Piecewise Bounds for Estimating Bernoulli-Logistic Latent Gaussian Models

Mohammad Emtiyaz Khan

Joint work with Benjamin Marlin, and Kevin Murphy

University of British Columbia

June 29, 2011

Modeling Binary Data

Main Topic of our paper

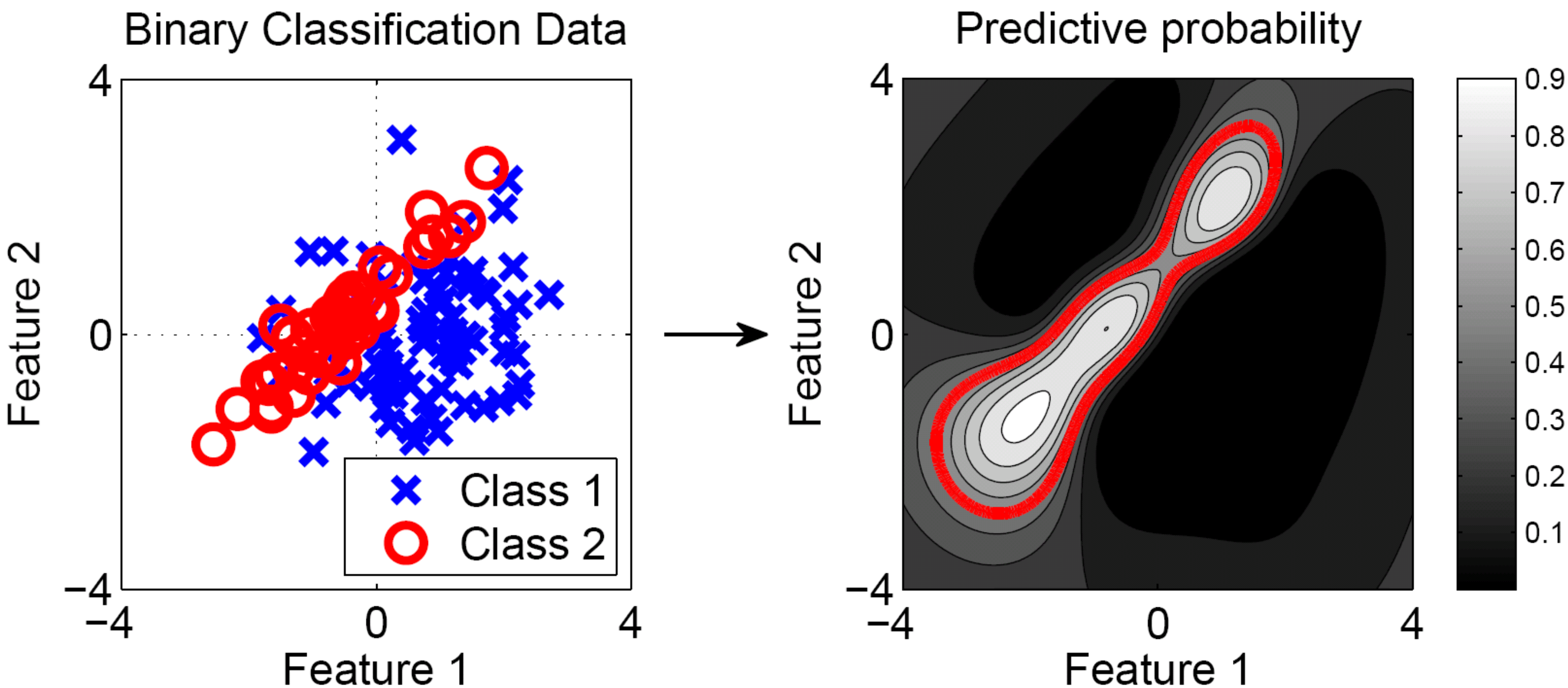
Bernoulli-Logistic Latent Gaussian Models (bLGMs)

Image from http://thenextweb.com/in/2011/06/06/india-to-join-the-open-data-revolution-in-july/

bLGMs - Classification Models

Bayesian Logistic Regression and Gaussian Process Classification

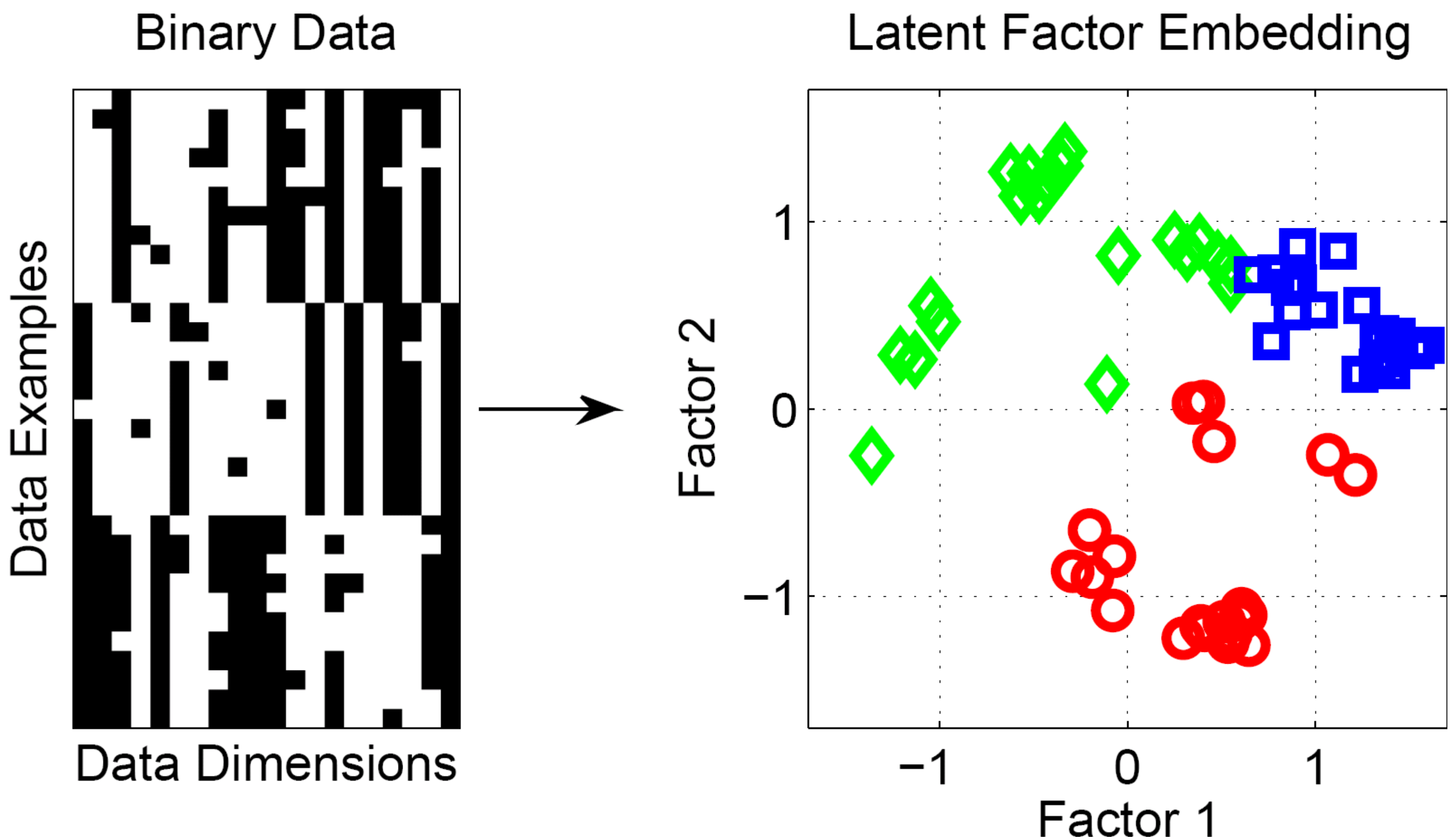
(Jaakkola and Jordan 1996, Rasmussen 2004, Gibbs and Mackay 2000, Kuss and Rasmussen 2006, Nickisch and Rasmussen 2008, Kim and Ghahramani, 2003).



Figures reproduced using GPML toolbox

bLGMs - Latent Factor Models

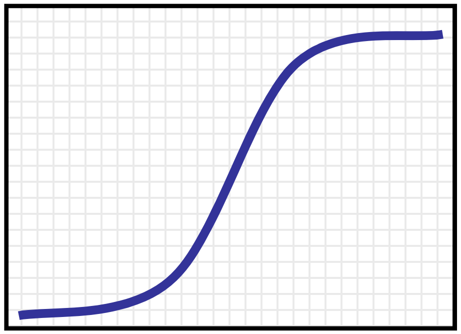
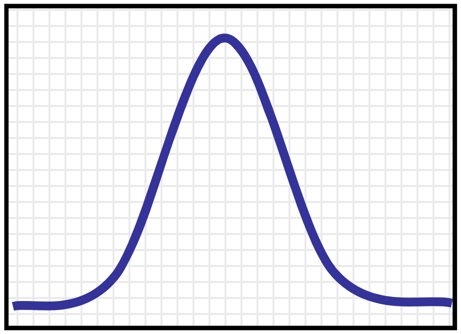
Probabilistic PCA and Factor Analysis models (Tipping 1999, Collins, Dasgupta and Schapire 2001, Mohammed, Heller, and Ghahramani 2008, Girolami 2001, Yu and Tresp 2004).



Parameter Learning is Intractable

Logistic Likelihood is not conjugate to the Gaussian prior.

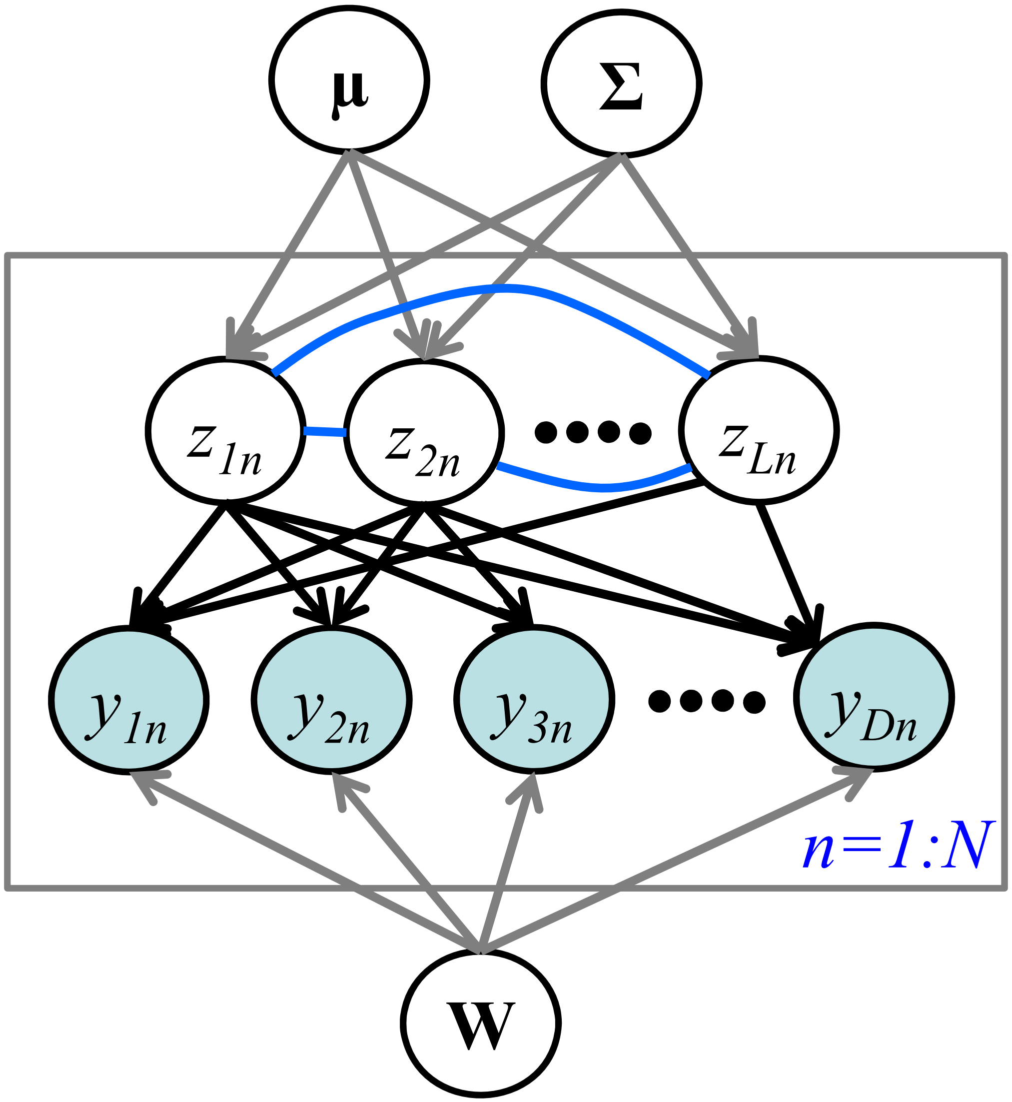
$(1+\exp(z))^{-1}$

$\int$  x  $d\mathbf{z}$

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

Learning in bLGMs

Bernoulli-Logistic Latent Gaussian Models



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

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

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

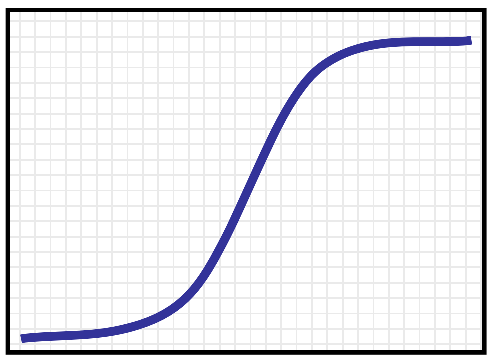
Parameter Set

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

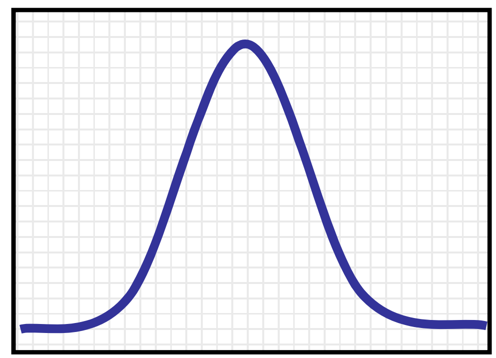
Learning Parameters of bLGMs

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

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



x



$$d\mathbf{z}$$

Variational Lower Bound (Jensen's)

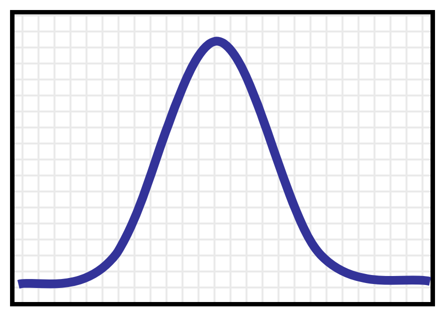
$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) = \log \int \prod_{d=1}^{D} p(y_d|\mathbf{z},\boldsymbol{\theta})\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{z}$$

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

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

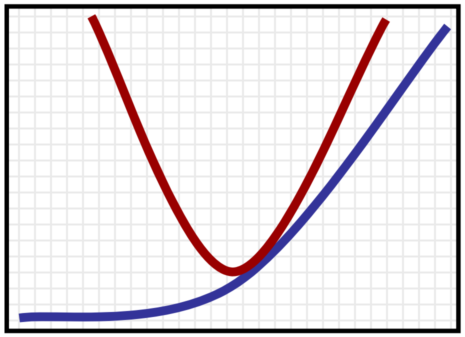
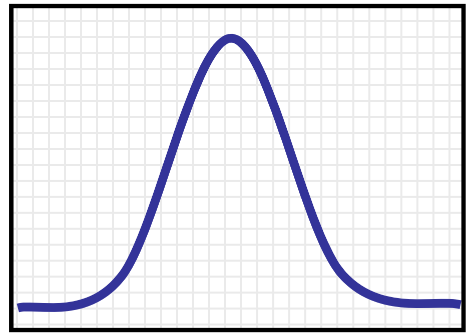
$$= \max_{\mathbf{m},\mathbf{V}} \sum_{d=1}^{D} \int [-\log(1+e^{x_d})]\mathcal{N}(\tilde{m}_d,\tilde{v}_d)dx_d$$

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

$-\int$  x  dx_d

Quadratic Bounds

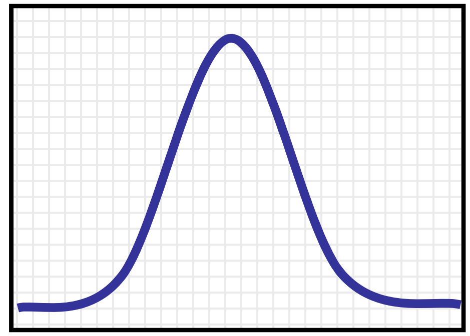
$\log(1+\exp(x_d))$ $\mathcal{N}(\tilde{m}_d, \tilde{v}_d)$

$-\int$  x  dx_d

- Bohning's bound (Bohning, 1992)

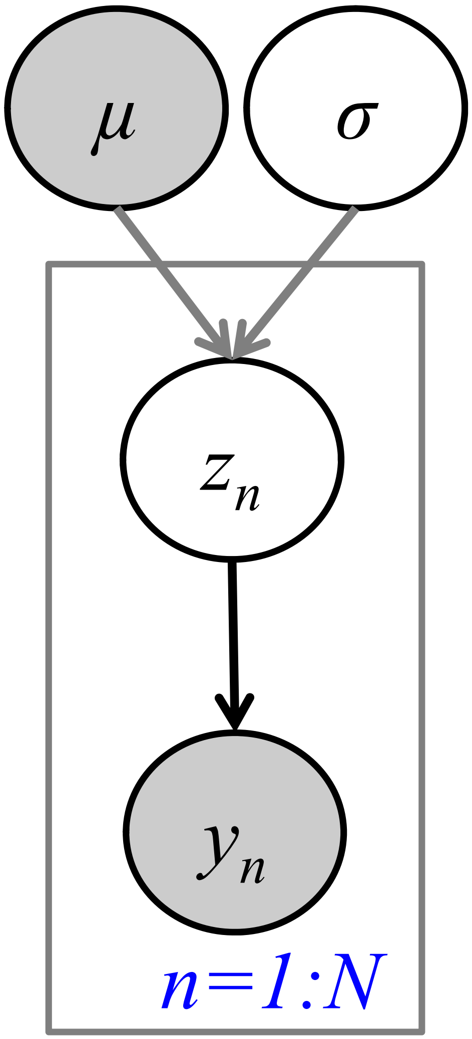
Quadratic Bounds

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

$-\int$  x  dx_d

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

Problems with Quadratic Bounds



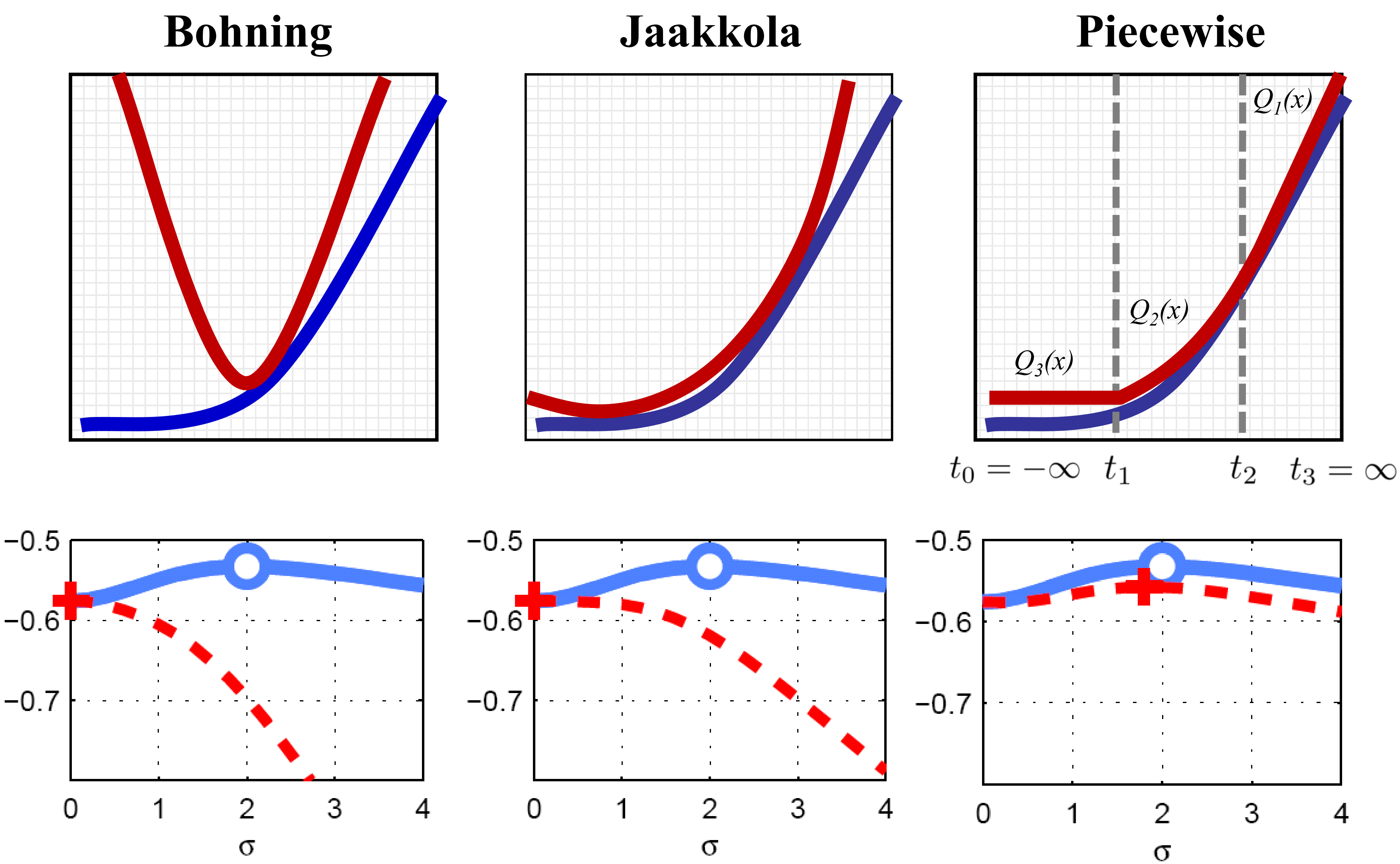
1-D example with $\mu = 2$, $\sigma = 2$

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

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

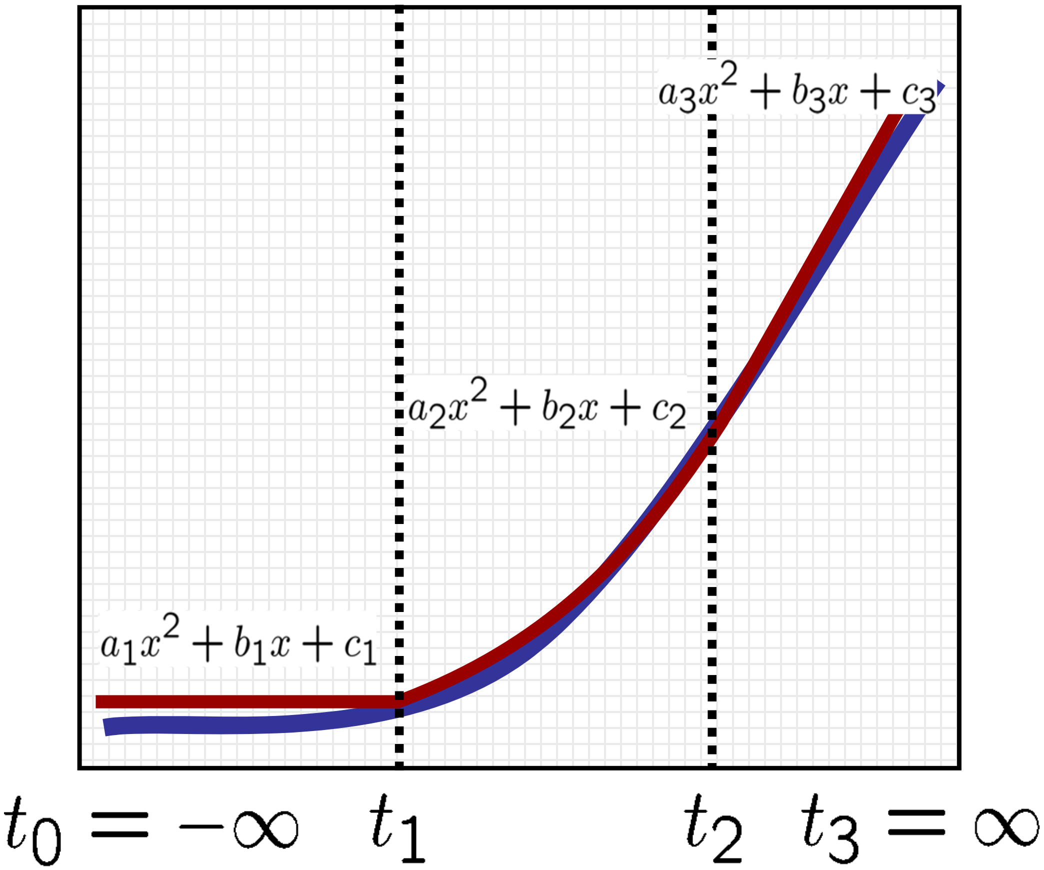
As this is a 1-D problem, we can compute lower bounds without Jensen's inequality. So plots that follow have errors only due to error in bounds.

Problems with Quadratic Bounds



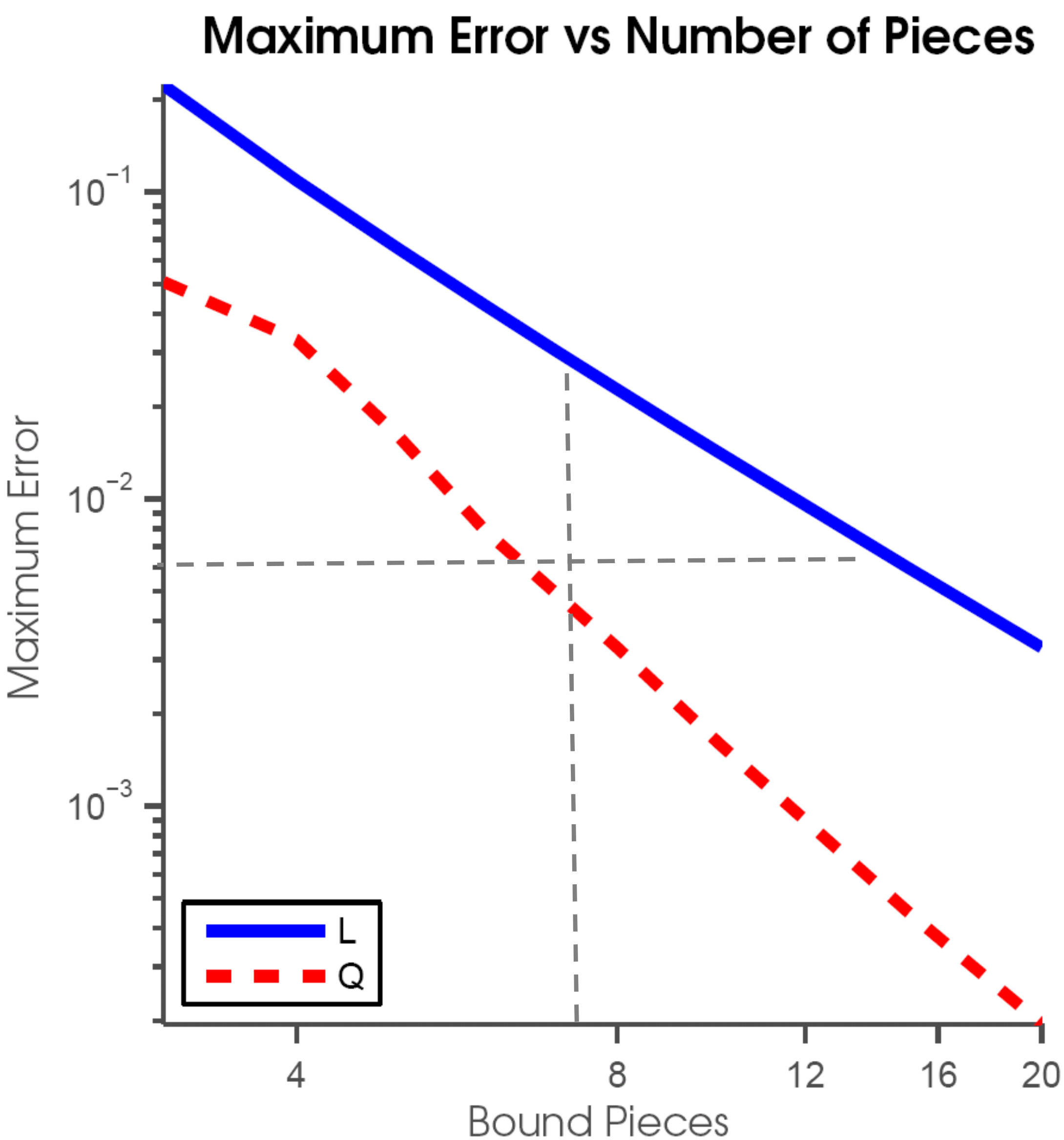
Piecewise Bounds

Finding Piecewise Bounds



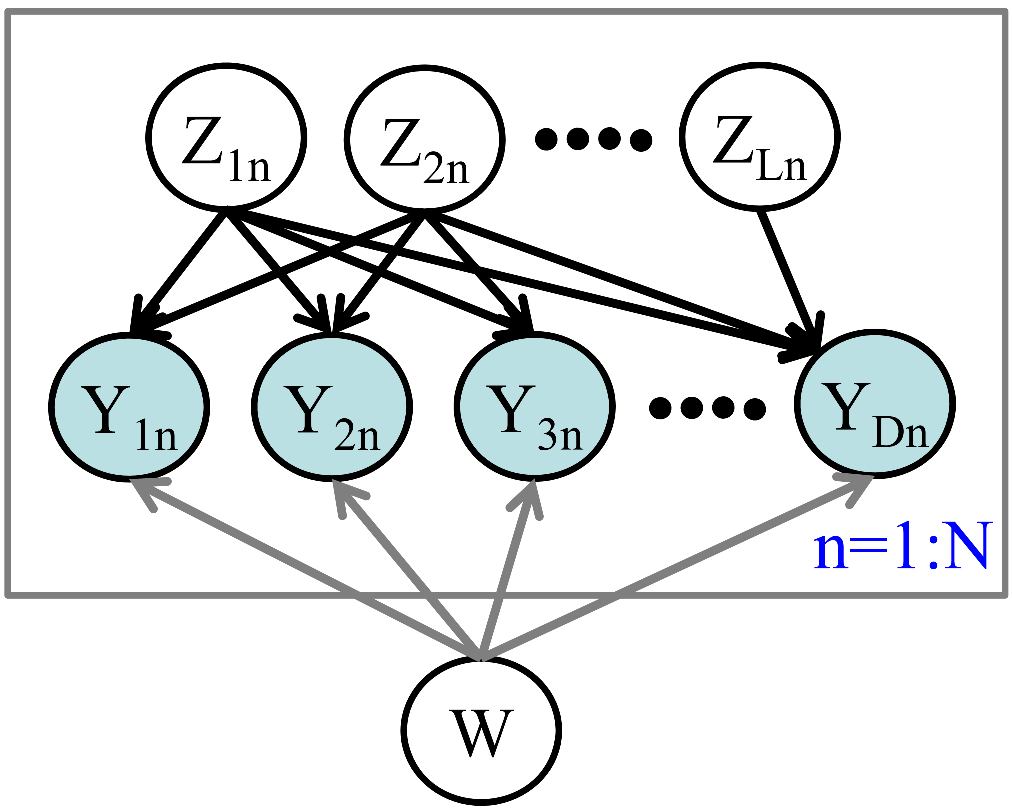
- Find Cut points, and parameters of each pieces by minimizing maximum error.
- Linear pieces (Hsiung, Kim and Boyd, 2008)
- Quadratic Pieces (Nelder-Mead method)
- Fixed Piecewise Bounds!
- Increase accuracy by increasing number of pieces.

Linear Vs Quadratic



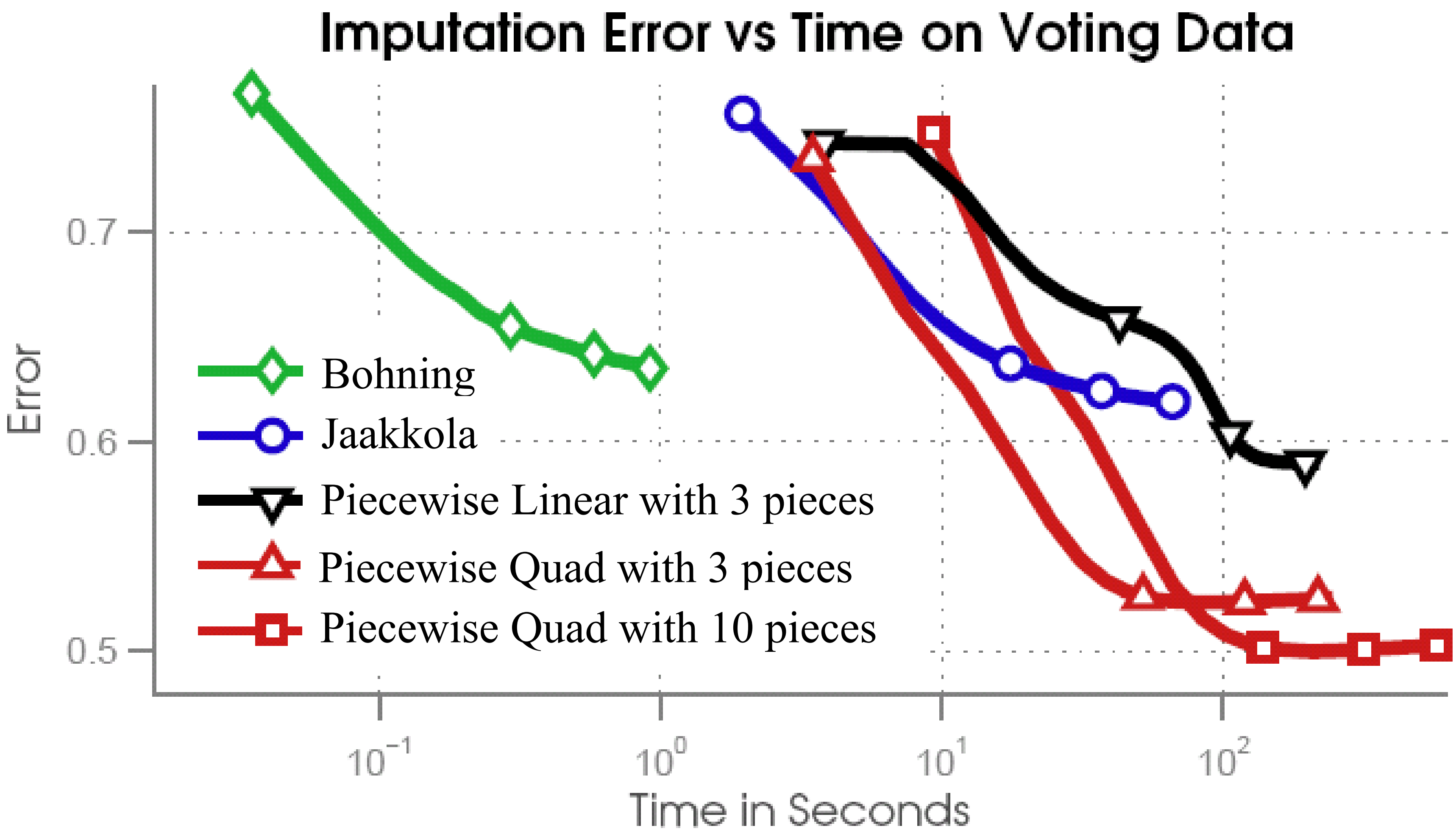
Results

Binary Factor Analysis (bFA)

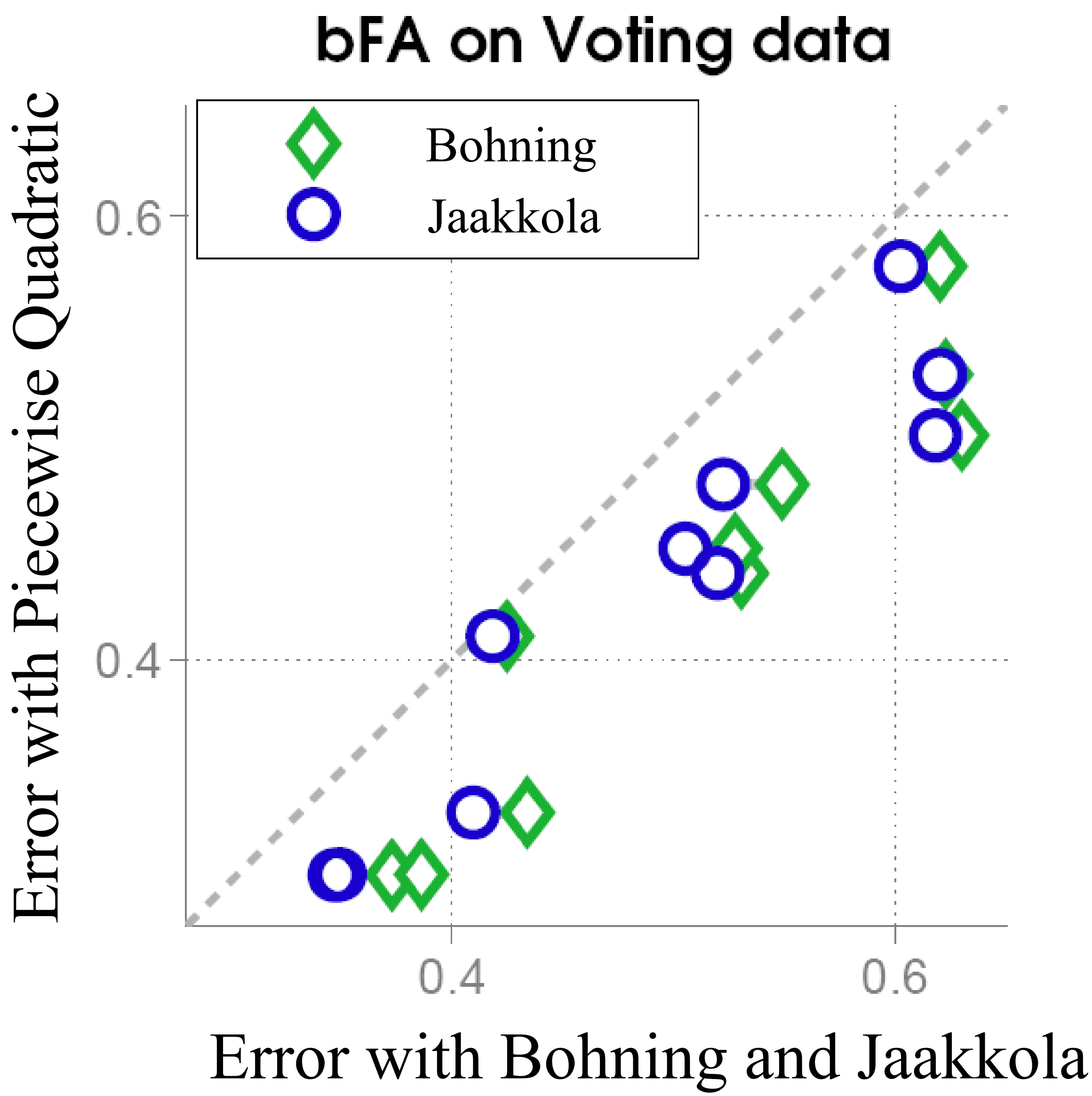


- UCI voting dataset with D=15, N=435.
- Train-test split 80-20%
- Compare cross-entropy error on missing value prediction on test data.

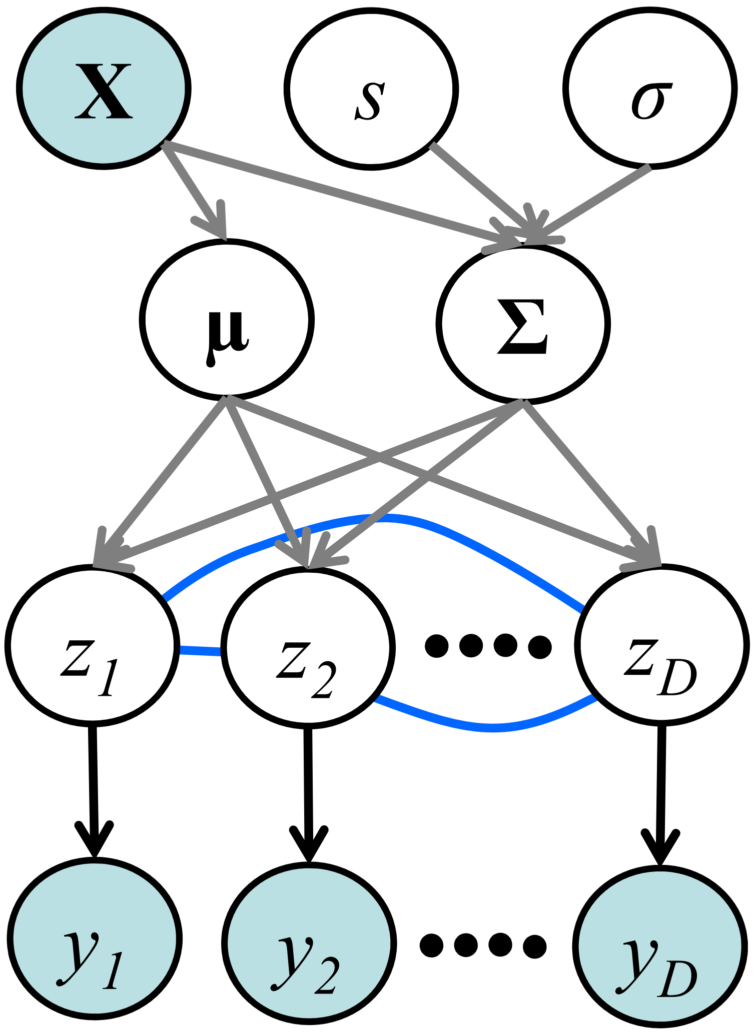
bFA – Error vs Time



bFA – Error Across Splits



Gaussian Process Classification

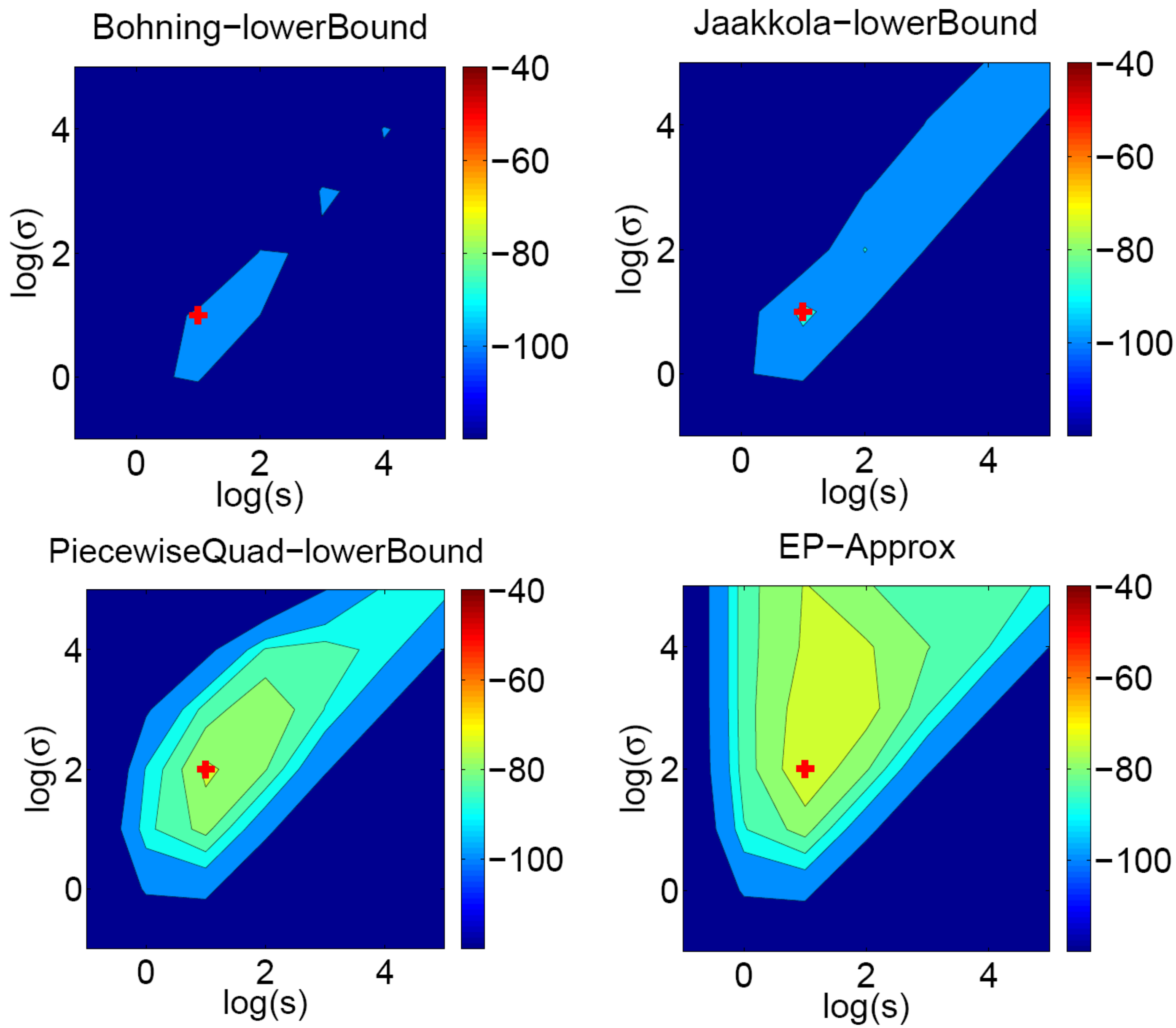


- We repeat the experiments described in Kuss and Rasmussen, 2006
- We set μ =0 and squared exponential Kernel

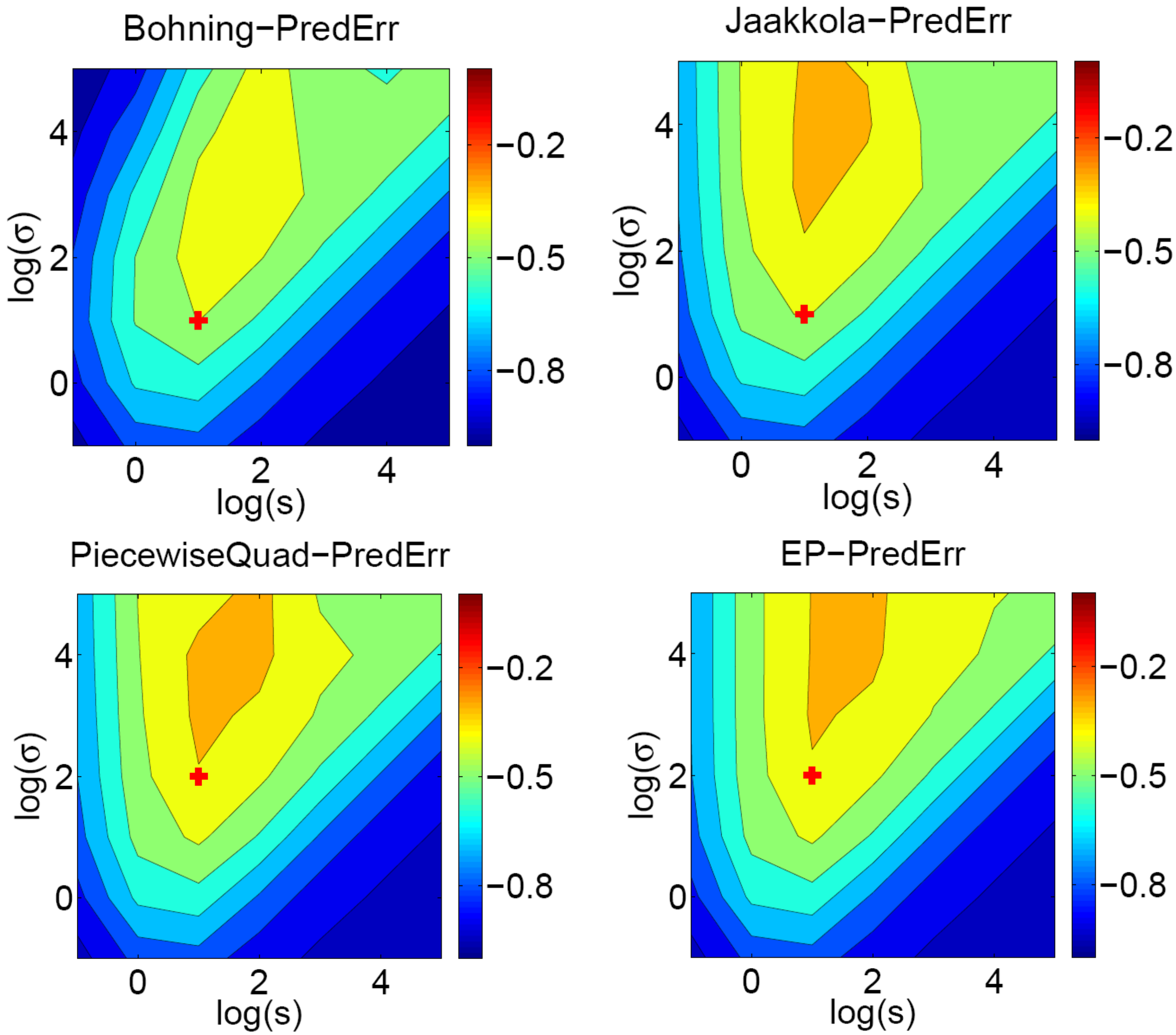
$\Sigma_{ij} = \sigma\ exp[(x_i\text{-}x_j)\text{^}2/s]$

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

GP – Marginal Likelihood



GP – Prediction Error



EP vs Variational

- We see that the variational approach underestimates the marginal likelihood in some regions of parameter space.
- However, both methods give comparable results for prediction error.
- In general, the variational EM algorithm for parameter learning is guaranteed to converge when appropriate numerical methods are used,
- Nickisch and Rasmussen (2008) describe the variational approach as more principled than EP.

Conclusions

Conclusions

- Fixed piecewise bounds can give a significant improvement in estimation and prediction accuracy relative to variational quadratic bounds.
- We can drive the error in the logistic-log-partition bound to zero by letting the number of pieces increase.
- This increase in accuracy comes with a corresponding increase in computation time.
- Unlike many other frameworks, we have a very fine grained control over the speed-accuracy trade-off through controlling the number of pieces in the bound.

Thank You

Piecewise-Bounds: Optimization Problem

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

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

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

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

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

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

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

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

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

Algorithm 1 bLGM Generalized EM Algorithm

E-Step:

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

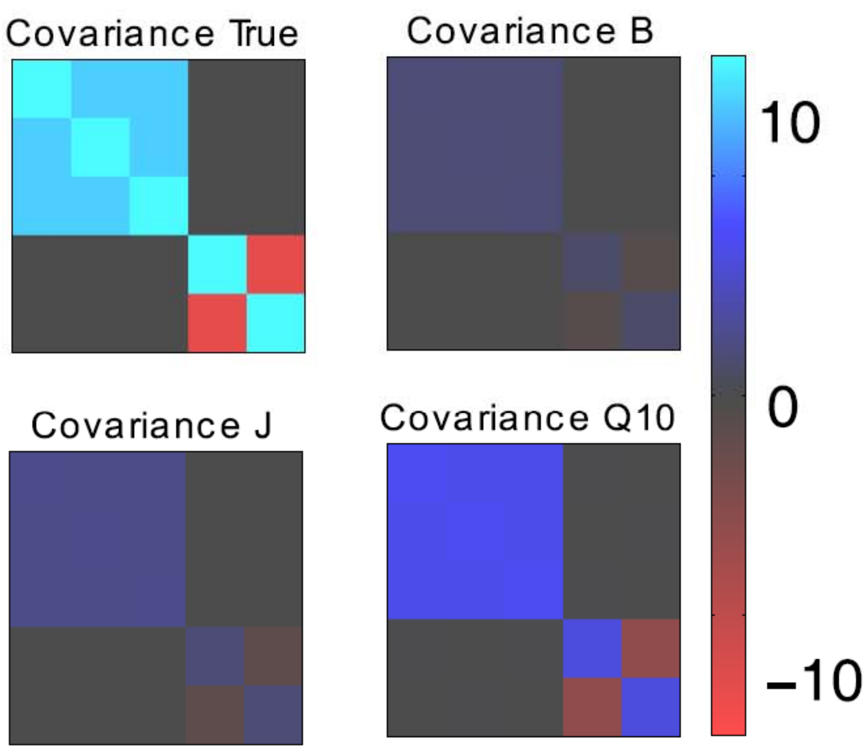
$$\frac{\partial \mathcal{L}_{QJ}}{\partial \mathbf{V}_{kl}} \leftarrow \frac{1}{2}(\boldsymbol{\Sigma}^{-1})_{kl} - \frac{1}{2}(\mathbf{V}_n^{-1})_{kl} - \sum_{r=1}^{R}\sum_{d=1}^{D}\mathbf{W}_{dk}\mathbf{W}_{dl}\frac{\partial f_r(\tilde{m}_{dn}, \tilde{v}_{dn}, \boldsymbol{\alpha})}{\partial \tilde{v}_{dn}}$$

M-Step:

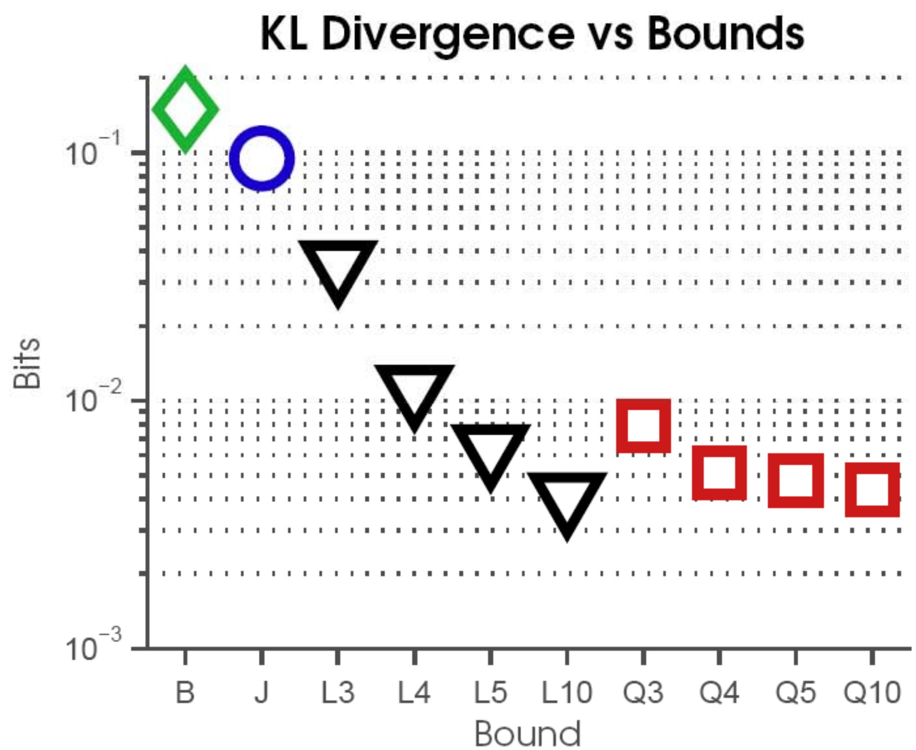
$$\boldsymbol{\mu} \leftarrow \frac{1}{N}\sum_{n=1}^{N}\mathbf{m}_n$$

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

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



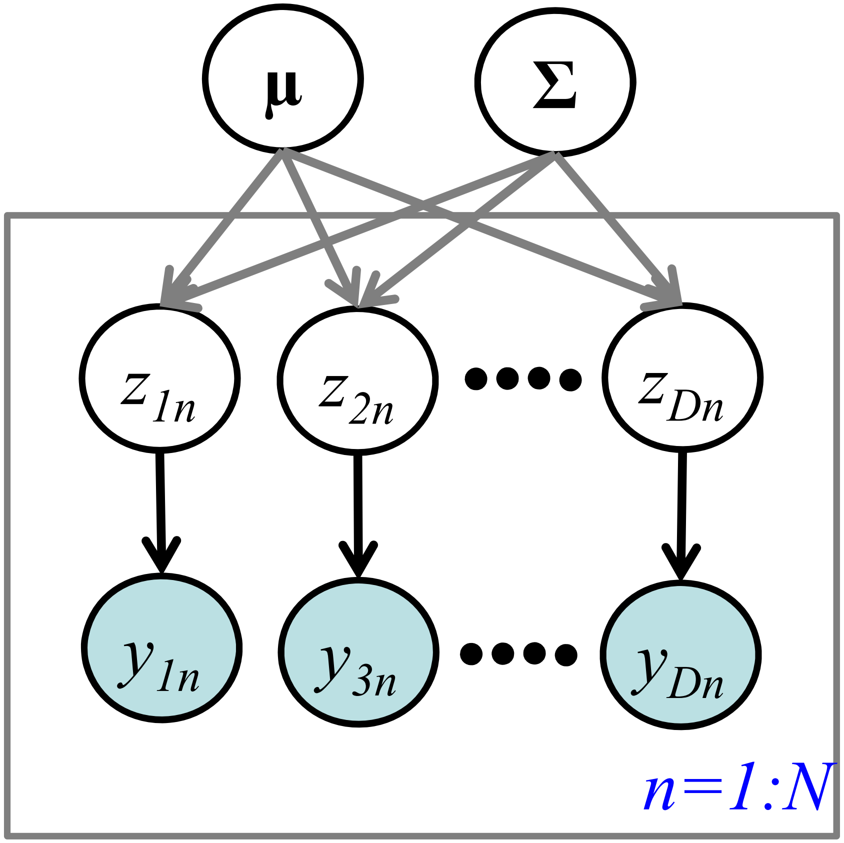
(b) 5D bLGGM Covariance

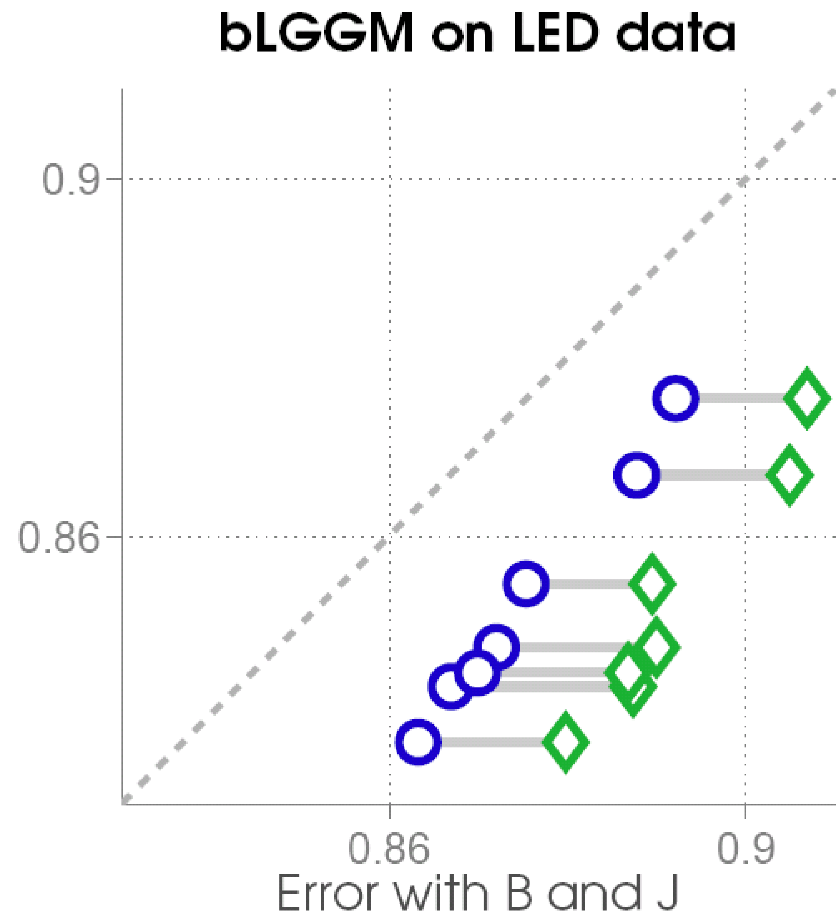


(c) 5D bLGGM KL Divergence

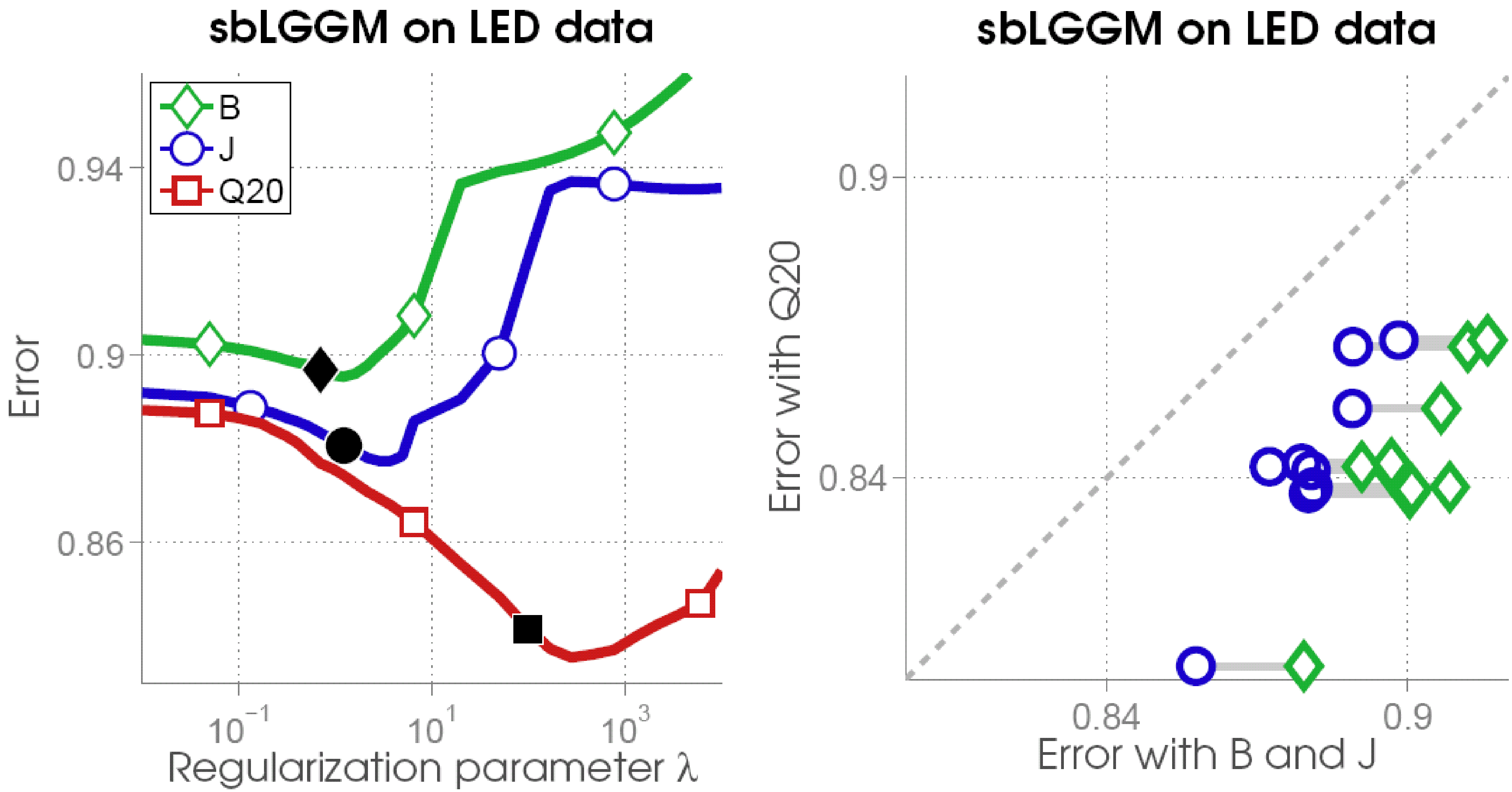
Latent Gaussian Graphical Model

LED dataset, 24 variables, N=2000

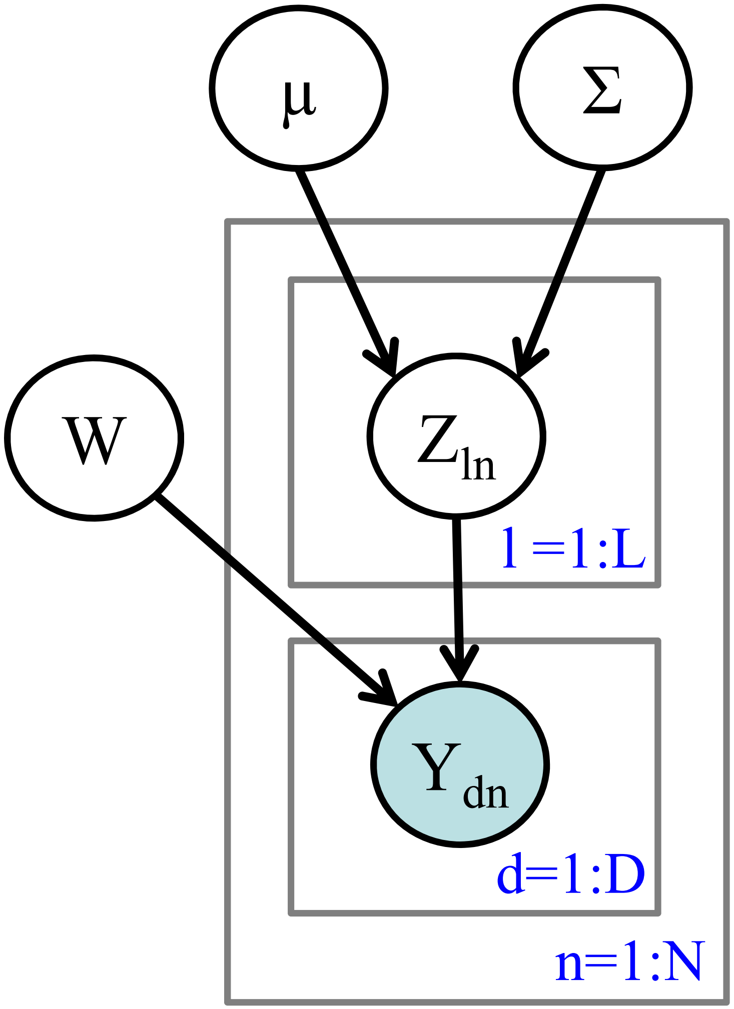




Sparse Version



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

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