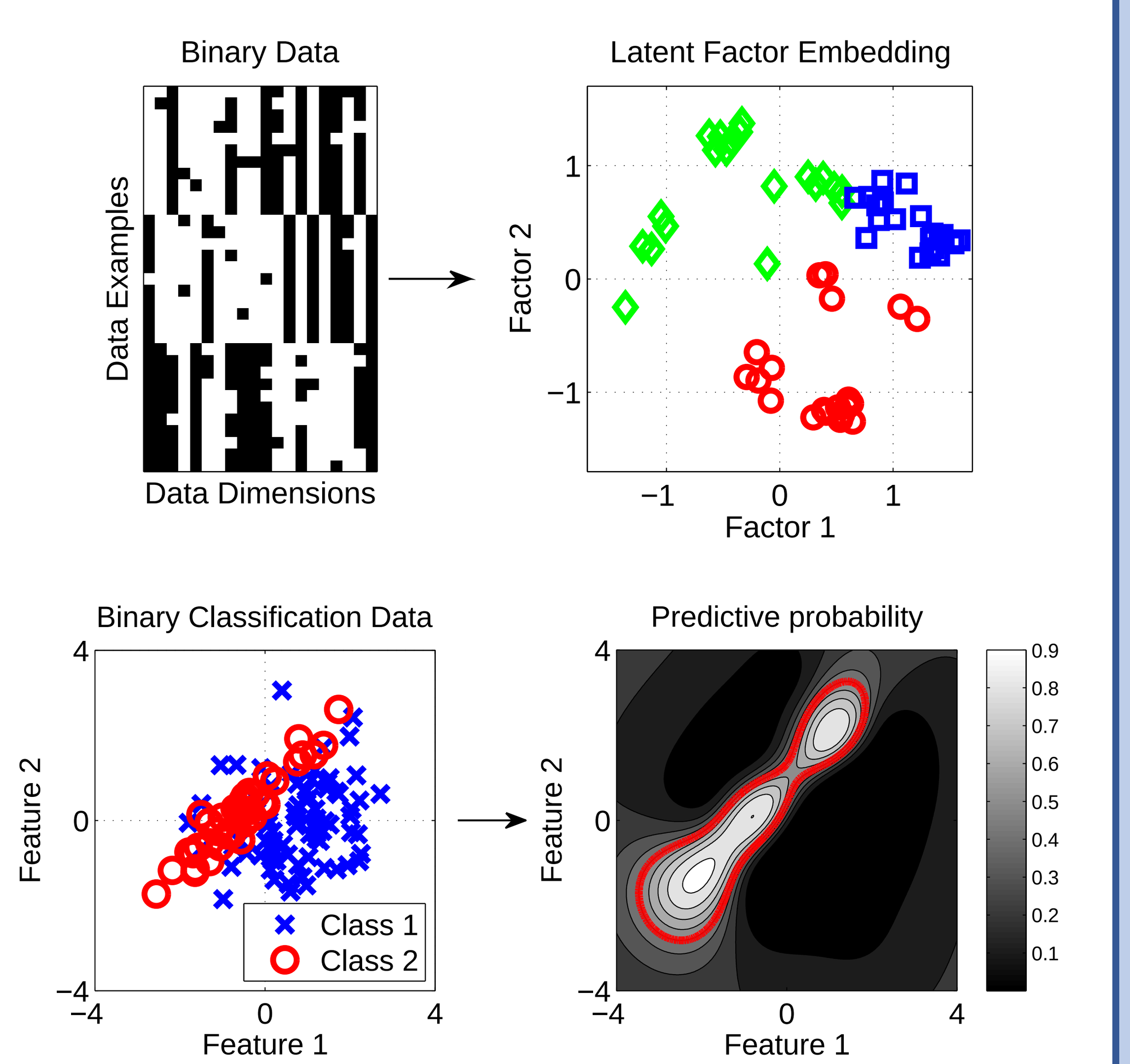
Piecewise Bounds for Estimating Bernoulli-Logistic Latent Gaussian Models

Benjamin M. Marlin, Mohammad Emtiyaz Khan, and Kevin P. Murphy
Department of Computer Science, University of British Columbia

Introduction

Motivation: Bernoulli-logistic Latent Gaussian Models (bLGMs) are an important class of probabilistic models that includes models such as binary factor analysis and binary probabilistic principal components analysis, as well as Bayesian logistic regression and Gaussian process classification.

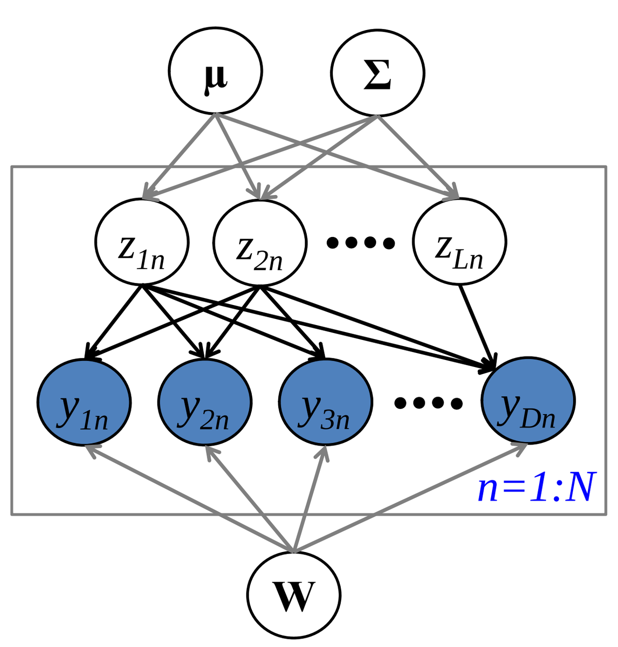


Problem: Parameter learning is difficult in bLGMs due to an intractable integral in the marginal likelihood.

Solution: We propose to solve the intractable integral through the application of piecewise linear and quadratic bounds to the logistic-log-partition function. Piecewise bounds have the important property that their maximum error is bounded and can be driven to zero by increasing the number of pieces. Resulting algorithms achieve significant improvements over the existing variational quadratic bounds.

Bernoulli-Logistic LGMs

Our model uses latent Gaussian variables to model the distribution of binary observations. To obtain n'th data vector, we first sample a latent Gaussian vector $\mathbf{z}_n \in \mathbb{R}^L$ and then take a linear combination of $\mathbf{z}_n$ to obtain the parameter η_{dn} for the d'th Bernoulli-logistic distribution. Distribution of the binary vector $\mathbf{y}_n \in \{0,1\}^D$ is the product of individual Bernoulli-logistic distribution. Our goal is to learn the maximum likelihood estimate of parameter θ given $\mathbf{y}_1, \ldots, \mathbf{y}_N$.



$$p(\mathbf{z}_n|\boldsymbol{\theta}) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$\eta_{dn} = \mathbf{W}_d \mathbf{z}_n + b_d$$
$$p(\mathbf{y}_n|\mathbf{z}_n, \boldsymbol{\theta}) = \prod_{d=1}^{D_d} \frac{\exp(y_{dn}\eta_{dn})}{1+\exp(\eta_{dn})}$$
$$\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{W}, \mathbf{b}\}$$

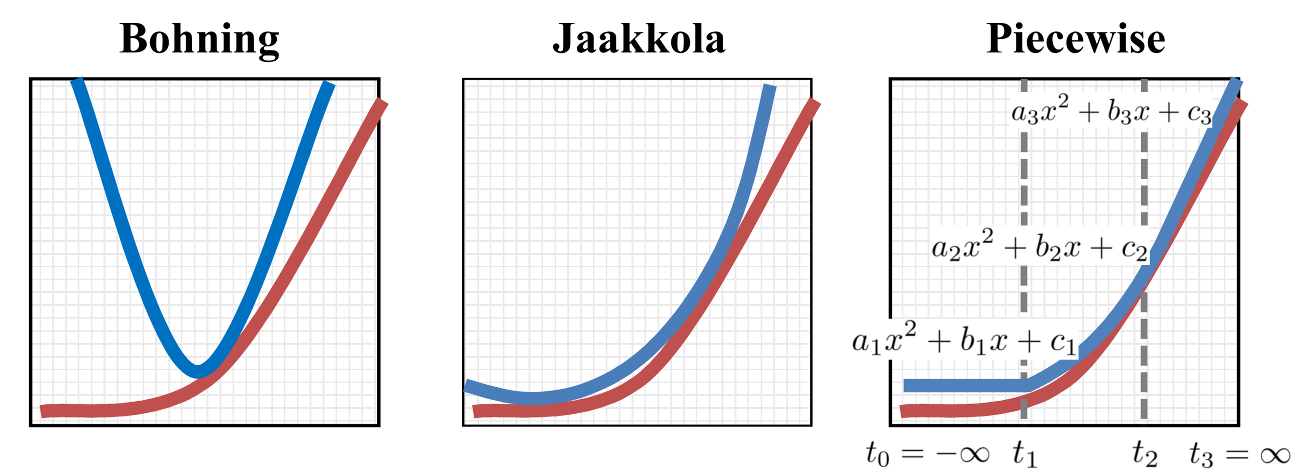
Parameter Learning

Variational Lower Bound to the Marginal Likelihood: Computation of the marginal likelihood is intractable as the Bernoulli-logistic likelihood is not conjugate to the Gaussian prior. Using Jensen's inequality, we can obtain a lower bound to the marginal likelihood, which can be computed as sum of many one dimensional integrals.

$$l(\boldsymbol{\theta}) = \log \int \prod_{d=1}^{D} p(y_d|\mathbf{z}, \boldsymbol{\theta})\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\mathbf{z} = \log \int \frac{\prod_{d=1}^{D} p(y_d|\mathbf{z}, \boldsymbol{\theta})\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V})}\mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V})d\mathbf{z}$$
$$\geq \sum_{d=1}^{D} \int \log p(y_d|\mathbf{z}, \boldsymbol{\theta})\mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{V})d\mathbf{z} - KL\left[\mathcal{N}(\mathbf{m}, \mathbf{V})||\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\right]$$
$$\underline{l}(\boldsymbol{\theta}) = \max_{\mathbf{m},\mathbf{V}} \sum_{d=1}^{D} -\int \log(1+e^{x_d})\mathcal{N}(\tilde{m}_d, \tilde{v}_d)dx_d + \text{Tractable terms in } \mathbf{m} \text{ and } \mathbf{V}$$

These one-dimensional integrals are still intractable but a tractable lower bound can be computed using bounds on $\log(1+\exp(x))$, the logistic-log-partition (LLP) function.

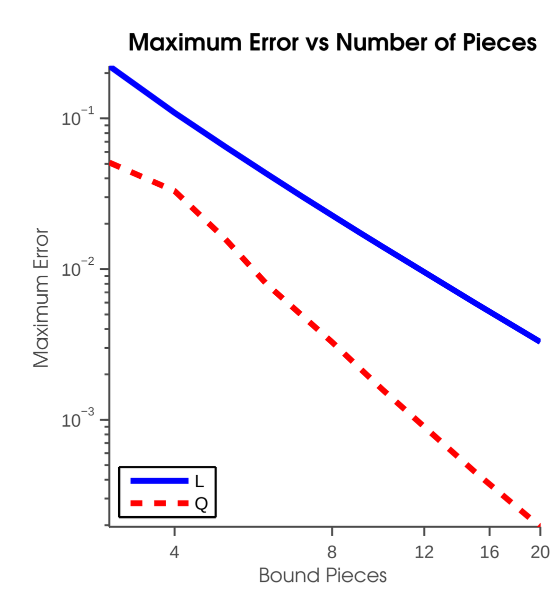
Quadratic Bounds vs Piecewise Bounds: Two existing quadratic bounds are due to Bohning and Jaakkola. Both of these have infinite maximum error, which can cause severe bias in the parameter estimates. Piecewise bounds have bounded error.



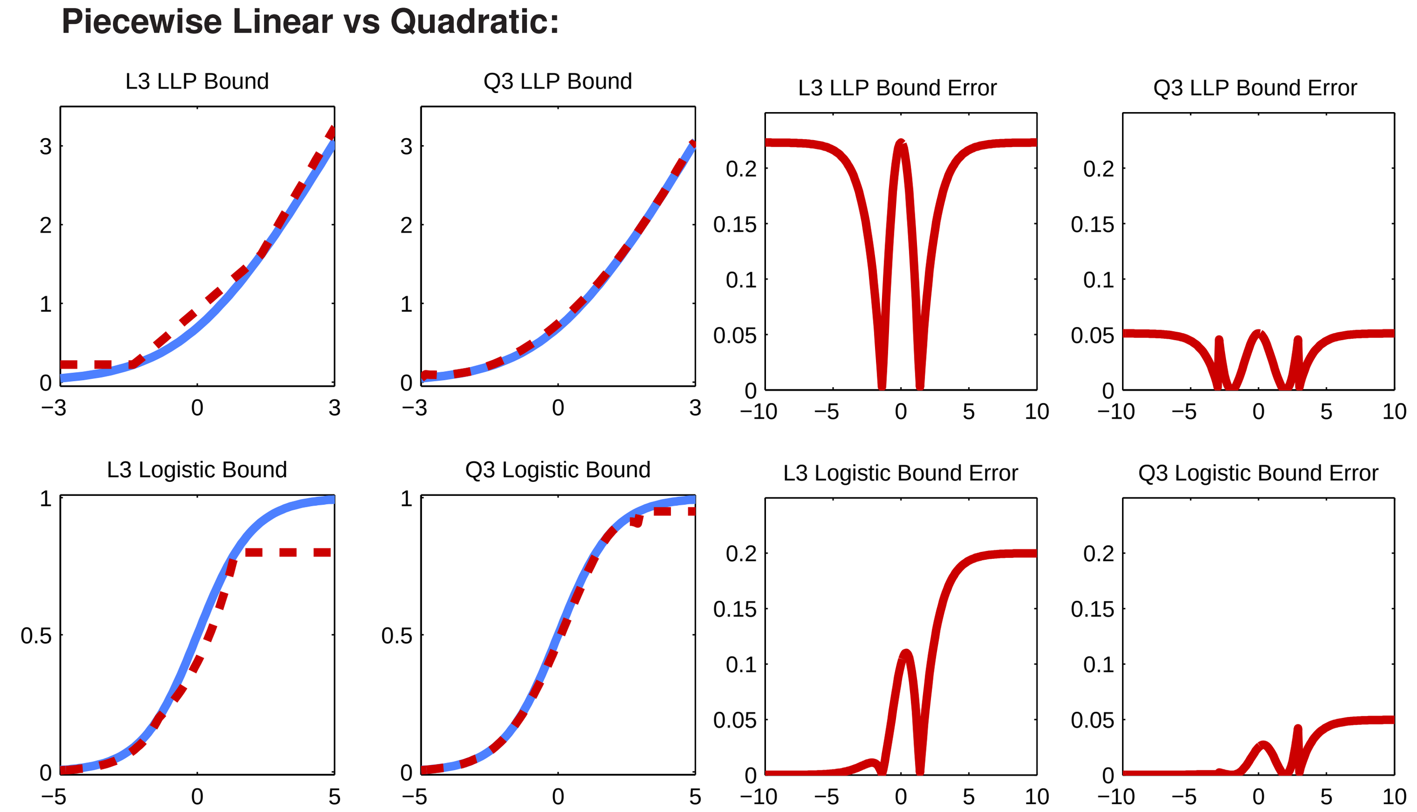
Piecewise Bounds

The Optimization Problem: We can find the parameters of piecewise bounds by minimizing the maximum error with constraints to enforce the upper bound property and to make sure that the intervals are ordered.

$$\min_{\alpha} \max_{r\in\{1,..,R\}} \max_{t_{r-1}\leq x<t_r} a_r x^2 + b_r x + c_r - \text{llp}(x)$$
$$a_r x^2 + b_r x + c_r - \text{llp}(x) \geq 0 \;\; \forall\, r, x \in [t_{r-1}, t_r]$$
$$t_r - t_{r-1} > 0 \qquad \forall r \in \{1, .., R\}$$
$$a_r \geq 0 \qquad \forall r \in \{1, .., R\}$$



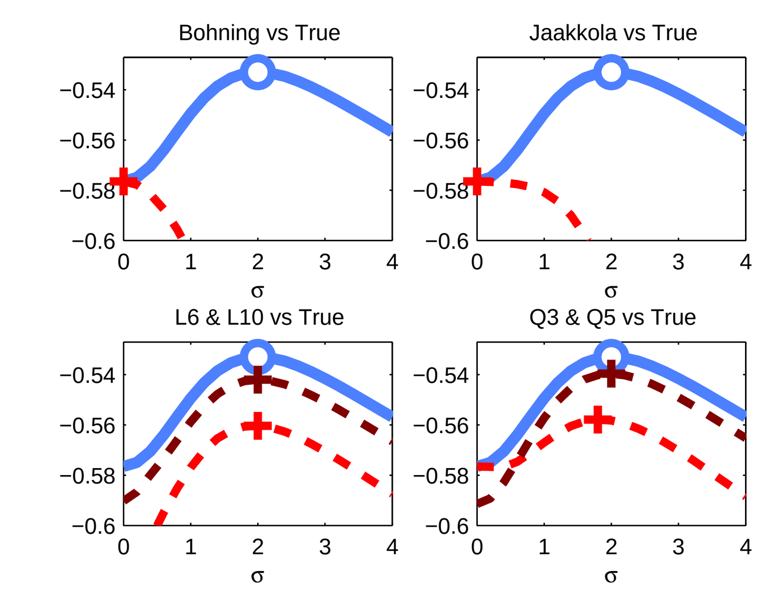
Piecewise Linear vs Quadratic:



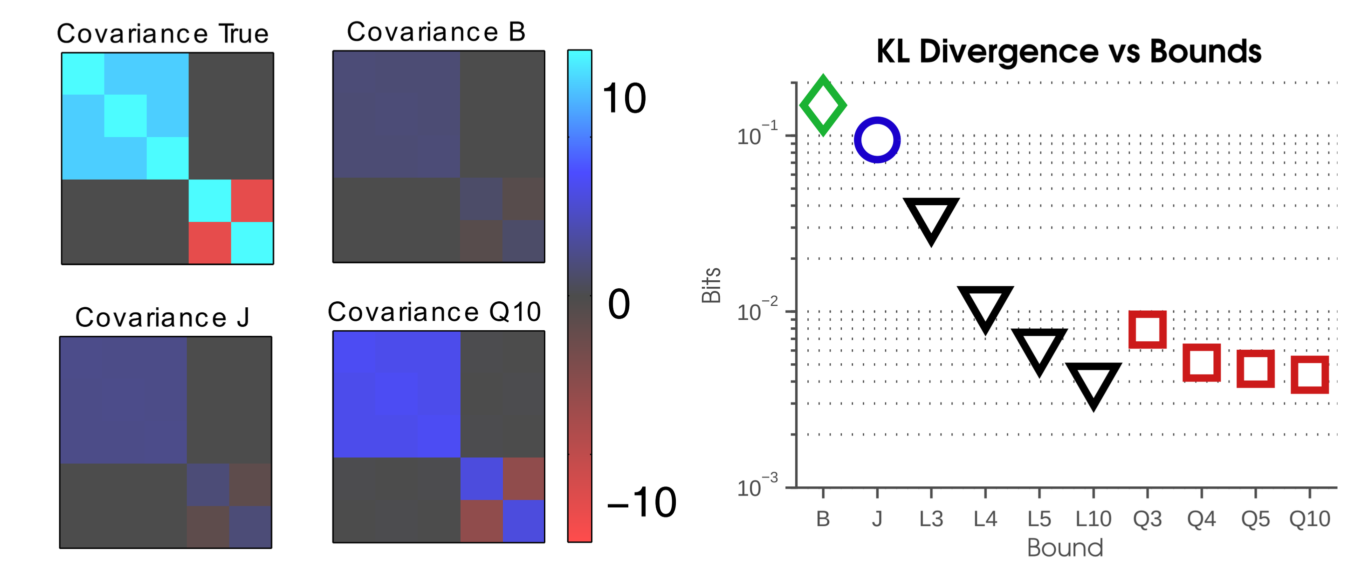
Results

In all experiments, we refer to the Bohning bound as 'B', the Jaakkola bound as 'J', piecewise linear (quadratic) bound with n pieces as 'Ln' ('Qn').

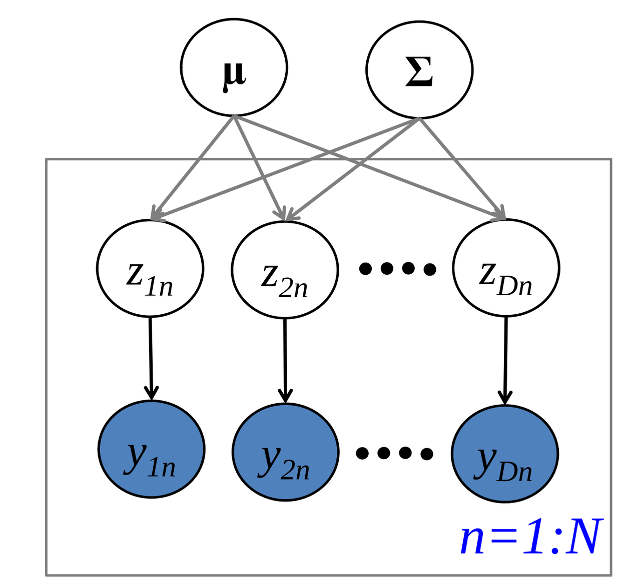
1-D Synthetic Data Experiments: On a 1-D bLGGM with $\mu = 2$ and $\sigma = 2$, we compare the true likelihood and the lower bounds.



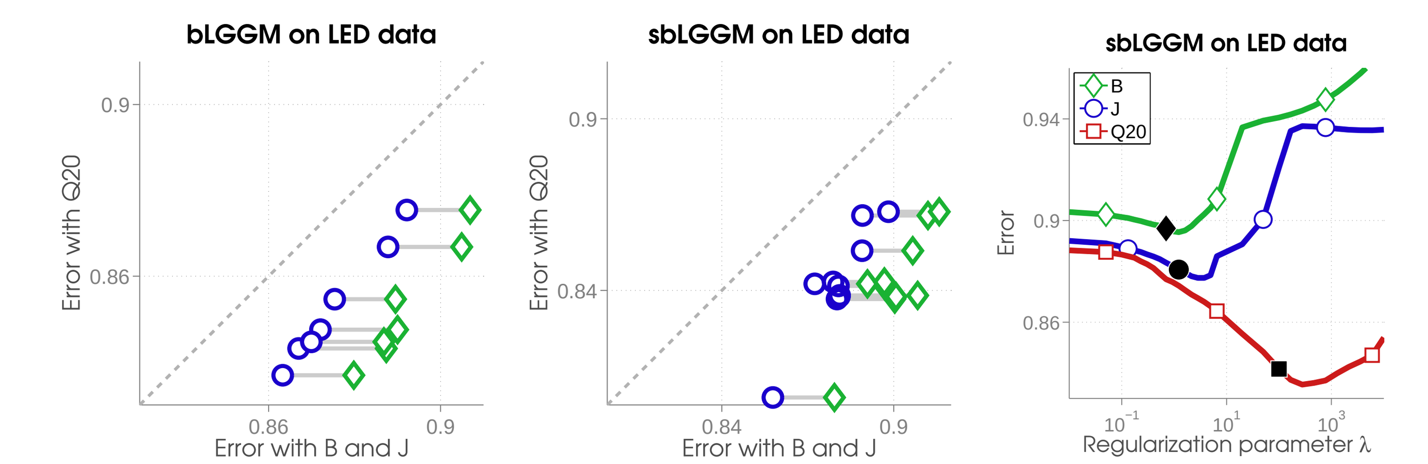
5-D Synthetic Data Experiments: On a 5-D bLGGM, we compare covariance estimates and KL divergences between the true discrete distribution and the estimated discrete distribution.



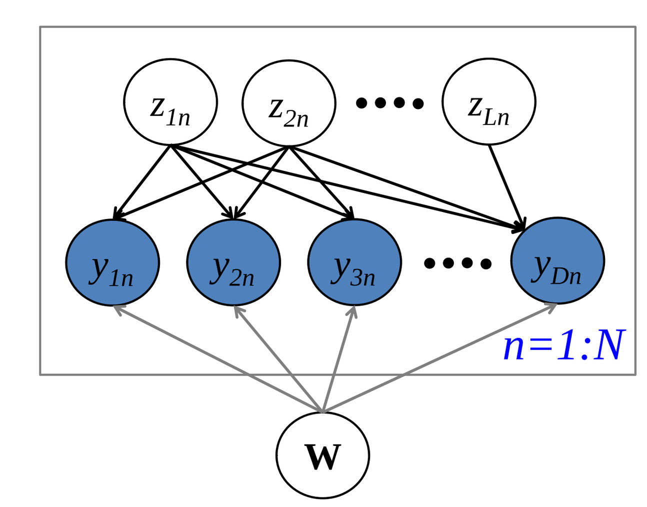
Binary Latent Gaussian Graphical Model



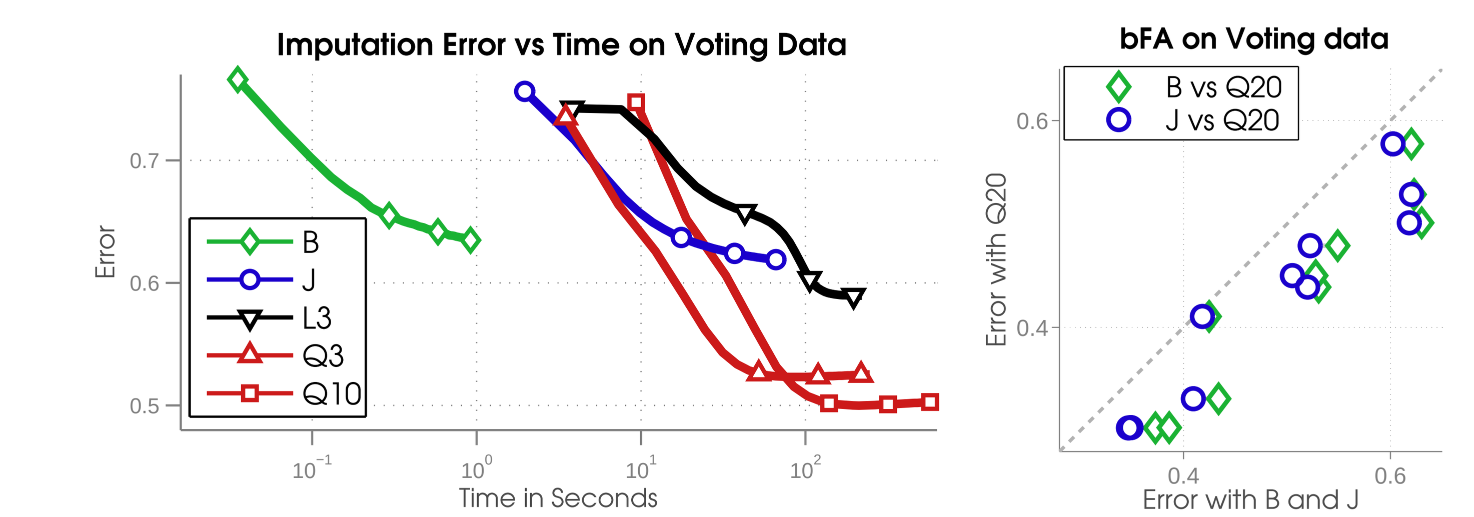
The LED dataset ($D = 24$ and $N = 2000$) First plot shows the imputation error for the LED dataset on the bLGGM. Each point is a random train-test split. A point below dashed line shows that the error in the piecewise bound is lower than other bounds. Second plot shows the same for the sparse bLGGM ($\boldsymbol{\Sigma}$ is sparse). Third plot shows the error with respect to regularization parameters.



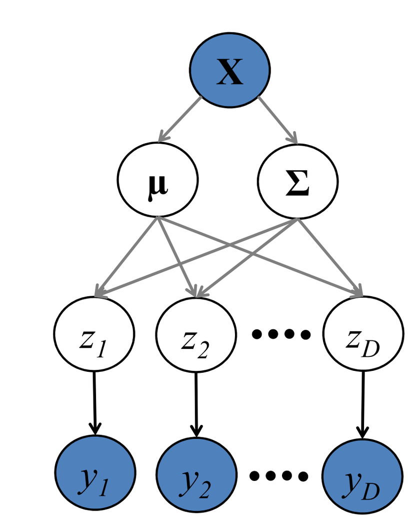
Binary Factor Analysis



The Voting Dataset ($D = 24$ and $N = 2000$) First plot shows the imputation error vs time for binary factor analysis (bFA). Second plot shows the imputation error for random train-test split.



Gaussian Process Classification



The Ionosphere Dataset ($D = 200$) First two plots show the cross-entropy prediction error obtained with out algorithm and Expectation Propagation (EP). Next two plots show the lower bound obtained by our algorithm and approximation to the marginal likelihood obtained by EP.