Expressing model uncertainty in Bayesian variable selection using credible sets

Jim Griffin
University College London

Introduction

Variable selection is often needed when there are a large number of potential covariates that could explain the variation in a response variable.

Introduction

Variable selection is often needed when there are a large number of potential covariates that could explain the variation in a response variable.

This often motivated by

- Avoiding overfitting.
- Understanding factors which affect the response variable.

Introduction

Variable selection is often needed when there are a large number of potential covariates that could explain the variation in a response variable.

This often motivated by

- Avoiding overfitting.
- Understanding factors which affect the response variable.

There are a number of classical approaches: subset selection, stepwise selection, penalized maximum likelihood (Lasso, elastic net, MCP, etc.).

Bayesian variable selection

Assume a parametric model $y \sim f(x^\gamma, \theta)$ where x^γ is a subset of included variables indexed by γ ($\gamma_i = 1$ if the i-th variable is included and 0 otherwise).

Put a prior on γ. For example, $\gamma_i \sim \text{Bernoulli}(\pi)$, $\pi \sim \text{Be}(a, b)$ then a and b can be chosen to encourage sparsity.

This leads to a posterior distribution $p(\gamma \mid \text{data})$, which expresses our uncertainty about γ.

Bayesian variable selection

Good theoretical properties (Castillo et al., 2015) and performance (Porwal and Raftery, 2022)

Bayesian variable selection

Good theoretical properties (Castillo et al., 2015) and performance (Porwal and Raftery, 2022)

Recent work on high-dimensional problems

- MCMC methods – Importance Tempering (Zanella and Roberts, 2019), ASI (Griffin et al., 2021), PARNI (Liang et al., 2022)
- Stochastic search – SVEN (Li et al., 2023)

Outputs from Bayesian variable selection

Bayesian model averaged predictions are provided by weighting predictions from each model by their posterior probability.

Outputs from Bayesian variable selection

Bayesian model averaged predictions are provided by weighting predictions from each model by their posterior probability.

To understand the relative importance of different variables, there are summaries

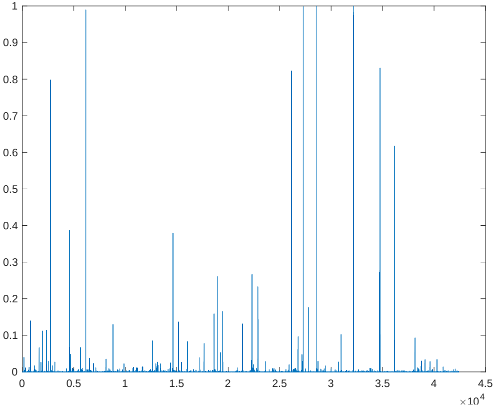
- Posterior inclusion probabilities (PIPs): $p(\gamma_i \mid \text{Data})$
- Maximum a posterior (MAP) model: the mode of $\gamma \mid \text{Data}$.
- Median model $\hat{\gamma}$ where $\hat{\gamma}_i = \mathrm{I}(p(\gamma_i \mid \text{Data}) > 0.5)$.

GWAS example: Systemic Lupus Erythematosus (case-control study)

chromosome 3
$n = 10,995$
Cases: 4,036
Controls: 6,959

$p = 42,430$

Median model has 13 variables



These are summaries of importance but don't represent any relationships between variables included in models.

These are summaries of importance but don't represent any relationships between variables included in models.

Under an independent prior on γ, uncorrelated variables in linear models $\iff$ independence of γ_i's.

These are summaries of importance but don't represent any relationships between variables included in models.

Under an independent prior on γ, uncorrelated variables in linear models $\iff$ independence of γ_i's.

These relationships are due to multi-collinearity (*i.e.* correlation between variables). For example, due to linkage disequilibrium in GWAS.

Simulated example (George and McCulloch, 1997)

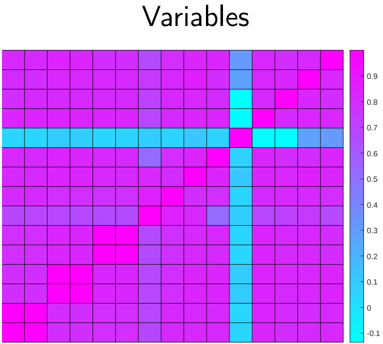
Linear regression example with $n = 180$ and $p = 15$.

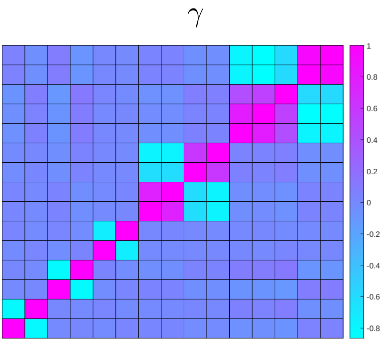
Non-zero regression coefficients are 1, 3, 5, 7, 8, 11, 12, 13.

Strong multicollinearity between variables:

- 1 and 2
- 3 and 4
- 5 and 6
- 7, 8, 9, 10
- 11, 12 13, 14, 15

Simulated example - Correlation





What's this got to do with Bayesian nonparametrics?

Variable selection in BART or Gaussian process regression.

What's this got to do with Bayesian nonparametrics?

Variable selection in BART or Gaussian process regression.

Bayesian variable selection leads to a posterior distribution on a high-dimensional discrete space. The same is true of a lot of Bayesian nonparametric methods (*e.g.* clustering, feature allocation).

What's this got to do with Bayesian nonparametrics?

Variable selection in BART or Gaussian process regression.

Bayesian variable selection leads to a posterior distribution on a high-dimensional discrete space. The same is true of a lot of Bayesian nonparametric methods (*e.g.* clustering, feature allocation).

There are summarisations methods for some problems (particularly clustering). How to represent uncertainty?

Credible sets

Let Γ be the set of all possible combination of variables then $A \subset \Gamma$ is a $100\alpha\%$ credible set (CS) if $p(A \mid \text{Data}) \geq \alpha$.

Credible sets

Let Γ be the set of all possible combination of variables then $A \subset \Gamma$ is a $100\alpha\%$ credible set (CS) if $p(A \mid \text{Data}) \geq \alpha$.

The smallest $100\alpha\%$ CS can be found by

1. Rank the models by decreasing probability,

$$p(\gamma^{(1)} \mid \text{Data}) \geq p(\gamma^{(2)} \mid \text{Data}) \geq p(\gamma^{(3)} \mid \text{Data}) \geq \cdots \geq p(\gamma^{(2^p)} \mid \text{Data})$$

2. Find the smallest K such that $\sum_{k=1}^{K} p(\gamma^{(k)} \mid \text{Data}) \geq \alpha$ then $\left\{\gamma^{(1)}, \gamma^{(2)}, \ldots, \gamma^{(K)}\right\}$ is the smallest 100% CS.

Simulated example (smallest 50% credible set)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Prob
1	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0.0849
1	0	1	0	1	0	0	0	1	1	0	0	0	1	1	0.0817
1	0	0	1	1	0	0	0	1	1	1	1	1	0	0	0.0424
1	0	0	1	1	0	0	0	1	1	0	0	1	1	1	0.0396
0	1	1	0	1	0	0	0	1	1	0	0	0	1	1	0.0345
0	1	0	1	1	0	0	0	1	1	0	0	0	1	1	0.0338
1	0	1	0	1	0	0	0	1	1	0	0	1	1	1	0.0279
0	1	0	1	1	0	0	0	1	1	1	1	1	0	0	0.0264
0	1	0	1	1	0	0	0	1	1	0	0	1	1	1	0.0248
1	0	1	0	1	0	0	0	1	1	1	1	1	0	0	0.0218
1	0	1	0	0	1	0	0	1	1	0	0	0	1	1	0.0218
1	0	0	1	0	1	0	0	1	1	0	0	0	1	1	0.0202
0	1	1	0	1	0	0	0	1	1	0	0	1	1	1	0.0183
1	0	1	1	1	0	0	0	1	1	0	0	0	1	1	0.0176
0	1	1	0	1	0	0	0	1	1	1	1	1	0	0	0.0142

Simulated example (smallest 50% credible set)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Prob
1	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0.0849
1	0	1	0	1	0	0	0	1	1	0	0	0	1	1	0.0817
1	0	0	1	1	0	0	0	1	1	1	1	1	0	0	0.0424
1	0	0	1	1	0	0	0	1	1	0	0	1	1	1	0.0396
0	1	1	0	1	0	0	0	1	1	0	0	0	1	1	0.0345
0	1	0	1	1	0	0	0	1	1	0	0	0	1	1	0.0338
1	0	1	0	1	0	0	0	1	1	0	0	1	1	1	0.0279
0	1	0	1	1	0	0	0	1	1	1	1	1	0	0	0.0264
0	1	0	1	1	0	0	0	1	1	0	0	1	1	1	0.0248
1	0	1	0	1	0	0	0	1	1	1	1	1	0	0	0.0218
1	0	1	0	0	1	0	0	1	1	0	0	0	1	1	0.0218
1	0	0	1	0	1	0	0	1	1	0	0	0	1	1	0.0202
0	1	1	0	1	0	0	0	1	1	0	0	1	1	1	0.0183
1	0	1	1	1	0	0	0	1	1	0	0	0	1	1	0.0176
0	1	1	0	1	0	0	0	1	1	1	1	1	0	0	0.0142

Strategy

Other 100α% CS may be easier to understand and calculable using MCMC output.

The strategy is

- Remove variables with low PIPs.
- Partition remaining variables into approximately uncorrelated blocks
- Approximate the distribution in each block.
- Construct the credible sets from the approximation.

Estimating the correlation structure

Calculate the correlation $\rho_{ij} = \text{Correlation}(\gamma_i, \gamma_j)$ under the posterior distribution.

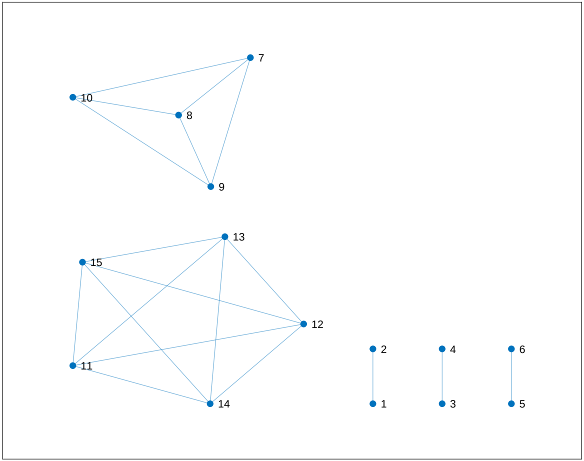
Estimating the correlation structure

Calculate the correlation $\rho_{ij} = \text{Correlation}(\gamma_i, \gamma_j)$ under the posterior distribution.

Define the matrix A by $A_{ij} = \mathrm{I}(|\rho_{ij}| > \tau)$ for some user-chosen threshold τ.

Find the components of the graph defined by the adjacency matrix A.

Simulated example



Choosing τ

Smaller τ leads to

- Larger components
- Smaller credible sets
- Harder to understand and compute the approximation

Multivariate Bernoulli distribution (Dai et al., 2013)

Let $\mathcal{D}$ be the set of non-empty subsets of $\{1, 2, \ldots, K\}$, *i.e.* $\mathcal{D} = \{\{1\}, \{2\}, \ldots, \{1, 2, \ldots, K\}\}$.

The K-dimensional multivariate Bernoulli distribution with parameters $\mathbf{f} = (f^{\epsilon} \in \mathbb{R} \mid \epsilon \in \mathcal{D})^T$ has the log probability mass function

$$\log p(y) = \sum_{r=1}^{K} \left(\sum_{1 \leq j_1 < j_2 < \cdots < j_r \leq K} f^{j_1 j_2 \cdots j_r} B^{j_1 j_2 \cdots j_r} \right) - b(\mathbf{f})$$

where $B^{j_1 j_2 \cdots j_r}(y) = y_{j_1} y_{j_2} \ldots y_{j_r}$ and $b(\mathbf{f})$ is the log normalizing constant.

Properties

- The multivariate Bernoulli distribution is a member of the exponential family and $\mathbf{f}$ are the natural parameters.
- These natural parameters can be linked to the general parameters using the relationship

$$\exp\{f^{j_1 j_2 \cdots j_r}\} = \frac{p\begin{pmatrix}\text{even \# zeros among } j_1, j_2, \ldots, j_r \text{ components} \\ \text{and other components are all zero}\end{pmatrix}}{p\begin{pmatrix}\text{odd \# zeros among } j_1, j_2, \ldots, j_r \text{ components} \\ \text{and other components are all zero}\end{pmatrix}}.$$

Properties

- For random vector $Y = (Y_1, \ldots, Y_K)$ following the multivariate Bernoulli distribution, suppose there are two blocks of nodes $Y' = (Y_1, Y_2, \ldots, Y_r)$ and $Y''. = (Y_{r+1}, Y_{r+2}, \ldots, Y_s)$, and denote index set $\tau_1 = \{1, 2, \ldots, r\}$ and $\tau_2 = \{r+1, r+2, \ldots, s\}$. Then Y' and Y'' are independent if and only if

$$f^{\tau} = 0, \forall \tau \cap \tau_1 = \emptyset \text{ and } \tau \cap \tau_2 = \emptyset.$$

Properties

- For random vector $Y = (Y_1, \ldots, Y_K)$ following the multivariate Bernoulli distribution, suppose there are two blocks of nodes $Y' = (Y_1, Y_2, \ldots, Y_r)$ and $Y''. = (Y_{r+1}, Y_{r+2}, \ldots, Y_s)$, and denote index set $\tau_1 = \{1, 2, \ldots, r\}$ and $\tau_2 = \{r+1, r+2, \ldots, s\}$. Then Y' and Y'' are independent if and only if

$$f^{\tau} = 0, \forall \tau \cap \tau_1 = \emptyset \text{ and } \tau \cap \tau_2 = \emptyset.$$

- Restricting the model to only first and second order terms, (*i.e.* $f^{j_1 j_2 \cdots j_r} = 0$ for all $j_1 j_2 \ldots j_r$ with $r > 2$) leads the quadratic exponential binary model (Cox and Weimurth, 1994).

Approximation

Suppose there are r blocks $\gamma_{m_1}, \ldots, \gamma_{m_r}$ and $q(\gamma|\mathbf{f})$ is the approximating multivariate Bernoulli distribution.

$$\begin{aligned}
\mathrm{KL} &= \sum p(\gamma \mid \text{Data}) \log p(\gamma) - \sum p(\gamma \mid \text{Data}) \log q(\gamma) \\
&= C - \sum p(\gamma \mid \text{Data}) \log q(\gamma \mid \mathbf{f}) \\
&= C - \sum_{j=1}^{q} p(\gamma_{m_j} | \text{Data}) \log q(\gamma_{m_j} \mid \mathbf{f})
\end{aligned}$$

Approximation

Suppose there are r blocks $\gamma_{m_1}, \ldots, \gamma_{m_r}$ and $q(\gamma|\mathbf{f})$ is the approximating multivariate Bernoulli distribution.

$$\begin{aligned}
\mathrm{KL} &= \sum p(\gamma \mid \mathrm{Data}) \log p(\gamma) - \sum p(\gamma \mid \mathrm{Data}) \log q(\gamma) \\
&= C - \sum p(\gamma \mid \mathrm{Data}) \log q(\gamma \mid \mathbf{f}) \\
&= C - \sum_{j=1}^{q} p(\gamma_{m_j}|\mathrm{Data}) \log q(\gamma_{m_j} \mid \mathbf{f})
\end{aligned}$$

If there is a sample $\gamma^{(1)}, \gamma^{(2)}, \ldots, \gamma^{(N)} \sim p(\gamma \mid \mathrm{Data})$ then the a Monte Carlo approximation to the KL divergence is used

$$-\frac{1}{N} \sum_{j=1}^{q} \sum_{i=1}^{N} \log q\left(\gamma_{m_j}^{(i)} \middle| \mathbf{f}\right)$$

Finding the credible set

Suppose there are r blocks and let Γ_i be all models formed from the variables in the i-th block. Let S_i be a subset of Γ_i.

Finding the credible set

Suppose there are r blocks and let Γ_i be all models formed from the variables in the i-th block. Let S_i be a subset of Γ_i.

The credible set $\mathcal{S}$ is a Cartesian product of $S_1, \ldots, S_r$ and then

$$p(\mathcal{S} \mid \text{Data}) = \prod_{i=1}^{r} p(Q_i \mid \text{Data}).$$

Finding the credible set

Suppose there are r blocks and let Γ_i be all models formed from the variables in the i-th block. Let S_i be a subset of Γ_i.

The credible set $\mathcal{S}$ is a Cartesian product of $S_1, \ldots, S_r$ and then

$$p(\mathcal{S} \mid \text{Data}) = \prod_{i=1}^{r} p(Q_i \mid \text{Data}).$$

This allows the derivation of algorithms which control $p(\mathcal{S} \mid \text{Data})$ by changing the elements of $S_1, \ldots, S_r$.

Example (3 variables / 2 blocks)

$$\Gamma_1 = \{0, 1\}, \Gamma_2 = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$$

	Block 1		Block 2			
	0	1	(0, 0)	(0, 1)	(1, 0)	(1, 1)
Prob	0.9	0.1	0	0.5	0.5	0

Example (3 variables / 2 blocks)

$$\Gamma_1 = \{0, 1\}, \Gamma_2 = \{(0,0), (0,1), (1,0), (1,1)\}$$

	Block 1		Block 2			
	0	1	(0, 0)	(0, 1)	(1, 0)	(1, 1)
Prob	0.9	0.1	0	0.5	0.5	0

$$S_1 = \{0, 1\} \Rightarrow p(S_1) = 1,$$
$$S_2 = \{(0,0), (0,1), (1,0), (1,1)\} \Rightarrow p(S_2) = 1,$$
$$\mathcal{S} = \{(0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0), (1,1,1)\} \Rightarrow p(\mathcal{S}) = 1.$$

Example (3 variables / 2 blocks)

$$\Gamma_1 = \{0, 1\}, \Gamma_2 = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$$

	Block 1		Block 2			
	0	1	(0, 0)	(0, 1)	(1, 0)	(1, 1)
Prob	0.9	0.1	0	0.5	0.5	0

$$S_1 = \{0, 1\} \Rightarrow p(S_1) = 1,$$
$$S_2 = \{(0, 0), (0, 1), (1, 0), (1, 1)\} \Rightarrow p(S_2) = 1,$$
$$\mathcal{S} = \{(0, 0, 0), (0, 0, 1), (0, 1, 0), (0, 1, 1), (1, 0, 0), (1, 0, 1), (1, 1, 0), (1, 1, 1)\} \Rightarrow p(\mathcal{S}) = 1.$$

$$S_1 = \{0\} \Rightarrow p(S_1) = 0.9,\ S_2 = \{(0, 1), (1, 0)\} \Rightarrow p(S_2) = 1$$
$$\mathcal{S} = \{(0, 0, 1), (0, 1, 0)\} \Rightarrow p(\mathcal{S}) = 0.9.$$

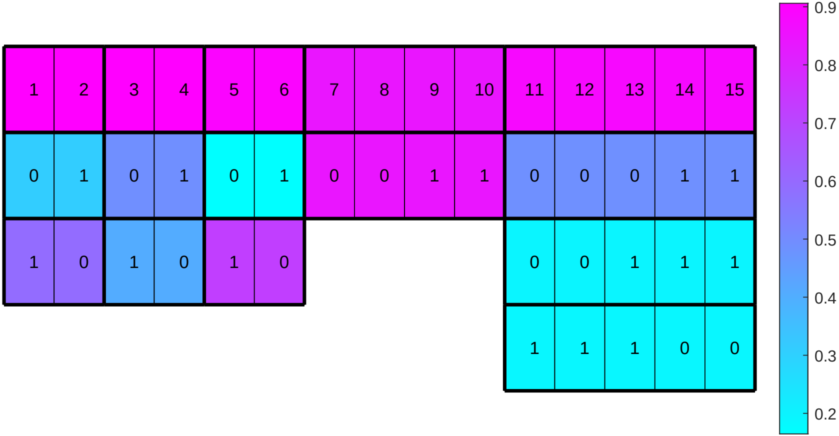
Algorithms

1. Calculate the probability of all possible credible sets (there are $2^{\#\Gamma_i}$). Find the smallest set with probability above the desired level.

Algorithms

1. Calculate the probability of all possible credible sets (there are $2^{\#\Gamma_i}$). Find the smallest set with probability above the desired level.
2. Let Δ_i be the smallest change in $p(S_i)$ by removing an element from S_i. Choose $k = \arg\min(\Delta_1, \ldots, \Delta_r)$ and remove the corresponding element from S_k. Continue until removing any element leads to a probability below the desired level.

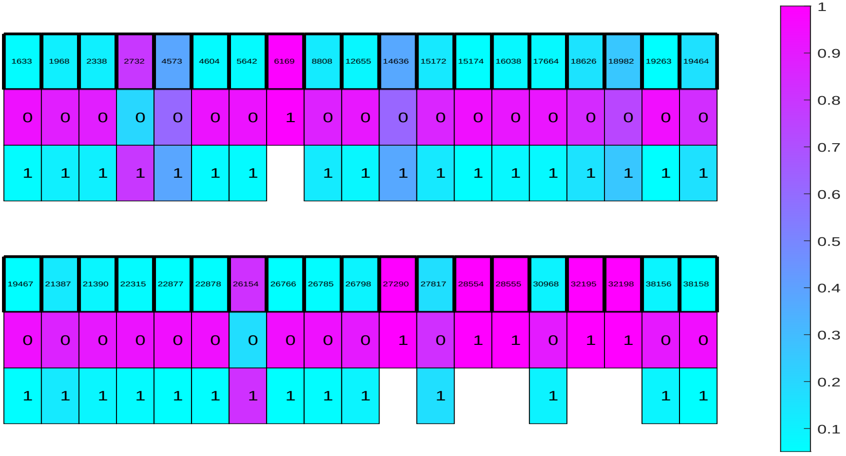
Simulated example (50 % credible set)



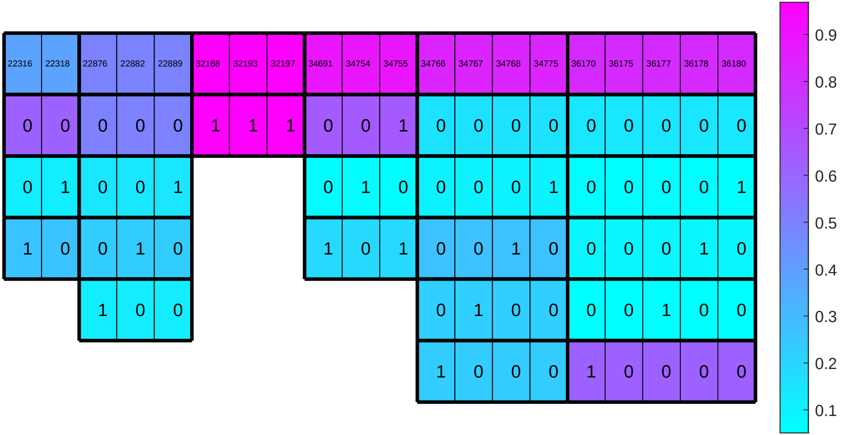
Simulated example (smallest 50% credible set)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Prob
1	0	0	1	1	0	0	0	1	1	0	0	0	1	1	0.0849
1	0	1	0	1	0	0	0	1	1	0	0	0	1	1	0.0817
1	0	0	1	1	0	0	0	1	1	1	1	1	0	0	0.0424
1	0	0	1	1	0	0	0	1	1	0	0	1	1	1	0.0396
0	1	1	0	1	0	0	0	1	1	0	0	0	1	1	0.0345
0	1	0	1	1	0	0	0	1	1	0	0	0	1	1	0.0338
1	0	1	0	1	0	0	0	1	1	0	0	1	1	1	0.0279
0	1	0	1	1	0	0	0	1	1	1	1	1	0	0	0.0264
0	1	0	1	1	0	0	0	1	1	0	0	1	1	1	0.0248
1	0	1	0	1	0	0	0	1	1	1	1	1	0	0	0.0218
1	0	1	0	0	1	0	0	1	1	0	0	0	1	1	0.0218
1	0	0	1	0	1	0	0	1	1	0	0	0	1	1	0.0202
0	1	1	0	1	0	0	0	1	1	0	0	1	1	1	0.0183
1	0	1	1	1	0	0	0	1	1	0	0	0	1	1	0.0176
0	1	1	0	1	0	0	0	1	1	1	1	1	0	0	0.0142

GWAS example



GWAS example



- One of 34766, 34767, 34768 and 34775 is included with probability 0.84 (individual PIPs are 0.10, 0.27, 0.22, 0.24)
- One of 22876, 22882, and 22889 is included with probability 0.50 (individual PIPs are 0.14, 0.23, 0.12)

Discussion

- Credible sets are useful way to explore uncertainty in the posterior distribution in Bayesian variable selection
- The method can identify blocks of highly correlated variables which can dilute marginal posterior inclusion probabilities
- The methods work with MCMC but could be easily extended to other inference frameworks (*e.g.* variational Bayes)
- These approaches could be extended to other discrete structures by finding a representation of the posterior with independence structure (*e.g.* factor models, etc.)

Castillo, I., J. Schidt-Hieber, and A. van der Vaart (2015). Bayesian linear regression with sparse priors. *Annals of Statistics 43*, 1986–2018.

Cox, D. R. and N. Weimurth (1994). A note on the quadratic exponential binary model. *Biometrika 81*, 403–408.

Dai, B., S. Ding, and G. Wahba (2013). Multivariate Bernoulli distribution. *Bernoulli 19*, 1454–1483.

George, E. I. and R. E. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica 7*, 339–373.

Griffin, J. E., K. G. Latuszynski, and M. F. J. Steel (2021). In search of lost mixing time: adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p. *Biometrika 108*, 53–69.

Li, D., S. Dutta, and V. Roy (2023). Model based screening embedded Bayesian variable selection for ultra-high dimensional settings. *Journal of Computational and Graphical Statistics 32*, 61 – 73.

Liang, X., S. Livingstone, and J. E. Griffin (2022). Adaptive random neighbourhood informed Markov chain Monte Carlo for high-dimensional Bayesian variable selection. *Statistics and Computing 32*, 84.

Porwal, A. and A. E. Raftery (2022). Comparing methods for statistical inference with model uncertainty. *Proceedings of National Academy of Sciences 119*, e2120737119.

Zanella, G. and G. Roberts (2019). Scalable importance tempering and Bayesian variable selection. *Journal of the Royal Statistical Society Series B 81*, 489–517.