Variational Bounds for Mixed-Data Factor Analysis

Mohammad Emtiyaz Khan[1], Benjamin M. Marlin[1], Guillaume Bouchard[2], and Kevin P. Murphy[1]

1. Department of Computer Science, University of British Columbia 2. Xerox Research Center Europe

Introduction

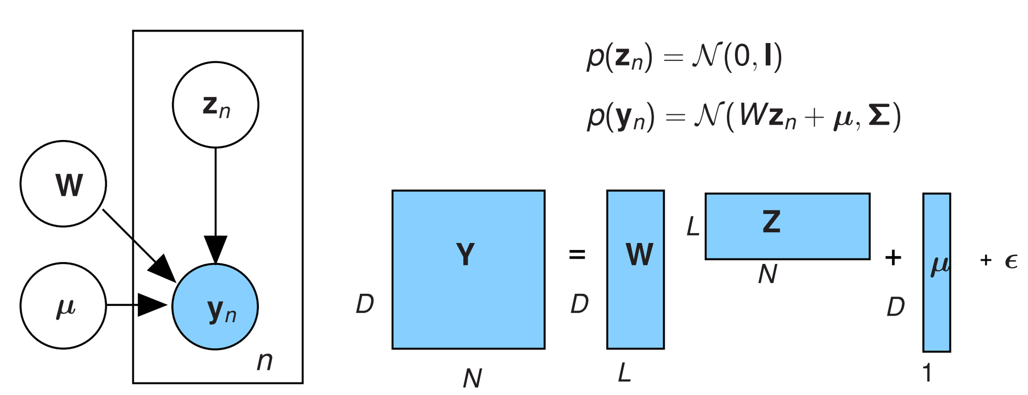
Motivation: Gaussian latent factor models, such as factor analysis (FA) and probabilistic principal components analysis (PPCA), are very commonly used density models for continuous-valued data. They have many applications including latent factor discovery, dimensionality reduction, and missing data imputation. In this work, we consider generalized FA models for mixed continuous and discrete data. These models are extremely useful since they allow for non-trivial dependencies between data variables with mixed types.

Problem: Unlike standard FA and PPCA, Gaussian latent factor models for discrete data have an intractable integral in the marginal likelihood that makes learning difficult.

Solution: We propose to solve the intractable integral through the application of a simple variational quadratic bound to the log-sum-exp function. The bound applies to both categorical and binary data. The resulting learning algorithm has advantages over other approaches to learning such models.

Factor Analysis Models

Gaussian Likelihood: Standard factor analysis models assume a Gaussian prior on the latent factor vector and a Gaussian likelihood on the observed data. The mean of the Gaussian on the observed data is modeled as a linear projection of the continuous latent factor.



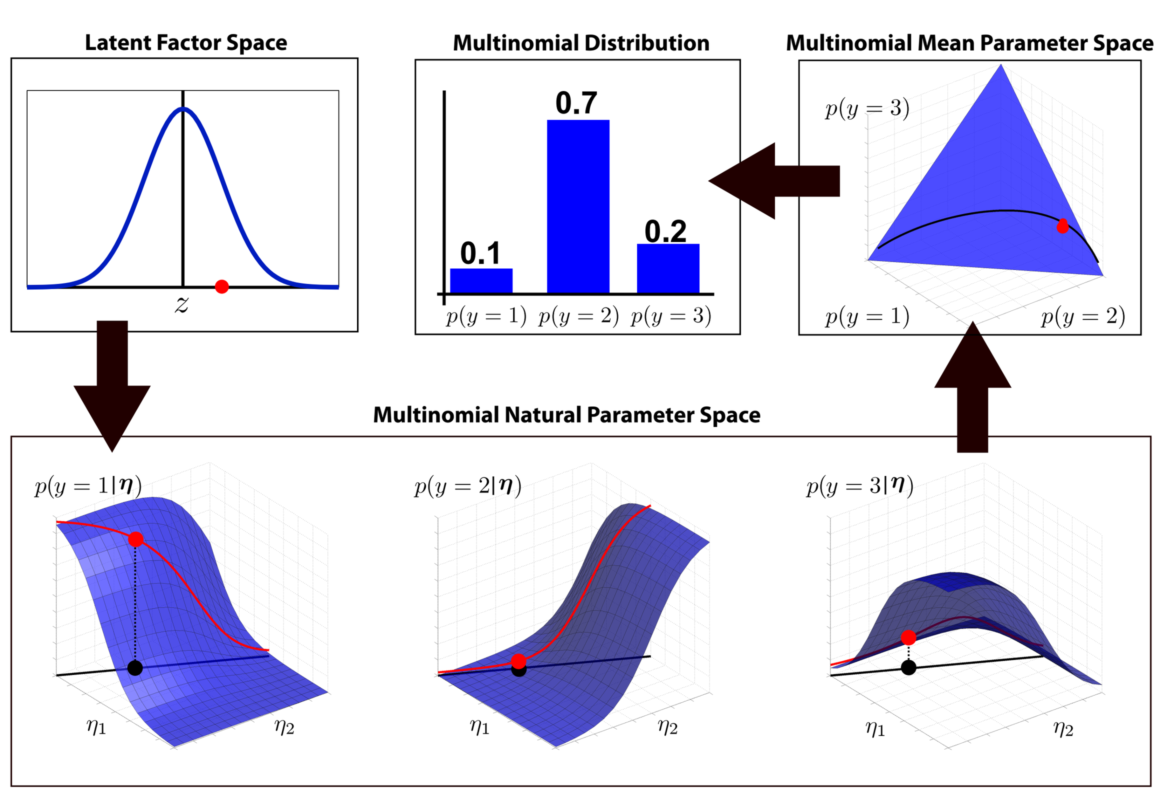
$$p(\mathbf{z}_n) = \mathcal{N}(0, \mathbf{I})$$
$$p(\mathbf{y}_n) = \mathcal{N}(W\mathbf{z}_n + \mu, \mathbf{\Sigma})$$

Such models are easy to fit since marginal likelihood is available in closed form,

$$p(\mathbf{y}_n|\theta) = \int_{\mathbf{z}_n} \mathcal{N}(\mathbf{y}_n|\mathbf{W}\mathbf{z}_n, \mathbf{\Sigma})\mathcal{N}(\mathbf{z}_n|0, \mathbf{I}) = \mathcal{N}(\mathbf{y}_n|\mathbf{W}\mathbf{W}^T + \mathbf{\Sigma})$$

Discrete Likelihood: Standard factor analysis can be generalized to any exponential family likelihood by modeling the natural parameters as a linear projection of a Gaussian-distributed continuous latent factor vector. In the case of discrete data, the mean parameters of the multinomial (Bernoulli) distribution are obtained through a softmax (logistic) transformation applied to the linear projection of the latent factor vector.

$$p(\mathbf{z}_n|\theta) = \mathcal{N}(\mathbf{z}_n|0, \mathbf{I}_L) \qquad \mathcal{S}_m(\eta) = \exp[\eta_m - \text{lse}(\eta)]$$
$$\eta_n = \mathbf{W}\mathbf{z}_n + \mu \qquad \text{lse}(\eta) = \log[\sum_{m=1}^{M+1}\exp(\eta_m)]$$
$$p(\mathbf{y}_n^D|\mathbf{z}_n, \theta) = \mathcal{M}(\mathbf{y}_n^D|\mathcal{S}(\eta_n))$$



Variational Bounds

Tractable Lower bound to the Marginal Likelihood: Computation of the marginal likelihood is intractable as the multinomial likelihood is not conjugate to the Gaussian prior. We use variational bounds to compute a tractable lower bound.

$$p(\mathbf{y}_n^D|\theta) = \int_{\mathbf{z}_n} p(\mathbf{y}_n^D|\eta_n)p(\mathbf{z}_n)d\mathbf{z}_n$$
$$= \int_{\mathbf{z}_n} \exp\left[\eta_n^T\mathbf{y}_n^D - \text{lse}(\eta_n)\right]\mathcal{N}(\mathbf{z}_n|0, \mathbf{I})d\mathbf{z}_n$$
$$\geq \max_{\psi} \int_{\mathbf{z}_n} \exp\left[\eta_n^T\mathbf{y}_n^D - \tfrac{1}{2}\eta_n^T\mathbf{A}_\psi\eta_n + \mathbf{b}_\psi^T\eta_n - c_\psi\right]\mathcal{N}(\mathbf{z}_n|0, \mathbf{I})d\mathbf{z}_n$$
$$\text{for all } \psi \in \mathbb{R}^M.$$

The Bohning Bound: We use a quadratic bound due to Bohning. This bound can be derived using a Taylor series expansion around $\psi \in \mathbb{R}^M$,

$$\text{lse}(\eta) = \text{lse}(\psi) + (\eta - \psi)^T\mathbf{g}(\psi) + \tfrac{1}{2}(\eta - \psi)^T\mathbf{H}(\chi)(\eta - \psi)$$

where $\mathbf{g}(\cdot)$ and $\mathbf{H}(\cdot)$ are the gradient and Hessian of lse$(\cdot)$, and $\chi \in \mathbb{R}^M$ is chosen such that the equality holds. An upper bound to lse(η) is found by replacing the Hessian matrix $\mathbf{H}(\chi)$ with a fixed matrix $\mathbf{A}$ such that $\mathbf{A} - \mathbf{H}(\chi)$ is positive definite for all χ. Bohning gives one such matrix $\mathbf{A}$, which we define below:.

$$\text{lse}(\eta) \leq \tfrac{1}{2}\eta^T\mathbf{A}\eta - \mathbf{b}_\psi^T\eta + c_\psi$$
$$\mathbf{A} = \tfrac{1}{2}[\mathbf{I}_M - \mathbf{1}_M\mathbf{1}_M^T/(M+1)]$$
$$\mathbf{b}_\psi = \mathbf{A}\psi - \mathcal{S}(\psi)$$
$$c_\psi = \tfrac{1}{2}\psi^T\mathbf{A}\psi - \mathcal{S}(\psi)^T\psi + \text{lse}(\psi)$$

where $\psi \in \mathbb{R}^M$ is the variational parameter vector, $\mathbf{I}_M$ is the identity matrix of size $M \times M$ and $\mathbf{1}_M$ is a vector of ones of length M.

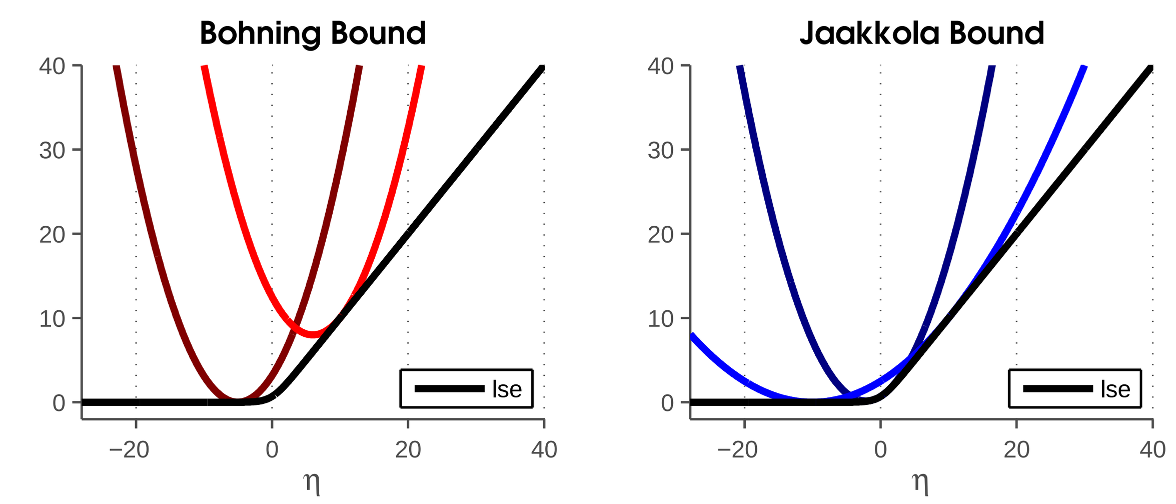
Bohning Bound:
- Less accurate.
- Faster.
- Fixed curvature.

$$A_\psi = 1/4$$
$$b_\psi = A\psi - (1 + e^{-\psi})^{-1}$$
$$c_\psi = \tfrac{1}{2}A\psi^2 - (1 + e^{-\psi})^{-1}\psi + \log(1 + e^{\psi})$$

Jaakkola Bound:
- More accurate.
- Slower.
- Variable curvature.

$$A_\psi = 2\lambda_\psi$$
$$b_\psi = -\tfrac{1}{2}$$
$$c_\psi = -\lambda_\psi\psi^2 - \tfrac{1}{2}\psi + \log(1 + e^{\psi})$$
$$\lambda_\psi = [(1 + e^{-\psi})^{-1} - \tfrac{1}{2}]/(2\psi)$$

Illustration of bounds: Variational bounds to $\log(1 + e^{\eta})$. The Bohning bound has a fixed curvature and is tight at one point, while the Jaakkola bound has a variable curvature and is tight at two points.



Posterior Inference and Parameter Estimation

Posterior Inference and Lower Bound to the Marginal Likelihood:

$$p(\mathbf{y}_n^D|\theta) \geq \max_{\psi} |\mathbf{V}_n|^{\frac{1}{2}}\exp\left[\tfrac{1}{2}\mathbf{m}_n^T\mathbf{V}_n^{-1}\mathbf{m}_n - c_\psi + \mu^T\mathbf{A}_\psi\mu + \mathbf{b}_\psi^T\mu + \mu^T\mathbf{y}_n^D\right]$$
$$\mathbf{V}_n = \left(\mathbf{W}^T\mathbf{A}_\psi\mathbf{W} + \mathbf{I}_L\right)^{-1}$$
$$\mathbf{m}_n = \mathbf{V}_n\mathbf{W}^T(\mathbf{y}_n^D + \mathbf{b}_\psi - \mathbf{A}\mu)$$

where $q(\mathbf{z}) = \mathcal{N}(\mathbf{m}_n, \mathbf{V}_n)$ is the approximate posterior distribution. The maximum with respect to ψ satisfies the following equation: $\psi = \mathbf{W}\mathbf{m}_n + \mu$.

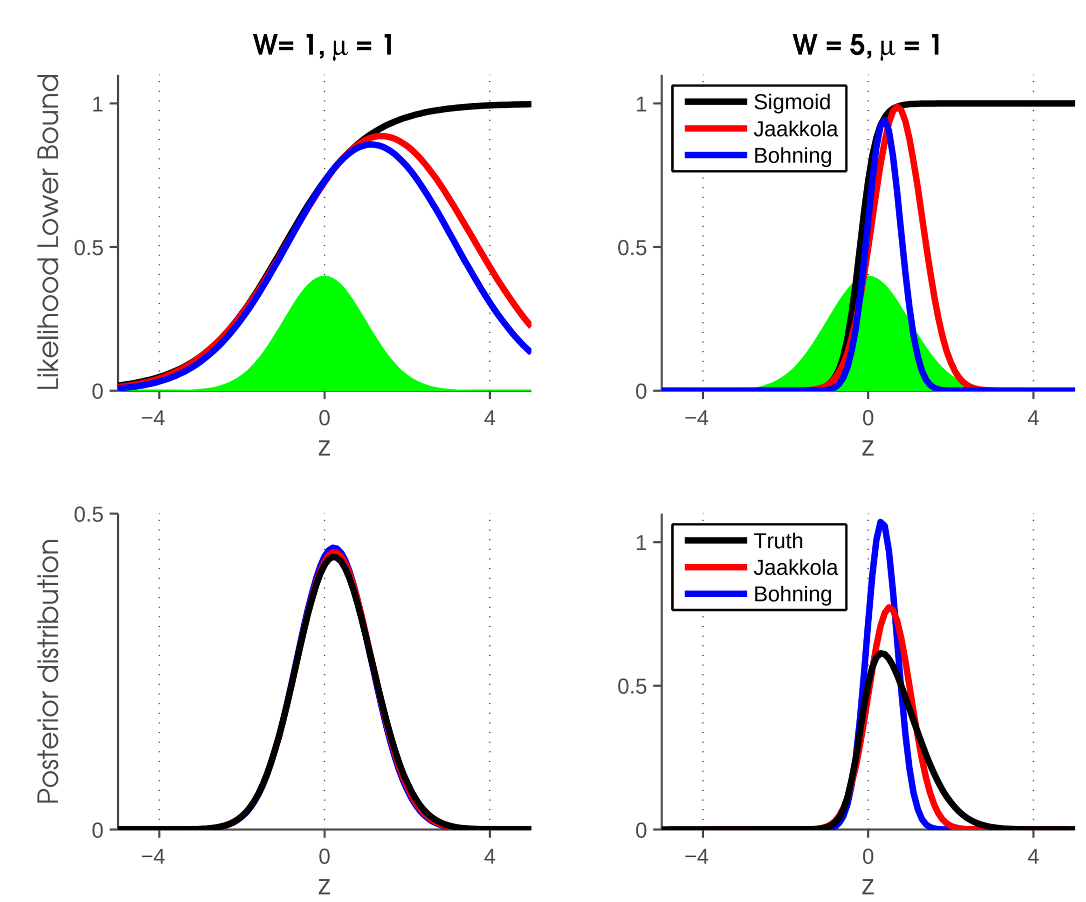
Parameter Estimation with EM algorithm: To get closed-form updates in the M step, we further lower bound the marginal likelihood using Jensen's inequality with the Gaussian variational posterior $q(\mathbf{z}_n)$

$$p(\mathbf{y}_n^D|\theta) \geq \max_{\psi} \mathbb{E}_q\left[\eta_n^T\mathbf{y}_n^D - \tfrac{1}{2}\eta_n^T\mathbf{A}_\psi\eta_n + \mathbf{b}_\psi^T\eta_n - c_\psi\right] + \mathbb{E}_q\mathcal{N}(\mathbf{z}_n|0, \mathbf{I}) + \mathbb{H}(q)$$

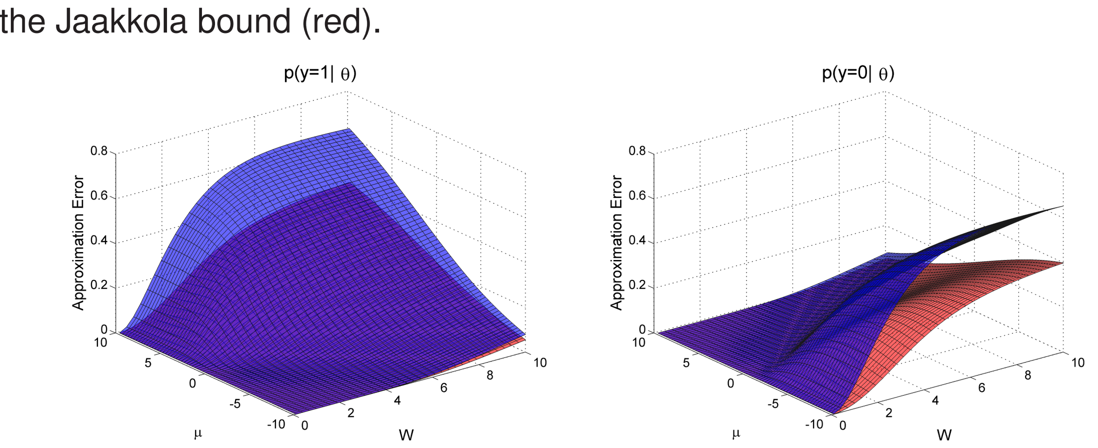
Fixed Curvature	Variable Curvature
Complexity: $O(L^2DNI)$ per iteration	Complexity: $O(L^3DNI)$ per iteration

Initialize $\mathbf{W}$ and μ.
repeat
$\mathbf{V} = \left(\mathbf{W}^T\mathbf{A}\mathbf{W} + \mathbf{I}_L\right)^{-1}$
for $n = 1, \ldots, N$ **do**
Initialize ψ.
for $i = 1, \ldots, I$ **do**
$\mathbf{m}_n = \mathbf{V}\mathbf{W}^T(\mathbf{y}_n^D + \mathbf{b}_\psi - \mathbf{A}\mu)$
Update ψ, $\mathbf{b}_\psi$ and c_ψ.
end for
$\tilde{\mathbf{y}}_n = \mathbf{A}^{-1}(\mathbf{y}_n^D + \mathbf{b}_\psi) - \mu$
end for
$\mu = \sum\tilde{\mathbf{y}}_n/N$.
$\mathbf{W} = \left(\sum\tilde{\mathbf{y}}_n\mathbf{m}_n^T\right)\left[\sum\mathbf{V} + \mathbf{m}_n\mathbf{m}_n^T\right]^{-1}$
until convergence
	Initialize $\mathbf{W}$ and μ.
	repeat
	for $n = 1, \ldots, N$ **do**
	Initialize ψ.
	for $i = 1, \ldots, I$ **do**
	$\mathbf{V}_n = \left(\mathbf{W}^T\mathbf{A}_\psi\mathbf{W} + \mathbf{I}_L\right)^{-1}$
	$\mathbf{m}_n = \mathbf{V}_n\mathbf{W}^T(\mathbf{y}_n^D + \mathbf{b}_\psi - \mathbf{A}_\psi\mu)$
	Update ψ, $\mathbf{A}_\psi$, $\mathbf{b}_\psi$ and c_ψ.
	end for
	$\tilde{\mathbf{y}}_n = \mathbf{A}_\psi^{-1}(\mathbf{y}_n^D + \mathbf{b}_\psi) - \mu$
	end for
	$\mu = \sum\tilde{\mathbf{y}}_n/N$.
	$\mathbf{W} = \left(\sum\tilde{\mathbf{y}}_n\mathbf{m}_n^T\right)\left[\sum\mathbf{V}_n + \mathbf{m}_n\mathbf{m}_n^T\right]^{-1}$
	until convergence

Inference Example: Top row shows the likelihood for a binary observation $y = 1$ along with lower bounds and the prior distribution. Bottom row show the true and approximate posterior distributions.



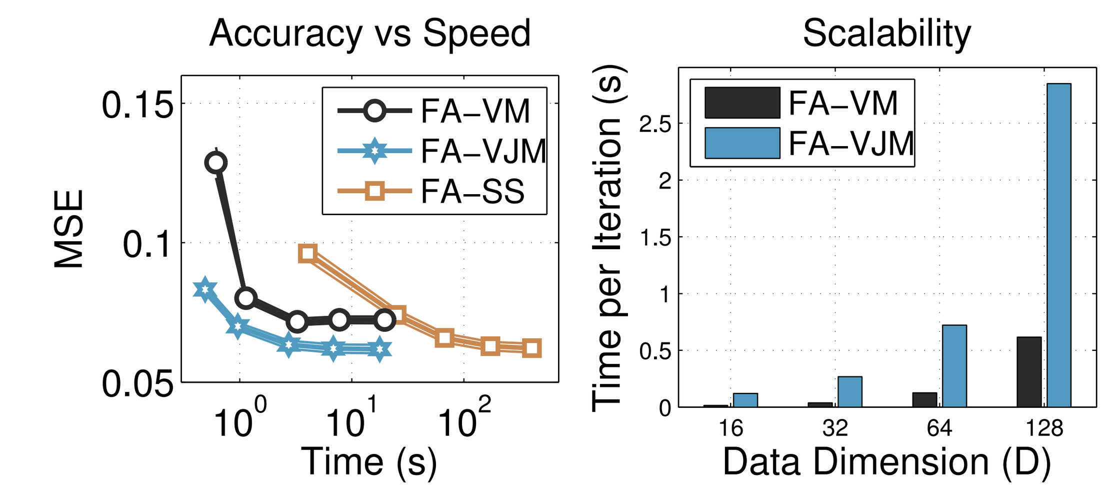
Error in Estimating the Marginal Likelihood: The Bohning bound (blue) and the Jaakkola bound (red).



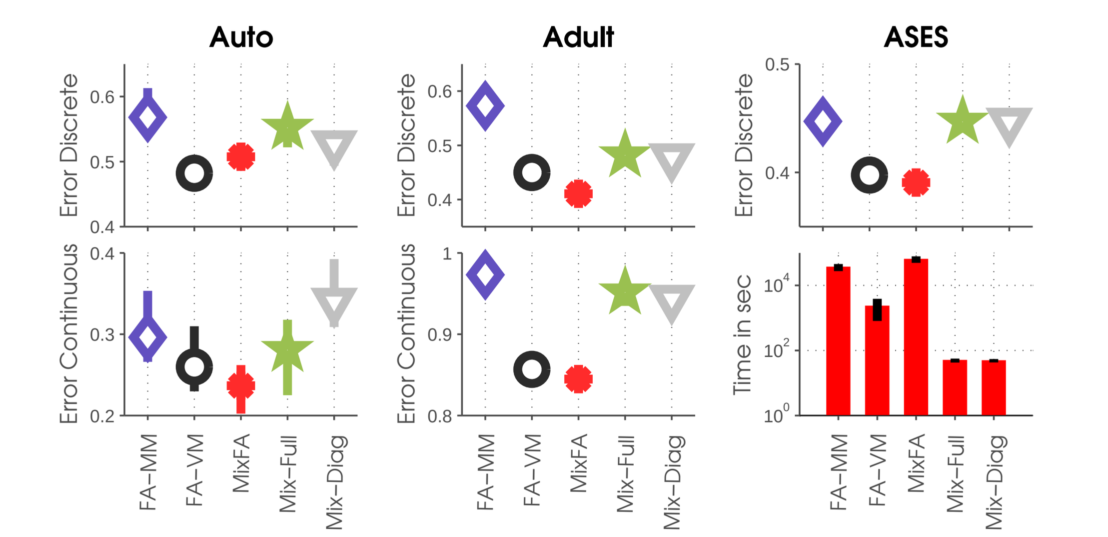
Results

Models and Methods:
- FA-VM FA model with the Bohning bound.
- FA-VJM FA model with the Jaakkola bound for binary data.
- Mix-FA Mixture of FA model with the Bohning bound.
- FA-MM FA model with the Maximize-maximize approach (Collins et. al. 2002).
- FA-SS FA model with the Sample-sample approach (Mohamed et. al. 2008).
- Mix-Full/Diag Mixture model with a full or diagonal covariance matrix.

Synthetic Data Experiment: MSE vs time on synthetic Binary data with $N = 600$, $D = 16$, $L = 10$ and 10% missing data.



Real Data Experiment: We compute imputation MSE and entropy on three datasets. We choose number of latent factors and number of mixture components using cross-validation.



Auto dataset has 392 observations of 3 continuous and 5 discrete variables with total of 21 categories. *Adult* dataset has 45,222 observations of 4 continuous and 5 discrete variables with total of 27 categories. *ASES* dataset has 16,815 observations of 42 discrete variables with total of 156 categories.

Continuous FA vs Mixed-Data FA: Latent factors for Auto data. Top row shows factors using only continuous variables. Bottom row shows factors obtained by including discrete variables.