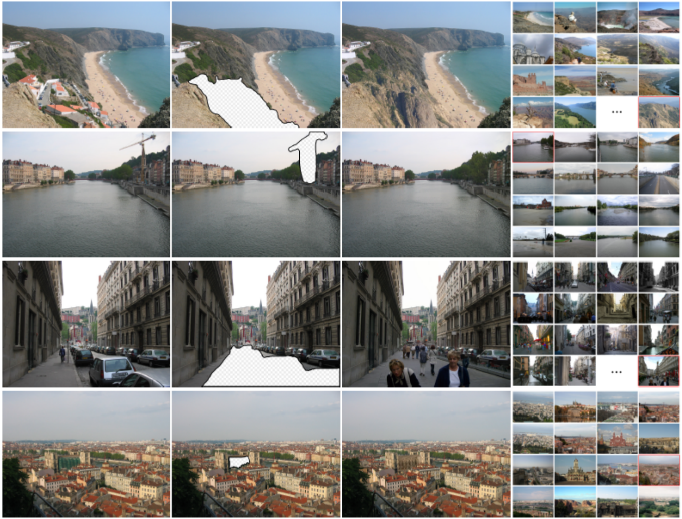
- Use representation that scales well (avoid curse of dimensionality)
- Unsupervised learning in non-parametric models (e.g. huge word frequency tables)

What you can do with 1M images

J. Hays and A.A. Efros, *Scene Completion Using Millions of Photographs*, Comm. ACM, 2008

Original Image Input Scene Matches Output

What you can do with 1M images



Original Image Input Output Matching Scene

Netflix prize

Netflix prize

- Netflix - online movie rental (DVDs).
- Collaborative filtering – predict user rating from past behavior of user.
- Improve Netflix own system by 10% to win.
- training.txt – $R = 10^8$ ratings, scale 1 to 5 for $M = 17.770$ movies and $N = 480.189$ users.
- qualifying.txt – 2.817.131 movie-user pairs, (continuous) predictions submitted to Netflix returns a RMSE.
- Rating matrix r_{mn} mostly missing values, 98.5%.

Some key numbers

Method	RMSE	% Improv.
Cinematch	0.9514	0%
Our Method	?	?
Best 13-5-2009	?	?
Grand prize	0.8563	10%

RMSE = root mean squared error

Collaborative filtering task

- Relatively large data set - 10^8 data points
- Very heterogeneous - viewers and movies with few ratings
- Ratings $\in \{1, 2, 3, 4, 5\}$ noisy (subjective use of scale, non-stationary,...)
- Complex model needed to capture latent structure
- Regularization! We use Bayesian averaging – easy to tune parameters.

matrix factorization

- Model taste of viewer n with a K-dimensional vector $\mathbf{v}_n$:

$$h_{mn} = \mathbf{u}_m \cdot \mathbf{v}_n + \epsilon_{mn} \qquad \mathcal{N}(\epsilon_{mn}|0, \gamma^{-1})$$

- Linear factor model $r_{mn} = h_{mn}$ or ordinal regression:

$$p(r_{mn}|h_{mn}) = \Phi(h_{mn} - b_{r_{mn}}) - \Phi(h_{mn} - b_{r_{mn}+1})$$

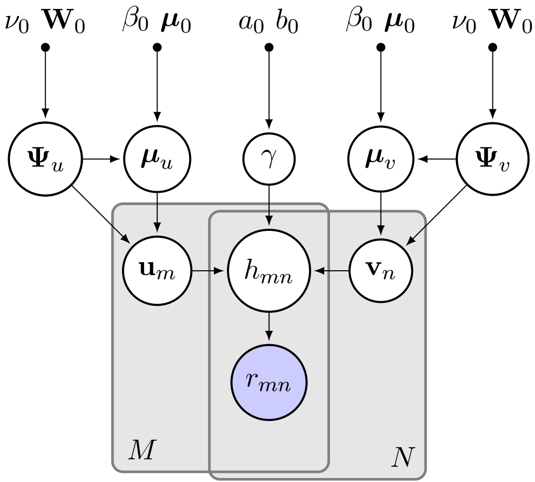
- Quadratic regularization of factors

$$p(\mathbf{u}_m|\boldsymbol{\mu}_u, \boldsymbol{\Psi}_u) = \mathcal{N}(\mathbf{u}_m|\boldsymbol{\mu}_u, \boldsymbol{\Psi}_u^{-1})$$

- Hierarchical Bayesian prior

$$p(\boldsymbol{\mu}_u, \boldsymbol{\Psi}_u) = \mathcal{N}(\boldsymbol{\mu}_u|\boldsymbol{\mu}_0, (\beta_0 \boldsymbol{\Psi}_u)^{-1})\, \mathcal{W}(\boldsymbol{\Psi}_u|\mathbf{W}_0, \nu_0)$$

matrix factorization



- Draw samples from distribution $p(\theta)$

$$\theta^{(1)}, \ldots, \theta^{(R)}$$

- Approximate average of $f(\theta)$ as

$$\langle f(\theta) \rangle = \int d\theta\, f(\theta)\, p(\theta) \approx \frac{1}{R} \sum_{r=1}^{R} f(\theta^{(r)})$$

- Sample $\{\theta^{(r)}\}_{r=1,\ldots,R}$ is called Markov chain because it is generated from a Markov process with transition kernel $T(\theta^{(r)}|\theta^{(r-1)})$.

- Markov chain sufficient and necessary condition: $p(\theta)$ must be stationary distribution, ergodicity and non-cyclic.
- Sufficient condition: Detailed balance

$$T(\theta'|\theta)\, p(\theta) = T(\theta|\theta')\, p(\theta')$$

- Important practical issue: convergence of Markov chain (burn-in).

Gibbs Sampling

- Just one example of a MCMC method.
- A special case of Metropolis-Hastings (the workhorse of MCMC).
- Split variables in a number of subsets for example $\theta = \{\theta_1, \theta_2\}$
- Many cases impossible to sample from $p(\theta_1, \theta_2)$ but easy to sample from conditionals:

$$p(\theta_1|\theta_2) \quad \text{and} \quad p(\theta_2|\theta_1)$$

Gibbs sampling: Alternate between drawing from each conditional

Detailed balance Gibbs sampling

- Detailed balance definition:

$$T(\theta'|\theta)\, p(\theta) = T(\theta|\theta')\, p(\theta')$$

- Transition kernel Gibbs for first sub-step:

$$T_1(\theta'|\theta) = p(\theta'_1|\theta_2)\delta(\theta'_2 - \theta_2)$$

- Detailed balance proof Gibbs - use that θ_2 remains unchanged in both directions:

$$\begin{aligned} T_1(\theta'|\theta)p(\theta) &= p(\theta'_1|\theta_2)\delta(\theta'_2 - \theta_2)p(\theta_1|\theta_2)p(\theta_2) \\ T_1(\theta|\theta')p(\theta') &= p(\theta_1|\theta_2)\delta(\theta'_2 - \theta_2)p(\theta'_1|\theta_2)p(\theta_2) \end{aligned}$$

- Easy to show $T = T_2 T_1$ obeys detailed balance if T_1 and T_2 do.

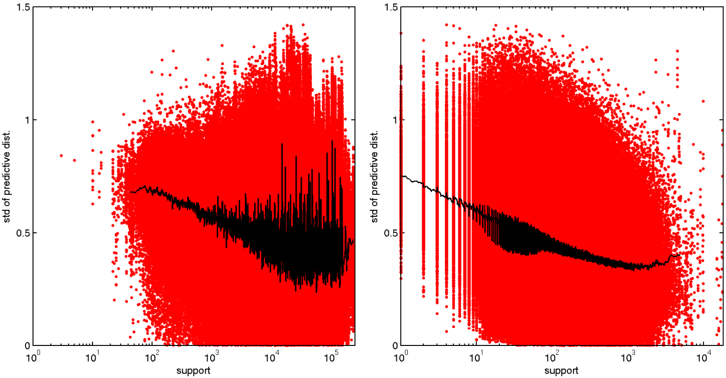
Gibbs sampling inference Netflix

- Draw samples from conditionals, e.g.

$$
\begin{aligned}
p(\mathbf{u}_m|\mathrm{rest}) &\propto \prod_{n\in\Omega(m)} p(h_{mn}|\mathbf{u}_m, \mathbf{v}_n, \gamma)\, p(\mathbf{u}_m|\boldsymbol{\mu}_u, \boldsymbol{\Psi}_u) \\
&= \prod_{n\in\Omega(m)} \mathcal{N}(h_{mn}|\mathbf{u}_m \cdot \mathbf{v}_n, \gamma^{-1})\, \mathcal{N}(\mathbf{u}_m; \boldsymbol{\mu}_u, \boldsymbol{\Psi}_u^{-1})
\end{aligned}
$$

- We have many parameters $K(M+N) + K(K+1)/2 + 1 = 10^8$ for $K = 200$!
- Convergence (prediction-wise): 20 burn-in steps and $S = 180$ samples!
- Predictive mean $\langle r_m n \rangle \approx \frac{1}{S} \sum_{s=1}^{S} \sum_{r=1}^{5} rp(r_{mn}|h_{mn}$
- Highly parallelizable!

Results



Predictive uncertainty: Standard deviation $\sqrt{\langle r_{mn}^2 \rangle - \langle r_{mn} \rangle^2}$ as a function of coverage, movie (left) and viewer (right).

Some performance numbers

Method	RMSE	Improv.
Cinematch	0.9514	0%
Our Method, $k = 50$	0.8958	5.84%
Our Method, $k = 100$	0.8930	6.14%
Our Method, $k = 200$	0.8917	6.27%
Best 13-5-2009	0.8590	9.71%
Grand prize	0.8563	10%

Our approach is to our knowledge best ‘single model’
Further improvements - model temporal effects.

Summary and reading

- Graphical models and Bayesian networks
- Machine learning – hypothesis generating and predictive approaches
- Large scale Bayesian inference for collaborative filtering (w Ulrich Paquet and Blaise Thomson, Cambridge)
- Books: C. Bishop, Pattern Recognition and Machine Learning, Springer; D. MacKay, Information Theory, Inference, and Learning Algorithms, Cambridge; J. Pearl, Causality: Models, Reasoning, and Inference, Cambridge; Gelman, Carlin, Stern & Rubin (Bayesian standard ref.)