# Large-scale Bayesian Inference for Collaborative Filtering


**Ole Winther**

**Informatics and Mathematical Modelling**
**Technical University of Denmark**
**DK-2800 Lyngby, Denmark**
owi@imm.dtu.dk

**Bioinformatics Centre, University of Copenhagen**

In collaboration with
**Ulrich Paquet** (Cambridge and Stellenbosch, S. Africa)
**Blaise Thomson** (Cambridge)


UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot·your knowledge partner

UNIVERSITY OF CAMBRIDGE

DTU

# Large scale approximative inference

Netflix prize

Solutions - some trends

Ordinal regression

Variational Bayes (VB)

VB predictive distribution

Expectation propagation

Our performance – work in progress

# Netflix prize

- 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%.

# Solution trends

- **Nearest neighborhood based**: for example KNN (Bell and Koren)

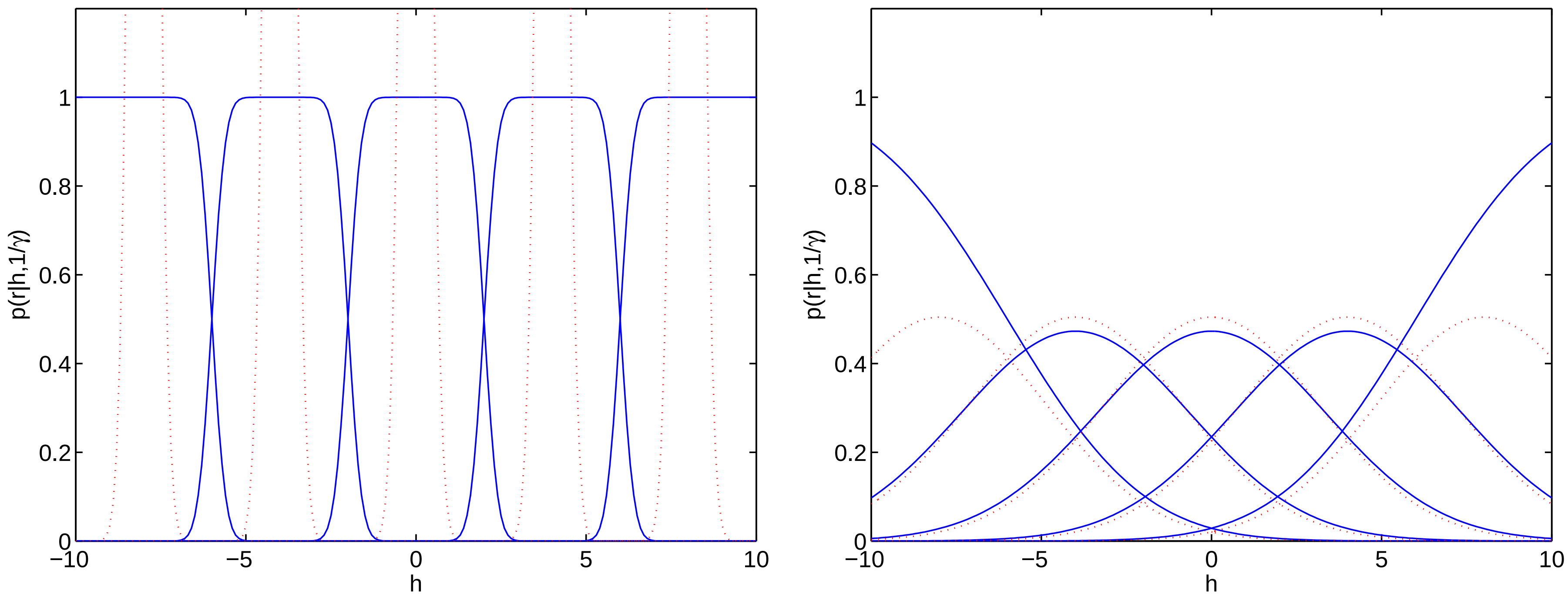
- **Low rank factorization:** regularized SVD (Funk, Patarak, Lim and Teh, Raiko, Ilin and Karhunen), low rank + ordinal regression

- **Linear combinations of predictors.**

# Bayesian Ordinal regression

Use “correct” likelihood model (Chu and Ghahramani)

$$p(r|h,\sigma^2) = \Phi\left(\frac{h-b_r}{\sigma}\right) - \Phi\left(\frac{h-b_{r+1}}{\sigma}\right)$$




Model $h_{mn}$ as factor model, GP, etc.

# Variational Bayes (VB)

Low rank decomposition for $h_{mn}$

$$h_{mn} = \mathbf{u}_m \cdot \mathbf{v}_n = \sum_{k=1}^{K} u_{mk} v_{nk}$$

Treat $h$ as latent variable $\sigma^2 = \sigma_0^2 + \sigma_1^2$

$$p(r_{mn}|\mathbf{u}_m, \mathbf{v}_n, \sigma^2) = \int p(r_{mn}|h_{mn}, \sigma_0^2)\, \mathcal{N}(h_{mn}|\mathbf{u}_m \cdot \mathbf{v}_n, \sigma_1^2)\, dh_{mn}$$

Variational distribution

$$q(\mathbf{H}, \mathbf{U}, \mathbf{V}) = \prod_{(m,n)} q(h_{mn}) q(\mathbf{U}) q(\mathbf{V})$$

Priors $p(\mathbf{U})$ and $p(\mathbf{V})$ can be Gaussian, Laplace, etc. and hierarchical.

# VB solution

We choose $p(u_{mk}) = \mathcal{N}(u_{mk}|0, 1/\alpha)$ and $p(v_{nk}) = \mathcal{N}(v_{nk}|0, 1/\beta)$ and fully factorized $q(\mathbf{U})$ and $q(\mathbf{V})$ free form optimization.

Run over all movies $m = 1, \ldots, M$:

- Run over all users having watched $m$, $n \in \Omega(m)$

$$q(h_{mn}) \propto p(r_{mn}|h_{mn}, \sigma_0^2)\mathcal{N}(h_{mn}| \langle \mathbf{u}_m \rangle \cdot \langle \mathbf{v}_n \rangle, \sigma_1^2)$$

- Run over components $k = 1, \ldots, K$:

$$q(u_{mk}) = \mathcal{N}\left(u_{mk}; \frac{\Sigma_{mk}}{\sigma_1^2} \sum_{n \in \Omega(m)} \langle v_{nk} \rangle \left(\langle h_{mn} \rangle - [\langle \mathbf{u}_m \rangle \cdot \langle \mathbf{v}_n \rangle]_{\backslash k}\right), \Sigma_{mk}\right)$$

$$\Sigma_{mk} = \left(\alpha + \sum_{n \in \Omega(m)} \frac{\left\langle u_{nk}^2 \right\rangle}{\sigma^2}\right)^{-1}$$

# VB solution cont.

- Run over all users $n = 1, \ldots, N$ in same fashion.
- h-updates "local" - no need to store them.
- Symmetry between $\sigma_0^2$ and $\sigma_1^2$ broken.
- Multivariate $q(\mathbf{U})$ and $q(\mathbf{V})$ – complexity increase $\mathcal{O}(K^2)$.

# Predictive distribution

The objective is to minimize RMSE so use predictive mean

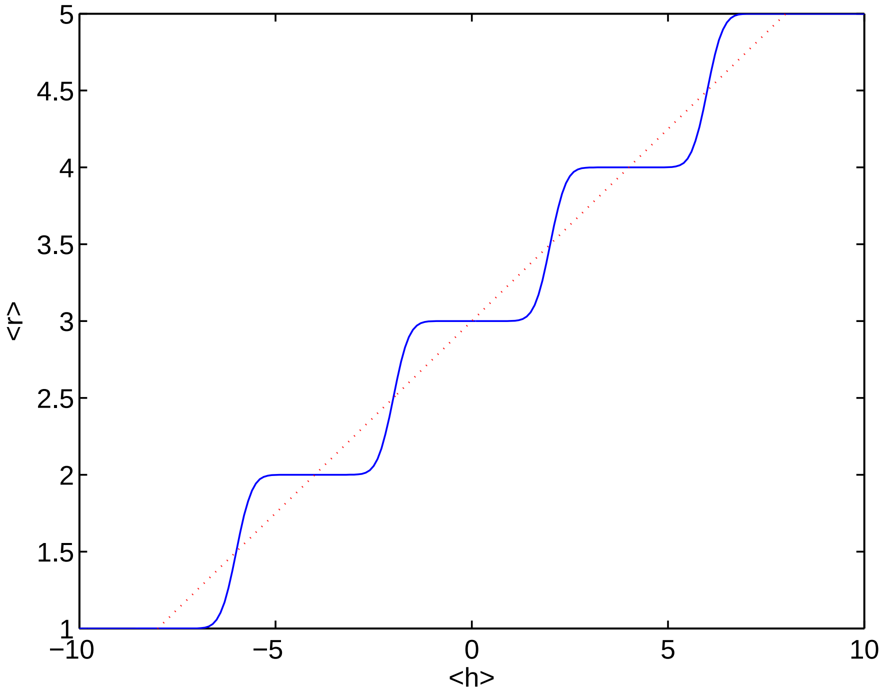
$$\langle r_{mn} \rangle \approx \sum_{r=1}^{5} \int r \, p(r|h, \sigma_0^2) \, \mathcal{N}(h|\mathbf{u}_m \cdot \mathbf{v}_n, \sigma_1^2) \, q(\mathbf{u}_m) q(\mathbf{v}_n) \, dh d\mathbf{u}_m d\mathbf{v}_n$$

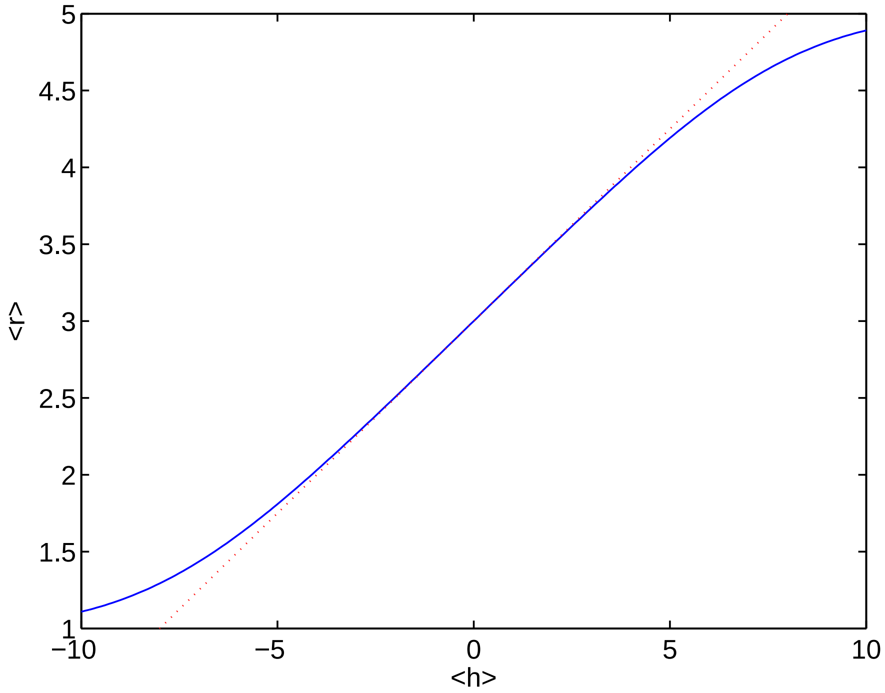
Not analytically tractable so replace by mean-field

$$\begin{aligned} \langle r_{mn} \rangle &\approx \sum_{r=1}^{5} \int r \, p(r|h, \sigma_0^2) \, q(h_{mn}) \, dh_{mn} \\ &= \sum_{r=1}^{5} \int r \, p(r|h, \sigma_0^2) \mathcal{N}(h_{mn}| \langle \mathbf{u}_m \rangle \cdot \langle \mathbf{v}_n \rangle , \sigma_1^2) \, dh_{mn} \end{aligned}$$

# Ordinal regression - soft clipping

Small and large $\sigma_1^2$ with $\sigma_0^2 = 0$



# Predictive distribution - better approximation

We can apply central limit theorem (CLT) to go beyond simple mean field:

$$
\begin{aligned}
\mathbf{u}_m \cdot \mathbf{v}_n &\sim \mathcal{N}(\langle \mathbf{u}_m \rangle \cdot \langle \mathbf{v}_n \rangle \,, \sigma_{uv}^2) \\
\sigma_{uv}^2 &= \left\langle (\mathbf{u}_m \cdot \mathbf{v}_n)^2 \right\rangle - (\langle \mathbf{u}_m \rangle \cdot \langle \mathbf{v}_n \rangle)^2
\end{aligned}
$$

Effective variance

$$
\sigma^2 + \sigma_{uv}^2
$$

Variance term for fully factorized

$$
\begin{aligned}
\sigma_{uv}^2 = \sum_{k=1}^{K} \Big[ & (\left\langle u_{mk}^2 \right\rangle - \langle u_{mk} \rangle^2)(\left\langle v_{nk}^2 \right\rangle - \langle v_{nk} \rangle^2) \\
& + \langle u_{nk} \rangle^2 (\left\langle v_{nk}^2 \right\rangle - \langle v_{nk} \rangle^2) + \langle v_{nk} \rangle^2 (\left\langle u_{mk}^2 \right\rangle - \langle u_{mk} \rangle^2) \Big]
\end{aligned}
$$

# Expectation propagation

Exponential family (Gaussian)

$$
\begin{aligned}
q(\mathbf{u}_m) &\propto \exp\left(\sum_{n\in\Omega(m)} \mathbf{a}_{mn}\cdot\boldsymbol{\phi}(\mathbf{u}_m)\right)\\
q(\mathbf{v}_n) &\propto \exp\left(\sum_{m\in\Pi(n)} \mathbf{b}_{mn}\cdot\boldsymbol{\phi}(\mathbf{v}_n)\right)\\
q_{mn}(\mathbf{u}_m,\mathbf{v}_m) &\propto p(r_{mn}|\mathbf{u}_m,\mathbf{v}_n,\sigma^2)\\
&\quad \exp\left(\sum_{n'\in\Omega(m)\backslash n} \mathbf{a}_{mn'}\phi(\mathbf{u}_m) + \sum_{m'\in\Pi(n)\backslash m} \mathbf{b}_{m'n}\phi(\mathbf{v}_n)\right)
\end{aligned}
$$

Expectation consistency between

$$
\langle\boldsymbol{\phi}(\mathbf{u}_m)\rangle_{q(\mathbf{u}_m)} = \langle\boldsymbol{\phi}(\mathbf{u}_m)\rangle_{q_{mn}(\mathbf{u}_m,\mathbf{v}_n)}
$$

and likewise for $\mathbf{u}_m$.

# Expectation propagation cont.

- $q_{mn}(\mathbf{u}_m, \mathbf{v}_m)$ not tractable – use CLT approximation

$$q_{mn}(\mathbf{u}_m, \mathbf{v}_m) \propto p(r_{mn}|\mathbf{u}_m, \mathbf{v}_n, \sigma^2) \exp\left( \sum_{n' \in \Omega(m) \backslash n} a_{mn'} \phi(\mathbf{u}_m) + \sum_{m' \in \Pi(n) \backslash m} b_{m'n} \phi(\mathbf{v}_n) \right)$$

- What is perhaps worse:

  We have to determine and store $\mathcal{O}(R*K)$ parameters $\{\mathbf{a}_{mn}, \mathbf{b}_{mn}\}$

- VB we have $K(M+N)$

# Simplifying EP

- First round of EP is ADF (Bayes online): find moments of

$$q_{mn}(\mathbf{u}_m, \mathbf{v}_m) = p(r_{mn}|\mathbf{u}_m, \mathbf{v}_n, \sigma^2)q(\mathbf{u}_m)q(\mathbf{v}_m)$$

  to update

$$q(\mathbf{u}_m)q(\mathbf{v}_m)\ .$$

- In subsequent sweeps, the contribution of observation $r_{mn}$ can be removed approximately (to linear order) before updating.

# Performance – work in progress

- $K = 20$, $\alpha = \beta \approx \sigma_1^2 \approx 10$

  0.9143

- VB linear low rank slightly worse (Lim and Teh, Raiko, Ilin and Karhunen)

- Best linear low rank special regularization $K = 96$ (Funk)

  0.8914

- Current leaders Bell and Koren neighbor + low rank

  0.8705

# Next steps

- **Low rank decompositions:** hierarchical and in general better priors.

- **Nearest neighbor:** GP ordinal regression with specially designed kernel functions on smaller sets relevant for prediction.

- **Model averaging.**

- and maybe **more accurate approximate inference**...