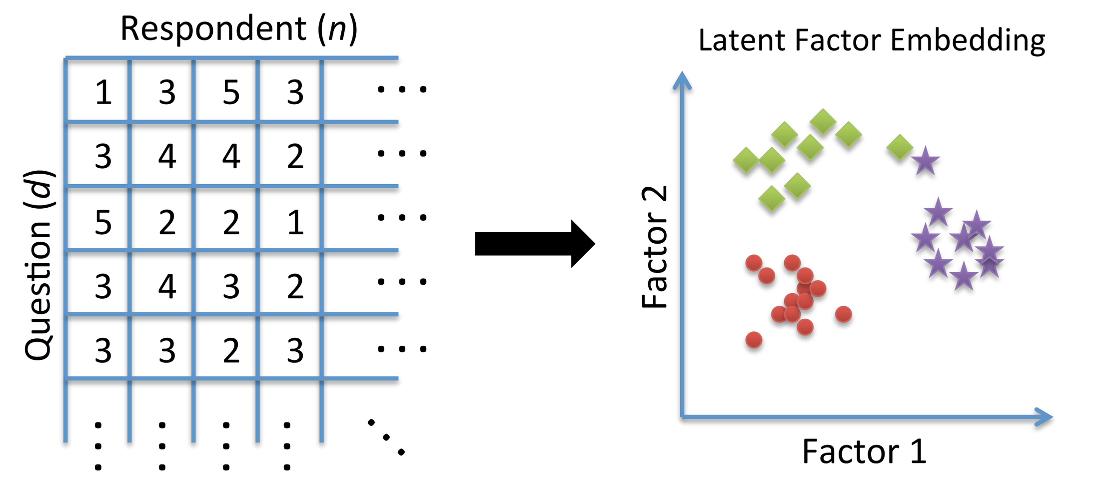
A Stick-Breaking Likelihood for Categorical Data Analysis with Latent Gaussian Models

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

[1]University of British Columbia, [2]University of Massachusetts, Amherst

Introduction

Motivation: Analysis of high-dimensional categorical data is essential in applications such as recommender systems, econometrics, social sciences, and medical diagnostics. Such analysis can be carried out using latent Gaussian models, which include multinomial logistic regression, multi-class Gaussian process classification, categorical factor analysis, etc.

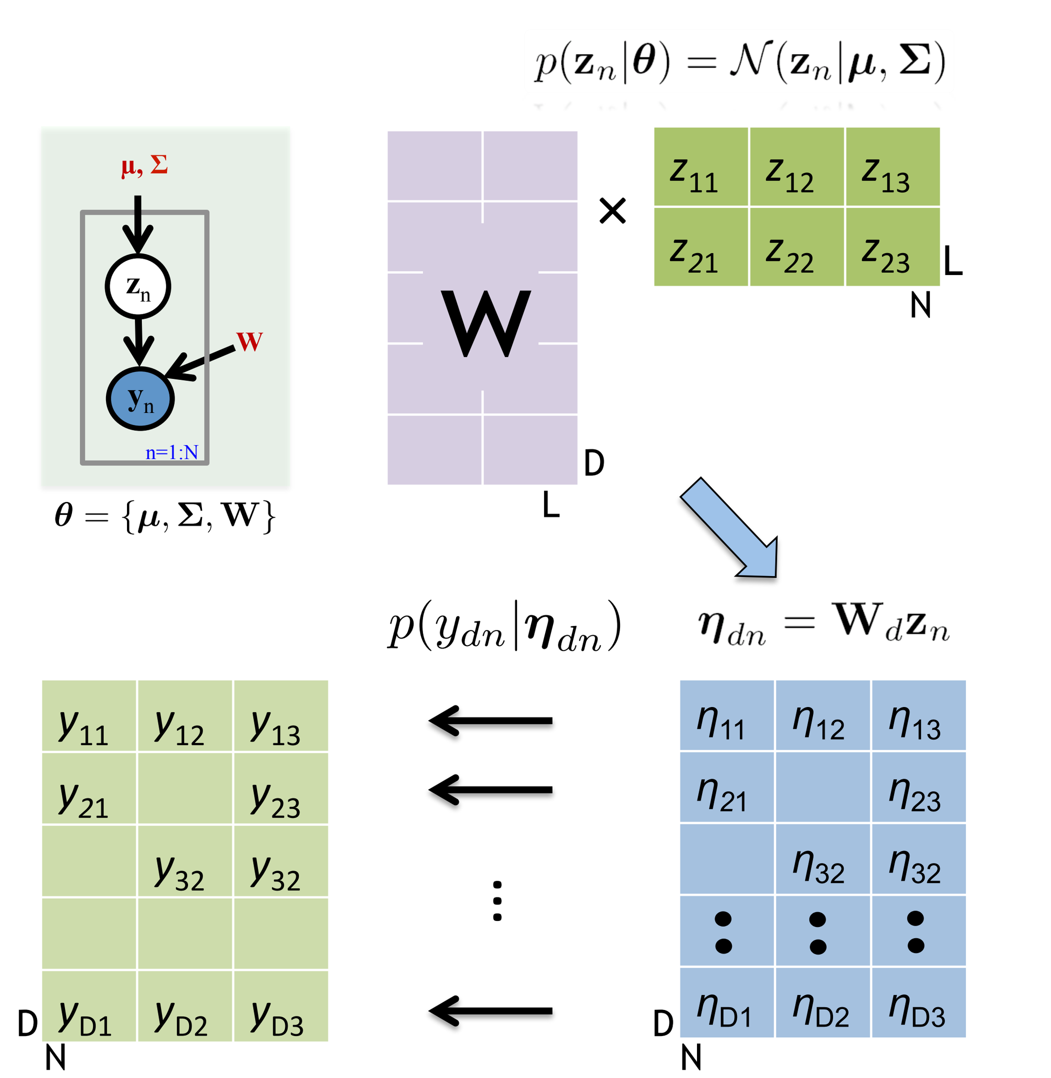


Problem: Parameter learning is difficult since the marginal likelihood contains an intractable integral, which arises due to the non-conjugacy between the likelihood and Gaussian prior on the latent variables.

Solution: We propose a novel stick-breaking likelihood for categorical data analysis and derive tractable and accurate lower bounds for the marginal likelihood. Our results demonstrate that the proposed stick-breaking model effectively captures correlation and is well suited to the analysis of categorical data.

Latent Gaussian Models

Our model uses latent Gaussian variables to model the distribution of categorical observations. For categorical data, each element y_{dn} of the observed vector $\mathbf{y}_n$ can take values from a finite discrete set $S_d = \{C_1, C_2, \ldots, C_K\}$. For the n'th data vector, the generative process is: (1) Sample latent Gaussian vectors $\mathbf{z}_n \in \mathbb{R}^L$. (2) Take a linear combination of $\mathbf{z}_n$ to obtain the predictor $\eta_{dn} \in \mathbb{R}^K$. (3) Draw data from a categorical distribution given η_{dn}. Our goal is to learn the model parameters θ given $\mathbf{y}_1, \ldots, \mathbf{y}_N$.



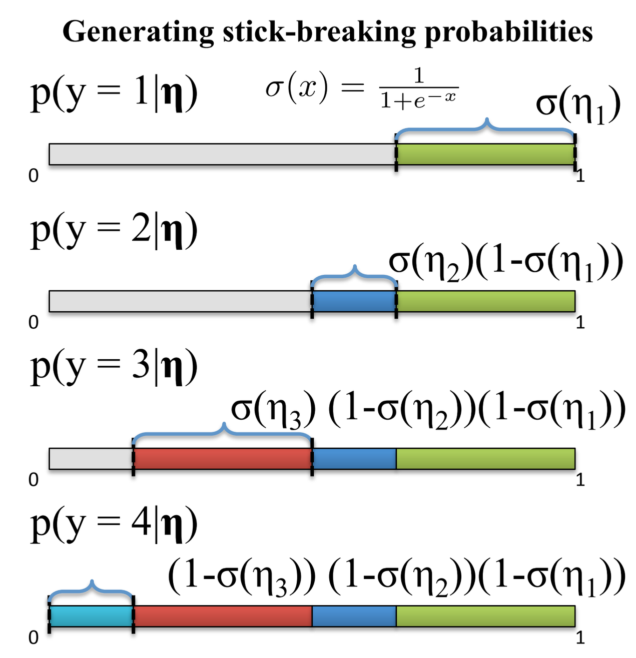
Stick-Breaking Likelihood

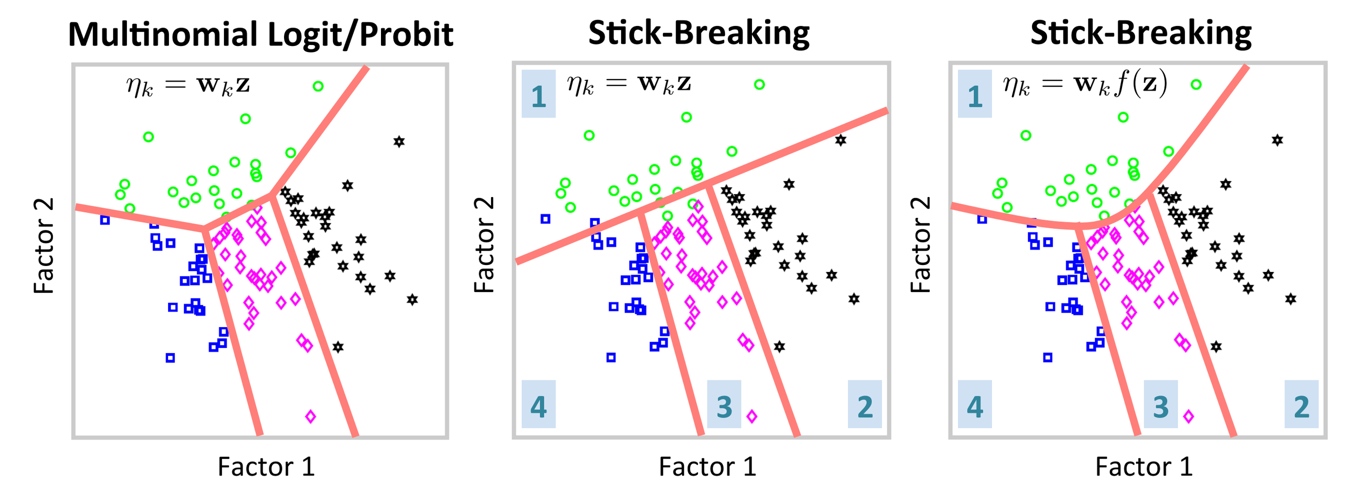
We propose a novel stick-breaking likelihood to model $p(y = C_k|\eta)$, defined in terms of the logistic log-partition function $llp(x) = log(1 + exp(x))$. This is simpler than the multinomial logit model which uses the log-sum-exp function $lse = \log \sum_j \exp(\eta_j)$, a difficult function to approximate.

Stick: $$\frac{\exp(\eta_k)}{\prod_{j\le k}[1+\exp(\eta_j)]} = \exp[\eta_k - \sum_{j\le k} llp(\eta_j)]$$

Logit: $$\frac{\exp(\eta_k)}{\sum_{j=1}^K \exp(\eta_j)} = \exp[\eta_k - lse(\boldsymbol{\eta})]$$

Probit: $$\int \prod_{j\ne k} \boldsymbol{\Phi}(\epsilon + \eta_k - \eta_j) d\epsilon$$





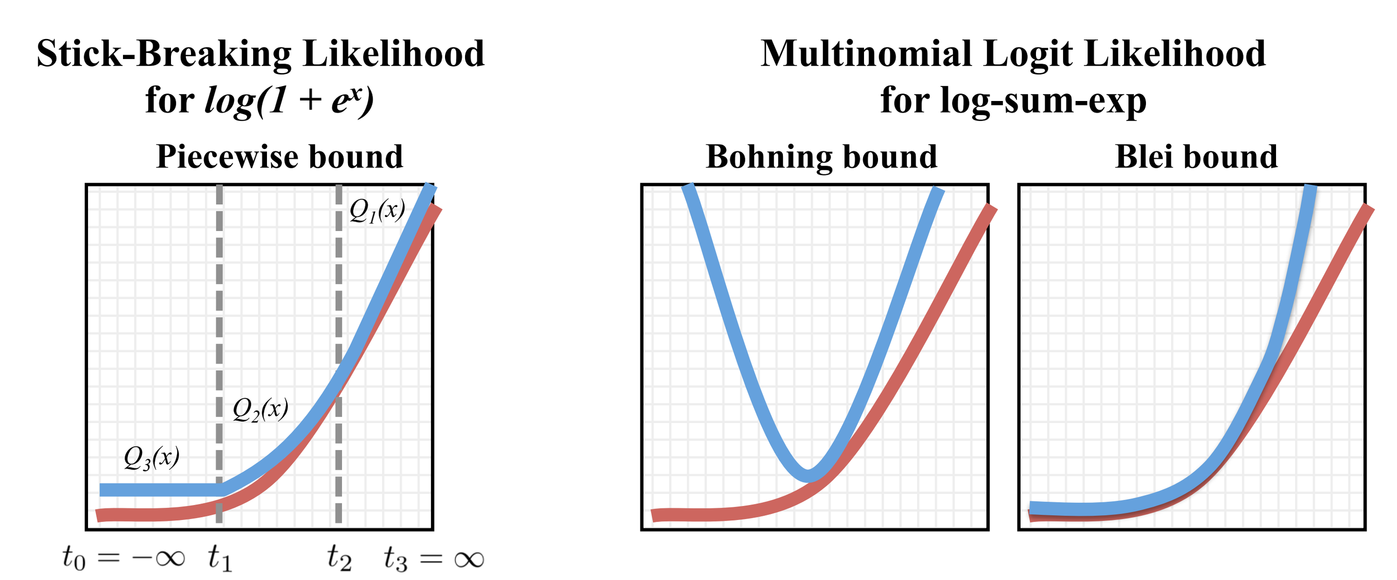
Variational Inference

Variational lower bound: Computation of the marginal likelihood is intractable since the categorical likelihood is not conjugate to the Gaussian prior. Using Jensen's inequality, we can obtain a lower bound to the marginal likelihood.

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

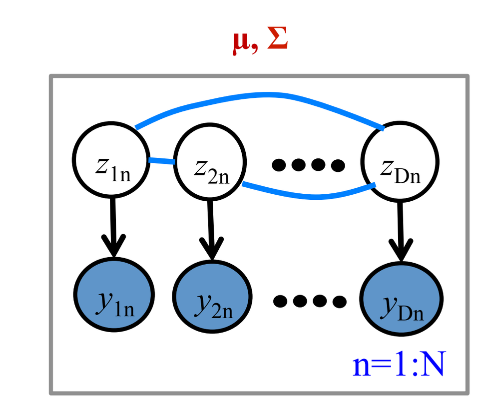
Piecewise bounds for the stick-breaking likelihood: These expectations of the likelihood w.r.t a Gaussian, are still intractable due to the presence of $\log(1 + \exp(x))$, the logistic-log-partition (LLP) function. This is made tractable by using piecewise lower bounds for the LLP, for which the expectation of each piece w.r.t. a Gaussian is tractable.

Existing bounds for the multinomial logit likelihood: Existing bounds for the log-sum-exp term have unbounded error and can cause severe bias in parameter estimates. Piecewise bounds have bounded error.

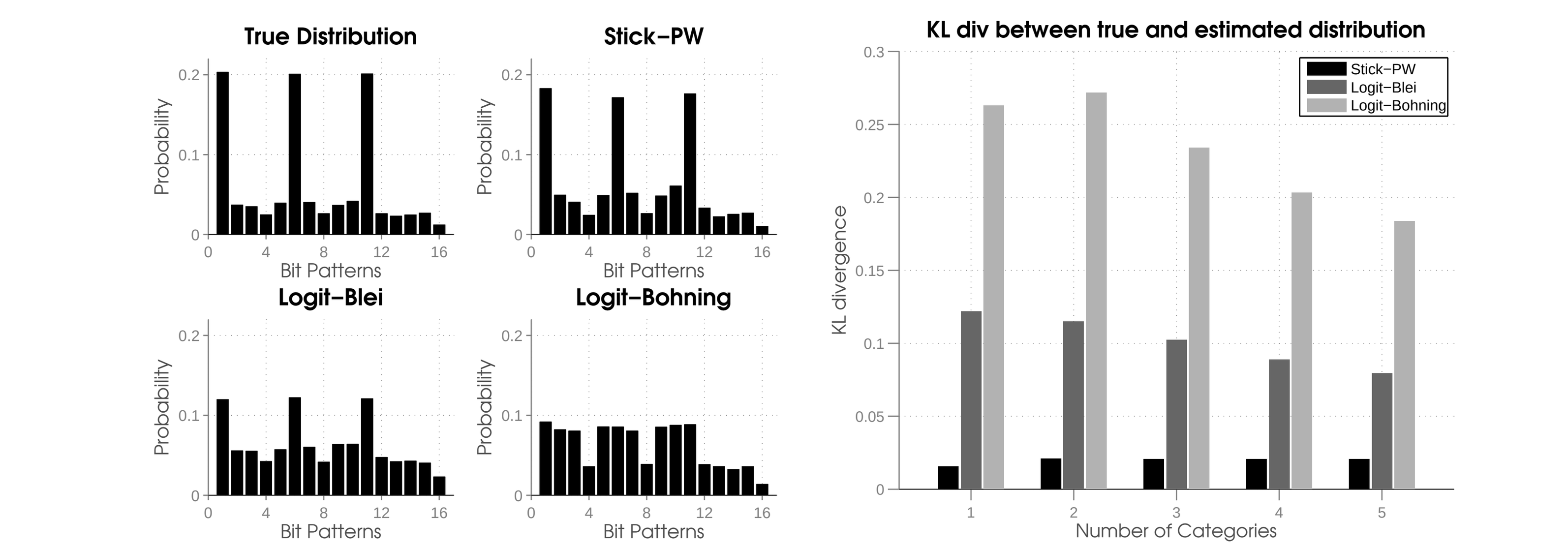


Results

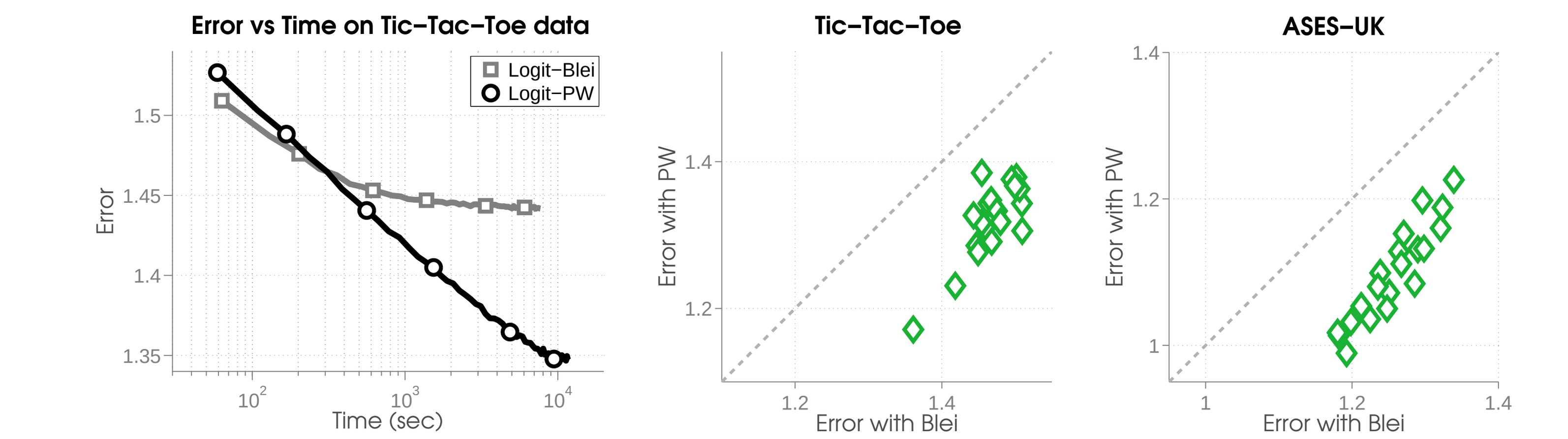
Categorical Latent Gaussian Graphical Model (cLGGM)



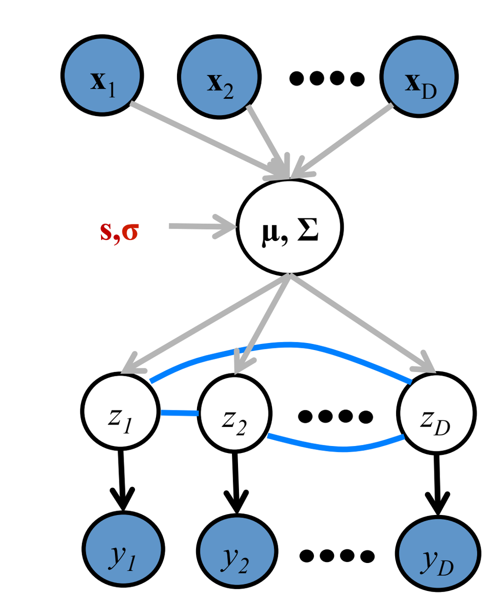
Highly correlated synthetic data: $(D = 2, K = 4, N = 10,000)$ We compare the true discrete distribution with the estimated distribution using various methods and models.



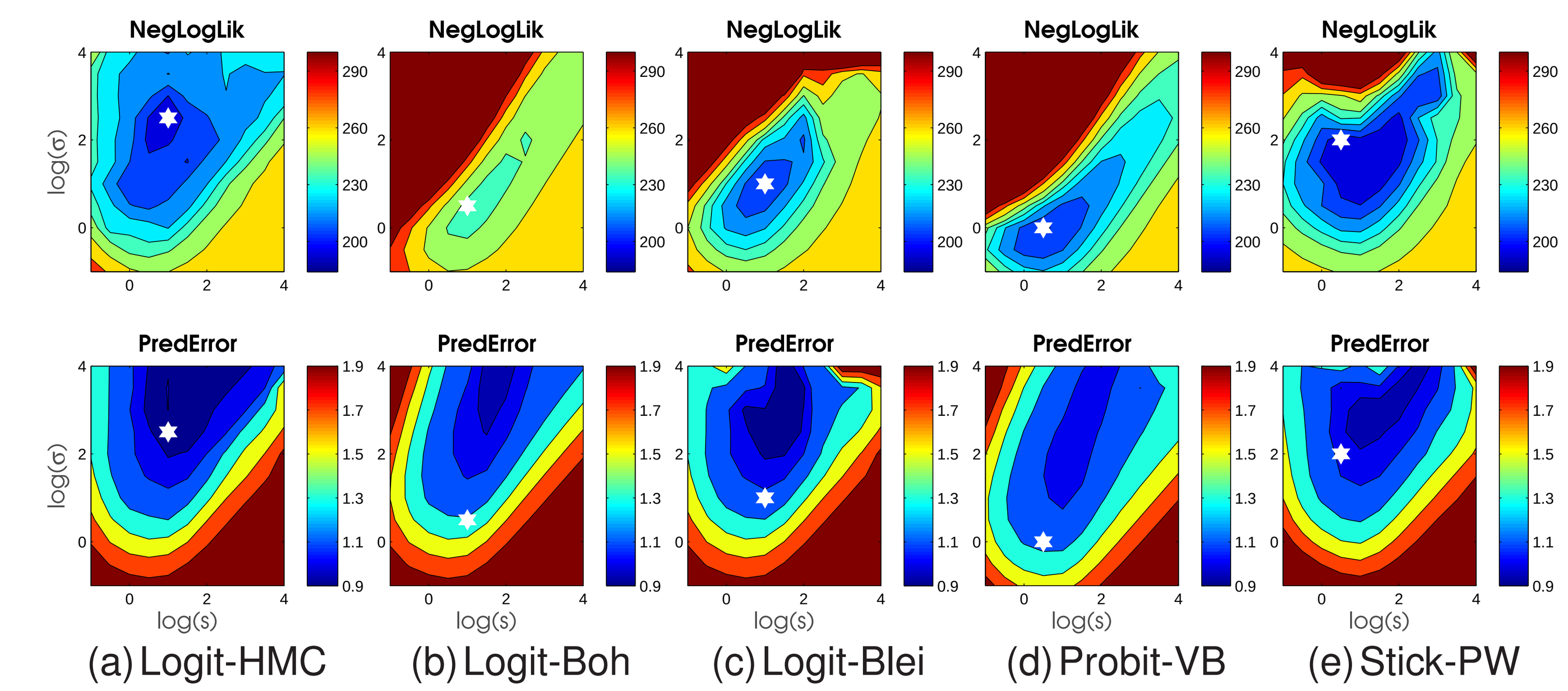
Real data: We compare missing value imputation error for two datasets:
Tic-tac-toe ($\sum_d K_d = 29, D = 10, N = 968$),
ASES-UK ($\sum_d K_d = 60, D = 17, N = 913$)



Multi-class Gaussian Process Classification
$\Sigma_{ij} = \sigma^2 \exp\left[||\mathbf{x}_i - \mathbf{x}_j||_2^2/(2s)\right]$



Forensic glass data: $(D = 214, K = 6)$. The top row shows contour plots of negative log-likelihood for the training data obtained over various hyperparameter settings; the bottom row shows the prediction error on the test set. The star indicates the hyperparameter setting at the minimum negative log-likelihood.



Illustrating LGM on voting dataset $(D = 16, K = 2, N = 435)$. We plot the posterior mean of the latent factors with size proportional to the log-likelihood. We also plot the probability of two variables (votes) taking the same value and the probability of voting 'yes' given the party.

	Description
V1	budget-resolution
V2	aid-nicaraguan-contras
V3	mx-missile
V4	duty-free-exports
V5	anti-satellite-test-ban
V6	handicapped-infants
V7	export-admin-south-africa
V8	immigration
V9	synfuels-corp-cutback
V10	water-project-cost-sharing
V11	physician-fee-freeze
V12	El-salvador-aid
V13	religious-groups-in-schools
V14	education-spending
V15	superfund-right-to-sue
V16	crime