Piecewise Bounds for Estimating Discrete-Data Latent Gaussian Models

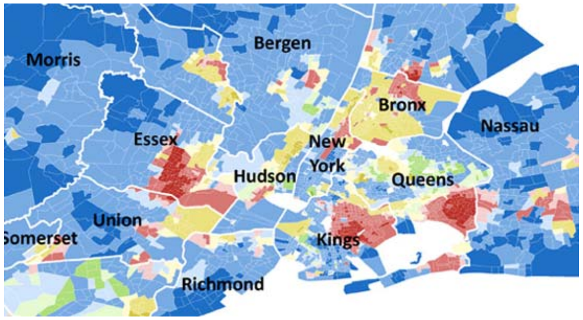
Mohammad Emtiyaz Khan

Joint work with Benjamin Marlin, and Kevin Murphy

University of British Columbia

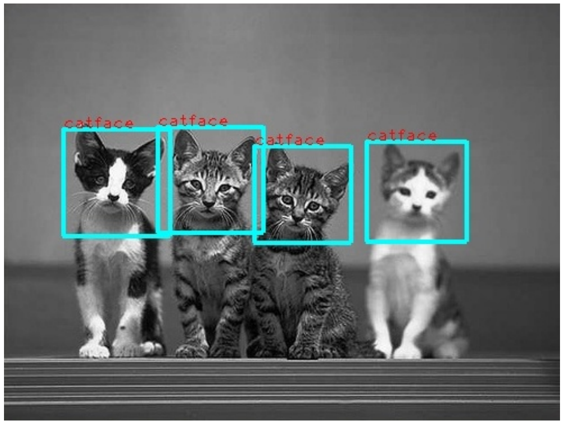
September 29, 2011

Survey/voting data and blogs for sentiment analysis



User rating data

Object detection, classification, tag correlation.



Consumer choice data

Sports/game data

Health data

Modeling Discrete Data

We would like to learn correlations in data, so that we can make useful predictions.

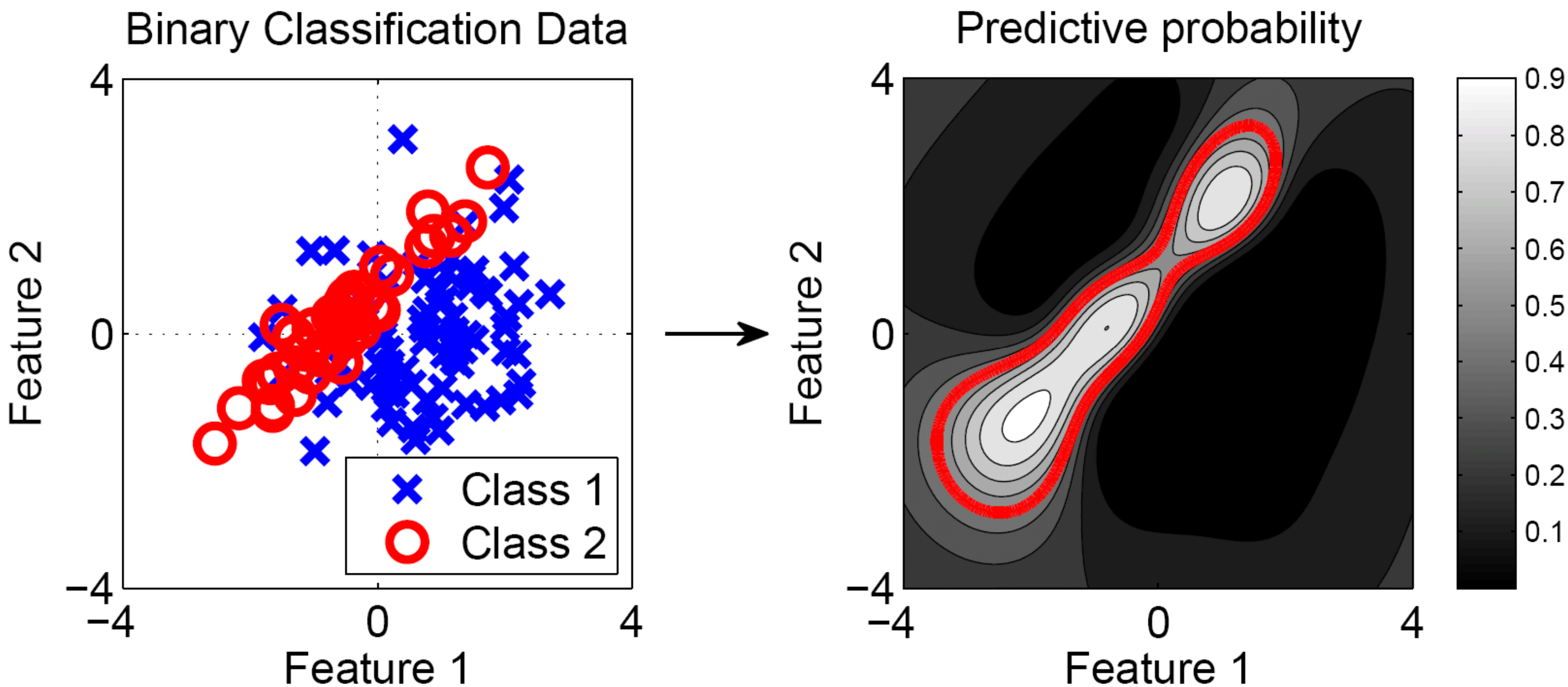
This talk, we focus on an important class of models,

- Latent Gaussian Model (LGM)
- Likelihoods based on the logit link
- Binary, categorical, and ordinal data

LGMs - Classification Models

Bayesian Logistic Regression and Gaussian Process Classification

Jaakkola and Jordan 1996, Rasmussen 2004, Gibbs and Mackay 2000, Kuss and Rasmussen 2006, Nickisch and Rasmussen 2008, Kim and Ghahramani, 2003, Girolami and Rogers 2006, Seeger and Jordan 2004, William and Barber 1998, Minka 2001, Albert and Chib 1993, Holmes and held 2004, Scott 2010, Braum and McAulliffe 2010, Rue and Held 2009, Cseke and Heskes 2010.

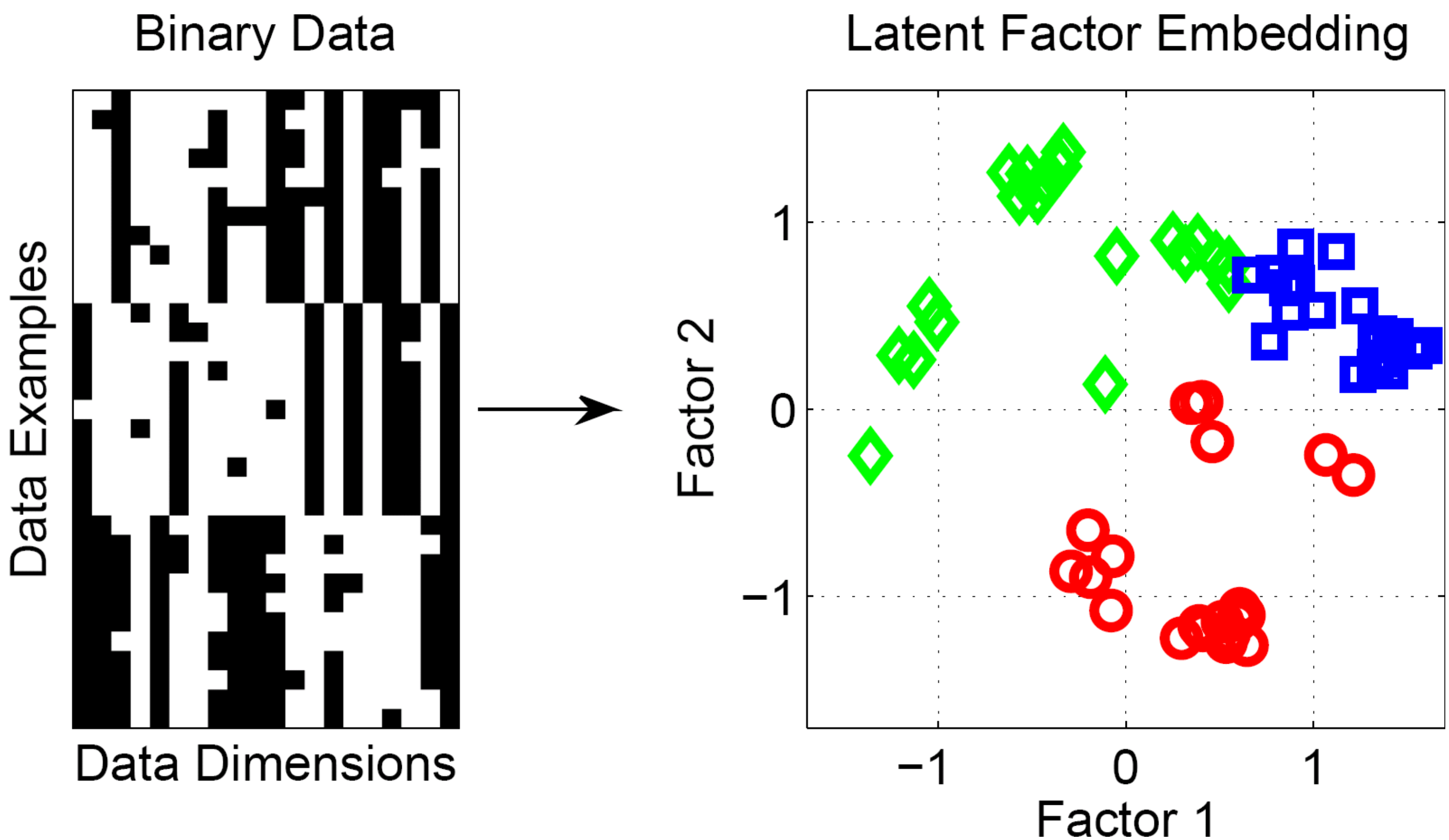


Figures reproduced using GPML toolbox

LGMs - Latent Factor Models

Probabilistic PCA and Factor Analysis Models

Tipping 1999, Collins, Dasgupta and Schapire 2001, Mohammed, Heller, and Ghahramani 2008, Girolami 2001, Yu and Tresp 2004, Khan, Marlin, and Murphy 2010.

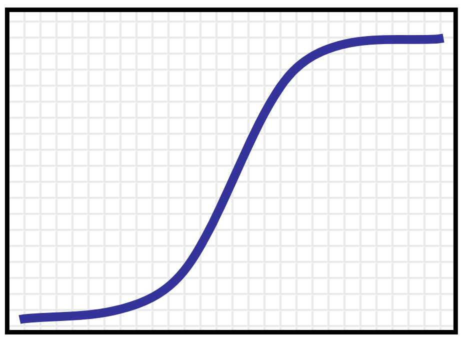


Parameter Learning is Intractable

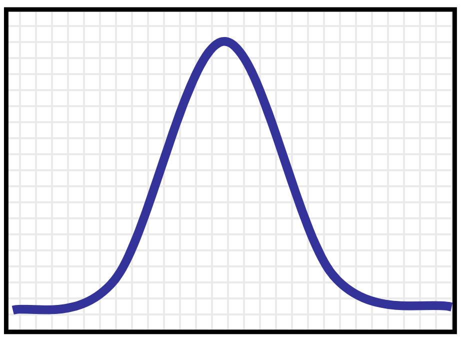
Likelihood based on logit function is not conjugate to the Gaussian prior.

$(1+\exp(z))^{-1}$

$\int$



x



$d\mathbf{z}$

We propose piecewise bounds to obtain tractable lower bounds to marginal likelihood.

Outline

Binary Data LGMs ICML 2011

Difficulty in parameter learning - Jensen's inequality is insufficient - Existing bounds can be bad - Piecewise bounds – Results

Categorical Data LGMs Work in Progress

Multinomial Logit model - Existing bounds can be bad - A new model Stick-breaking LGM - Use of piecewise bounds – Results

Ordinal Data LGMs

Application of piecewise bounds to Proportional-Odds model

Conclusions

Outline

Binary Data LGMs ICML 2011

Difficulty in parameter learning - Jensen's inequality is insufficient - Existing bounds can be bad - Piecewise bounds – Results

Categorical Data LGMs Work in Progress

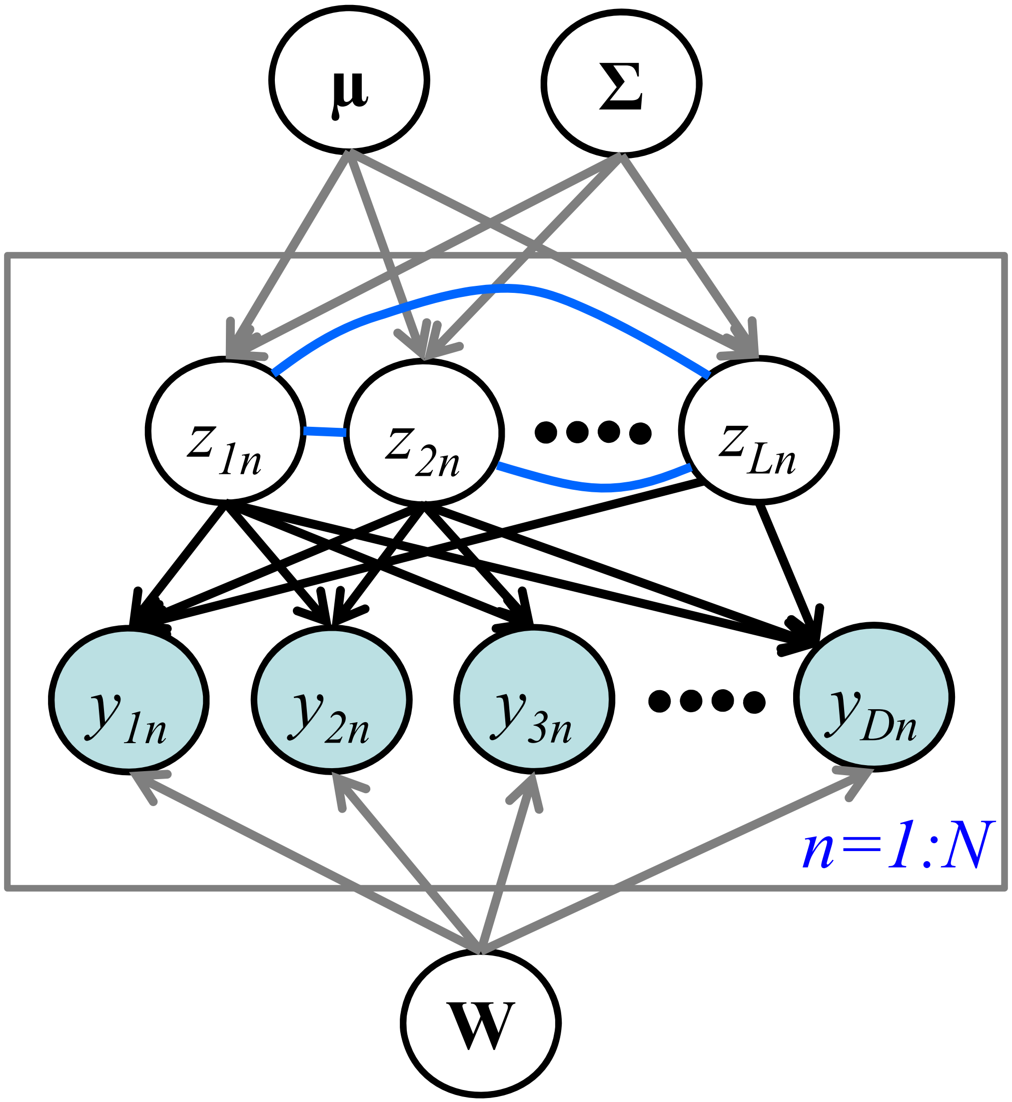
Multinomial Logit model - Existing bounds can be bad - A new model Stick-breaking LGM - Use of piecewise bounds – Results

Ordinal Data LGMs

Application of piecewise bounds to Proportional-Odds model

Conclusions

Latent Gaussian Models (LGMs)



Sample Gaussians

$$p(\mathbf{z}_n|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

Linear transform

$$\boldsymbol{\eta}_{dn} = \mathbf{W}_d \mathbf{z}_n$$

Likelihood Examples

Binary

$$p(y=1|\mathbf{z},\boldsymbol{\theta}) = \sigma(\eta)$$

Categorical

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

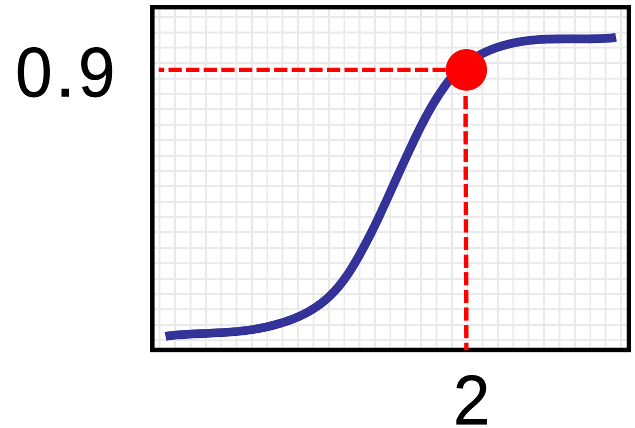
Parameter Set

$$\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{W}\}$$

Bernoulli-Logistic LGM

$$\eta = \mathbf{w}_d^T \mathbf{z}_n$$

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

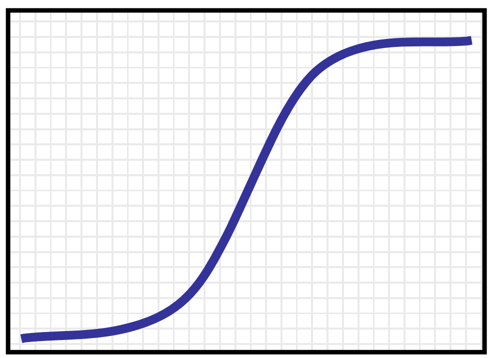


$$\boxed{\log p(y = 1|\eta) = \eta - \log(1 + e^\eta)}$$

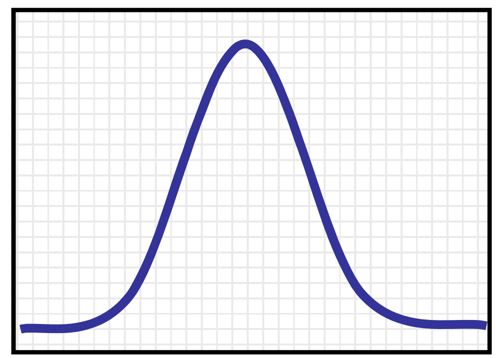
Parameter Estimation

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N) = \sum_{n=1}^{N} \log \int \prod_{d=1}^{D} p(y_{dn}|\mathbf{z}, \boldsymbol{\theta}) \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{z}$$

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



x



$$d\mathbf{z}$$

Variational Lower Bound (Jensen's)

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

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

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

Variational Lower Bound (Jensen's)

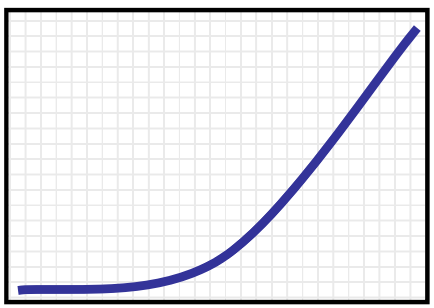
$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) \geq \max_{\mathbf{m},\mathbf{V}} \sum_{d=1}^{D} \int [\log p(y_d|\mathbf{z},\boldsymbol{\theta})] \mathcal{N}(\mathbf{z}|\mathbf{m},\mathbf{V}) d\mathbf{z}$$

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

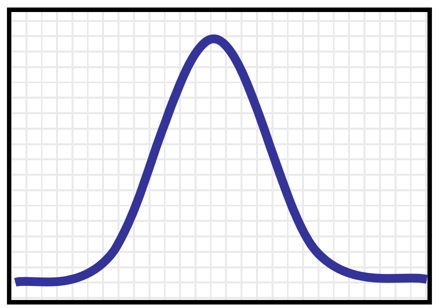
$$= \max_{\mathbf{m},\mathbf{V}} \sum_{d=1}^{D} \int [-\log(1+e^{\eta})] \mathcal{N}(\tilde{m}_d, \tilde{v}_d) d\eta \quad +$$

some other tractable terms in **m** and **V**

$-\int$



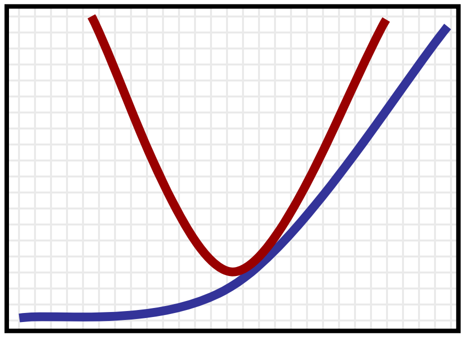
x



$d\eta$

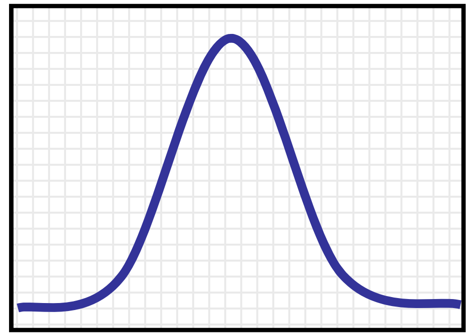
Quadratic Bounds

$\log(1 + \exp(x_d))$



x

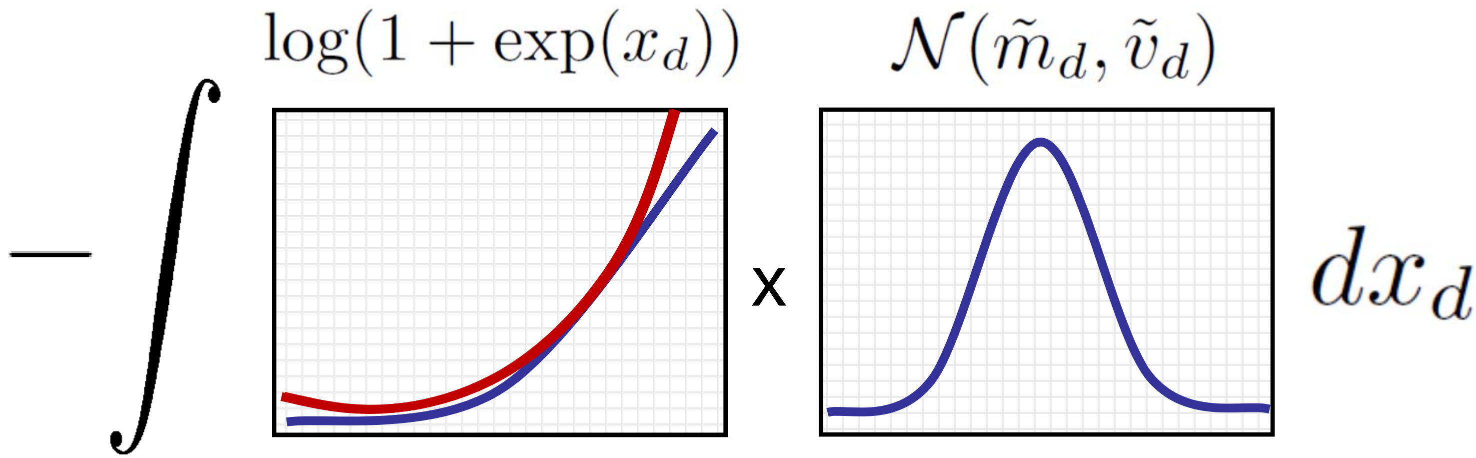
$\mathcal{N}(\tilde{m}_d, \tilde{v}_d)$



$$- \int \log(1 + \exp(x_d)) \times \mathcal{N}(\tilde{m}_d, \tilde{v}_d) \, dx_d$$

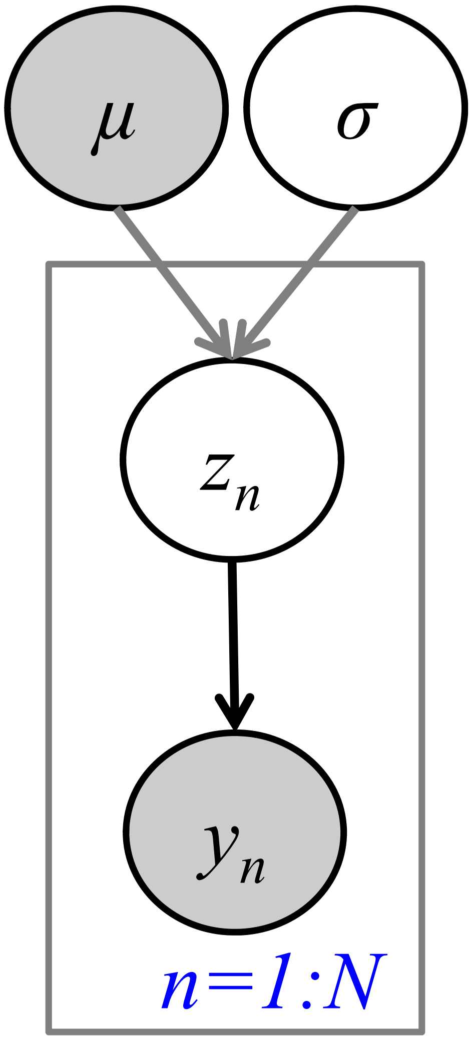
- Bohning's bound (Bohning, 1992)

Quadratic Bounds



- Bohning's bound (Bohning, 1992)
- Jaakkola's bound (Jaakkola and Jordan, 1996)
- Both bounds have unbounded error.

Problems with Quadratic Bounds



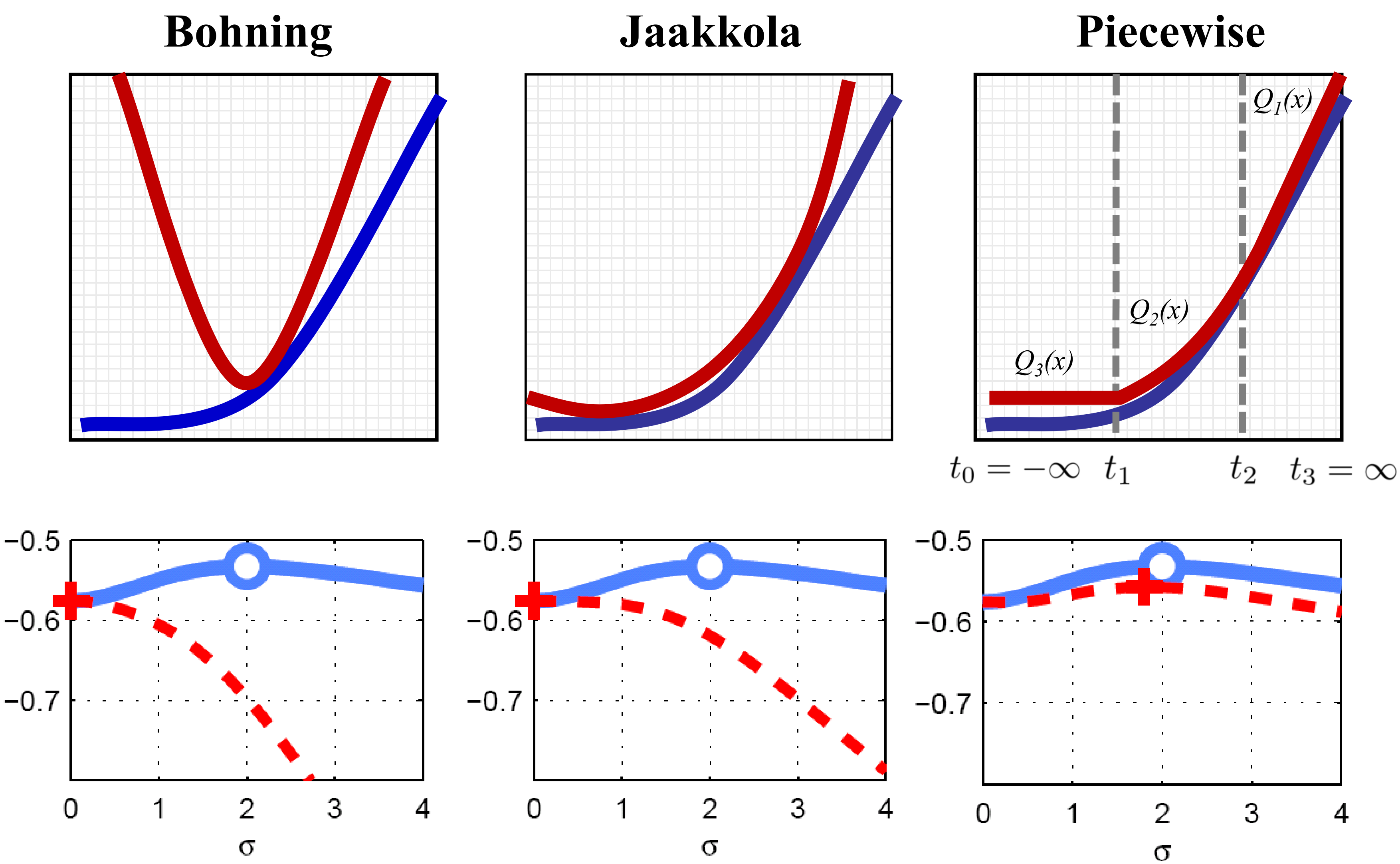
1-D example with $\mu = 2$, $\sigma = 2$

$$p(y = 1|\mu, \sigma^2) = \int (1 + \exp(z))^{-1} \mathcal{N}(z|\mu, \sigma^2) dz$$

Generate data, fix $\mu = 2$, and compare marginal likelihood and lower bound wrt σ

As this is a 1-D problem, we can compute lower bounds without Jensen's inequality. So plots that follow have errors only due to error in bounds.

Problems with Quadratic Bounds



Outline

Binary Data LGMs ICML 2011

Difficulty in parameter learning - Jensen's inequality is insufficient - Existing bounds can be bad - Piecewise bounds – Results

Categorical Data LGMs Work in Progress

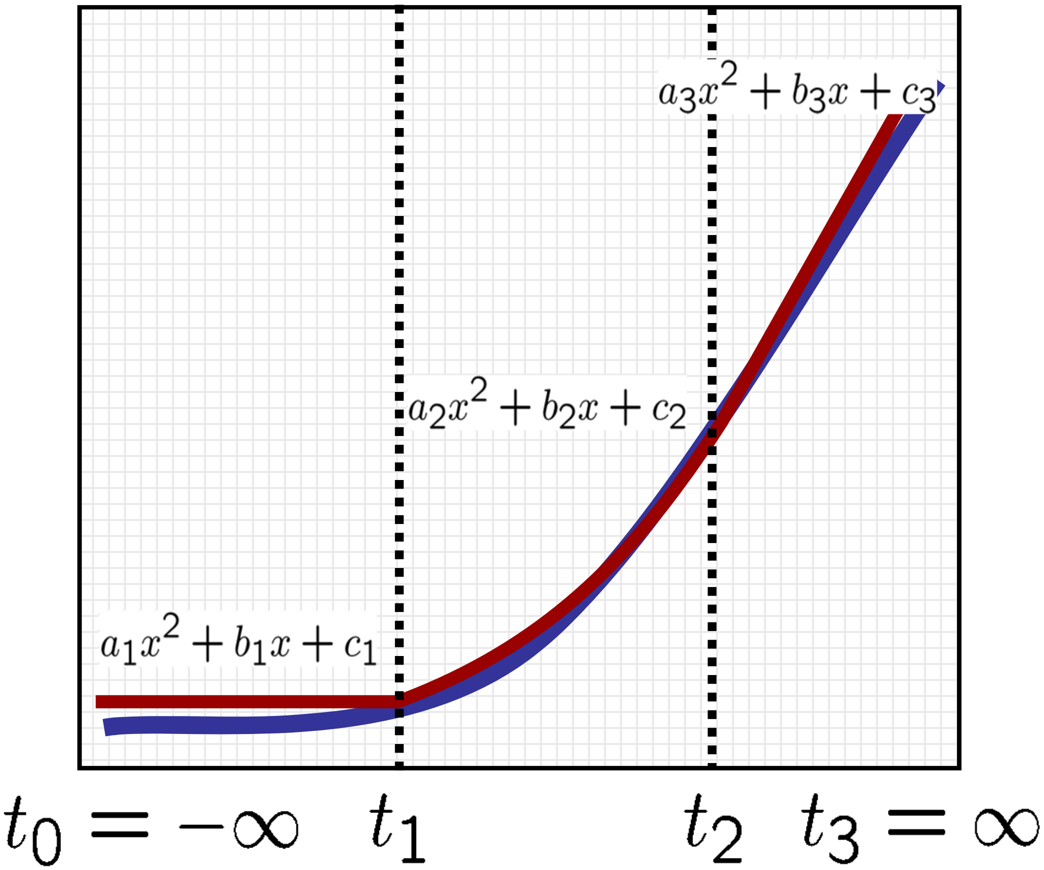
Multinomial Logit model - Existing bounds can be bad - A new model Stick-breaking LGM - Use of piecewise bounds – Results

Ordinal Data LGMs

Application of piecewise bounds to Proportional-Odds model

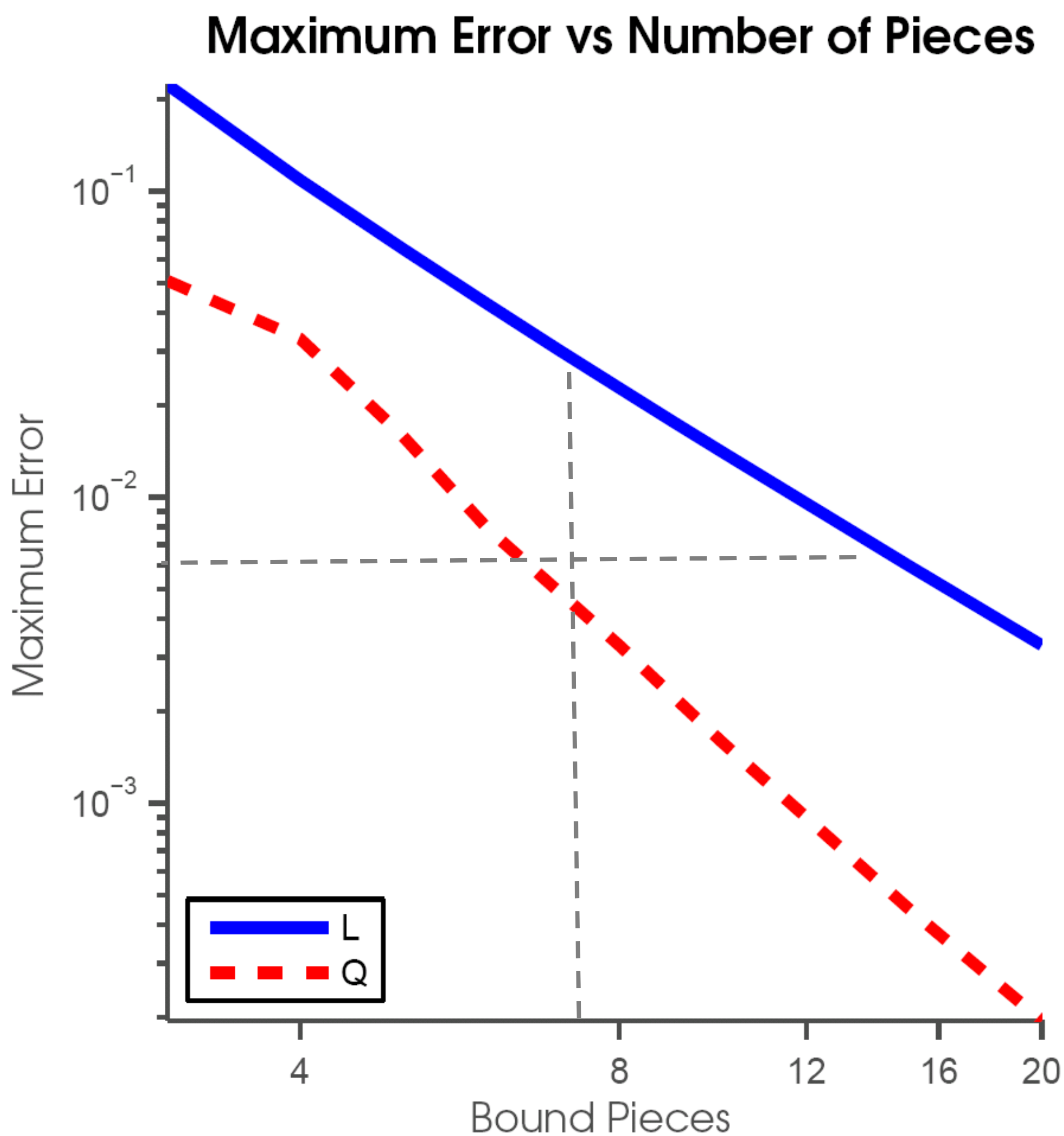
Conclusions

Finding Piecewise Bounds



- Find cut points and parameters of each pieces 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.

Linear Vs Quadratic



Outline

Binary Data LGMs ICML 2011

Difficulty in parameter learning - Jensen's inequality is insufficient - Existing bounds can be bad - Piecewise bounds – Results

Categorical Data LGMs Work in Progress

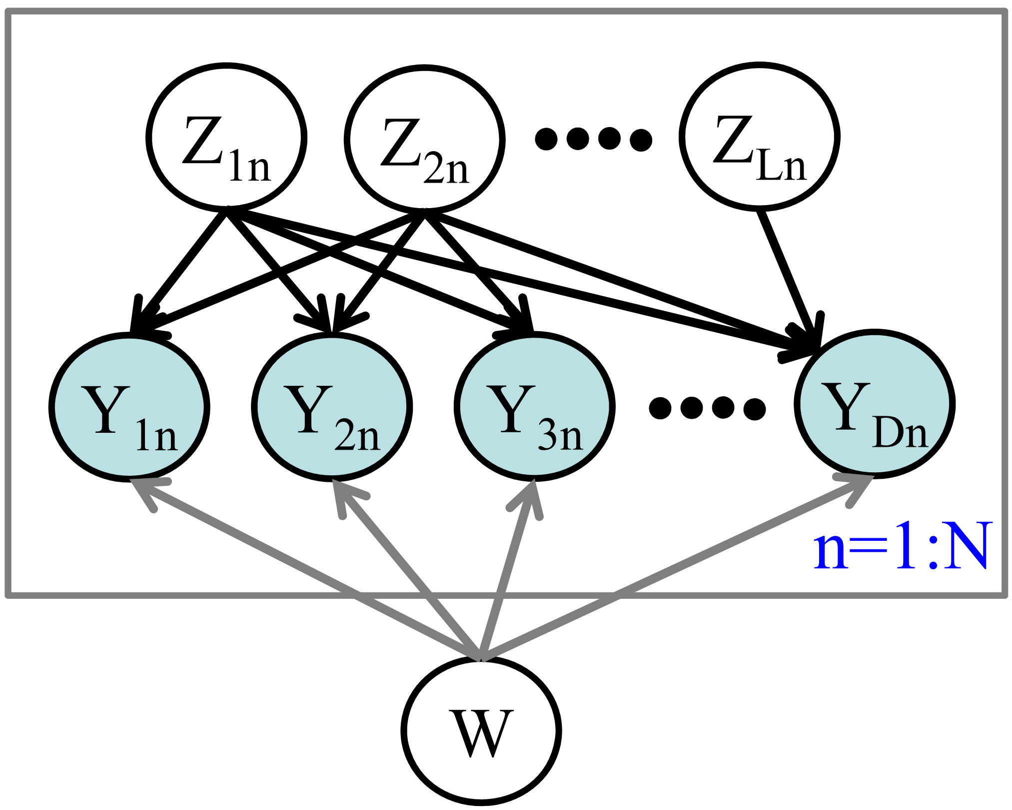
Multinomial Logit model - Existing bounds can be bad - A new model Stick-breaking LGM - Use of piecewise bounds – Results

Ordinal Data LGMs

Application of piecewise bounds to Proportional-Odds model

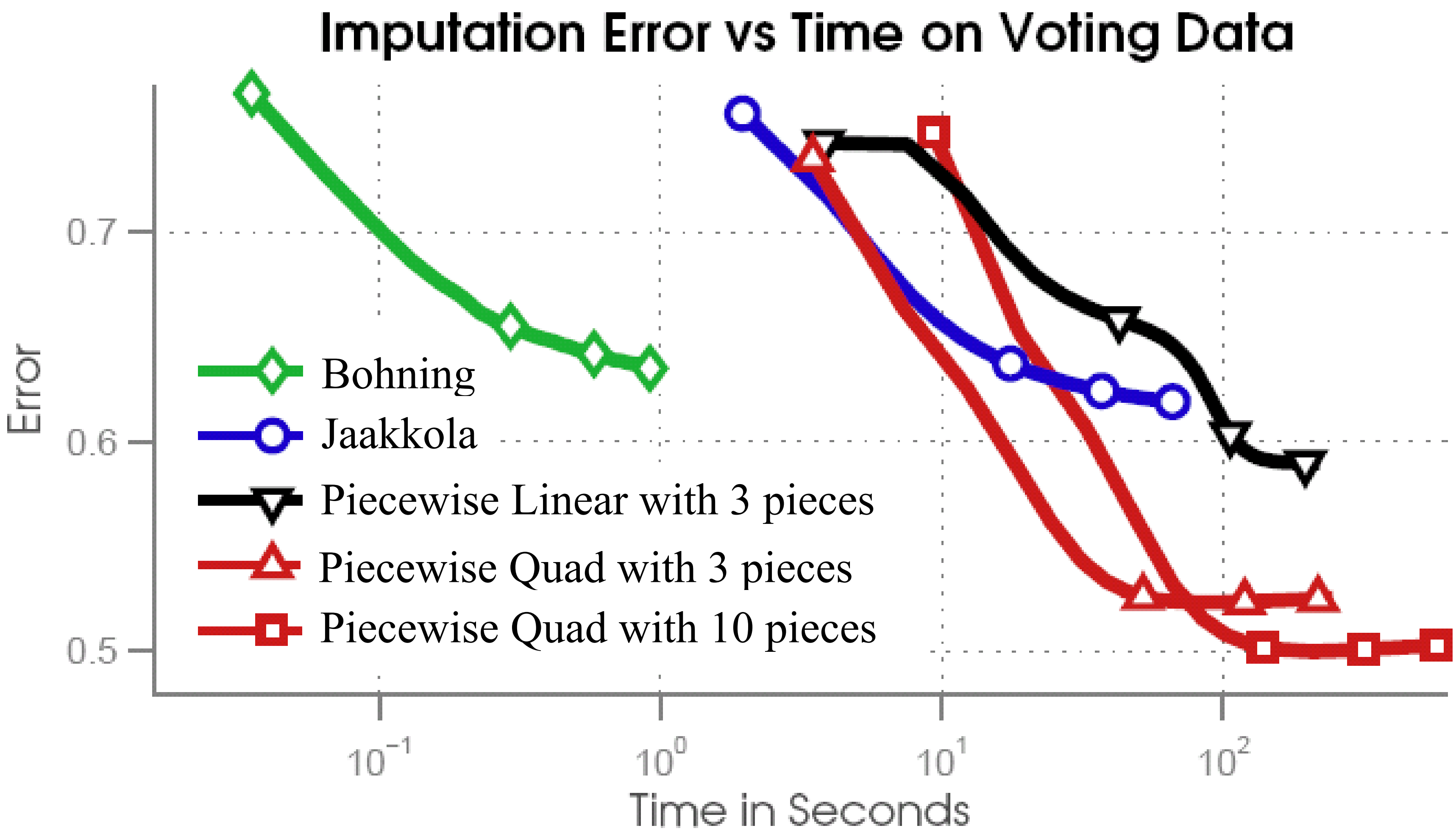
Conclusions

Binary Factor Analysis (bFA)

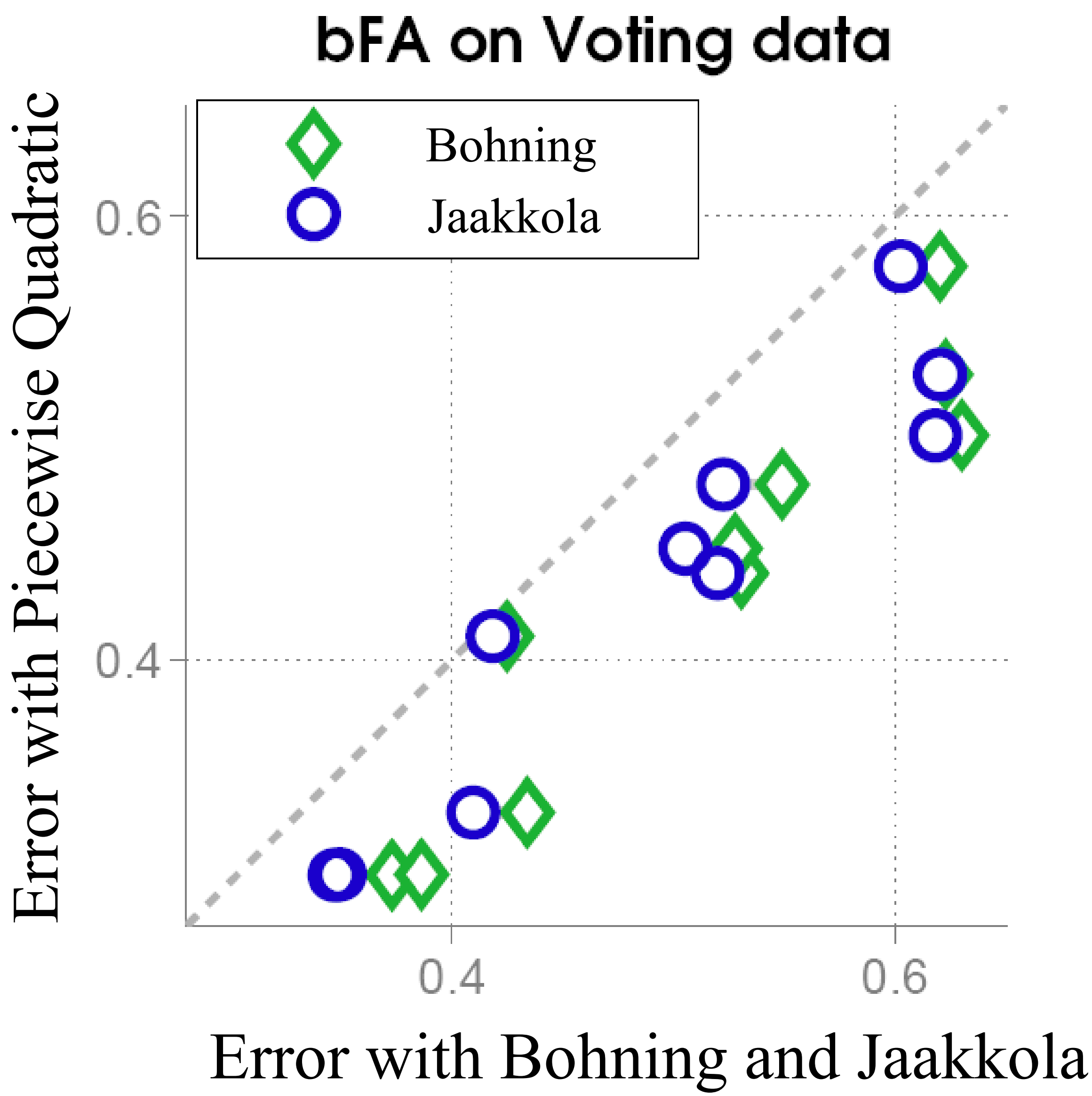


- UCI voting dataset with D=15, N=435.
- Train-test split 80-20%
- Compare cross-entropy error on missing value prediction on test data.

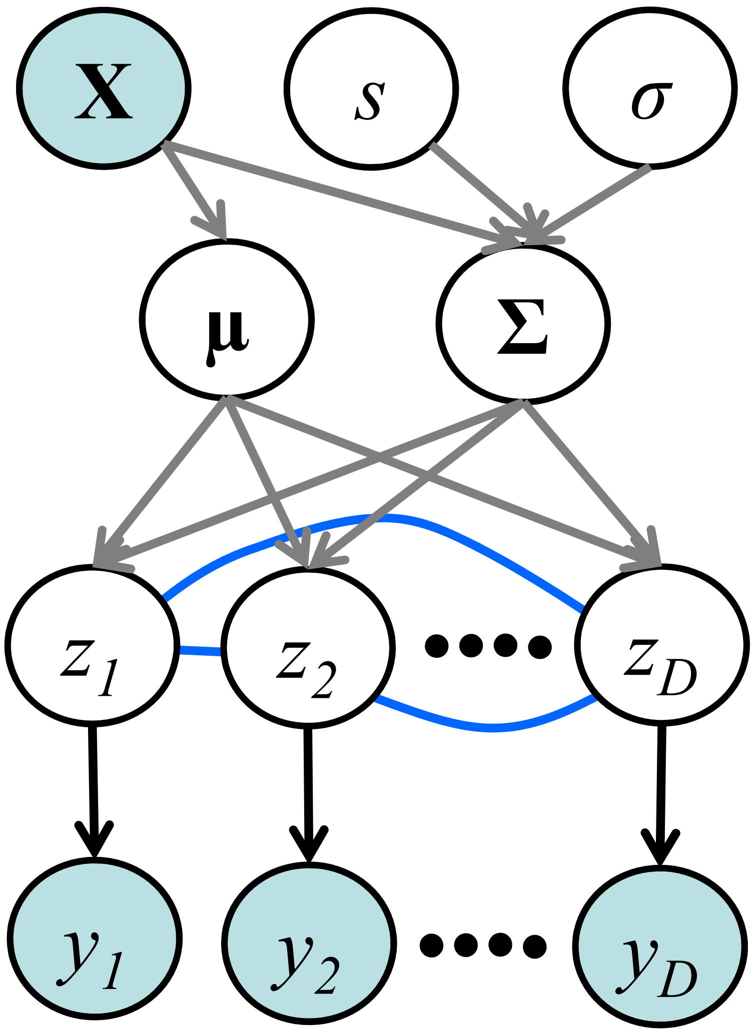
bFA – Error vs Time



bFA – Error Across Splits



Gaussian Process Classification

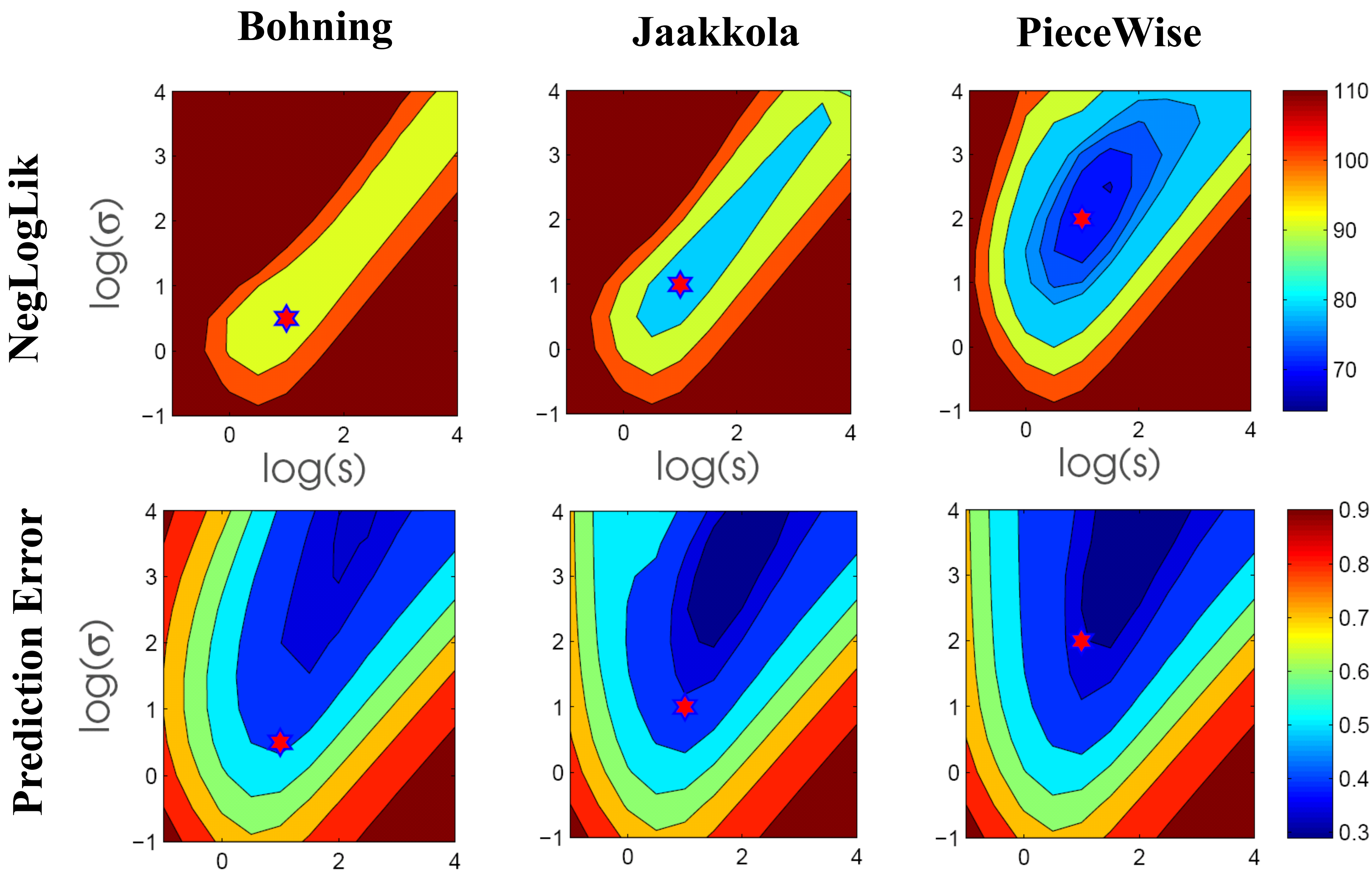


- We repeat the experiments described in Kuss and Rasmussen, 2006
- We set μ = 0 and squared exponential Kernel

$\Sigma_{ij} = \sigma \, exp[(x_i - x_j)\hat{}2/s]$

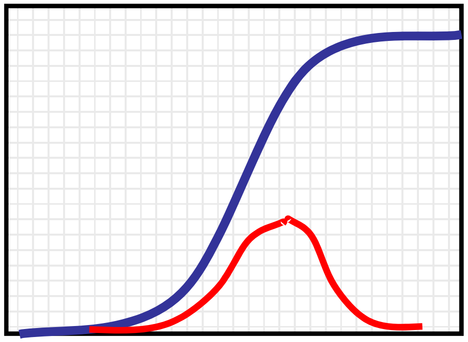
- Estimate σ and s.
- We run experiments on Ionoshphere dataset (D = 200)
- Compare Cross-entropy Prediction Error for test data.

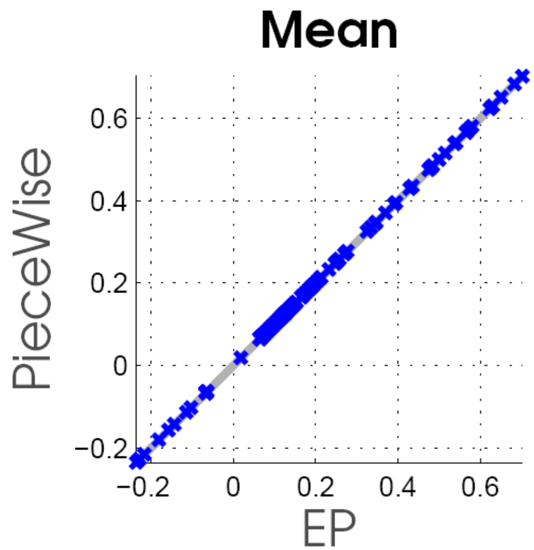
Binary GP Classification

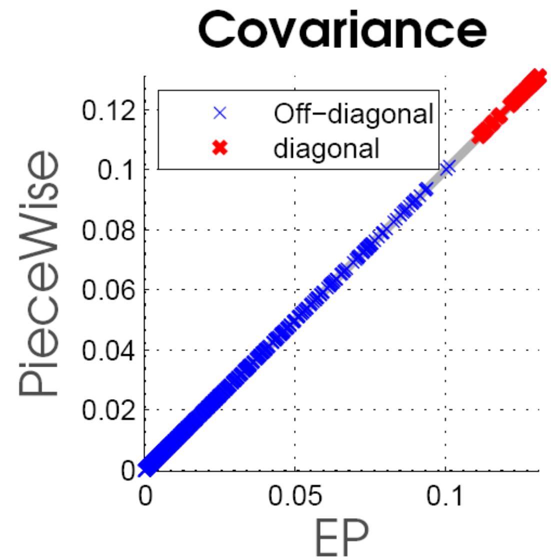


Comparison with EP – Posterior Distribution

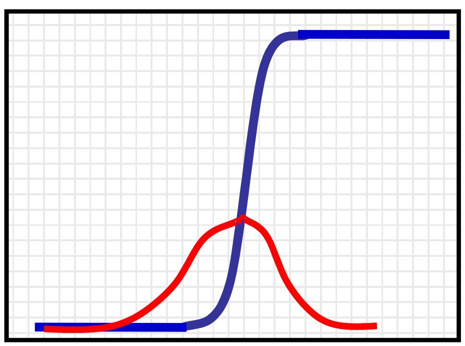
log(σ) = -1, log(*s*) = *-1*

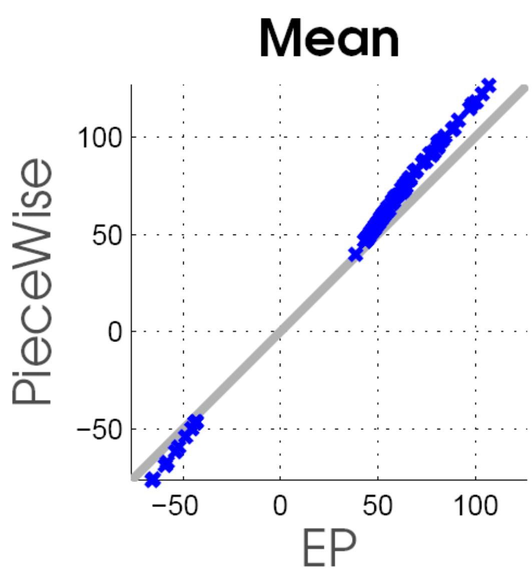


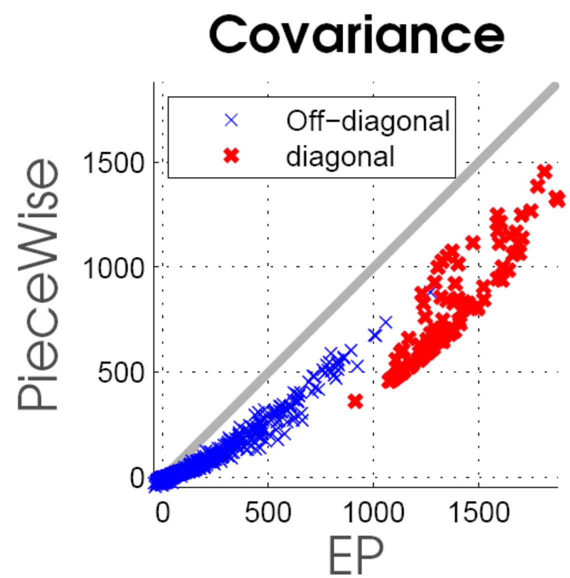




log(σ) = 4, log(*s*) = *-1*

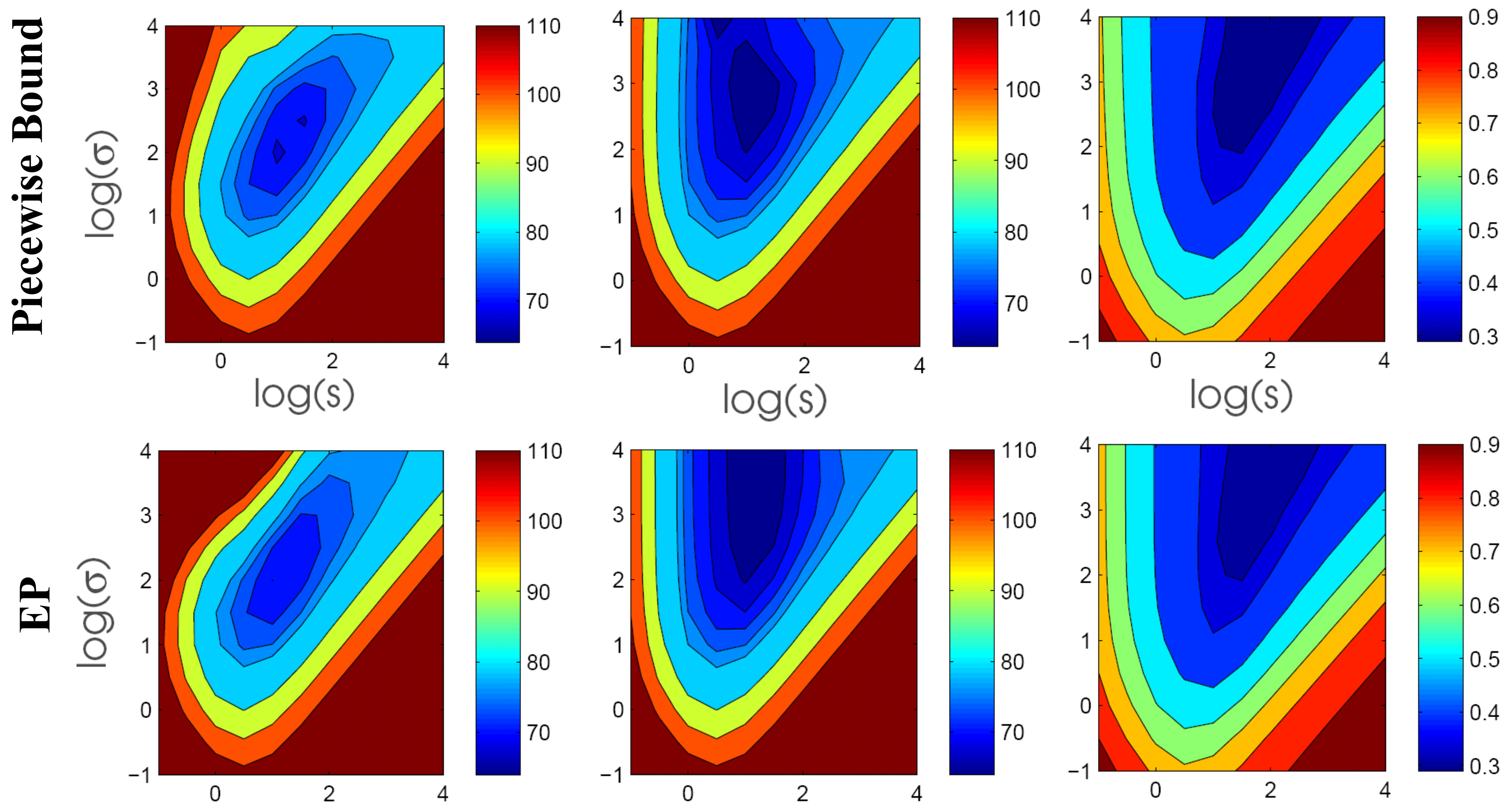






So which one is better?

(Neg) KL-Lower Bound to MargLik | **Approximation To Marglik** | **Pred Error**



Comparison with EP

- Both methods give very similar results.
- For parameter learning, variational EM algorithm based on the piecewise bound has a well-defined objective function and hence the algorithm is guaranteed to converge when appropriate numerical methods are used.
- Nickisch and Rasmussen (2008) describe the variational approach as more principled than EP.

Outline

Binary Data LGMs ICML 2011

Difficulty in parameter learning - Jensen's inequality is insufficient - Existing bounds can be bad - Piecewise bounds – Results

Categorical Data LGMs Work in Progress

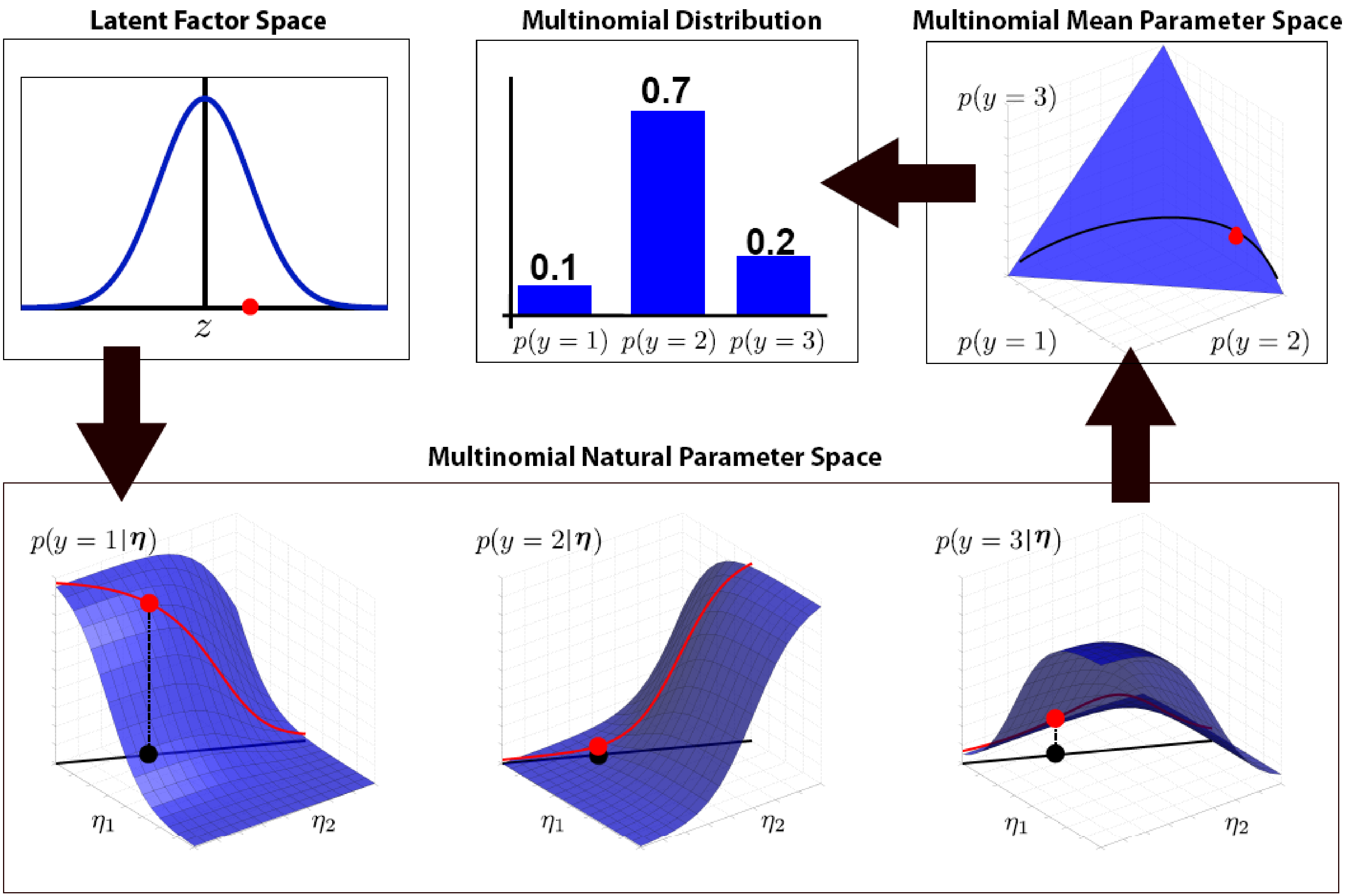
Multinomial Logit model - Existing bounds can be bad - A new model Stick-breaking LGM - Use of piecewise bounds – Results

Ordinal Data LGMs

Application of piecewise bounds to Proportional-Odds model

Conclusions

Multinomial-Logit LGM



Multinomial-Logit LGM

$$y \in \{C_1, C_2, C_3, \ldots, C_K\}$$

Let $\boldsymbol{\eta} \in \mathbb{R}^{K-1}$ and defined as follows:

$$\eta_k = \mathbf{w}_{dk} \mathbf{z}_n \qquad p(y = k|\boldsymbol{\eta}) = \frac{e^{\eta_k}}{\sum_{j=1}^{K} e^{\eta_j}}$$

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

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

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

Stick (breaking) LGMs

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

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

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

$$\vdots$$

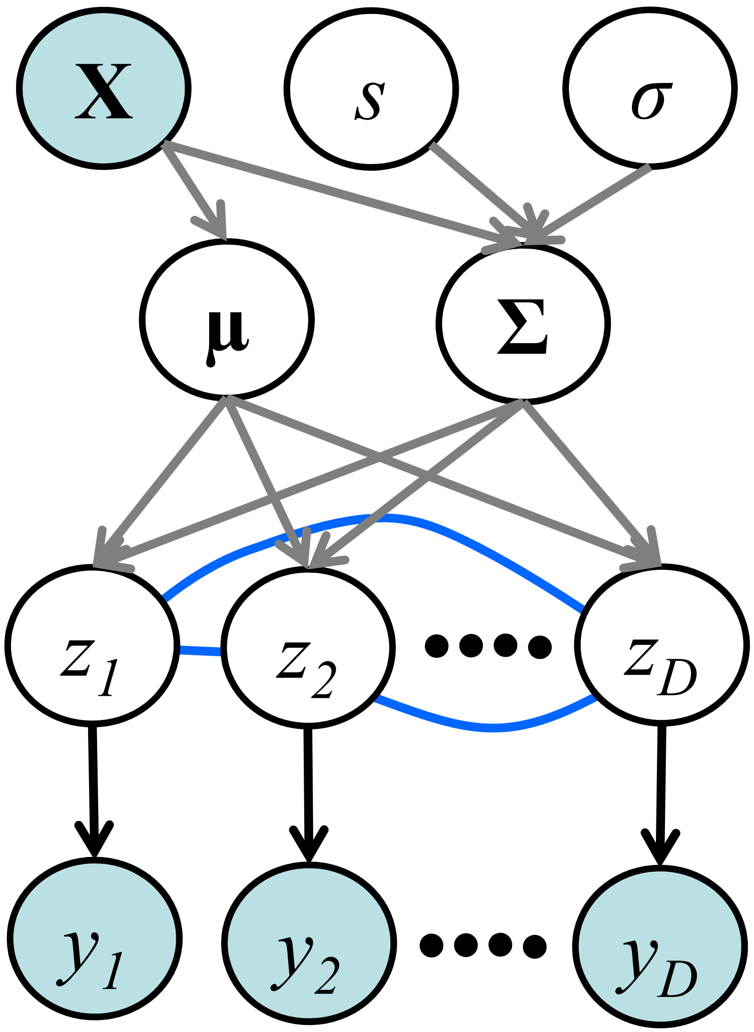
$$p(y=K|\boldsymbol{\eta}) = \prod_{j=1}^{K-1}\left(1 - \frac{e^{\eta_j}}{1+e^{\eta_j}}\right)$$

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

Multinomial Logit vs Stick

- Stick model depends on the order the categories are chosen, hence it is not easy to interpret the weights.
- However, for parameter estimation the ordering may not matter and we could still use the model for cases we do not care about interpretability.
- The advantage is that fitting the stick model is more accurate than multinomial logit model!
- I came across this model during a discussion with Guillaume Bouchard, XRCE, France. It has also been used in Mnih and Hinton 2009 for language modeling.

Gaussian Process Classification

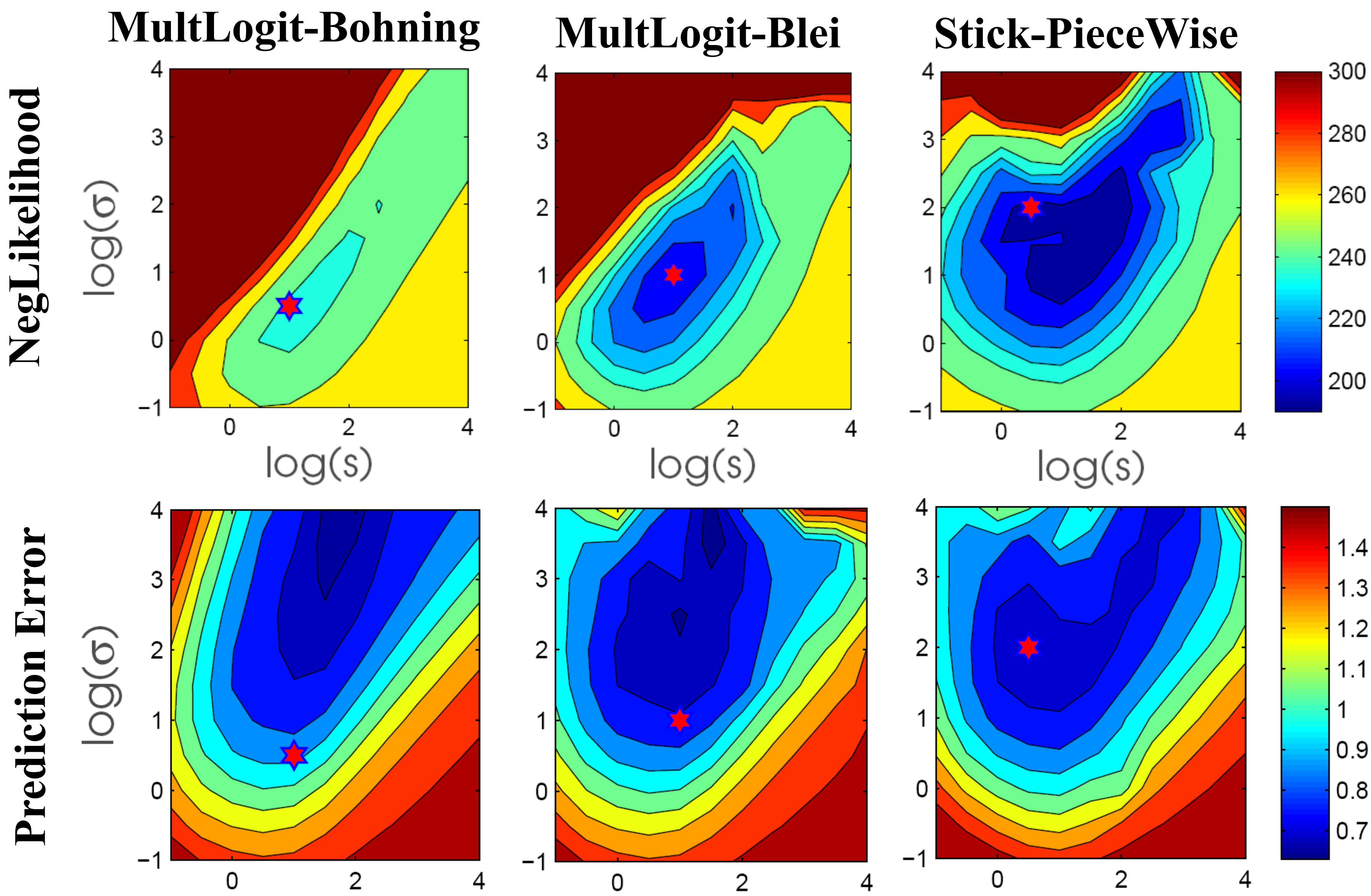


- We repeat the experiments described in Girolami and Rogers, 2006
- We set μ =0 and squared exponential Kernel

$\Sigma_{ij}\,(k) = \sigma\; exp[(x_i - x_j)^\wedge 2/s]$

- Estimate σ and *s*.
- We run experiments on Glass dataset (D = 143)
- Compare Cross-entropy Prediction Error for test data (D = 41)

Multi-Class GP



Work in progress

- Comparison with ground truth using MCMC (joint work with Dr. Shakir Mohamed) and Variational-Bayes Message Passing (VBMP) due to Girolami and Rogers 2006.
- Comparison on the factor analysis model to get time vs accuracy plots.

Outline

Binary Data LGMs ICML 2011

Difficulty in parameter learning - Jensen's inequality is insufficient - Existing bounds can be bad - Piecewise bounds – Results

Categorical Data LGMs Work in Progress

Multinomial Logit model - Existing bounds can be bad - A new model Stick-breaking LGM - Use of piecewise bounds – Results

Ordinal Data LGMs

Application of piecewise bounds to Proportional-Odds model

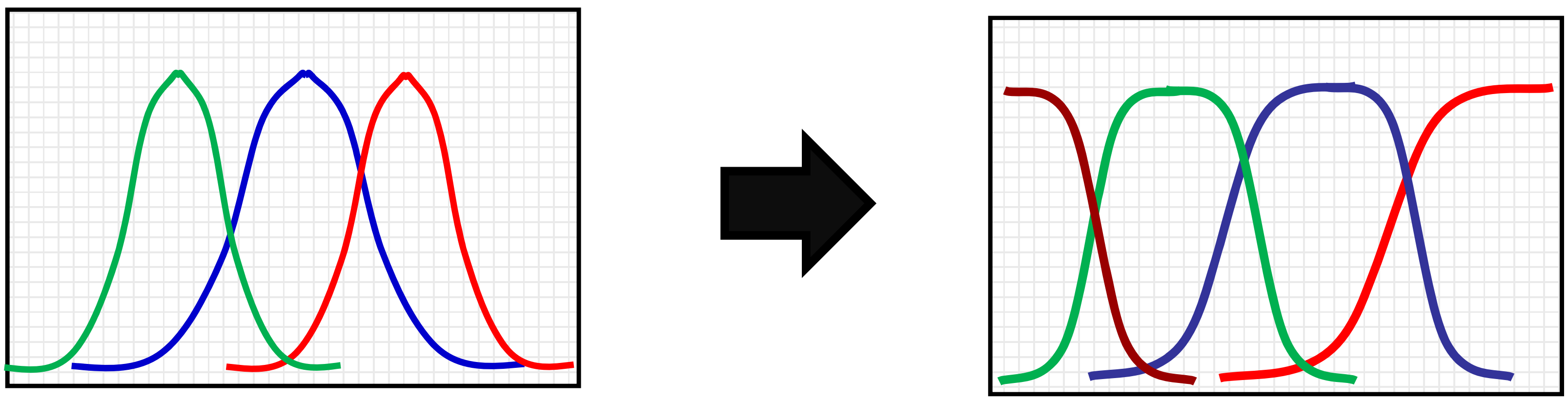
Conclusions

LGM for Ordinal Data

$y \in \{1, 2, 3, \ldots, K\}$

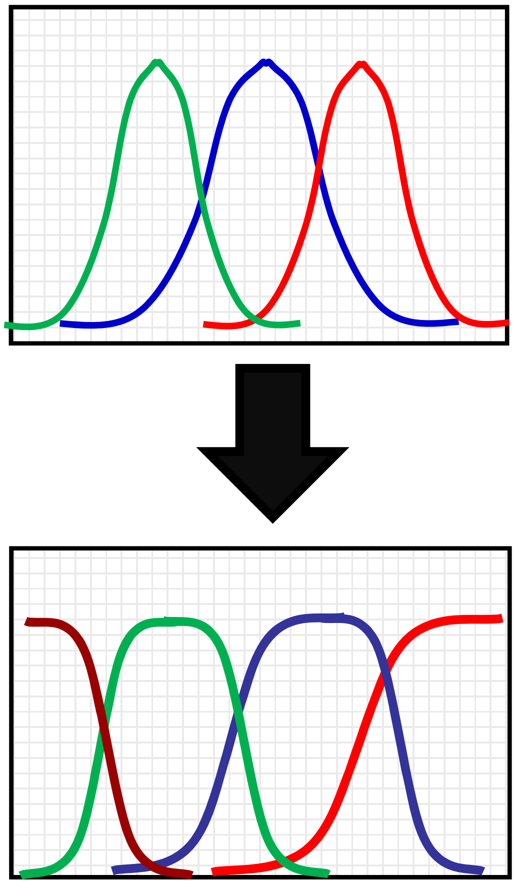
Let $\boldsymbol{\eta} \in \mathbb{R}^{K-1}$ and defined as follows:

$\eta_k = \mathbf{w}_d \mathbf{z}_n + w_{0k}$



Proportional Odds Model

$$p(y = k|\boldsymbol{\eta}) = p(y \le k|\boldsymbol{\eta}) - p(y \le k-1|\boldsymbol{\eta})$$



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

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

$$\vdots$$

$$p(y = K-1|\boldsymbol{\eta}) = \frac{1}{1+e^{\eta_{K-1}}} - \frac{1}{1+e^{\eta_{K-2}}}$$

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

Variational Lower Bound

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

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

$$\log p(y=k|\boldsymbol{\eta}) = \log\left[\frac{1}{1+e^{\eta_k}} - \frac{1}{1+e^{\eta_{k-1}}}\right]$$

$$= \log\left[\frac{e^{\eta_{k-1}} - e^{\eta_k}}{(1+e^{\eta_k})(1+e^{\eta_{k-1}})}\right] = \log\left[\frac{e^{\eta_{k-1}}(1-e^{\eta_k-\eta_{k-1}})}{(1+e^{\eta_k})(1+e^{\eta_{k-1}})}\right]$$

$$= \eta_{k-1} - \log(1+e^{\eta_k}) - \log(1+e^{\eta_{k-1}}) + \text{cnst}$$

Conclusions

Binary Data LGMs Variational inference can perform badly if bounds have unbounded error. In piecewise bounds, we can drive error in the bound to zero by increasing the number of pieces. This leads to improved performance. We also get a fine control over speed vs accuracy.

Categorical Data LGMs the new Stick-breaking LGM is much easier to fit than the multinomial logit model. Preliminary experiment show promising results.

Ordinal Data LGMs Application of piecewise bounds to Proportional-Odds model.

Other Work

- Variational bounds and approximation for binary, categorical and ordinal data.
- Theoretical analysis of errors made by various bounds and derivation of sufficient conditions under which one bound is superior than the others.
- Design guidelines to choose a particular approximation based on speed-accuracy trade-offs.
- Application to many real-world data.

Thank You

Piecewise-Bounds: Optimization Problem

$$\min_{\mathbf{t},\mathbf{a}} \max_{r\in\{1,..,R\}} \max_{t_{r-1}\leq x<t_r} a_r x^2 + b_r x + c_r - \mathrm{lse}(x)$$

$$\begin{aligned} & a_r x^2 + b_r x + c_r - \mathrm{lse}(x) \geq 0 && \forall r \in \{1,..,R\}, \forall x \in [t_{r-1}, t_r] \\ s.t. \quad & t_r - t_{r-1} > 0 && \forall r \in \{1,..,R\} \\ & a_r \geq 0 && \forall r \in \{1,..,R\} \end{aligned}$$

$$\min_{\mathbf{t},\mathbf{a}} \max_{r\in\{1,..,R\}} \left(\max_{t_{r-1}\leq x<t_r} a_r x^2 + b_r x - \mathrm{lse}(x) \right) - \left(\min_{t_{r-1}\leq x<t_r} a_r x^2 + b_r x - \mathrm{lse}(x) \right)$$

$$E_{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)}[\log p(\mathbf{y}_n|\mathbf{z},\boldsymbol{\theta})]$$

$$\geq \sum_{d=1}^{D} \left(y_{dn}\mathbf{W}_d^T\mathbf{m}_n - E_{q_n(\mathbf{z}|\boldsymbol{\gamma}_n)}[B_{\boldsymbol{\alpha}}(\mathbf{W}_d^T\mathbf{z})] \right)$$

$$= \sum_{d=1}^{D} \left(y_{dn}\mathbf{W}_d^T\mathbf{m}_n - E_{q_n(\eta|\tilde{\boldsymbol{\gamma}}_{dn})}[B_{\boldsymbol{\alpha}}(\eta)] \right)$$

$$\tilde{\gamma}_{dn} = \{\tilde{m}_{dn}, \tilde{v}_{dn}\},\ \tilde{m}_{dn} = \mathbf{W}_d^T\mathbf{m}_n,\ \tilde{v}_{dn} = \mathbf{W}_d^T\mathbf{V}_n\mathbf{W}_d$$

$$E_{q_n(\eta_{dn}|\tilde{\boldsymbol{\gamma}}_{dn})}[B_{\boldsymbol{\alpha}}(\eta)] = \sum_{r=1}^{R} f_r(\tilde{m}_{dn}, \tilde{v}_{dn}, \boldsymbol{\alpha})$$

$$= \sum_{r=1}^{R} \int_{t_{r-1}}^{t_r} (a_r\eta^2 + b_r\eta + c_r)\mathcal{N}(\eta|\tilde{m}_{dn}, \tilde{v}_{dn})d\eta$$

Algorithm 1 bLGM Generalized EM Algorithm

E-Step:

$$\frac{\partial \mathcal{L}_{QJ}}{\partial \mathbf{m}_{kn}} \leftarrow \sum_{d=1}^{D} y_{dn} \mathbf{W}_{dk} - \sum_{l=1}^{K} (\boldsymbol{\Sigma}^{-1})_{lk} (\mathbf{m}_{ln} - \boldsymbol{\mu}_l) - \sum_{r=1}^{R} \sum_{d=1}^{D} \mathbf{W}_{dk} \frac{\partial f_r(\tilde{m}_{dn}, \tilde{v}_{dn}, \boldsymbol{\alpha})}{\partial \tilde{m}_{dn}}$$

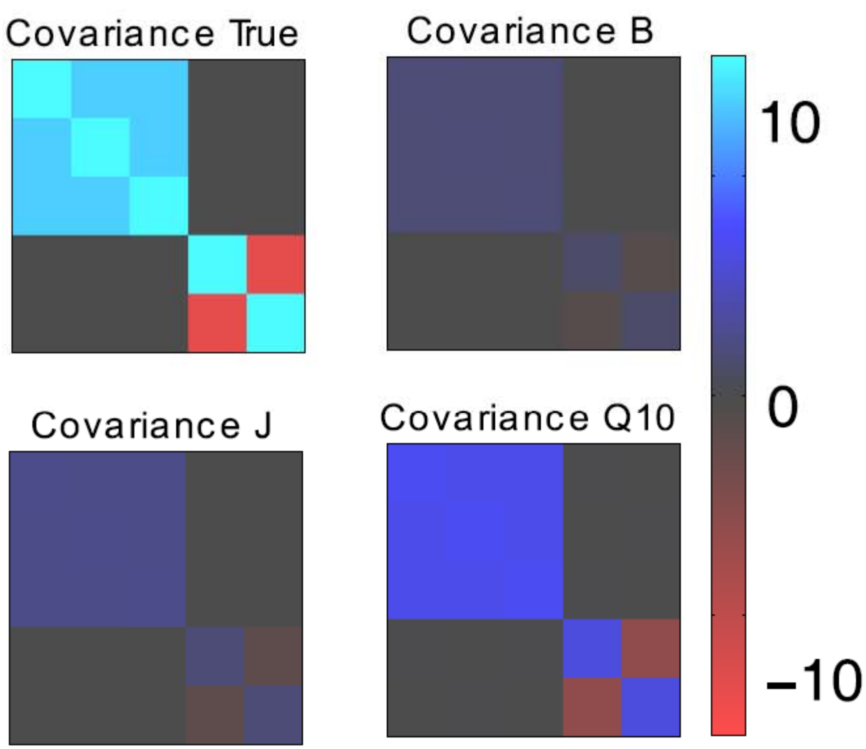
$$\frac{\partial \mathcal{L}_{QJ}}{\partial \mathbf{V}_{kl}} \leftarrow \frac{1}{2} (\boldsymbol{\Sigma}^{-1})_{kl} - \frac{1}{2} (\mathbf{V}_n^{-1})_{kl} - \sum_{r=1}^{R} \sum_{d=1}^{D} \mathbf{W}_{dk} \mathbf{W}_{dl} \frac{\partial f_r(\tilde{m}_{dn}, \tilde{v}_{dn}, \boldsymbol{\alpha})}{\partial \tilde{v}_{dn}}$$

M-Step:

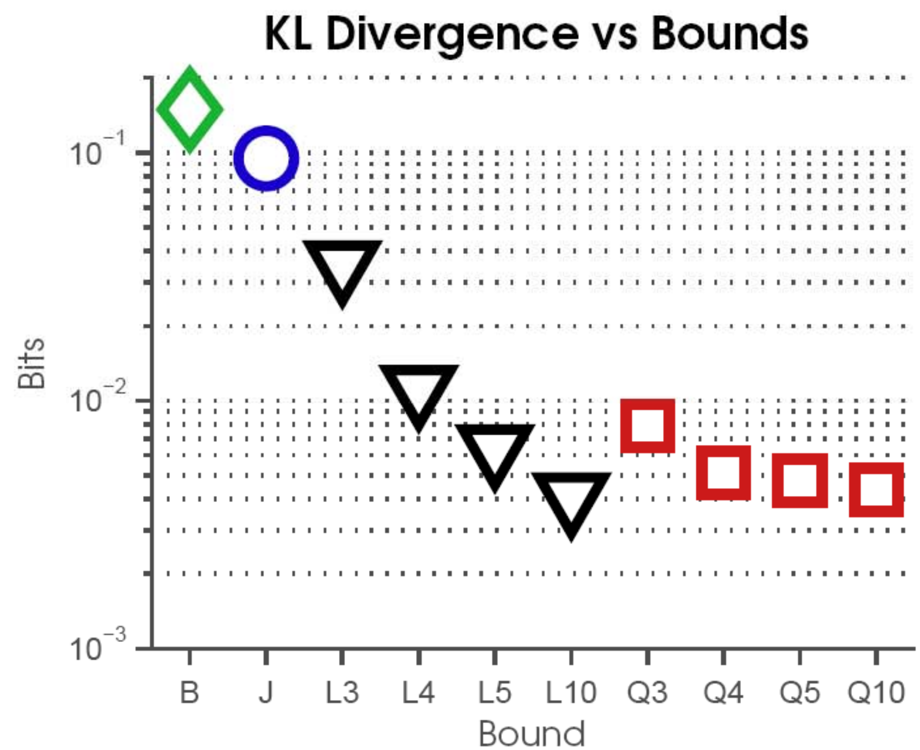
$$\boldsymbol{\mu} \leftarrow \frac{1}{N} \sum_{n=1}^{N} \mathbf{m}_n$$

$$\boldsymbol{\Sigma} \leftarrow \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{V}_n + (\mathbf{m}_n - \boldsymbol{\mu})(\mathbf{m}_n - \boldsymbol{\mu})^T \right)$$

$$\frac{\partial \mathcal{L}_{QJ}}{\partial \mathbf{W}_{dk}} \leftarrow \sum_{n=1}^{N} \left[\mathbf{m}_{kn} \left(y_{dn} - \sum_{r=1}^{R} \frac{\partial f_r(\tilde{m}_{dn}, \tilde{v}_{dn}, \boldsymbol{\alpha})}{\partial \tilde{m}_{dn}} \right) - \left(2 \sum_{l=1}^{K} \mathbf{V}_{kln} \mathbf{W}_{dk} \right) \sum_{r=1}^{R} \frac{\partial f_r(\tilde{m}_{dn}, \tilde{v}_{dn}, \boldsymbol{\alpha})}{\partial \tilde{v}_{dn}} \right]$$



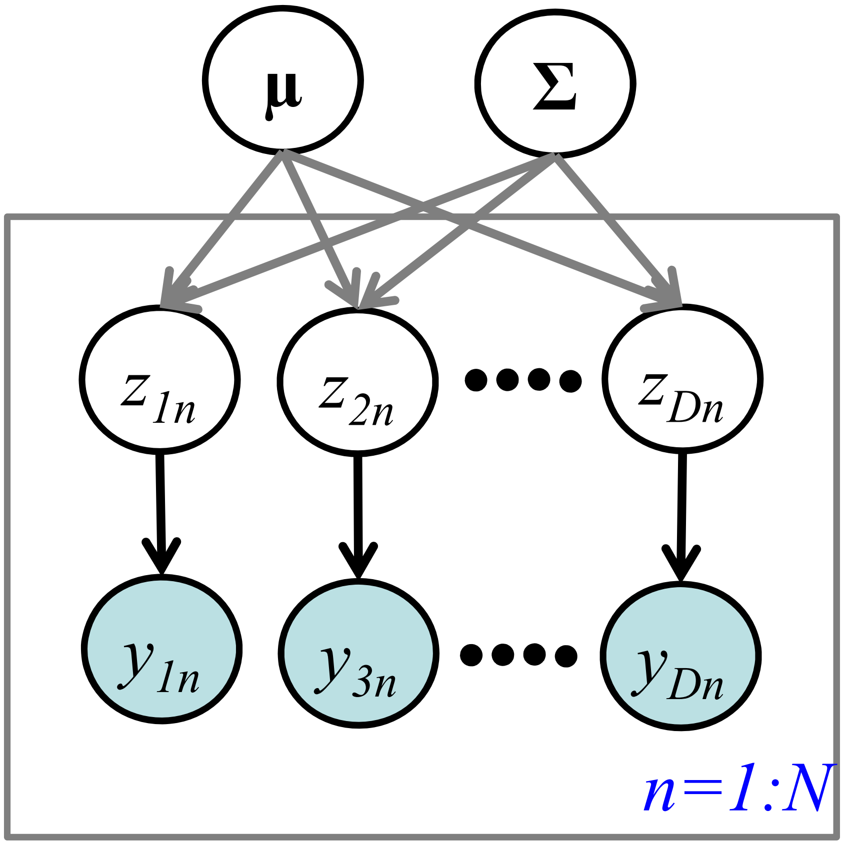
(b) 5D bLGGM Covariance

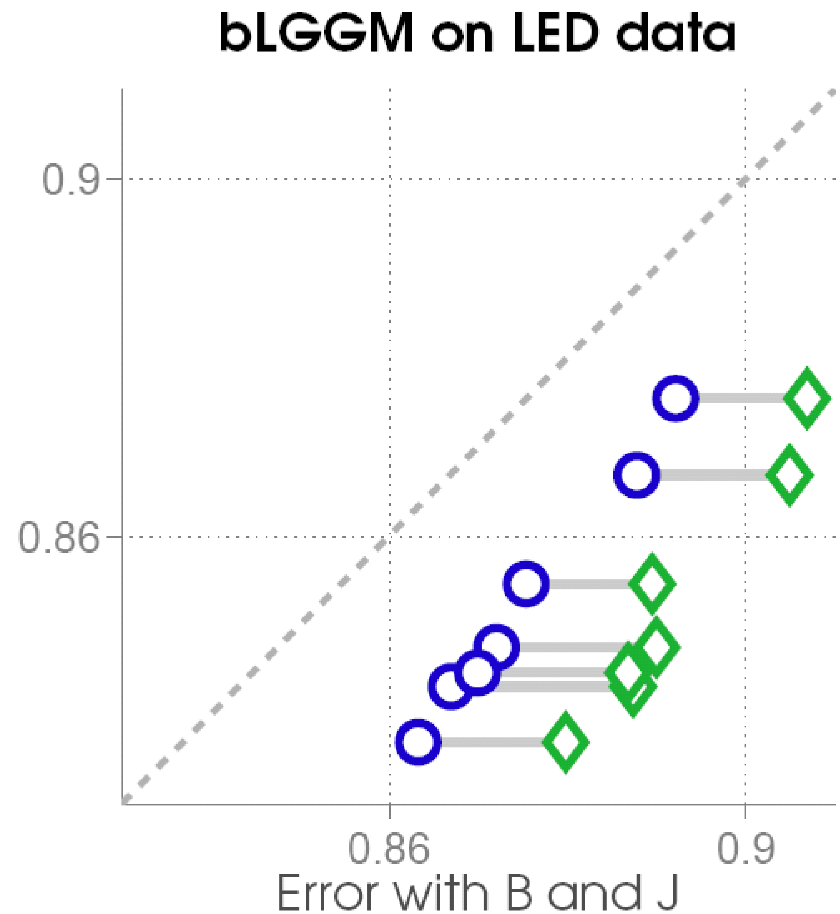


(c) 5D bLGGM KL Divergence

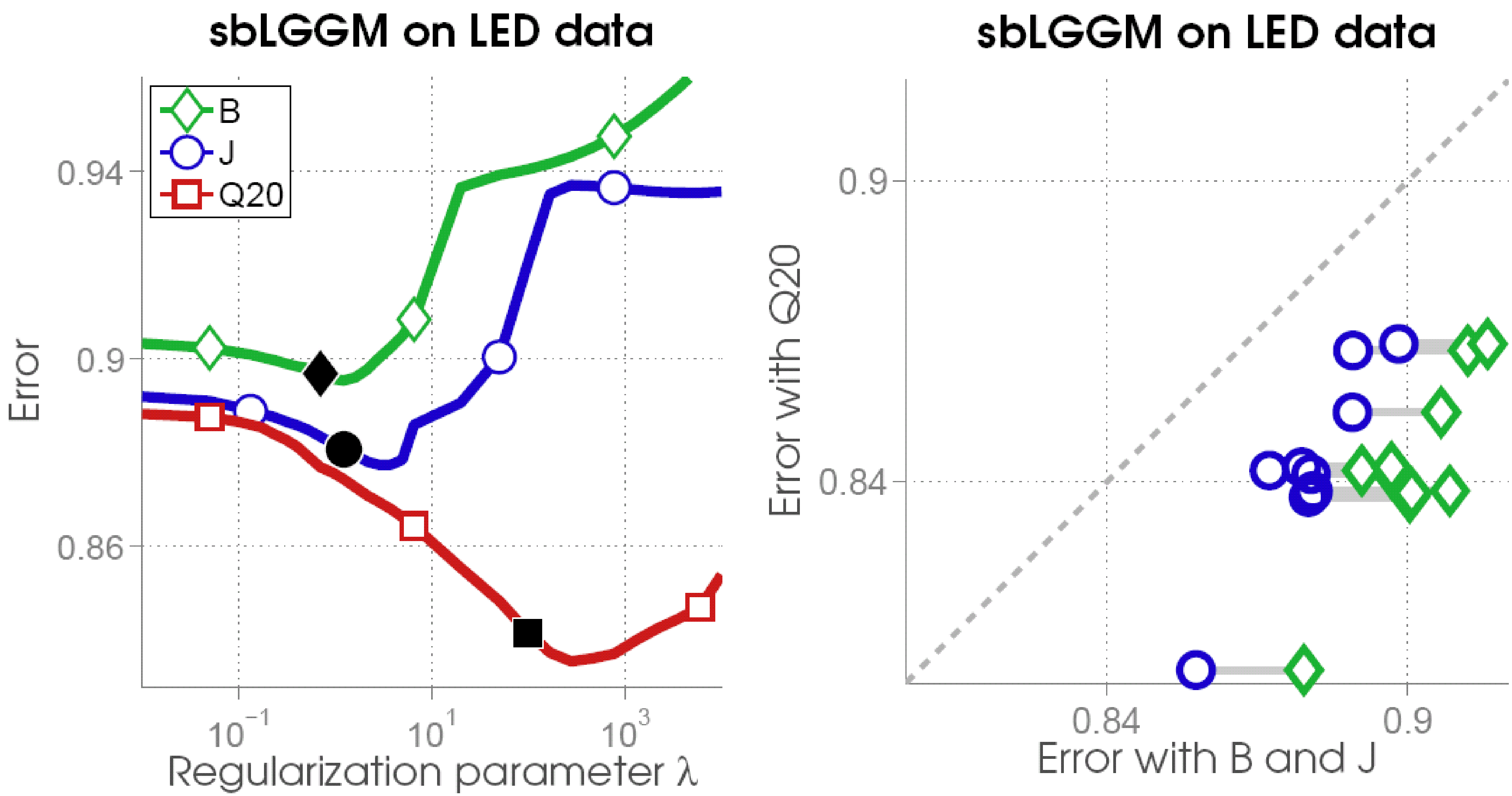
Latent Gaussian Graphical Model

LED dataset, 24 variables, N=2000

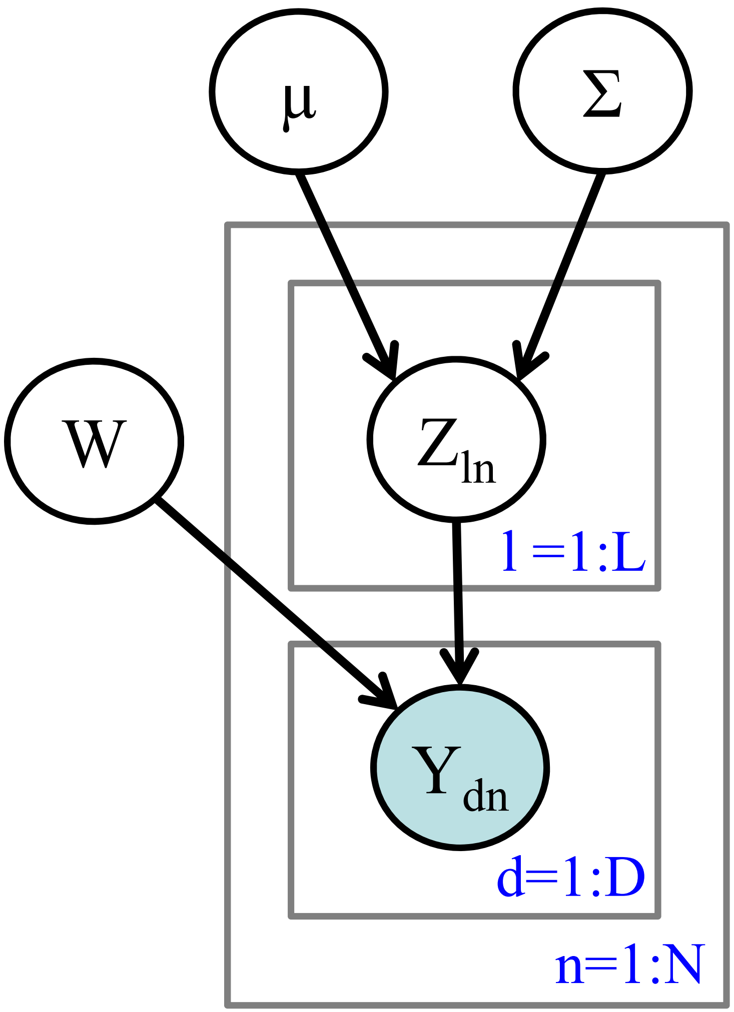




Sparse Version



Binary Latent Gaussian Models



$$p(\mathbf{z}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$p(y_d = 1) = \sigma(\mathbf{w}_d^T \mathbf{z})$$

$$\sigma(x) = (1 + \exp(x))^{-1}$$

We are interested in maximum likelihood estimate of parameters

$$\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{W}\}$$