Variational Methods for Discrete-Data Latent Gaussian Models

Mohammad Emtiyaz Khan

University of British Columbia
Vancouver, Canada

March 6, 2012

The Big Picture

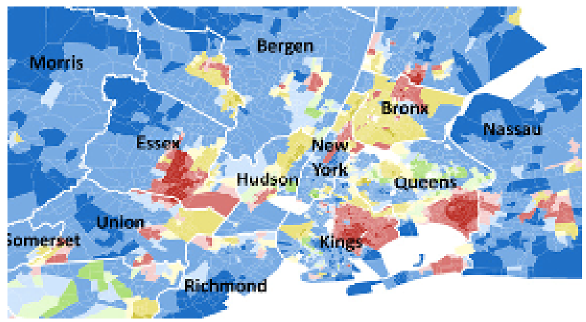
- Joint density models for data with mixed data types
- Bayesian models – principled and robust approach
- Algorithms that are not only accurate and fast, but are also easy to tune, implement, and intuitive (speed-accuracy tradeoffs)

Variational Methods for Discrete-Data Latent Gaussian Models

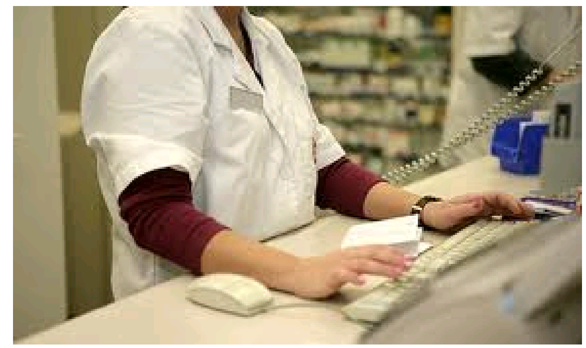
Sources of Discrete Data

User rating data

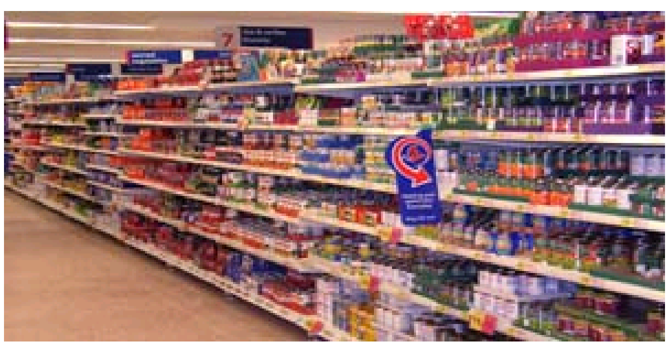
Survey/voting data and blogs for sentiment analysis



Health data

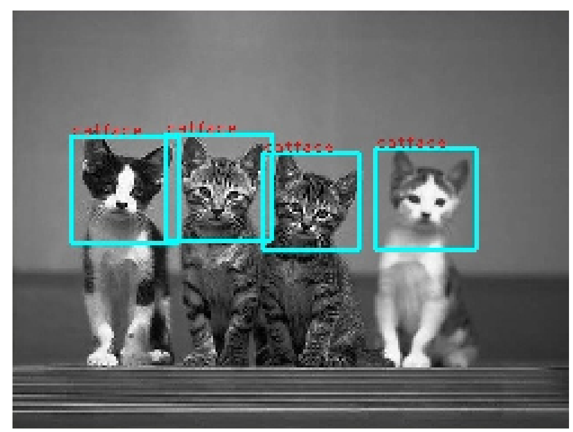


Consumer choice data



Sports/game data

tag correlation.



Motivation: Recommendation system

Movie rating dataset – Missing values – Different types of data

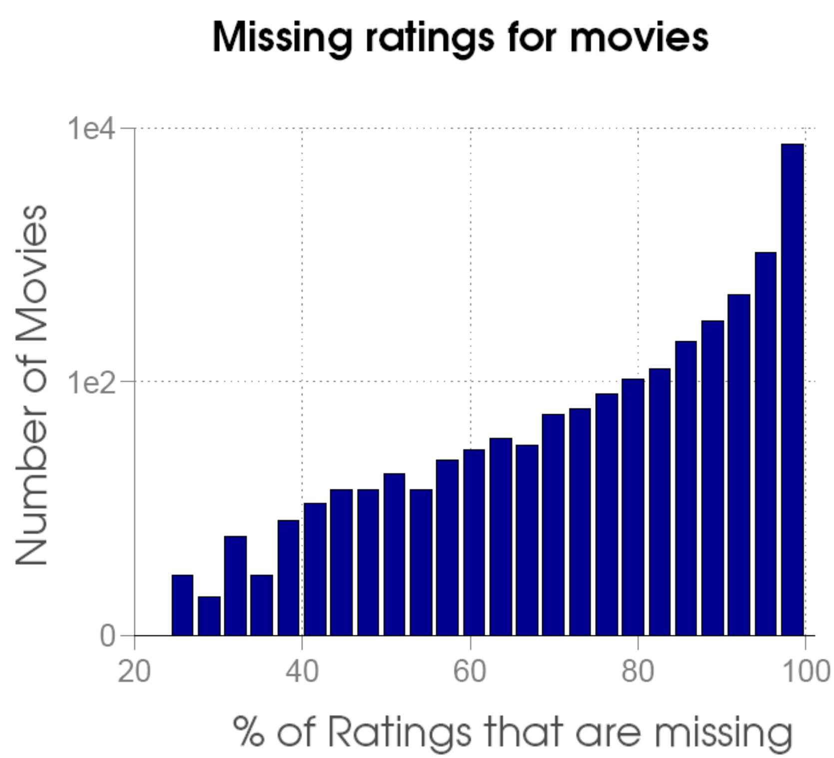
	User1	User2	User3	User4	User5	User6	....
Movie1	9	2	3		9		....
Movie2	8			8		2	....
Movie3		2		8			....
Movie4	3	8	8			1	....
Movie5	2		7		1		....
Movie6		7		2		1	....

Missing Ratings

From Wikipedia on Netflix-prize dataset

"The training set is such that the average user rated over 200 movies, and the average movie was rated by over 5000 users. But there is wide variance in the data—some movies in the training set have as few as 3 ratings, while one user rated over 17,000 movies."

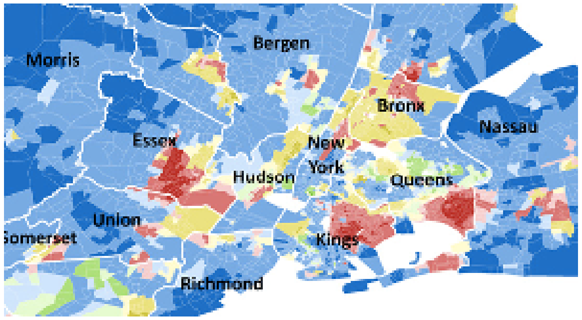
Movielens Dataset



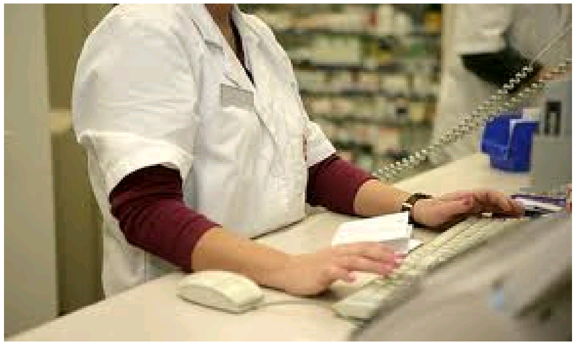
Sources of Discrete Data

User rating data

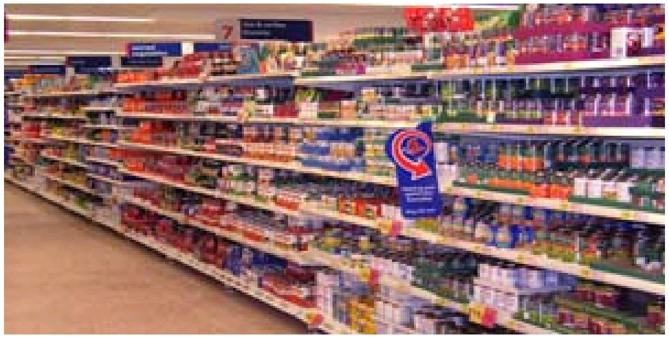
Survey/voting data and blogs for sentiment analysis



Health data

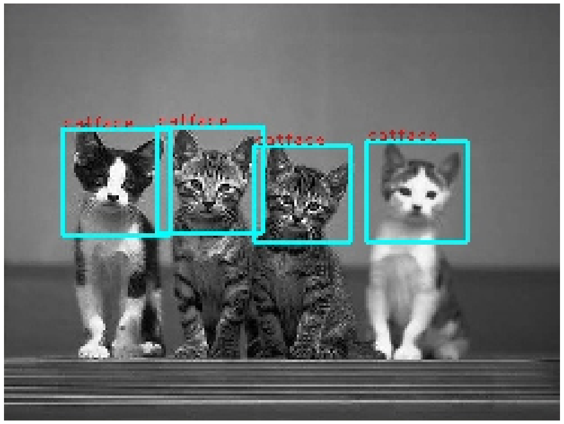


Consumer choice data



Sports/game data

tag correlation.

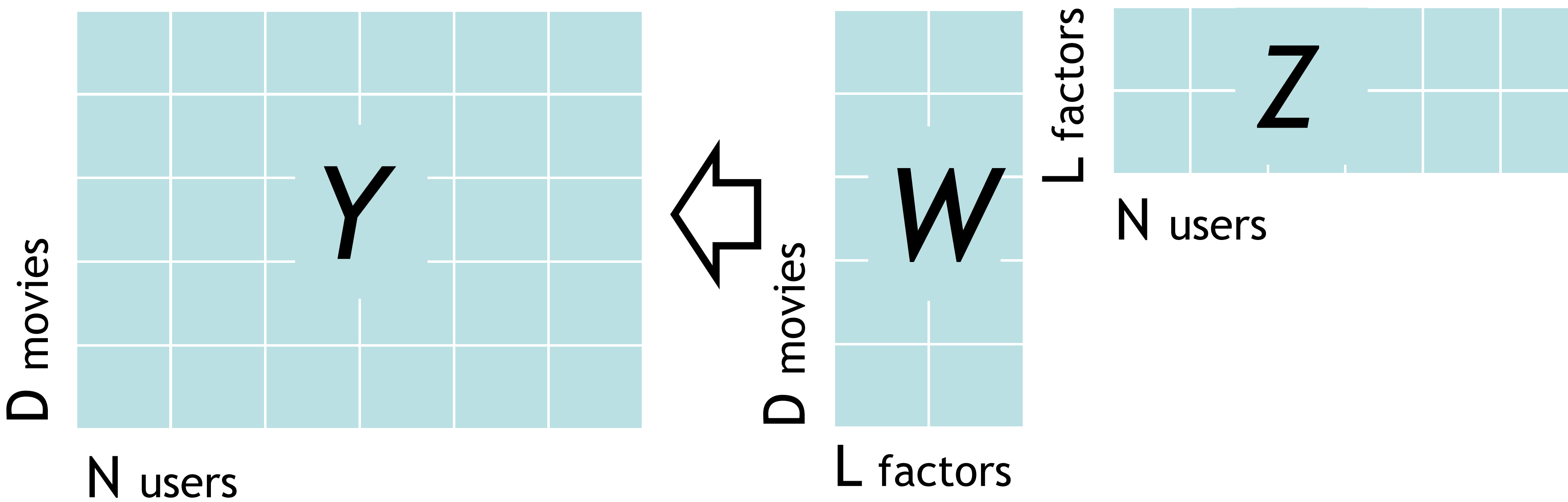


What we need!

For these datasets, we need a method of analysis which

- Handles missing values efficiently
- Makes efficient use of the data by weighting "reliable" data vectors more than the "unreliable" ones
- Makes efficient use of the data by "fusing" different types of data efficiently (binary, ordinal, categorical, count, text)

Factor Model



ExpFamily: $$\min_{\mathbf{W}} \sum_{n=1}^{N} \min_{\mathbf{z}_n} -\log p(\mathbf{y}_n|\mathbf{z}_n, \mathbf{W})$$

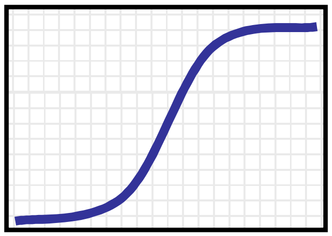
Gaussian: $$\min_{\mathbf{W}} \sum_{n=1}^{N} \min_{\mathbf{z}_n} ||\mathbf{y}_n - \mathbf{W}\mathbf{z}_n||_2^2$$

Collins et. al. 2002, Khan et. al.2010, Yu et.al. 2009

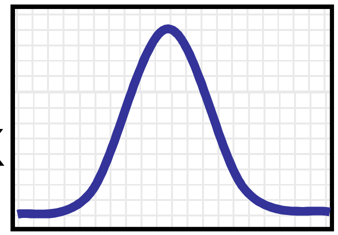
Bayesian Learning

$$\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \log \int p(\mathbf{y}_n|\mathbf{z}_n, \mathbf{W}) \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{z}_n$$

$$\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \log \int$$



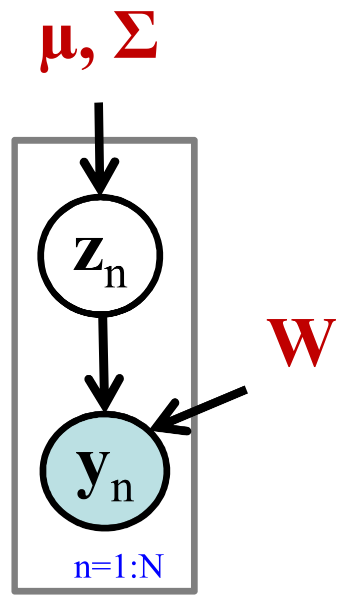
x



$$d\mathbf{z}$$

$$\geq \max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \max_{\boldsymbol{\psi}_n} \underline{\mathcal{L}}_n(\boldsymbol{\theta}, \boldsymbol{\psi}_n)$$

This talk: Lower bound maximization



Variational Methods

$$\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \log \int p(\mathbf{y}_n|\mathbf{z}_n, \mathbf{W}) \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{z}_n \quad \geq \max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \max_{\boldsymbol{\psi}_n} \underline{\mathcal{L}}_n(\boldsymbol{\theta}, \boldsymbol{\psi}_n)$$

- Design tractable bounds to reduce approximation error
- Efficient optimization since lower bounds are concave : good convergence rates and easy convergence diagnostics
- Efficient expectation-maximization (EM) algorithms for parameter leaning
- Comparable performance to MCMC, but much faster
- Algorithms with a wide range of speed-accuracy trade-offs

Outline

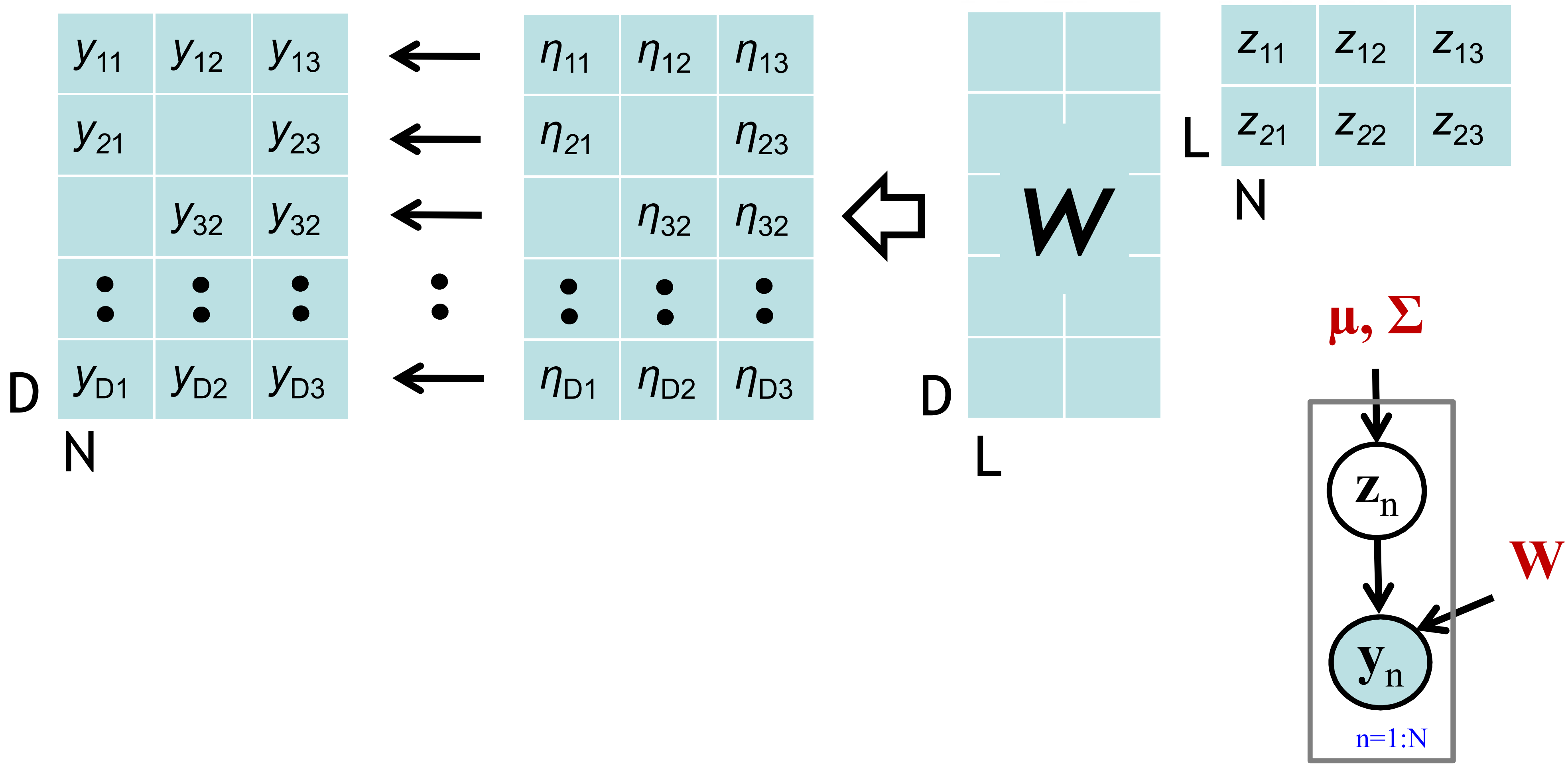
- Latent Gaussian models
- Bounds for binary data
- Bounds for categorical data
- Results
- Future work and conclusions

Outline

- Latent Gaussian models
  - Definition and examples
  - Problem with parameter learning
- Bounds for binary data
- Bounds for categorical data
- Results
- Future work and conclusions

Latent Gaussian Model (LGM)

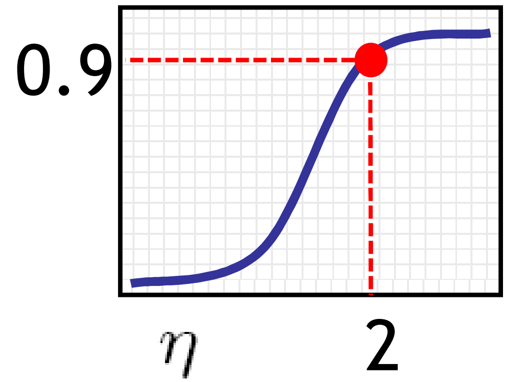
$$p(y_{dn}|\boldsymbol{\eta}_{dn}) \qquad \boldsymbol{\eta}_{dn} = \mathbf{W}_d\mathbf{z}_n \qquad p(\mathbf{z}_n|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$



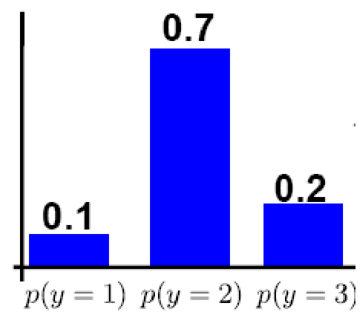
Likelihood Examples

Data type	Distribution
Real	Gaussian
Count	Poisson
Binary	Bernoulli-Logit
Categorical	Multinomial-Logit
Ordinal	Proportional-odds

$$p(y=1|\eta) = \frac{e^{\eta}}{1+e^{\eta}}$$

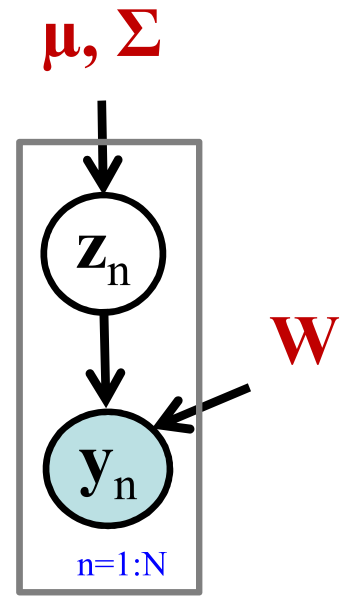


$$p(y=k|\boldsymbol{\eta}) = \frac{e^{\eta_k}}{\sum_{j=1}^{K} e^{\eta_j}}$$



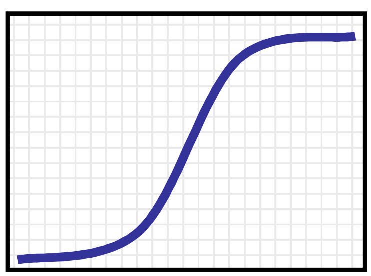
Parameter Estimation

$$\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \log \int \prod_{d=1}^{D} p(y_{dn}|\mathbf{z},\boldsymbol{\theta})\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})\,d\mathbf{z}$$

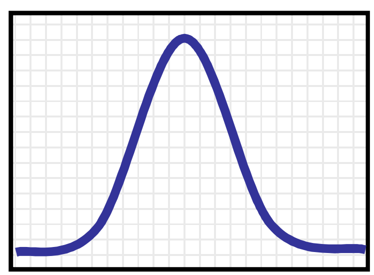


$$\log \int \prod_{d=1}^{D}$$

Bernoulli-logit



x



$d\mathbf{z}$

Jensen's Lower Bound

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) = \log \int \prod_{d=1}^{D} p(y_d|\mathbf{z},\boldsymbol{\theta})\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})dz$$

$$= \log \int \frac{\prod_{d=1}^{D} p(y_d|\mathbf{z},\boldsymbol{\theta})\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})}\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})dz$$

$$\boxed{\begin{aligned}\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) \geq \max_{\mathbf{m},\mathbf{V}} \sum_{d=1}^{D} \int [\log p(y_d|\mathbf{z},\boldsymbol{\theta})]\mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V})dz \\ - KL\left[\mathcal{N}(\mathbf{m},\mathbf{V})||\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})\right]\end{aligned}}$$

Variational Lower Bound

$$\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \max_{\mathbf{m}_n, \mathbf{V}_n} \sum_{d=1}^{D} \int \log p(y_{dn}|\boldsymbol{\eta}_{dn}) \mathcal{N}(\mathbf{z}|\mathbf{m}_n, \mathbf{V}_n) d\mathbf{z} - KL[\mathcal{N}(\mathbf{m}_n, \mathbf{V}_n)||\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})]$$

- Generalized EM algorithm
- E-step involves minimizing convex function
- Early stopping in E-step
- (Almost) no tuning parameters
- Easy convergence diagnostics

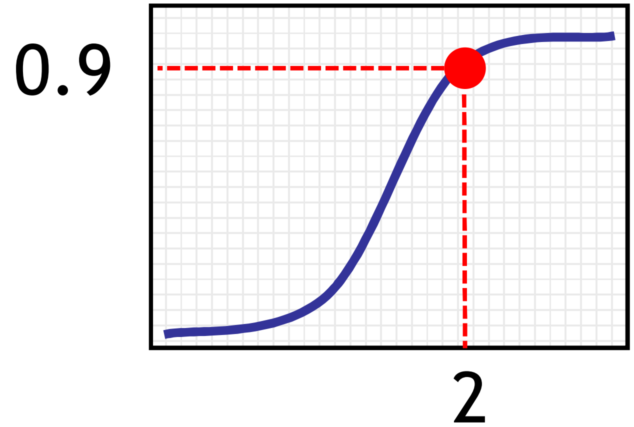
Outline

- Latent Gaussian models
- Bounds for binary data
  - Bernoulli-logistic likelihood
  - The Bohning bound (Khan, Marlin, Bouchard, Murphy, NIPS 2010)
  - Piecewise bounds (Marlin, Khan, Murphy, ICML 2011)
- Bounds for categorical data
- Results
- Future work and conclusions

Bernoulli-Logit Likelihood

$$p(y=1|\eta) = \frac{e^\eta}{1+e^\eta}$$

$$\log p(y=1|\eta) = \eta - \log(1+e^\eta)$$



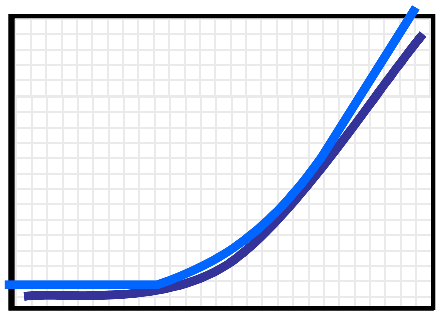
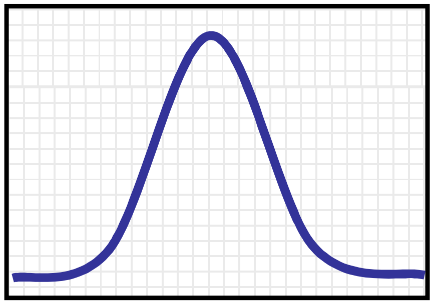
$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) \geq \max_{\mathbf{m},\mathbf{V}} \sum_{d=1}^{D} \int [\log p(y_d|\mathbf{z},\boldsymbol{\theta})] \mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V}) d\mathbf{z}$$

$$- KL\left[\mathcal{N}(\mathbf{m},\mathbf{V}) || \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})\right]$$

$$= \max_{\mathbf{m},\mathbf{V}} \sum_{d=1}^{D} \int [-\log(1+e^\eta)] \mathcal{N}(\tilde{m}_d, \tilde{v}_d) d\eta \quad + \quad \text{some other tractable terms in } \mathbf{m} \text{ and } \mathbf{V}$$

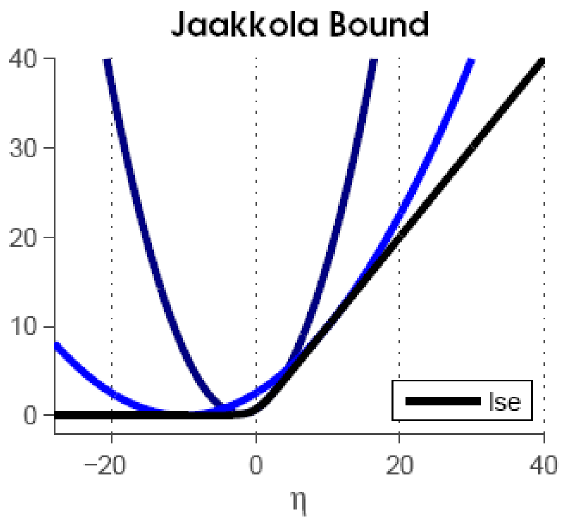
Local Variational Bounds

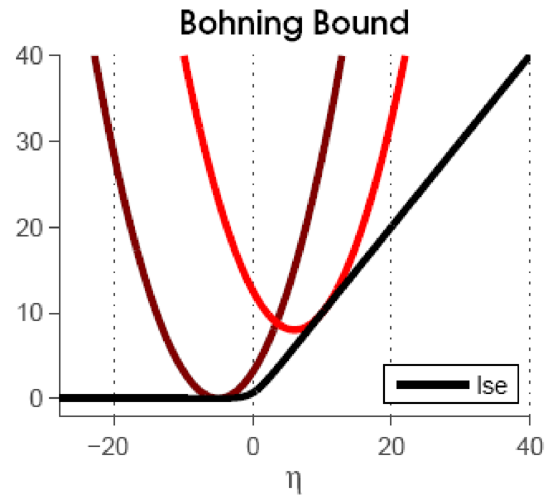
$$\max_{\mathbf{m},\mathbf{V}} \sum_{d=1}^{D} \int [-\log(1+e^{\eta})] \mathcal{N}(\tilde{m}_d, \tilde{v}_d) d\eta \quad + \quad \text{some other tractable terms in } \mathbf{m} \text{ and } \mathbf{V}$$

$-\int$  x  $d\eta$

- Bohning's bound (Khan, Marlin, Bouchard, Murphy 2010)
- Jaakola's bound (Jaakkola and Jordan1996)
- Piecewise quadratic bounds (Marlin, Khan, Murphy 2011)

Bohning Bound is Faster





$O(L^3ND)$

For n = 1:N

$V_n = (W^T A_n W + I)^{-1}$

$m_n = \ldots..$

end

$O(L^2ND)$

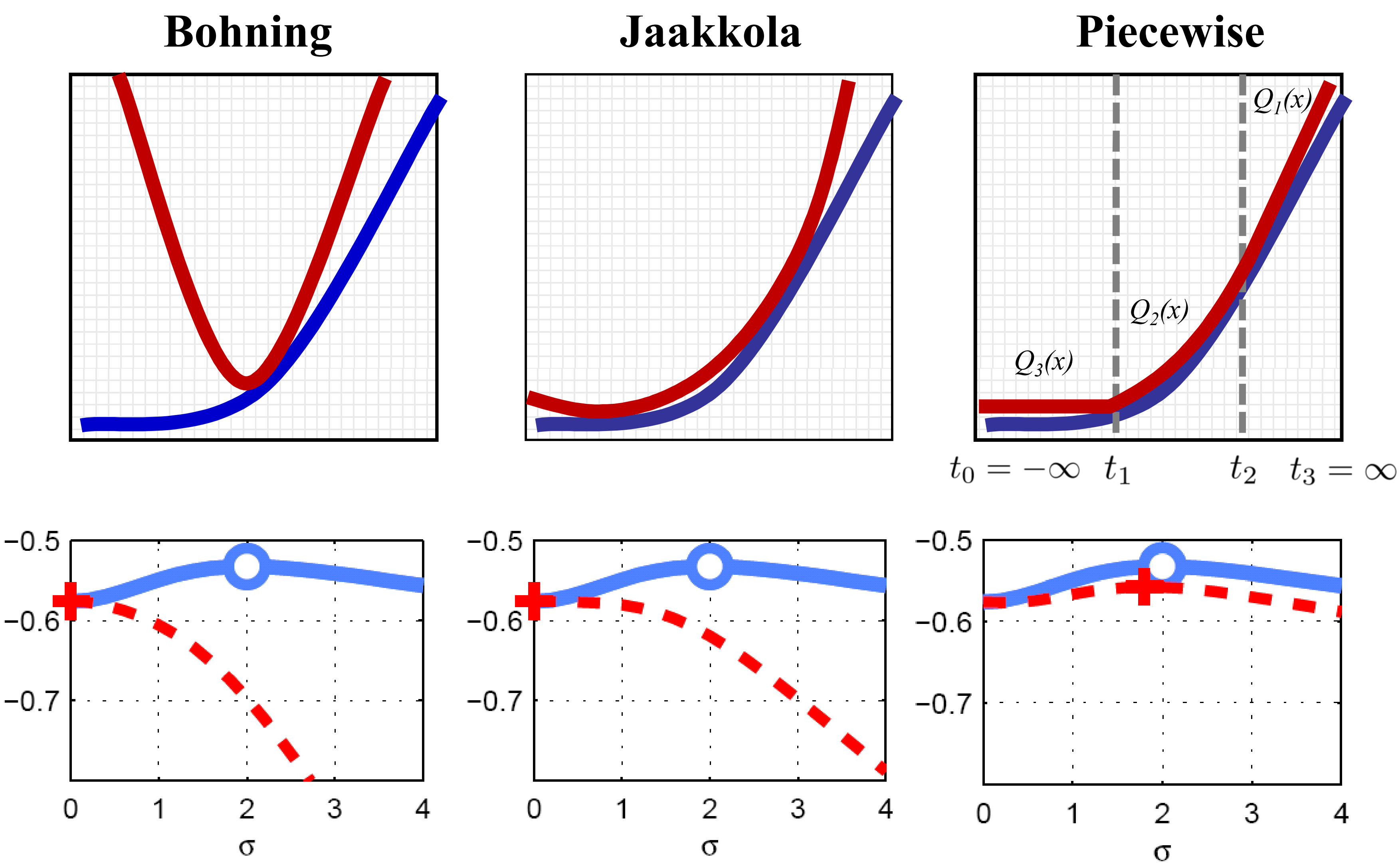
$V = (W^T A W + I)^{-1}$

For n = 1:N

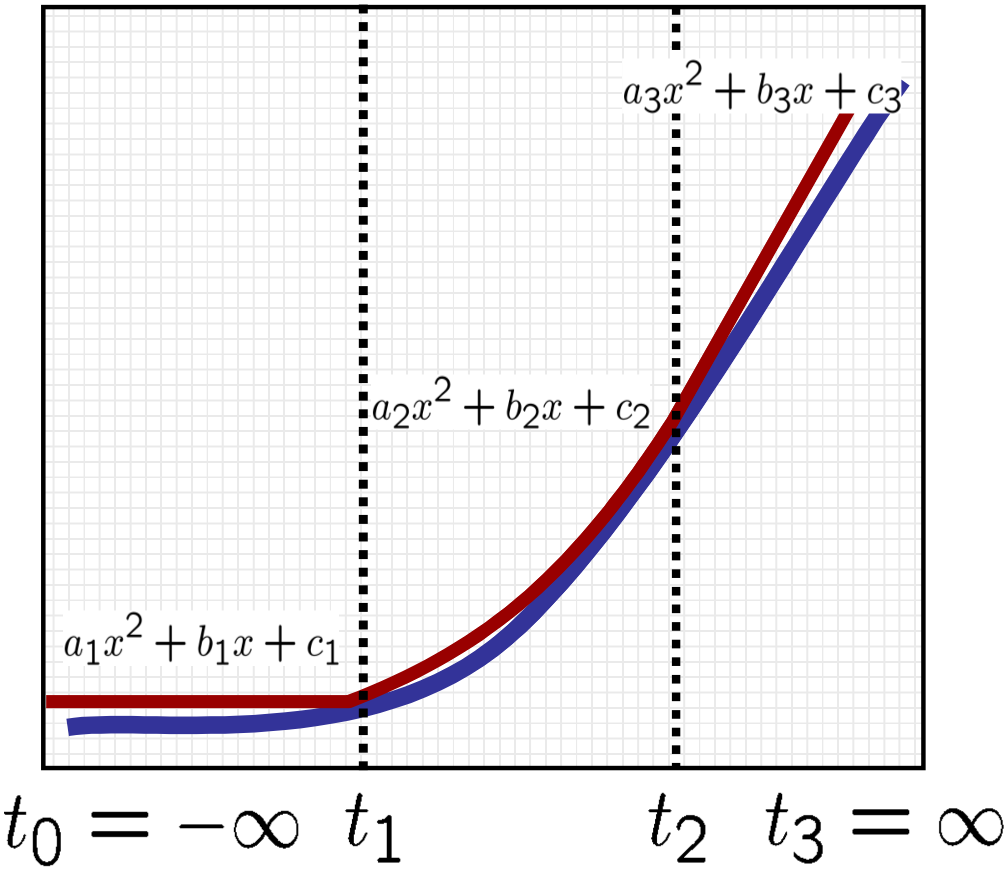
$m_n = \ldots..$

end

Piecewise bounds are more accurate



Details of Piecewise bounds



- Find cut points and parameters of each piece by minimizing maximum error
- Linear pieces (Hsiung, Kim and Boyd, 2008)
- Quadratic Pieces (Nelder-Mead method)
- Fixed Piecewise Bounds!
- Increase accuracy by increasing the number of pieces

Mohammad Emtiyaz Khan

Outline

- Latent Gaussian models
- Bounds for binary data
- Bounds for categorical data
  - Multinomial-logistic likelihood and local variational bounds
  - Stick-breaking likelihood (Khan, Mohamed, Marlin, Murphy, AI-Stats 2012)
- Results
- Future work and conclusions

Multinomial-Logit Likelihood

$$p(y = k|\boldsymbol{\eta}) = \frac{e^{\eta_k}}{\sum_{j=1}^{K} e^{\eta_j}}$$

$$\log p(y = k|\boldsymbol{\eta}) = \eta_k - \log \sum_{j=1}^{K} e^{\eta_j}$$

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) \;\; \geq \;\; & \max_{\mathbf{m},\mathbf{V}} \sum_{d=1}^{D} \int [\log p(y_d|\mathbf{z}, \boldsymbol{\theta})] \mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V}) d\mathbf{z} \\ & - KL\left[\mathcal{N}(\mathbf{m}, \mathbf{V})||\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\right]\end{aligned}$$

Local Variational bounds

- The Bohning bound
  - Fast and closed form updates
- The log bound (Blei and Lafferty 2006)
  - More accurate than the Bohning bound, but slower
- The product of sigmoid bound (Bouchard 2007)

Stick-Breaking Likelihood

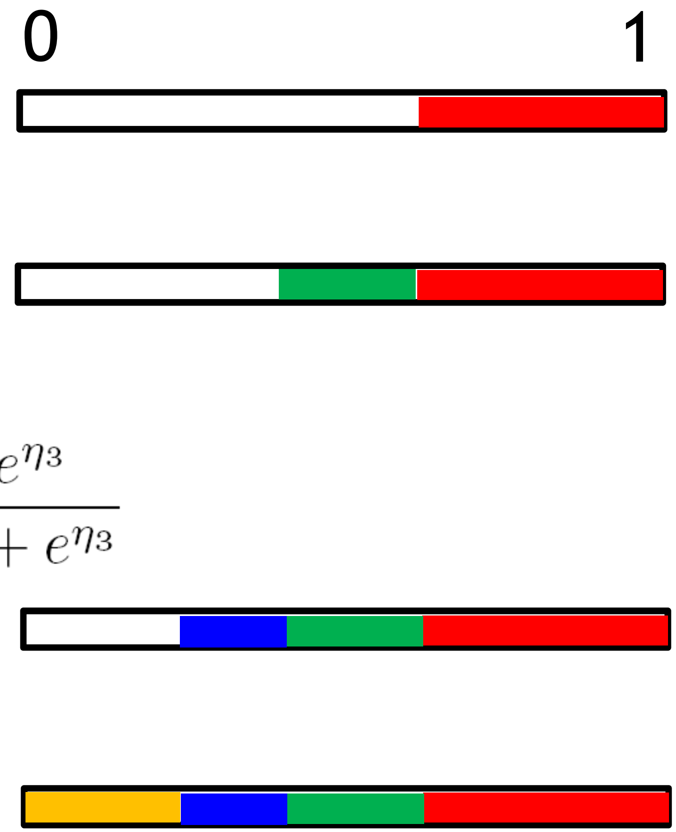
$$p(y=1|\boldsymbol{\eta}) = \frac{e^{\eta_1}}{1+e^{\eta_1}}$$

$$p(y=2|\boldsymbol{\eta}) = \left(1 - \frac{e^{\eta_1}}{1+e^{\eta_1}}\right)\frac{e^{\eta_2}}{1+e^{\eta_2}}$$

$$p(y=3|\boldsymbol{\eta}) = \left(1 - \frac{e^{\eta_1}}{1+e^{\eta_1}}\right)\left(1 - \frac{e^{\eta_2}}{1+e^{\eta_2}}\right)\frac{e^{\eta_3}}{1+e^{\eta_3}}$$

$$\vdots$$

$$p(y=K|\boldsymbol{\eta}) = \prod_{j=1}^{K-1}\left(1 - \frac{e^{\eta_j}}{1+e^{\eta_j}}\right)$$



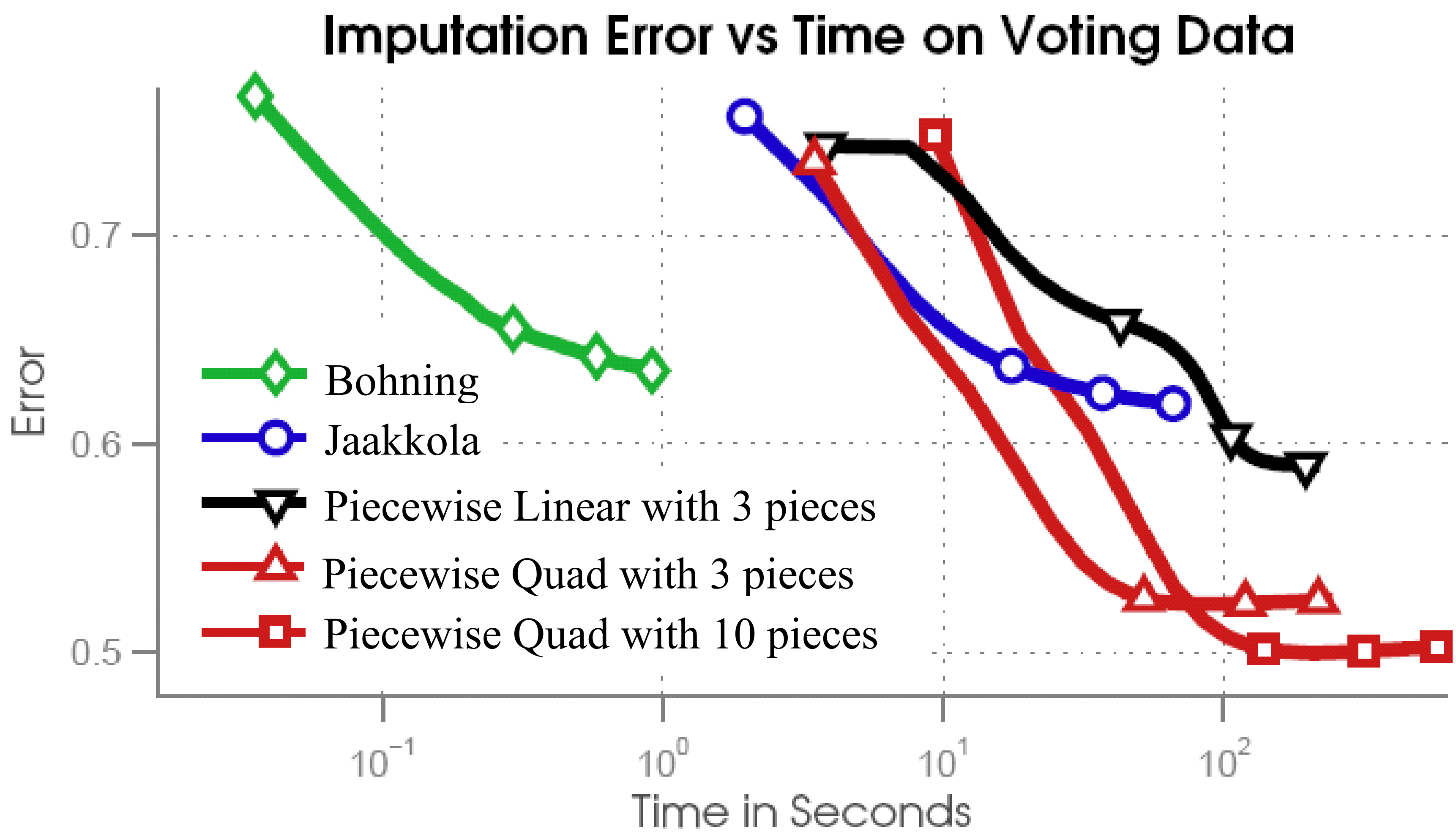
$$\boxed{\log p(y=k|\boldsymbol{\eta}) = \eta_k - \sum_{j=1}^{K-1}\mathbf{I}(j \le k)\,\boxed{\log(1+e^{\eta_j})}}$$

Outline

- Latent Gaussian models
- Bounds for binary data
- Bounds for categorical data
- Results
- Future work and conclusions

Speed Accuracy Trade-offs

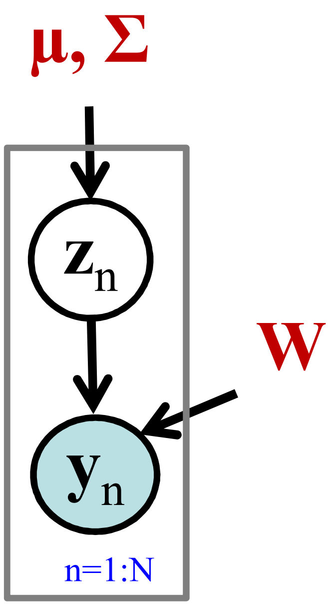
Binary FA : UCI voting dataset (D=15, N=435)



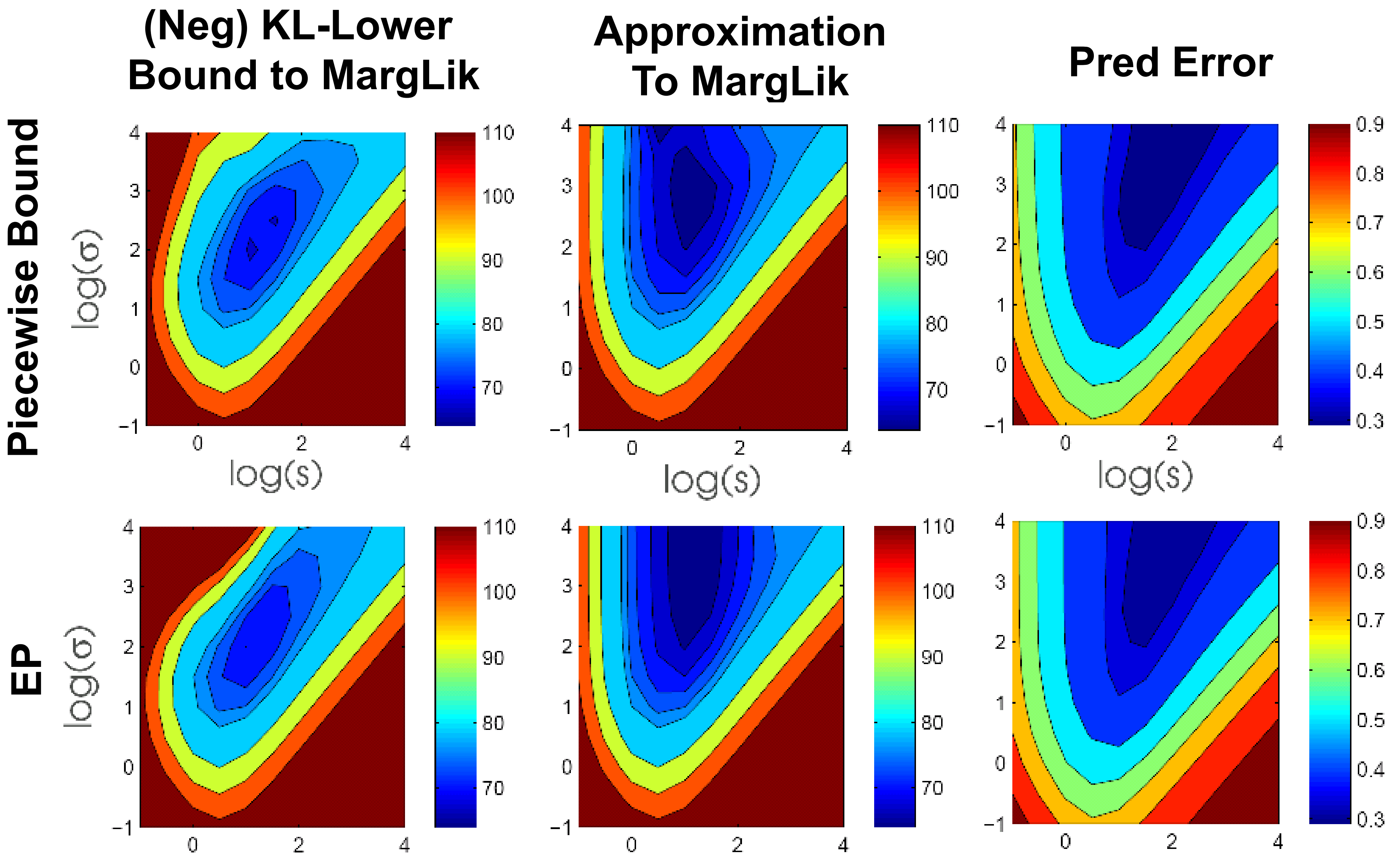
Comparison with EP

Binary Gaussian Process : Ionosphere dataset (D=200)

$\Sigma_{ij} = \sigma \exp[-||x_i - x_j||^2/s]$



EP vs PW : Posterior Distribution

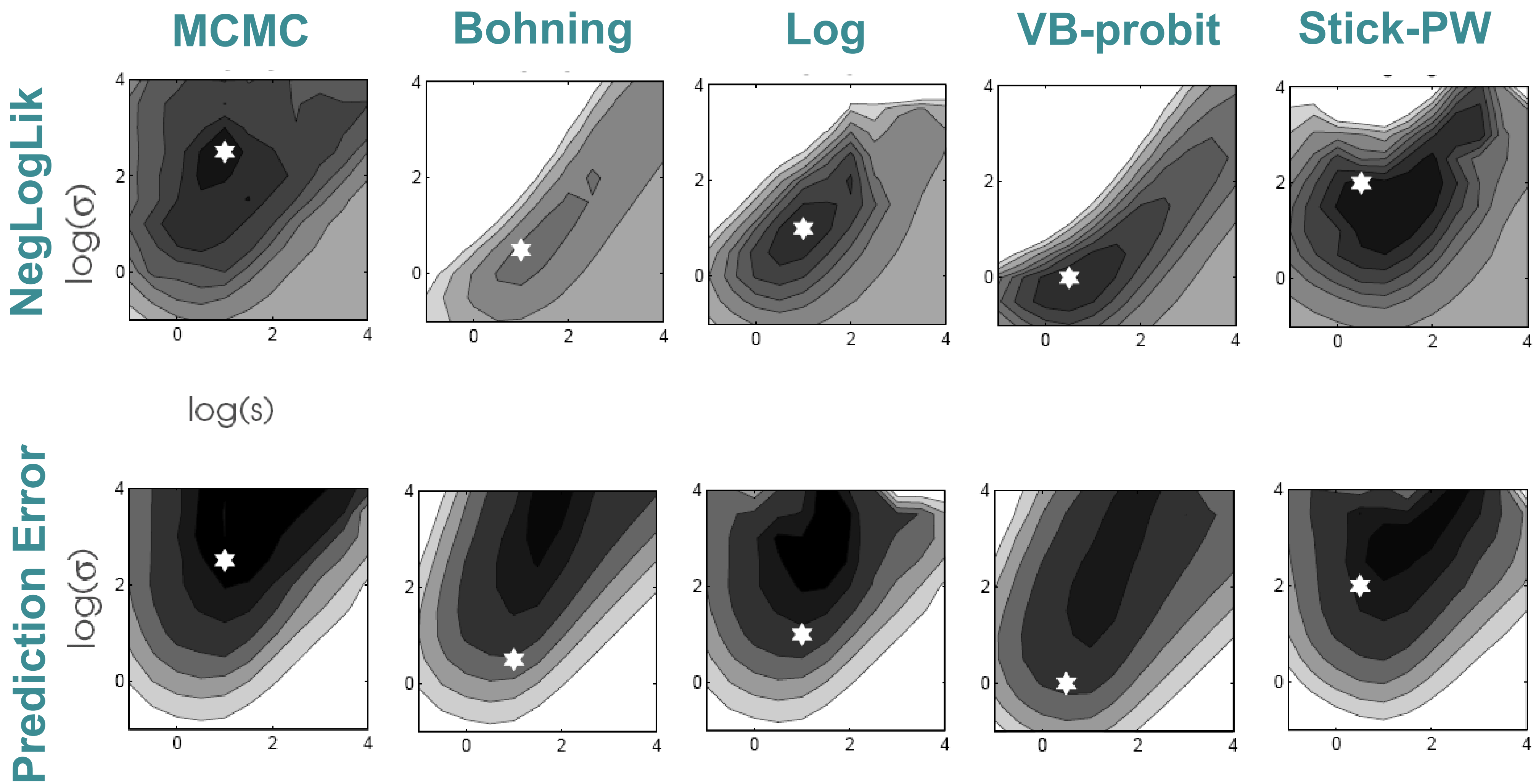


Comparison with EP

- Both methods give very similar results for GPs
- Our approach can be easily extended to factor models
- Variational EM objective function is well-defined and can be obtained by solving minimization of convex functions
- Numerically stable

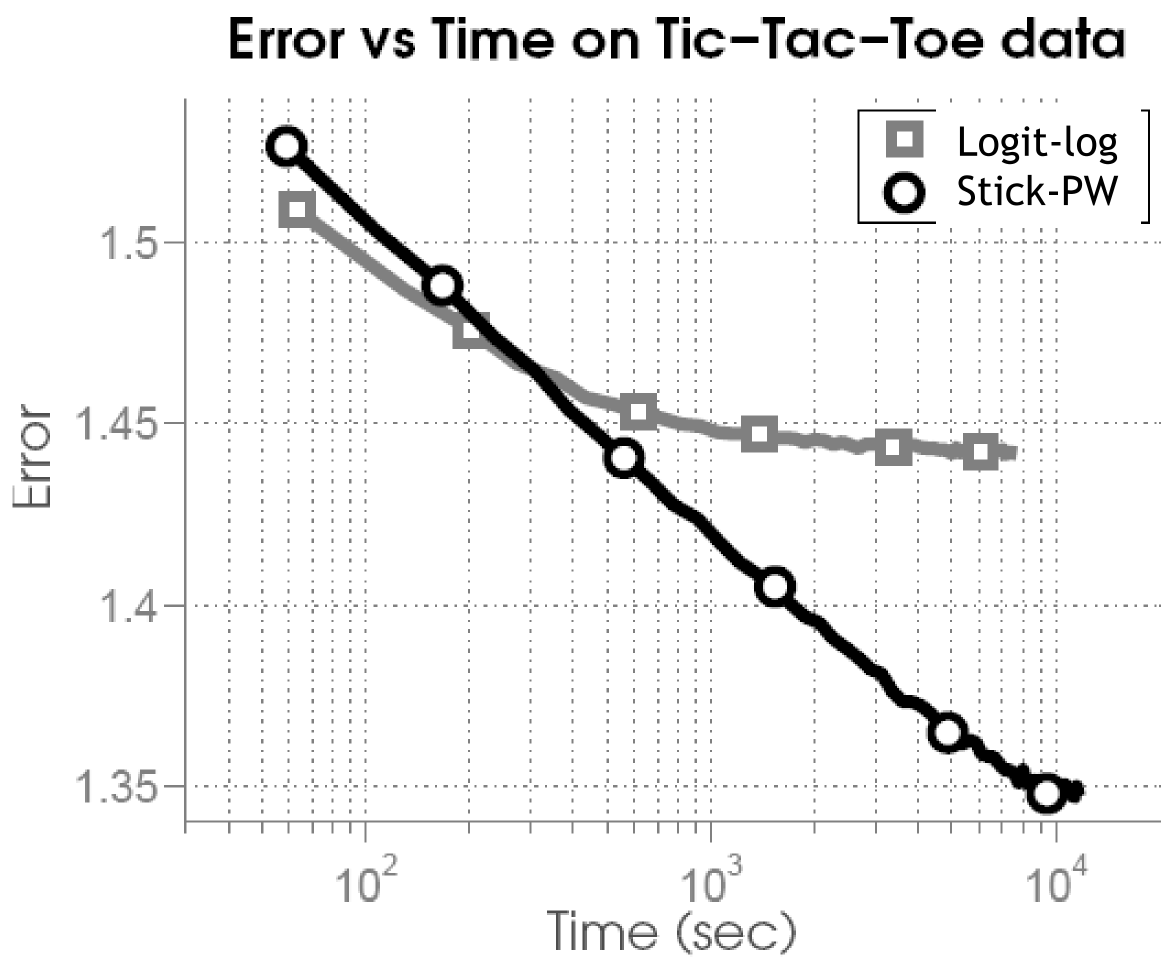
MultiClass Gaussian Process

Glass dataset (D=143, K=6)



Categorical Factor Analysis

Glass dataset (D = 10, N = 958, sum of K = 29)



Mohammad Emtiyaz Khan

Outline

- Latent Gaussian models
- Bounds for binary data
- Bounds for categorical data
- Results
- Future work and conclusions

Future Work

- Large-scale collaborative filtering
- Use convexity to design approximate gradient methods
- Sparse Gaussian Posterior Distribution
- Tuning HMC using Bayesian optimization methods
- Latent Sparse-factor model
- Conditional models (e.g. to model for tag-image correlation)

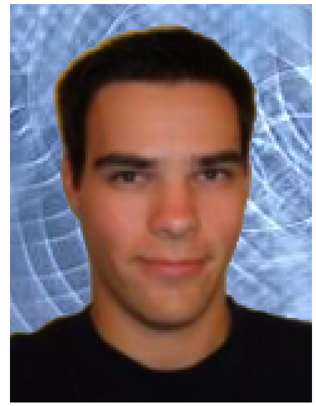
Conclusions

- Variational methods show comparable performance with existing approaches
  - The main sources of errors is the bounding error
  - Design of piecewise bounds to control these errors
  - A good control over speed-accuracy trade-offs can be obtained
- Variational lower bounds can be optimized efficiently
  - Use of convex optimization methods to get fast convergence rates and easy convergence diagnostics
  - Design of efficient expectation-maximization (EM) algorithms

Collaborators

Kevin Murphy
UBC

Benjamin Marlin
U. Mass-Amherst



Guillaume Bouchard
XRCE, France

Shakir Mohamed
U. Cambridge, now at UBC

Thank You