Group-sparse Variational Bayes

Michael Komodromos

Imperial College London

2022

Problem setting

We've been working on algorithms for group sparse regression. This a setting where there is group structure in the data.

Occurs for example in

- genomics: where allele frequencies / SNPs
- transcriptomics: where the functional groups of genes can be learnt
- radiomics: where features often correspond to different decompositions

Why?

If we believe that the groups represent the data sufficiently well, then we can boost the efficiency of our models.

i.e. learn models that can predict better with less data.

Within the lab this may pose to be useful. Current published methods include:

- group LASSO
- sparse-group LASSO
- group SCAD
- group spike-and-slab LASSO (works well for linear regression)

Summary

We've created a new algorithms for group sparse regression, available for:

- Linear model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$
- Logistic regression $y \in \{0, 1\}$
- Poisson regression $y \in \{0, 1, \ldots\}$

We've reduced the compute time by orders of magnitude (x300 faster), and preserved many useful features.

Summary con.

Additionally, we've approached the problem through the Bayesian paradigm.

Meaning, on top of the speed up over traditional methods, we are able to use our models for prediction with uncertainty.

A Brief Overview

Model

We are going to be considering models of the form

$$y = f\left(\sum_{k=1}^{M} X_{G_k}\beta_{G_k}\right) + \epsilon \qquad (1)$$

where there are $k = 1, \ldots, M$ groups.

In the linear regression setting, $f(x) = x$.

For other models

- logistic model, f is the logistic function
- Poisson, $f(x) = \exp(x)$

Formulation

We consider a group spike-and-slab (GSpSL) prior, which has a hierarchical representation,

$$\begin{aligned} \beta_{G_k}|z_k &\overset{\text{ind}}{\sim} z_k \Psi(\beta_{G_k}; \lambda) + (1 - z_k)\delta_0(\beta_{G_k}) \\ z_k|\theta_k &\overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_k) \\ \theta_k &\overset{\text{iid}}{\sim} \text{Beta}(a_0, b_0) \end{aligned} \tag{2}$$

where $\delta_0(\beta_{G_k})$ is the multivariate Dirac mass on zero with dimension m_k and $\Psi(\beta_{G_k})$ is the multivariate double exponential distribution with density

$$\psi(\beta_{G_k}; \lambda) = C_k \lambda^{m_k} \exp\left(-\lambda \|\beta_{G_k}\|\right) \tag{3}$$

where $\|\cdot\|$ is the ℓ_2-norm and C_k the normalizing constant.

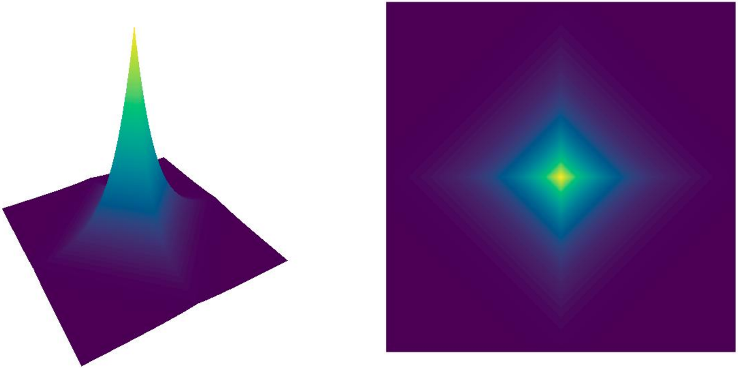

Figure: Double Exponential distribution

Rather than using MCMC to obtain a sample from the posterior, approximated it with an element from a tractable family of distributions. The variational posterior is given by solving,

$$\tilde{\Pi} = \arg\min_{Q \in \mathcal{Q}} \mathrm{KL}\left(Q \| \Pi(\cdot|\mathcal{D})\right) \tag{4}$$

where $\mathcal{Q}$ is the variational family and KL is the KL divergence

Results

Method	l2-error	AUC	Runtime
GSVB	0.987 (0.70, 1.31)	1.000 (1.00)	1.7s
GSVB	1.008 (0.74, 1.53)	1.000 (1.00)	2.4s
MCMC	0.993 (0.73, 1.30)	1.000 (1.00)	3m 23s
SpSLasso	1.039 (0.69, 4.45)	1.000 (0.90, 1.00)	7.3s

Table: (n=200, p=1,000, g=5, s=5)

Perform just as well as the other methods including MCMC.

Note the runtime for our method is the best (roughly x100 faster than MCMC)

Method	Coverage $\beta \neq 0$	Predictive Coverage
GSVB	0.780 (0.60, 0.96)	0.950 (0.90, 0.98)
GSVB	0.920 (0.76, 1.00)	0.960 (0.90, 0.99)
MCMC	0.960 (0.84, 1.00)	0.960 (0.91, 0.98)

Table: (n=200, p=1,000, g=5, s=5)

Coverage (1:best, 0: worst) for our method is comparable to MCMC.

Note: the posterior predictive coverage (uncertainty on prediction) is just as good as MCMC!

Large scale studies

Method	l2-error	AUC	Runtime
GSVB	1.702 (1.42, 2.02)	1.000 (1.00)	7m 47s
GSVB	1.702 (1.47, 2.00)	1.000 (1.00)	7m 17s
SpSLasso	1.719 (1.42, 2.01)	1.000 (1.00)	1m 14s

Table: (n=500, p=5,000, g=10, s=10)

Perform as well / slightly better than other methods in a very high dimensional setting.

Note: unlike the other method (SpSLasso) we are able to provide uncertainty around the coefficient estimates and the prediction!

Available Extensions

Extensions

We've also extended the method to logistic regression and Poisson regression.

We've seen similar performance for both these types of regression.

Meaning our method provides state-of-the-art performance for both binary classification and count regression

We may also provide a further extension to Cox regression

Packages

Currently, the project is being written up. However there are packages available for R.

These include functions to fit models, make predictions and produce credible intervals

In practice

In practice

In our previous presentation to the group we showed how group using the group structure in regression can improve performance.

We are yet to run our own analysis on real data, but for a great paper see

A sparse-group Lasso by Noah Simon, Jerome Friedman, Trevor Hastie and Rob Tibshirani

Great Exhibition Road Festival

GERF

Before concluding, I've submitted a proposal to GERF (for June 2023).

The theme is cancer and bias in machine learning.

The overarching goal is to engage with people about ML and cancer research and convey that it can be a really useful tool, made more useful through public engagement and interaction.

Questions?