Sparse Bayesian Logistic Regression with Hierarchical Prior and Variational Inference

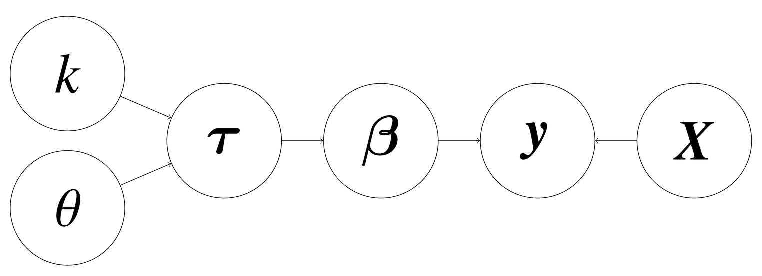
Bayesian Sparse Classifiers

Logistic regression model:

$$p(y_j|\boldsymbol{\beta}) = \sigma(\boldsymbol{x}_j^T\boldsymbol{\beta})^{y_j}(1-\sigma(\boldsymbol{x}_j^T\boldsymbol{\beta}))^{1-y_j},$$
$$\sigma(x) = 1/\left(1+e^{-x}\right).$$

Existing Bayesian sparse classifiers:

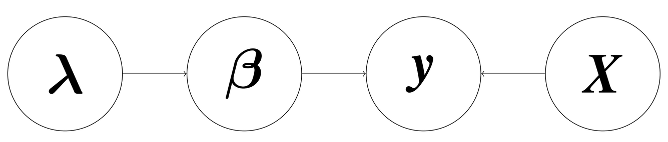
- Relevance Vector Machine [Tipping, 2001]



$$\tau_i \sim \mathcal{IG}(\tau_i|k,\theta)$$
$$\beta_i \sim \mathcal{N}(\beta_i|0,\tau_i)$$

  - Learning: ML or MAP estimate for γ and Laplace approximation

- Sparse Representation Prior [Serra et al., 2016]



$p(\beta_i|\lambda)$: Sparsity inducing prior (such as Laplace)

  - Learning: Majorize minimization and variational inference

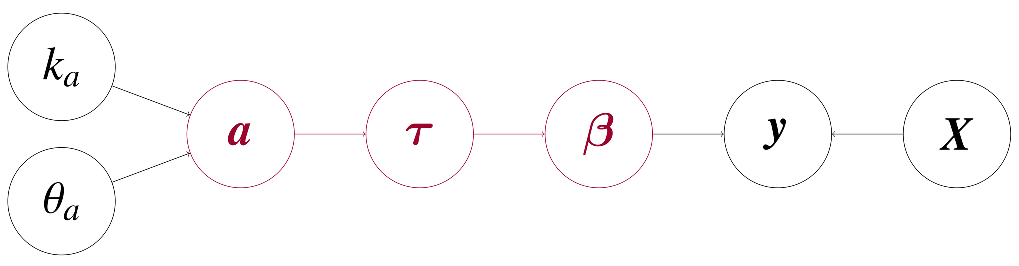
Proposed Model

- Hierarchical prior

$$\beta_i \sim \mathcal{N}(\beta_i|0,\tau_i),\ \ \tau_i \sim \mathcal{GIG}(\tau_i|a_i,b_i,\rho)$$
$$(a_i,b_i) \sim \text{Conjugate prior for } p(\tau_i|a_i,b_i,\rho)$$

$\mathcal{GIG}$ is a **generalized inverse Gaussian**, which is exponential when $b_i = \rho = 0$ and inverse gamma when $a_i = 0, \rho < 0$.

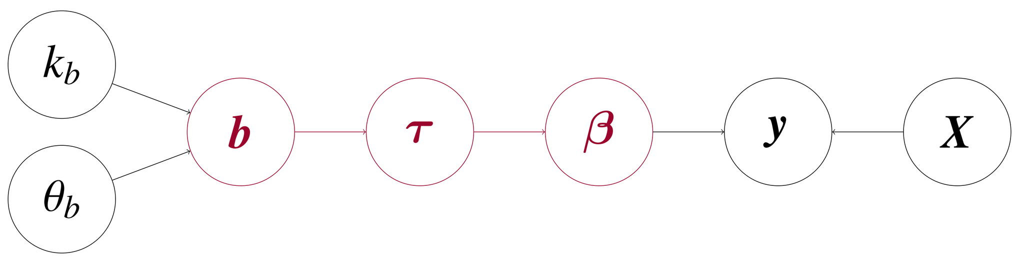
  - Exponential mixing:



$$\tau_i \sim \mathcal{EXP}(\tau_i|a_i)$$
$$a_i \sim \mathcal{G}(a_i|k_a,\theta_a)$$

The marginal distribution $p(\beta_i)$ is a **Laplace** distribution.

  - Inverse gamma mixing:



$$\tau_i \sim \mathcal{IG}(\tau_i|b_i,\rho)$$
$$b_i \sim \mathcal{G}(b_i|k_b,\theta_b)$$

The marginal distribution $p(\beta_i)$ is a **Student's t** distribution.

Learning Algorithm

- **Variational inference** with Mean-field approximation + Majorize Minimization
  - Extension of [Jaakkola and Jordan, 1997]

$h(\boldsymbol{\beta},\boldsymbol{\xi})$: a lower bound of $p(\boldsymbol{y}|\boldsymbol{\beta})$ and a quadratic function of $\boldsymbol{\beta}$

Algorithm

1. $q^*(\boldsymbol{\beta}) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{\tau})}\left[\ln h(\boldsymbol{\beta},\boldsymbol{\xi})p(\boldsymbol{\beta}|\boldsymbol{\tau})\right]\right)$
2. $q^*(\boldsymbol{\tau}) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{\beta})q(\boldsymbol{a},\boldsymbol{b})}\left[\ln p(\boldsymbol{\beta}|\boldsymbol{\tau})p(\boldsymbol{\tau}|\boldsymbol{a},\boldsymbol{b})\right]\right)$
3. $q^*(\boldsymbol{a},\boldsymbol{b}) \propto \exp\left(\mathbb{E}_{q(\boldsymbol{\tau})}\left[\ln p(\boldsymbol{\tau}|\boldsymbol{a},\boldsymbol{b})p(\boldsymbol{a},\boldsymbol{b})\right]\right)$
4. $\boldsymbol{\xi}^* = \mathrm{argmax}_{\boldsymbol{\xi}}\,\mathbb{E}_{q(\boldsymbol{\beta})}\left[\ln h(\boldsymbol{\beta},\boldsymbol{\xi})\right]$

Experiments on synthetic data

- True parameter $\boldsymbol{\beta}^* = (\boldsymbol{0}_{90}, \boldsymbol{2}_{10})$
- Parameter setting for the proposed algorithm:
  - Exponential mixing ($b_i = \rho = 0$)
  - $k_a = \theta_a = 10^{-6}$ (very flat prior)
- Compare with L_1 regularized logistic regression and L_2 regularized logistic regression

Table: MSE and prediction accuracy for synthetic data.

	MSE	Accuracy
Proposed	**0.1589 ± 0.1133**	**0.8195 ± 0.0477**
L_1	0.3974 ± 0.2939	0.7750 ± 0.0456
L_2	0.4391 ± 0.2597	0.7112 ± 0.0242

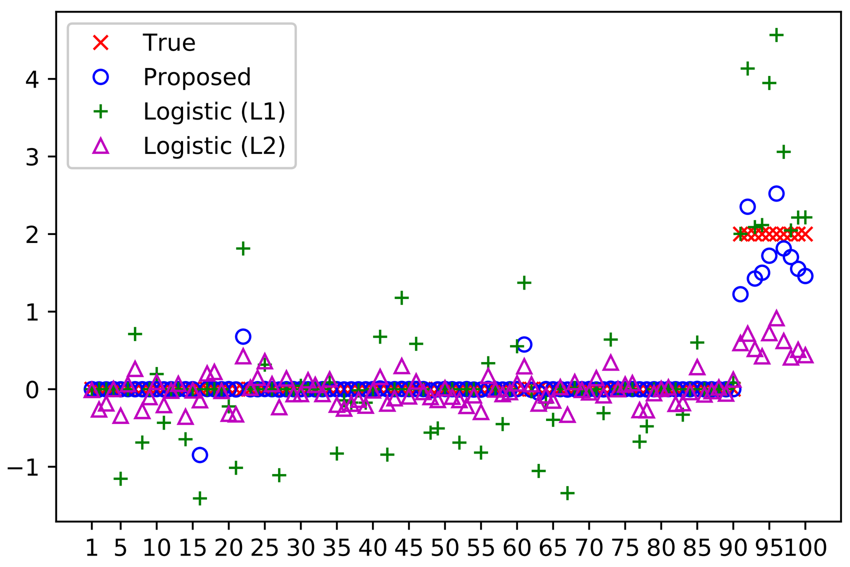


Figure: Estimation results on synthetic data.

Experiments on real world data

- Parameter setting is the same with the case of synthetic data

Table: Prediction accuracy for real world data.

	a1a	w1a	covtype
Proposed	**0.8400**	0.9757	**0.7436**
L_1	**0.8400**	0.9702	0.7412
L_2	0.8386	**0.9779**	0.7398

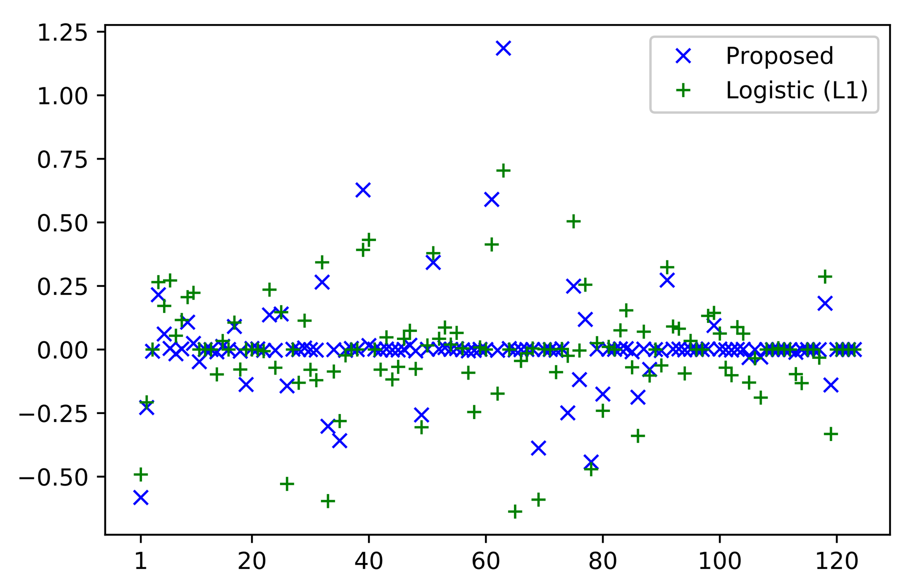


Figure: Estimation results on 'a1a' data.

References

M. E. Tipping. **Sparse bayesian learning and the relevance vector machine.** *Journal of machine learning research*, pp. 211-244, 2001.

J. G. Serra, P. Ruiz, R. Molina, and A. K. Katsaggelos. **Bayesian logistic regression with sparse general representation prior for multispectral image classification.** In *Image Processing (ICIP), 2016 IEEE International Conference on*, pp. 1893-1897, 2016.

T. Jaakkola and M. Jordan. **A variational approach to bayesian logistic regression models and their extensions.** In *Sixth International Workshop on Artificial Intelligence and Statistics*, Vol. 82, page 4, 1997.

Shunsuke Horii
s.horii@aoni.waseda.jp
Waseda University