Bayesian Variable Selection with Group-wise Screening

Eunjee Lee

Chungnam National University

eunjee.cnu@gmail.com

May 28, 2021

Overview

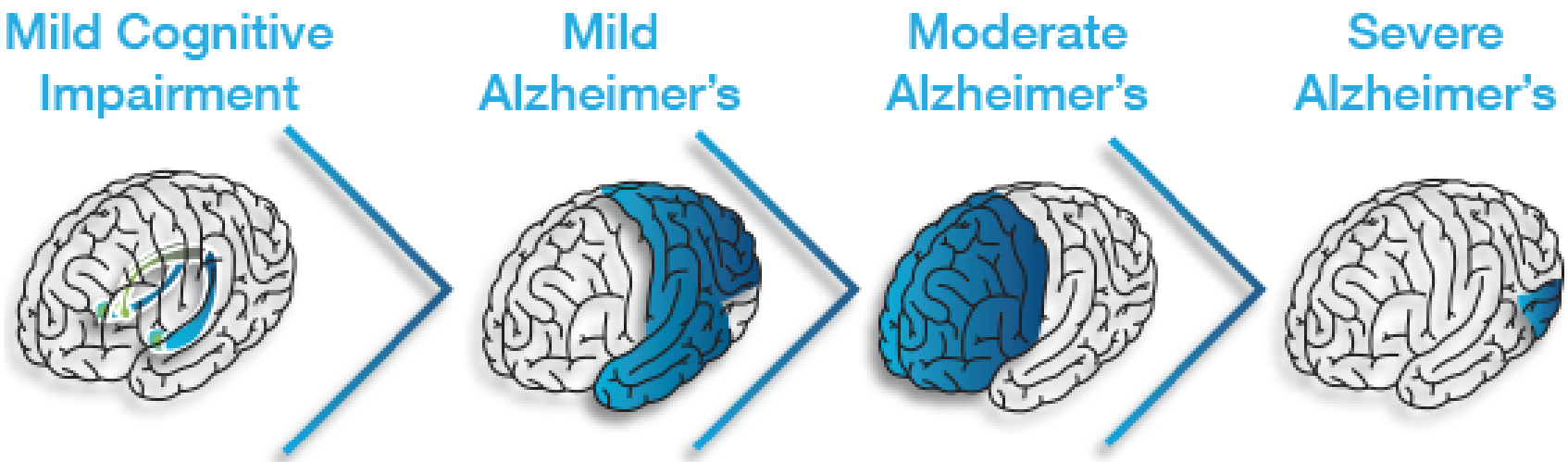
Motivation

Bayesian Bi-level Variable Selection

- Method
- Simulation Study
- Real Data Analysis

Motivation

Background



https://www.mccare.com/education/alzprogression.html

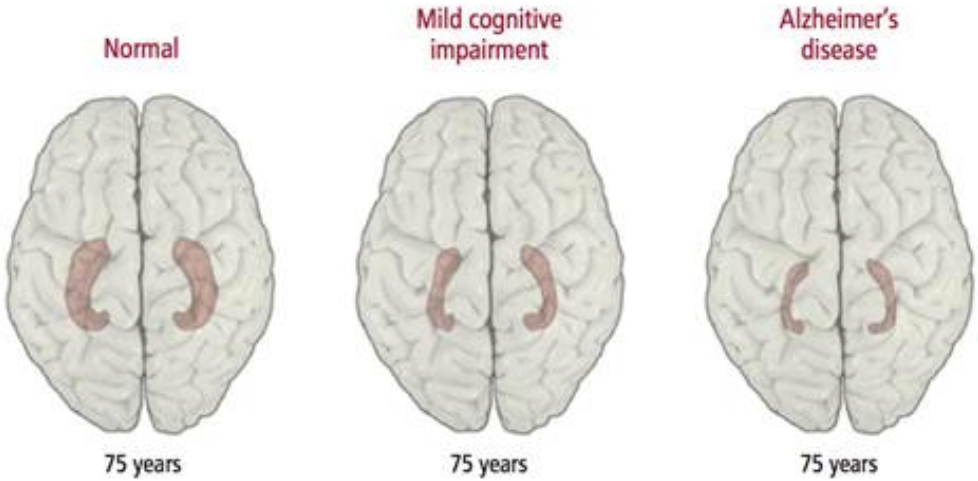
Early Detection of AD

- Improve access to medical and support services.
- Provide an opportunity to plan for the future.
- By beginning treatment early, preserve daily functioning.
- Future treatments could be timely given.

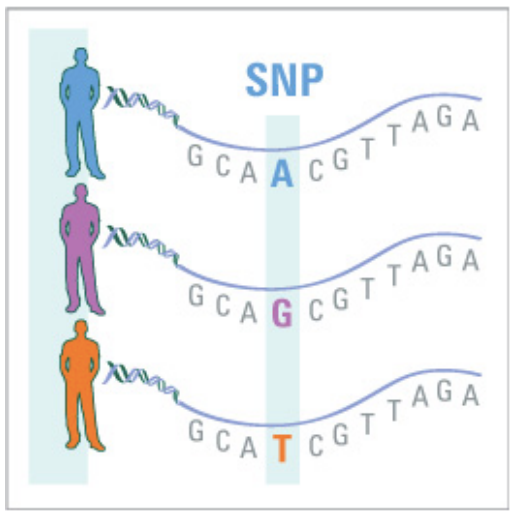
Aim of Study

Development of survival models to predict the time to conversion from MCI to AD by using the following risk factors:

- Hippocampal morphology
- Genetic variants



(a) The shrinking hippocampus



(b) SNP

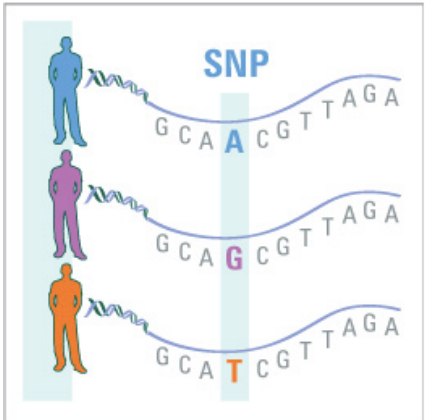
Figure: (a): http://www.helpguide.org , (b): https://www.broadinstitute.org

Bayesian Bi-level Variable Selection

Motivation

GWASs aim to identify of important SNPs to relate to clinical outcomes.

- Conduct multiple comparison.
- Discard a joint structure among SNPs.



Remedies

- Variable selection: Penalization, sure independence screening (SIS), Bayesian approach. $\rightarrow$ No grouping information.
- Zhang and Shen (2012): iBVS for gene and SNP level selection. $\rightarrow$ A prescreening step using marginal tests.

Aim of study

- Detect sparse signals associated with continuous clinical outcomes.
- Consider all the covariates simultaneously especially in (ultra) high-dimensional settings.
- Make use of pre-specified grouping information among the covariates.

Method

Bayesian Bi-level Variable Selection

We proposed a Bayesian bi-level variable selection method.

1. Identify important groups of variables.
   - The number of predictors can be significantly reduced.
   - Censored times will be imputed.
2. Conduct elementwise variable selection.
   - Shrinkage priors can be employed on regression parameters.
   - We extend Dirichlet-Laplace shrinkage priors proposed by Bhattacharya et al. (2014).

Accelerated failure time (AFT) model

Specifies that predictors act multiplicatively on the time to event:

$$Y_i = \exp(-\boldsymbol{x}_i'\boldsymbol{\beta})v_i, i = 1, \cdots, n,$$

which becomes the linear model in log scale

$$\log Y_i = -\boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i, i = 1, \cdots, n,$$

where $Y_1, Y_2, \cdots, Y_n$ are failure times, and ϵ_i is the error term.

Model setting

With predefined G blocks,

$$\log T_i = \boldsymbol{x}_{0i}'\boldsymbol{\beta}_0 + \sum_{g=1}^{G} \gamma_g \boldsymbol{x}_{ig}'\boldsymbol{\beta}_g + \epsilon_i, i = 1, \cdots, n. \quad (1)$$

- $x_{01}, \cdots, x_{0p_0}$: clinical/demographic information
- $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_G$: grouped covariates with coef $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_G)$.
- Group inclusion indicator γ_g
- The error term ϵ_i are iid $N(0, \sigma^2)$.
- Let $w_i = \log(t_i)$ be the augmented data such that

$$\begin{aligned} w_i &= \log(y_i) \quad \text{if } \nu_i = 1 \text{ (non-censored)}, \\ w_i &> \log(y_i) \quad \text{if } \nu_i = 0 \text{ (censored)}. \end{aligned} \quad (2)$$

Grouplevel selection: Priors

$$
\begin{aligned}
\boldsymbol{\beta}_0|\sigma^2 &\sim N(\mathbf{0}, \sigma^2 h_0 \mathbf{I}_{\mathbf{p_0}}) \qquad (3)\\
\boldsymbol{\beta}_g|\sigma^2 &\sim N(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_g),\ \boldsymbol{\Sigma}_g = \begin{cases} (\boldsymbol{X}_g'\boldsymbol{X}_g)^{-1}, \text{ if } k_g \leq n; \\ (\boldsymbol{X}_g'\boldsymbol{X}_g + \lambda \boldsymbol{I}_{k_g})^{-1}, \text{ if } k_g > n, \end{cases} \\
& \quad g = 1, \cdots, G \\
\sigma^2 &\sim \text{IG}(\nu_0/2, \nu_0\sigma_0^2/2) \\
\gamma_g &\sim \text{Bernoulli}(p_g) \\
p_g &\sim \text{Beta}(a, b)
\end{aligned}
$$

We assume IM/IMR priors on the regression coefficients $\boldsymbol{\beta}_g$, $g = 1, \cdots, G$.

Grouplevel selection: Full posteriors

The full posterior distribution of $(\beta_0, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2)$ is given by

$$
\begin{aligned}
L(\beta_0, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2 | \boldsymbol{w}, \boldsymbol{X}) &\propto L(\boldsymbol{w} | \boldsymbol{X}, \beta_0, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}) \pi(\beta_0 | \sigma^2) \pi(\boldsymbol{\beta} | \sigma^2) \pi(\boldsymbol{\gamma}) \pi(\sigma^2) \\
&\propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(w_i - \boldsymbol{x}'_{i0}\beta_0 - \sum_{g=1}^{G} \gamma_g \boldsymbol{x}'_{ig} \boldsymbol{\beta}_g \right)^2 \right\} \\
&\quad \times \exp\left\{ -\frac{1}{2h_0\sigma^2} \beta'_0 \beta_0 \right\} \times \prod_{g=1}^{G} \exp\left\{ -\frac{1}{2c_0\sigma^2} \boldsymbol{\beta}'_g \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\beta}_g \right\} \\
&\quad \times (\sigma^2)^{-\nu_0/2-1} \exp\left(-\frac{\nu_0 \sigma_0^2}{2\sigma^2} \right) \\
&\quad \times \prod_{g=1}^{G} p_g^{\gamma_g} (1 - p_g)^{1-\gamma_g} \times \prod_{g=1}^{G} \frac{1}{B(a,b)} p_g^{a-1} (1 - p_g)^{b-1}.
\end{aligned}
$$

Grouplevel selection: MCMC sampling

1. Given $\boldsymbol{\gamma}_{(g)}$, update γ_g from Bernoulli distribution with success probability $\frac{A}{A+B}$, where

$$\begin{aligned} A &= f_t(\boldsymbol{w}|\nu_0, \sigma_0(A_{\gamma_{(g)}} + c_0\gamma_g \boldsymbol{X}_g \boldsymbol{\Sigma}_g \boldsymbol{X}_g')) \times p_g, \\ B &= f_t(\boldsymbol{w}|\nu_0, \sigma_0 A_{\gamma_{(g)}}) \times (1 - p_g), \end{aligned}$$

   and $A_{\gamma_{(g)}} = \boldsymbol{I} + h_0 \boldsymbol{X}_0 \boldsymbol{X}_0' + c_0 \sum_{k \neq g}^{G} \gamma_k \boldsymbol{X}_k \boldsymbol{\Sigma}_k \boldsymbol{X}_k'$.

2. Update p_g from Beta$(a + \gamma_g, b + 1 - \gamma_g)$.

3. Update the censored element of $\boldsymbol{w}$ from

$$w_i|\boldsymbol{w}^{(i)}, \boldsymbol{X}, \boldsymbol{\gamma} \sim t_{n+\nu_0-1}(\mu_{w_i}, s_{w_i}), \quad w_i > \log(y_i).$$

Select groups whose posterior inclusion probabilities > 0.5.

Elementwise selection: Priors

Denote $\boldsymbol{X}_* = [\boldsymbol{X}_1^*, \cdots, \boldsymbol{X}_Q^*]$.

$$\log \tilde{Y}_i = \tilde{W} = \boldsymbol{x}_{0i}'\boldsymbol{\beta}_0 + \sum_{g=1}^{Q} \boldsymbol{x}_{ig}^{*\prime}\boldsymbol{\theta}_g + \epsilon_i, i = 1, \cdots, n. \quad (4)$$

For $g = 1, 2, \cdots, Q$, the priors are set to be

$$\begin{aligned}
\boldsymbol{\theta}_g|\sigma^2, \boldsymbol{\psi}_g, \boldsymbol{\phi}_g, \tau_g &\sim N(0, \sigma^2\boldsymbol{\Sigma}_g^*) \quad (5)\\
\sigma^2 &\sim \text{IG}(\nu_0/2, \nu_0\sigma_0^2/2)\\
\psi_{gj} &\sim \text{Exp}(1/2),\ j = 1, \cdots, q_g\\
(\phi_{g1}, \cdots, \phi_{gq_g}) &\sim \text{Dir}(a_g, \cdots, a_g)\\
\tau_g &\sim \text{gamma}(q_g a_g, 1/2)\\
a_g &\sim \text{Discrete uniform from } \frac{1}{q_g} \text{ to } 1/2 \text{ with length } 50,
\end{aligned}$$

where $\boldsymbol{\Sigma}_g^* = \text{diag}(\psi_{g1}\phi_{g1}^2\tau_g^2, \cdots, \psi_{gq_g}\phi_{gq_g}^2\tau_g^2)$.

Elementwise selection: MCMC sampling

1. Update $\boldsymbol{\beta}_0$ from

$$p(\boldsymbol{\beta}_0|-) \sim N_{p_0}\left(\left(\boldsymbol{X}_0'\boldsymbol{X}_0 + \frac{1}{h_0}\boldsymbol{I}\right)^{-1}\boldsymbol{X}_0'(\tilde{\boldsymbol{w}} - \boldsymbol{X}^*\boldsymbol{\theta}), \sigma^2\left(\boldsymbol{X}_0'\boldsymbol{X}_0 + \frac{1}{h_0}\boldsymbol{I}\right)^{-1}\right)$$

2. Update $\boldsymbol{\theta}_g$ from its full conditional distribution $N_{q_g}(\widetilde{\boldsymbol{\mu}}_g, \widetilde{\boldsymbol{\Sigma}}_g)$, where

$$\begin{aligned}\widetilde{\boldsymbol{\mu}}_g &= (\boldsymbol{X}_g^{*\prime}\boldsymbol{X}_g{}^* + \boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{X}_{\boldsymbol{g}}^{*\prime}(\tilde{\boldsymbol{w}} - \boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_{(g)}^*\boldsymbol{\theta}_{(g)}),\\ \widetilde{\boldsymbol{\Sigma}}_g &= \sigma^2(\boldsymbol{X}_g^{*\prime}\boldsymbol{X}_g^* + \boldsymbol{\Sigma}_g^{-1})^{-1}.\end{aligned}$$

3. Let $N = n + q + p_0 + \nu_0$ and $\boldsymbol{\eta} = \boldsymbol{X}_0'\boldsymbol{\beta}_0 + \boldsymbol{X}_*'\boldsymbol{\theta}$.
   Update σ^2 from

$$p(\sigma^2|-) \sim IG\left(\frac{N}{2}, \frac{1}{2}\left\{\nu_0\sigma_0^2 + ||\tilde{\boldsymbol{w}} - \boldsymbol{\eta}||^2 + \frac{\boldsymbol{\beta}_0'\boldsymbol{\beta}_0}{h_0} + \boldsymbol{\theta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}\right\}\right).$$

Elementwise selection: MCMC sampling

4. Independently sample ψ_{gj} from

$$p(\psi_{gj}|-) \sim IG\left(\frac{\phi_{gj}\tau_g\sigma}{|\theta_{gj}|}, 1\right).$$

5. Update τ_g from

$$p(\tau_g|-) \sim \text{giG}\left(q_g \times a_g - q_g, 1, 2\sum_{j=1}^{q_g}\frac{|\theta_{gj}|}{\phi_{gj}\sigma}\right).$$

6. Update ϕ_{gj}, where $\phi_{gj} = T_{gj}/T_g$ such that

$$p(T_{gj}|-) \sim \text{giG}\left(a_g - 1, 1, 2\frac{|\theta_{gj}|}{\sigma}\right).$$

7. Update a_g from $\text{MN}(1, \tilde{p}_1/\tilde{p}, \cdots, \tilde{p}_{50}/\tilde{p})$, where $\tilde{p} = \sum_{l=1}^{50}\tilde{p}_l$ and

$$\tilde{p}_l = \exp\left((u_l - 1)\sum_{j=1}^{q_g}\log(\phi_{gj}) + (q_g u_l - 1)\log(\tau_g) - \log 50\right).$$

Elementwise selection: After MCMC

The DL shrinkage prior does not give exactly zero coefficient values.

- A posterior sample of $\theta_j^{(m)}$'s consists of two groups: (1) nearly zero, and (2) far from zero.
- The absolute values of $\theta_1^{(m)}, \cdots, \theta_q^{(m)}$ are skewed to the right.
- At the m-th iteration, conduct k-means clustering on $\sqrt{|\theta_1^{(m)}|}, \cdots, \sqrt{|\theta_q^{(m)}|}$ with 2 groups and let h_m be the numbers of elements in the (2) group.
- Set the number of signals as $H = \text{Mode}(h_1, \cdots, h_M)$.

Simulation Study

Simulation setup: SNP data generation

SNP data is simulated from the Hapmap projects 2009-02 phaseIII data (Consortium et al., 2010).

- Formation of SNP sets: Form SNP-sets by LD blocks.
- Among blocks whose sizes are larger than 30, we randomly selected 2000 SNP sets.

Finally, we obtained about 45,000 SNPs with 2000 SNP sets.

Simulation setup: Model and parameters setting

$$\text{Model: } \log Y_i = \boldsymbol{x}_{0i}'\boldsymbol{\beta}_0 + \sum_{g=1}^{G} \gamma_g \boldsymbol{x}_{ig}'\boldsymbol{\beta}_g + \epsilon_i, i = 1, \cdots, n.$$

- Time to event outcome was generated from the above model, where $\gamma_j = 1, j = 1, \cdots, 10$ and $\gamma_{j'} = 0, j' = 11, \cdots, 2000$.
- Within the 10 relevant blocks, we randomly selected 10 SNPs.
- All the elements in $\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_{10} \sim N(-1, 0.5)$.
- $x_0 \sim Unif(0, 1)$ with $\beta_0 = 0.1$.
- Censoring times $\sim$ Uniform$(0, c^*)$.

We replicated the simulation 50 times.

Hyperparameters

- We assumed the inclusion indicator $\gamma_g \sim \text{Beta}(10, 190)$.
- $c_0 = h_0 = 10$: small impact of λ and X on the posterior analysis of $\boldsymbol{\theta}$.
- $\nu_0 = 3, \sigma_0 = 1$: relatively flat prior on σ^2.

Competing methods

We consider the following variable selection penalties:

- Group-level selection
  - Group Lasso (grLasso) by Yuan and Lin (2006)
  - Group MCP (grMCP) by Zhang (2010)
- Bi-level selection
  - Group bridge (gBridge) by Huang et al. (2009)
  - Group exponential lasso (gel) by Breheny (2015)
  - Composite MCP (cMCP) by Breheny and Huang (2009)

Performance measures

We consider the following performance measures:

- True positive rate (TPR or sensitivity)
- True negative rate (TNR or specificity)
- Positive predictive value (PPV)
- Negative predictive value (NPV)

		Estimated	
		Yes	No
Actual	Yes	TP	FN
	No	FP	TN

$$TPR = \frac{TP}{P}, TNR = \frac{TN}{N}, PPV = \frac{TP}{TP + FP}, NPV = \frac{TN}{TN + FN}$$

Under the true model, all the four rates are equal to one.

Simulation results: Non-censored case

Table: Bi-level variable selection methods include BBVS, gBridge, gel, cMCP , while grMCP, grSCAD, grLasso only enable group-level variable selection. When the group-level variable selection performs perfectly, TP=10, FP=0, TPR=TNR=PPV=NPV=1.

	TP	FP	TPR	TNR	PPV	NPV
BBVS	10.0 (0.0)	0.0 (0.0)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
gBridge	9.9 (0.1)	0.0 (0.0)	0.99 (0.01)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
gel	9.8 (0.1)	0.0 (0.0)	0.98 (0.01)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
grMCP	10.0 (0.0)	0.3 (0.1)	1.00 (0.00)	1.00 (0.00)	0.97 (0.01)	1.00 (0.00)
grSCAD	10.0 (0.0)	11.5 (0.4)	1.00 (0.00)	1.00 (0.00)	0.47 (0.01)	1.00 (0.00)
grLasso	10.0 (0.0)	19.2 (0.5)	1.00 (0.00)	0.99 (0.00)	0.35 (0.01)	1.00 (0.00)
cMCP	10.0 (0.0)	73.1 (4.3)	1.00 (0.00)	0.96 (0.00)	0.14 (0.01)	1.00 (0.00)

Simulation results: Non-censored case

Table: Bi-level variable selection methods include BBVS, gBridge, gel, cMCP. When the element-wise variable selection performs perfectly, TPR=TNR=PPV=NPV=1.

	TPR	TNR	PPV	NPV
BBVS	0.686 (0.012)	0.999 (0.000)	0.616 (0.009)	1.000 (0.000)
gBridge	0.643 (0.011)	0.999 (0.000)	0.503 (0.008)	1.000 (0.000)
gel	0.651 (0.012)	0.999 (0.000)	0.441 (0.009)	1.000 (0.000)
cMCP	0.306 (0.009)	0.998 (0.000)	0.165 (0.009)	0.999 (0.000)

Simulation results: Censored case

Table: It shows group-level selection results. When the group-level variable selection performs perfectly, TP=10, FP=0, TPR=TNR=PPV=NPV=1.

	TP	FP	TPR	TNR	PPV	NPV
BBVS	9.7 (0.1)	0.0 (0.0)	0.97 (0.01)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)

Table: It shows element-wise selection results.When the element-wise variable selection performs perfectly, TPR=TNR=PPV=NPV=1.

	TPR	TNR	PPV	NPV
BBVS	0.634 (0.012)	0.999 (0.000)	0.589 (0.011)	1.000 (0.000)

Real Data Analysis

ADNI-1 data analysis: Formation of SNP sets

We followed one of the strategies proposed by Wu et al. (2010):

1. Stringent quality control (QC) step on the raw genotype data.
2. Take all SNPs within 20kbp upstream and downstream of the gene as a single SNP set.
3. Merge several SNP sets that have overlapping bp regions.
4. Remove duplicated SNPs and the groups with size 1.

Finally, we obtained 6073 gene-based SNP sets with 280,587 SNPs.

ADNI-1 Data analysis: Framework

- A response, y_i: the time to conversion to AD from MCI
- Scalar covariates $\boldsymbol{X}_0$
  - clinical data: Gender, age at the baseline
  - The first 5 PCs for the whole SNP data to adjust for population stratification.
- Covriates with block information $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{6073}$: the SNP data consisting with 6073 groups.

ADNI-1 data analysis: Variable selection results

Table: BBVS identifies 13 SNP sets. It shows the list of selected SNP-sets associated with time to conversion from MCI to AD. 78 SNP were identified among the included 2126 SNPs of 13 SNP sets.

Chr	Start bp	End bp	# of SNPs	Gene	Related diseases
1	41924445	42521597	105	HIVEP3	Parkinson's disease hippocampal alterations
9	118896070	120197318	341	ASTN2	Cognitive decline, reduced hippocampal volume
13	96065852	97511817	193	DZIP1	Alzheimer's disease
13	23735059	24081604	96	SGCG	Type 2 diabetes
16	82640398	83866608	661	CDH13	Adult ADHD
16	78113309	79266565	422	WWOX	Alzheimer's disease
21	43599798	44021551	185	ABCG1	Alzheimer's disease

Discussion

- The BBVS was developed to enable bi-level variable selection .
- DL priors were adapted to reflect the grouping information in the element-wise variable selection.
- Our method can be directly used for any continuous/binary outcomes, such as brain MRI volumes, behavior score, and presence of disease.

Key References

1. Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2005). Bayesian Survival Analysis. Wiley Online Library. Verlag, New York.
2. Müller, H. G., Stadtmüller, U. (2005). Generalized functional linear models. Annals of Statistics, 774-805.
3. Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2014). Dirichlet-Laplace priors for optimal shrinkage. Journal of the American Statistical Association.
4. Sha, N., Tadesse, M. G., and Vannucci, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. Bioinformatics, 22(18), 2262-2268.
5. Luo, L., Boerwinkle, E., & Xiong, M. (2011). Association studies for next-generation sequencing. Genome research, 21(7), 1099-1108.

Thank you!