Bayesian shrinkage approach in variable selection for mixed effects models

Mingan Yang

GGI Statistics Conference, Florence, 2015
Bayesian Variable Selection

June 22-26, 2015

Outline

1. Introduction
2. model
3. Simulation
4. Application

Outline

1. Introduction
2. model
3. Simulation
4. Application

- Variable selection problem draws considerable attention. Penalized regression with shrinkage estimation and simultaneous variable selection seems promising
- Frequentist approaches: Tibshirani (1996), Fan and Li (2001), Zou (2006), Zou and Li (2008) etc.
- Bayesian approaches: Park and Casella (2008), Hans (2009), Carvalho et. al. (2009, 2010), Griffin and Brown (2007), and Armagan et al. (2013)
- None proposed for mixed effects models (LME). Even more challenging for LME coupled with random effects.

- Linear mixed effects (LME) model:

$$\begin{aligned} y_i &= X_i\beta + Z_i\zeta_i + \epsilon_i \\ \zeta_i &\sim N(0, \Psi) \\ \epsilon_i &\sim N(0, \sigma^2\Lambda) \end{aligned}$$

- Challenge: standard methods such as AIC, BIC, Bayes factors etc. do not work well different number of random effects. (4^p models for $p \times 1$ fixed and random vector)
- Very few approaches proposed so far: Chen and Dunson (2003), Cai and Dunson (2006), Kinney and Dunson (2007), Bondell et al. (2010), Ibrahim et al. (2010), Yang (2012, 2013).

Some current methods for random effects

- Chen and Dunson (2003): Stochastic search variable selection (SSVS) approach to the LMM.
- Kinney and Dunson (2007):Application of data augmentation algorithm, parameter expansion (Gelman, 2004, 2005) and model approximation for the GLMM.
- Ibrahim et. al. (2010):Maximum penalized likelihood (MPL) and smoothly clipped absolute deviation (SCAD) and ALASSO
- Bondell et. al. (2010): adaptive LASSO and EM
- Cai and Dunson (2006):Taylor series expansion for GLMM.
- All these models are parametric ones.

- We use Bayesian shrinkage priors for joint selection of both fixed and random effects
- Desirable properties of such priors: spikes at zero, student t-like tails, simple characterization as a scale mixture of normals to facilitate computaton
- Such shrinkage priors can shrink small coefficients to zero without over-shrink large coefficients.

Outline

1 Introduction

2 model

3 Simulation

4 Application

Shrinkage priors for fixed effects

Stochastic search variable selection (SSVS, George and McCulloch 1993): Introduce latent variable J, $\beta_k^J = (\beta_k, j_k = 1)'$
We use generalized double Pareto, horse shoe prior, and normal-exponential-gamma priors for the fixed effects.

Stochastic search variable selection:SSVS

- Proposed by George and McCulloch (1993), originally for subset selection in linear regression, searching for models with highest posterior probabilities
- starting with full models containing all candidate predictors
- choosing mixture priors that allow predictors to drop by zeroing the coefficients
- running Gibbs sampler with conditional conjugacy from the posterior distribution.

Mixed effects

- $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$, $\boldsymbol{J} = (j_1, \cdots, j_p)'$
  - $j_k = 0$ indicates $\beta_k = 0$
  - $j_k \neq 0$ for $\beta_k \neq 0$
- We take the Zellner g-prior (Zellner and Siow, 1980) $\beta^J \sim N(0, \sigma^2(X^{J'}X^J)^{-1}/g)$, $g \sim Gamma(.)$ and Bernoulli for j_k
- Generally $\zeta \sim N(0, \Omega)$. Application of Cholesky decomposition $\Omega = \boldsymbol{\Lambda\Gamma\Gamma'\Lambda}$ where $\boldsymbol{\Lambda}$ is a positive diagonal matrix with diagonal elements λ_l, $l = 1, \cdots, q$ proportional to the random effects standard deviations, and $\boldsymbol{\Gamma}$ is a lower triangular matrix with diagonal elements equal to 1.

- The generalized double Pareto (GDP) density is defined as $f(x|a,b) = 1/(2a)(1+|x|/(ab))^{-b+1}$, $x \sim GDP(a,b)$.
- The GDP resembles the Laplace density with similar sparse shrinkage properties, also has Cauchy-like tails, which avoids over-shrinkage
- However the posterior form is very complicated and the parameters are not in closed forms. Instead, it can be expressed as a scale mixture of normal distribution.
- $x \sim N(0,\tau)$, $\tau \sim Exp(\lambda^2/2)$, and $\lambda \sim Ga(\alpha,\eta)$, then $x \sim GDP(a=\eta/b, b)$.

$$
\begin{aligned}
\beta_k^j &\sim N(0, \sigma^2/g\kappa_k), \quad \kappa_k \sim \exp(\varphi_k^2/2), \quad \varphi_k \sim Gamma(\nu_1, \vartheta_1), \\
\beta_k^j &\sim N(0, \sigma^2/g\kappa_k^2), \quad \kappa_k \sim Ca^+(0, \nu_2), \quad \nu_2 \sim Ca^+(0, \vartheta_2), \\
\beta_k^j &\sim N(0, \sigma^2/g\kappa_k), \quad \kappa_k \sim \exp(\varsigma^2/2), \quad \varsigma^2 \sim Gamma(\nu_3, \vartheta_3),
\end{aligned}
$$

Model specification

- $$\begin{aligned} y_{ij} &= X'_{ij}\beta + Z'_{ij}\zeta_i + \epsilon_{ij}, \quad V(\zeta_i) = \Omega \\ i &= 1, \cdots, n, \quad j = 1, \cdots, n_i \\ X_{ij} &: q \times 1 \end{aligned}$$

- Cholesky decomposition

$$\begin{aligned} y_{ij} &= X'_{ij}\beta + Z'_{ij}\Lambda\Gamma b_i + \epsilon_{ij}, \quad \Lambda\Gamma V(b_i)\Gamma'\Lambda' = \Omega \\ \Lambda &: Diag(\lambda_1, \cdots, \lambda_q) \end{aligned}$$

Γ: lower triangular matrix $q \times q$ with diagonal element 1.

Prior for random effects

- The random effect with the generalized shrinkage prior is specified as follows:

$$\begin{aligned} b_{ik} &\sim G_k \\ G_k &\sim DP(\alpha G_0^k) \\ G_0^k &\sim GDP(1.0, 1.0) \end{aligned}$$

- The horse shoe prior and normal-exponential-gamma prior can be specified similarly.
- $\Lambda : Diag(\lambda_1, \cdots, \lambda_q)$ Γ: lower triangular matrix $q \times q$ with diagonal element 1.
- $\lambda_k \sim p_{k0}\delta_0 + (1 - p_{k0})N_+(0, \phi_k^2), \phi_k^2 \sim IG(1/2, 1/2)$

Parameter expansion (PX)

- Liu and Wu (1999), Gelman (2004, 2006)
  - $y|\theta, z \sim N(\theta + z, 1), \quad z|\theta \sim N(0, D)$
    θ unknown parameter, z missing
  - $y|\theta, \alpha, w \sim N(\theta - \alpha + w, 1), \quad w|\theta, \alpha \sim N(\alpha, D)$
  - $z \longrightarrow \alpha, w$
- PX breaks association, speeds up convergence.

Outline

1 Introduction

2 model

3 Simulation

4 Application

$$
\begin{aligned}
y_{ij} &= \boldsymbol{X_{ij}}\boldsymbol{\beta} + \boldsymbol{Z_{ij}}\boldsymbol{\zeta_i} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \\
\boldsymbol{X_{ij}} &= (1, x_{ij1}, \cdots, x_{ij15})', \quad x_{ijk} \sim Unif(-2, 2) \\
z_{ijk} &= x_{ijk}, k = 1, \cdots, 4 \\
\text{Case 1:}\boldsymbol{\varrho_i} &\sim N(0, \boldsymbol{\Omega}), \boldsymbol{\Omega} = \begin{pmatrix} 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.9 & 0.48 & 0.06 \\ 0.00 & 0.48 & 0.40 & 0.10 \\ 0.00 & 0.06 & 0.10 & 0.10 \end{pmatrix}. \\
\boldsymbol{\beta} &= (\beta_1, \cdots, \beta_{16})', \beta_1 = \beta_2 = \beta_3 = 1.5, \beta_k = 0, k = 4, ..., 16 \\
\text{Case 2:}q &= 8, (\varrho_{i1}, \cdots \varrho_{i4})' \sim N(0, \boldsymbol{\Omega_1}), \boldsymbol{\Omega} = \boldsymbol{\Omega 1} \text{ same as Case 1} \\
\text{Case 3:}\beta_4 &= \cdots = \beta_9 = 0.01, \text{everything else same as Case 2.}
\end{aligned}
$$

Case	Method	PF	PR	PM	MSE	S.E.
1	GDPP	1.00	0.63	0.63	0.12	0.18
	HSPP	1.00	0.64	0.63	0.11	0.19
	NEGP	0.98	0.59	0.58	0.13	0.19
	KDNP	0.86	0.57	0.46	0.18	0.25
2	GDPP	1.00	0.57	0.55	0.11	0.011
	HSPP	0.97	0.56	0.54	0.12	0.013
	NEGP	0.96	0.54	0.53	0.13	0.014
	KDNP	0.82	0.53	0.44	0.16	0.39
3	GDPP	1.00	0.60	0.60	0.18	0.10
	HSPP	0.99	0.61	0.59	0.20	0.11
	NEGP	0.93	0.58	0.56	0.26	0.15
	KDNP	0.40	0.55	0.23	0.35	0.24

captionComparisons of parameters estimates in the simulation study. S.E. is the standard error of the MSE.PM is the percentage of times the true model was selected, PF and PR are the percentage of times the correct fixed and random effects were selected respectively.

Table: Models with highest posterior probability in the simulation study for Case 3: $*$ is the true model

Model	GDPP	HSPP	NEGP	KDNP
$x_{ij1}, x_{ij2}, x_{ij3}, z_{ij2}, z_{ij3}, z^*_{ij4}$	0.62	0.61	0.59	0.23
$x_{ij1}, x_{ij2}, x_{ij3}, z_{ij1}, z_{ij2}, z_{ij3}, z_{ij4}$	0.37	0.38	0.39	0.15
$x_{ij1}, x_{ij2}, x_{ij3}, x_{ij6}, x_{ij9}, z_{ij2}, z_{ij3}, z_{ij4}$				0.082
$x_{ij1}, x_{ij2}, x_{ij3}, x_{ij5}, x_{ij7}, x_{ij9}, z_{ij2}, z_{ij3}, z_{ij4}$				0.060
$x_{ij1}, x_{ij3}, x_{ij5}, z_{ij2}, z_{ij3}, z_{ij4}$				0.051
$x_{ij1}, x_{ij2}, x_{ij3}, x_{ij9}, z_{ij2}, z_{ij3}, z_{ij4}$				0.031
$x_{ij1}, x_{ij2}, x_{ij3}, x_{ij6}, z_{ij1}, z_{ij2}, z_{ij3}, z_{ij4}$				0.027
$x_{ij1}, x_{ij2}, x_{ij3}, x_{ij8}, x_{ij9}, z_{ij2}, z_{ij3}, z_{ij4}$				0.025
$x_{ij1}, x_{ij2}, x_{ij3}, x_{ij8}, z_{ij1}, z_{ij2}, z_{ij3}, z_{ij4}$				0.023
$x_{ij1}, x_{ij2}, x_{ij3}, x_{ij8}, z_{ij2}, z_{ij3}, z_{ij4}$				0.022

- The shrinkage approaches provide better results than the KDNP.
- Based on extensive simulations, we find that the KDNP is not stable and suffers from numerical issues and fails easily.
- Of all the approaches, GDPP and HSPP perform better than the others in variable selection and parameter estimation. GDPP can be implemented in a more efficient Gibbs sampling algorithm than HSPP, which uses Metropolis-Hasting algorithm.

Outline

1 Introduction

2 model

3 Simulation

4 Application

- A bioassay study conducted to identify the in vivo estrogenic responses to chemicals.
- Data consisted of 2681 female rats, experimented by 19 participating labs in 8 countries from OECD.
- 2 protocols on immature female rats with administration of doses by oral gavage (protocol A) and subcutaneous injection (protocol B).
- 2 other protocols on adult female rats with oral gavage (protocol C) and subcutaneous injection (protocol D).
- Each protocol was comprised of 11 groups, each with 6 rats. 1 untreated group, 1 control group, 7 groups administered with varied doses of ethinyl estradiol (EE) and ZM.

- Goal: how the uterus weights interacts with the in vivo activity of chemical compounds under different experimental conditions.
- There is concerns about heterogeneity across different labs.
- Kannol et. al (2001) analyzed the data with simple parametric methods.
- The number of possible competing models is over 4,000.

Table: Top 10 models with highest posterior probabilities by GDPP, HSPP and NEGP for real data

Model	GDPP	HSPP	NEGP
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij3}, z_{ij4}$	0.098	0.093	0.072
$x_{ij1}, x_{ij3}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij2}$	0.063	0.049	0.097
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij2}, z_{ij4}$	0.047	0.039	0.038
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij4}$	0.038	0.023	0.033
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij3}, z_{ij4}, z_{ij6}$	0.026		
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij2}, z_{ij6}$	0.023		
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij2}, z_{ij5}$	0.020	0.018	0.034
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}$	0.020	0.013	0.046
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij2}, z_{ij3}, z_{ij4}$	0.020	0.025	
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij3}, z_{ij4}, z_{ij5}$	0.019		
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij3}, z_{ij4}, z_{ij6}$		0.015	
$x_{ij1}, x_{ij3}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij3}, z_{ij4}, z_{ij5}$			0.032
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij2}$		0.028	
$x_{ij1}, x_{ij3}, x_{ij4}, x_{ij5}, x_{ij6}, z_{ij1}, z_{ij2}, z_{ij4}, z_{ij5}$		0.015	

Table: Comparisons of parameter estimates in the real data study

Parameter	GDPP	HSPP	NEGP	*LMER*
β_1	3.61[a](3.31, 3.76)[b]	3.60[a](3.35, 3.87)[b]	3.60[a](3.28, 3.96)[b]	3.61[a](3.16, 4.10)[c]
β_2	NA[d]	NA[d]	NA[d]	0.06 (0.02, 0.12)
β_3	1.21 (1.04, 1.47)	1.18 (1.06, 1.53)	1.22 (0.97, 1.49)	1.20 (1.03, 1.69)
β_4	1.20 (0.89, 1.64)	1.21 (0.87, 1.70)	1.19 (0.79, 1.68)	1.27 (0.85, 1.64)
β_5	0.14 (0.13, 0.16)	0.14 (0.13, 0.16)	0.14 (0.13, 0.16)	0.14 (0.13, 0.15)
β_6	-0.48 (-0.57, -0.35)	-0.47 (-0.56, -0.39)	-0.47(-0.59, -0.24)	-0.49 (-0.57, -0.37)

[a] Mean
[b] 95% credible interval
[c] 95% confidence interval
[d] Not selected in the model

- We try to fit the model with KDNP but it fails with too many numerical issues.
- Fixed effect of protocol B is not selected for any model, but the corresponding random effects is, which indicates heterogeneity.
- Both fixed and random effects of protocol C and D are selected in the leading models.
- The fixed effects of EE dose and ZM dose are selected in the leading models, but seldom for the corresponding random effects.

References

- Chen and Dunson Biometrics, 2003
- Cai and Dunson Biometrics, 2006
- Kinney and Dunson Biometrics, 2007
- Bondell et al. Biometrics, 2010
- Ibrahim et al. Biometrics, 2010
- Yang, M, Computational Statistics and Data Analysis, 2012
- Yang, M, Biometrical Journal, 2013
- Yang, M, Under review, 2015