Prior distributions for SNP selection in genome–wide association studies

Clive Hoggart, Maria De Iorio and David Balding

Centre for Biostatistics, Imperial College

Introduction

- Problem: Detect/localize potentially multiple causal variants from dense genotype data of case control samples
  - do not look for interactions
- Typically more variables than observations $(p >> n)$, 2^p potential models
- Sequence data will soon be available – p even larger
- Classical forward-backward variable selection
  - computationally too slow
  - and inherently unstable for $p >> n$
- Bayesian stochastic search variable selection
  - Search procedure induces stability
  - Utilizes MCMC - computationally too slow for size of our problem

The model

- We fit a logistic regression model

$$p(y_i = 1) = \psi\left(\sum_j^p \beta_j x_{ij}\right)$$

  where ψ is the logit link function, $i = 1, \ldots, n$ indexes individuals and $j = 1, \ldots, p$ index SNPs.

- This talk concerns choosing prior distributions for β that optimize the process of picking causal variants, we consider two types of prior
  - Continuous shrinkage priors
  - Slab and spike sparsity prior.

Overview

- Shrinkage priors
  - Distributions
  - Optimization
  - Inference
- Bayesian variable selection
  - Model
  - Search algorithm
  - Inference
- Results
  - Simulated data sets
  - Analyses

Shrinkage priors

- Priors with a discontinuity in the derivative at zero (spike at zero) can have posterior modes such that $\beta_j = 0$ for some j.
  - however, the posterior distribution is continuous.
- Problem of variable selection can be simplified using such shrinkage priors by limiting inferential process to that of finding posterior modes rather than full posterior analysis via MCMC.
- We consider two shrinkage priors; double exponential distribution and a generalisation of it.
- Penalised likelihood

$$\log p(\beta \mid y) = \log p(y \mid \beta) + \log p(\beta)$$

- Tibshirani (1996) Lasso is equivalent to posterior mode inference with double exponential prior.

Double exponential prior

- The double exponential distribution commonly used shrinkage prior
- One parameter model – scale
- Over-shrinks informative variables – greedy
- It has the following representation as a scale mixture of a normal distribution;

$$DE(\beta \mid \lambda) = \int_0^\infty \mathsf{N}(\beta \mid 0, \sigma^2)\mathsf{Ga}(\sigma^2 \mid 1, \lambda^2/2)\, d\sigma^2$$
$$= \frac{\lambda}{2} \exp\{-\lambda|\beta|\}$$

Generalising exponential prior - normal exponential gamma distribution

- The normal exponential gamma distribution (NEG) is generated by generalising the scale mixture representation of the double exponential distribution

$$\text{NEG}(\beta \mid \lambda, \gamma) = \int_0^\infty \int_0^\infty \text{N}(\beta \mid 0, \sigma^2)\text{Ga}(\sigma^2 \mid 1, \psi)\text{Ga}(\psi \mid \lambda, \gamma^2)\, d\sigma^2 d\psi$$

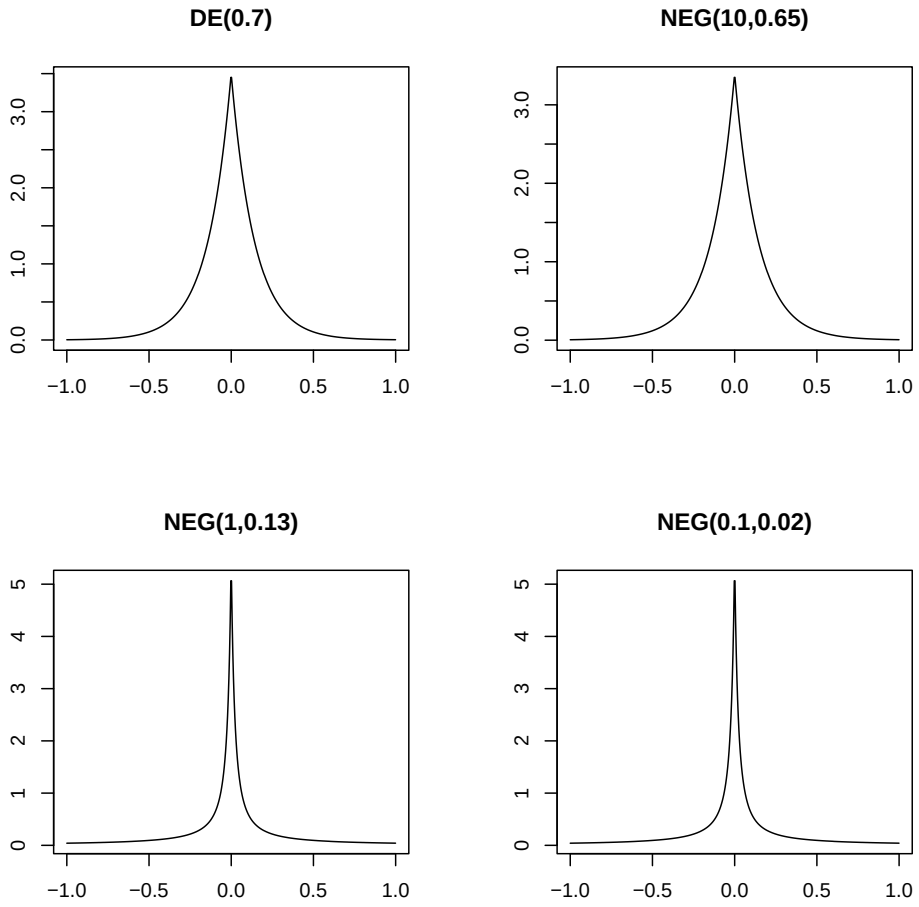
$$= \kappa \exp\left\{\frac{\beta^2}{4\gamma^2}\right\} D_{-2\lambda-1}\left(\frac{|\beta|}{\gamma}\right)$$

  where $D_v(z)$ is the parabolic cylinder function and κ is the integrating constant

- λ shape parameter, γ scale parameter
- fast algorithms for computing the parabolic cylinder function and its derivatives

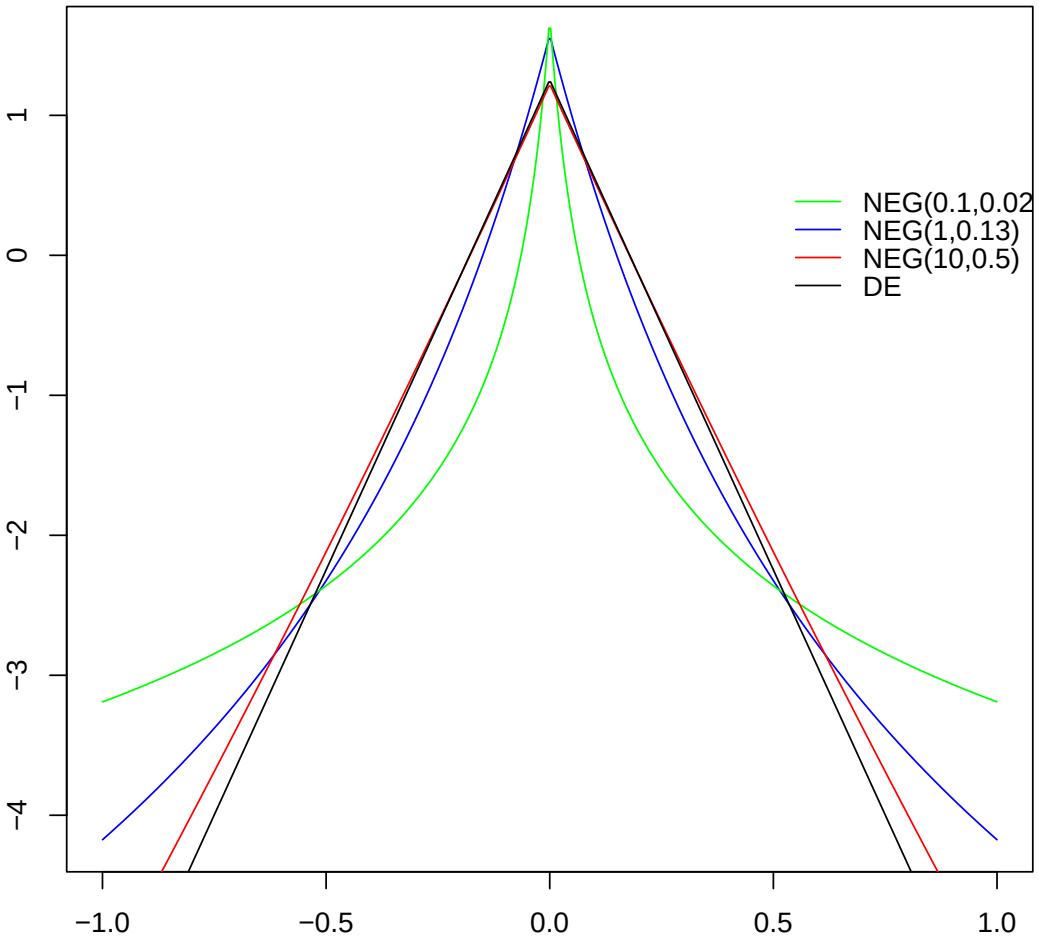
Griffin and Brown (2005).

Shapes of the NEG



- Double exponential recovered as $\lambda \to \infty$ and $\gamma \to \infty$

Tail behaviour of NEG



Finding posterior modes

- Griffin and Brown (2005) used the EM algorithm
  - Found algorithm converged slowly for binary regression.
- We use an adaptation of CLG - cyclic coordinate descent optimization algorithm - by Genkin et al. (2005).
- Genkin et al. (2005) applied algorithm to logistic regression with normal and double exponential prior

The CLG algorithm

CLG - cyclic coordinate descent optimization algorithm - Bazaraa and Shetty (1979)

- One dimensional optimization problem
- successively sets each variable in turn to minimize objective function – minus log posterior
- multiple passes over the variables are made until convergence
- for logistic regression no closed form solution for minimization
  - Zhang and Oles (2001) implement optimization using an adaptation of Newton's method (assumes quadratic function)

$$\Delta\beta_j = \beta_j^{(new)} - \beta_j = -\frac{g'(\beta_j)}{g''(\beta_j)}$$

where g' and g'' are the derivative and 2nd derivative of the log-posterior.

The CLG algorithm - adaptation of Newton's method

- For robustness avoid taking large steps where the quadratic approximation is poor by
  - limiting the size of step that can be taken
  - Replacing g'' with a function which is an upper bound on g''
- Details in Genkin et al. (2005)

Space of solutions

- With NEG prior, posterior is highly multimodal
- Smaller λ (shape parameter of NEG) the more multimodal
- Choose mode which maximises posterior – MAP
- Also with double exponential prior posterior is multimodal
- Posterior modes found and their complexity dependent on starting position;
  - closer to origin simpler solution
  - closer to mle more complex solution
- Also, with the CLG algorithm mode found is dependent on order of search
- We investigated accuracy of SNP selection for the different search methods

Setting prior parameters - controlling type-I error

- If search is started at origin we can chose parameters to control type-I error of selection via univariate regression – not conditional on other SNPs.
- Starting the search at the origin a parameter is set to zero if there is no turning point, ie. when the derivative of the log-posterior is monotone decreasing in $|\beta|$.
- From log-posterior we can get a constraint on $\hat{\beta}$ s.t $\tilde{\beta} = 0$.
- Under the null of no association

$$\hat{\beta} \sim \mathrm{N} \quad 0, \frac{2}{nf(1-f)}$$

where n is the sample size and f is the allele frequency

- parameters of the prior can be set to control the type-I error rate.

Alternative search algorithm

Bayesian variable selection

- Standard Bayesian variable selection
- Propose a slab and spike sparsity prior

$$p(\beta \mid \theta, \sigma) = \theta \mathsf{N}(\beta \mid 0, \sigma^2) + (1-\theta)\delta_{\beta=0}$$

  where $\delta_{x=0}$ is Dirac function with point mass at $x = 0$

- Hans et al (2005) proposed the shotgun stochastic search (SSS) algorithm to search this model space

Alternative search algorithm

The search algorithm

- SSS is an alternative to MCMC for exploring space
- Explores model space using Laplace approximation of the marginal likelihood

$$p(y \mid \gamma, X) = \int p(y \mid X, \beta) p(\beta \mid \gamma) d\beta$$

$$\approx (2\pi)^{k/2} |\tilde{\Sigma}|^{1/2} p(y \mid \tilde{\beta}, \gamma) p(\tilde{\beta} \mid \gamma)$$

where γ is a $p \times 1$ indicator vector s.t. $\gamma_j = 1(0)$ if SNP j is (not) in the model, k is the number of variables in the model, $\tilde{\beta}$ is the MAP estimate and

$$\tilde{\Sigma} = - \left(\frac{\partial^2 \log[p(y \mid \tilde{\beta}, \gamma) p(\tilde{\beta} \mid \gamma)]}{\partial \theta_i \partial \theta_j} \right)^{-1}, \quad k \times k \text{ matrix}$$

- Compare model probabilities, $p(\gamma \mid y) \propto p(y \mid \gamma) p(\gamma)$.

Alternative search algorithm

The search algorithm - moving around the space

- Similar moves to reversible jump MCMC which explores all local models; for a given model
  - deletions - all models formed by deleting a variable from current model
  - replacements - all models formed by replacing a variable in the current model with a variable not in the current model
  - insertions - all models formed by inserting a variable into the current model
  - Stochastic move to new model dependent on model probability
  - Repeat...
- Records model probabilities of all models visited
- Still liable to get stuck in local modes
- Variable selection;
  - Variables in best fitting model
  - Weighted average of models visited

Alternative search algorithm

Deterministic search

- We modified this search algorithm to a deterministic search
  - Find all one dimensional models that improve on the model with just the intercept
  - all two dimensional models that improve on their parent one dimensional model
  - and so on until we have a set of n dimensional models which cannot be improved by adding another variable.
- Do not make inference on weights θ, select variables in best fitting model.
- Will only miss model if there is no $k+1$ dimensional model that improves on a parent k dimensional model but there is a $k+2$ dimensional model that improves on a parent k dimensional model
- Main purpose was to evaluate SNP selection via shrinkage priors

Alternative search algorithm

Controlling type-I error

- Control type-I error by considering one dimensional models
- Variable selection criteria, variable i will be chosen if

$$\log p(y \mid \gamma_i) + \log p(\gamma_i) > \log p(y \mid \gamma_0) + \log p(\gamma_0)$$

  where γ_i indicates the model with just variable i, γ_0 is the model with just the intercept and $p(\gamma)$ is the prior on model γ.

- Can approximate distribution of differences in marginal likelihood by assuming asymptotic normality of $\hat{\beta}$.
- Prior is geometric on the number of variables in the model
- Set parameter of prior to control type-I error.

Simulated data

- Simulated sequence data from 20Mb region for 10,000 individuals
- SNP ascertainment
  - One SNP every 5kb s.t. MAF distribution approximately uniform, 4,000 SNPs
- 30 data sets sampled from population each with 1,000 cases and 1,000 controls
- Disease models
  - Each data set had ten causal SNPs (not necessarily genotyped) all with approximately the same allele frequency
  - 10 data sets for each of 3 allele frequencies of causal variant; 1%, 5% and 15%
  - Risk ratios respectively 6, 2.37 and 1.8
  - Chosen such that 70 percentile of test statistic were approximately equal.
- Scoring rule;
  - True positive if within 50kb of a causal variant, otherwise false.

SNP selection methods compared

- p-value threshold with regionwide false positive rate of 5% – Bonferonni correction
- Shrinkage priors
  - Double exponential
  - NEG with $\lambda = 0.1, 0.5, 1.0, 5.0, 10.0, 20.0$
  - γ set for a regionwide false positive rate of 5%
  - Both normalised and unnormalised SNP data for both priors
  - Also explored power of exploiting posterior multimodality
- Bayesian variable selection
  - Difuse prior of $\beta \sim \mathrm{N}(0, 100)$
  - Prior of geometric set for regionwide false positive rate of 5%
  - Both normalised and unnormalised x's

Overview of results - 1

- Results with double exponential prior much better with standardized data.
  - Approximate thresholding rule with non-standardized data $\hat{\beta} < \frac{2\lambda}{nf(1-f)}$
  - Approximate thresholding rule with standardized data $\hat{\beta} < \frac{4\lambda}{n}$
- With NEG prior best results with;
  - standardised data and $\lambda = 1$ and $\lambda = 5$
  - no benefit starting search away from origin and exploiting multimodality
  - little benefit of exploiting multimodality when search is started from origin

Results

- False positives and false negatives counted for the following methods
  - p-value threshold
  - Double exponential with standardised data
  - NEG $(\lambda = 1, 5)$ with standardised data.
  - Deterministic search

Type-I error rates

Validation of parameter settings controlling type-I error

Selection method	Average number of errors per 20Mb
Bayesian variable selection	0.07
Shrinkage prior MAP:	
Double exponential	0.04
NEG $(\lambda = 1)$	0.04
NEG $(\lambda = 5)$	0.04

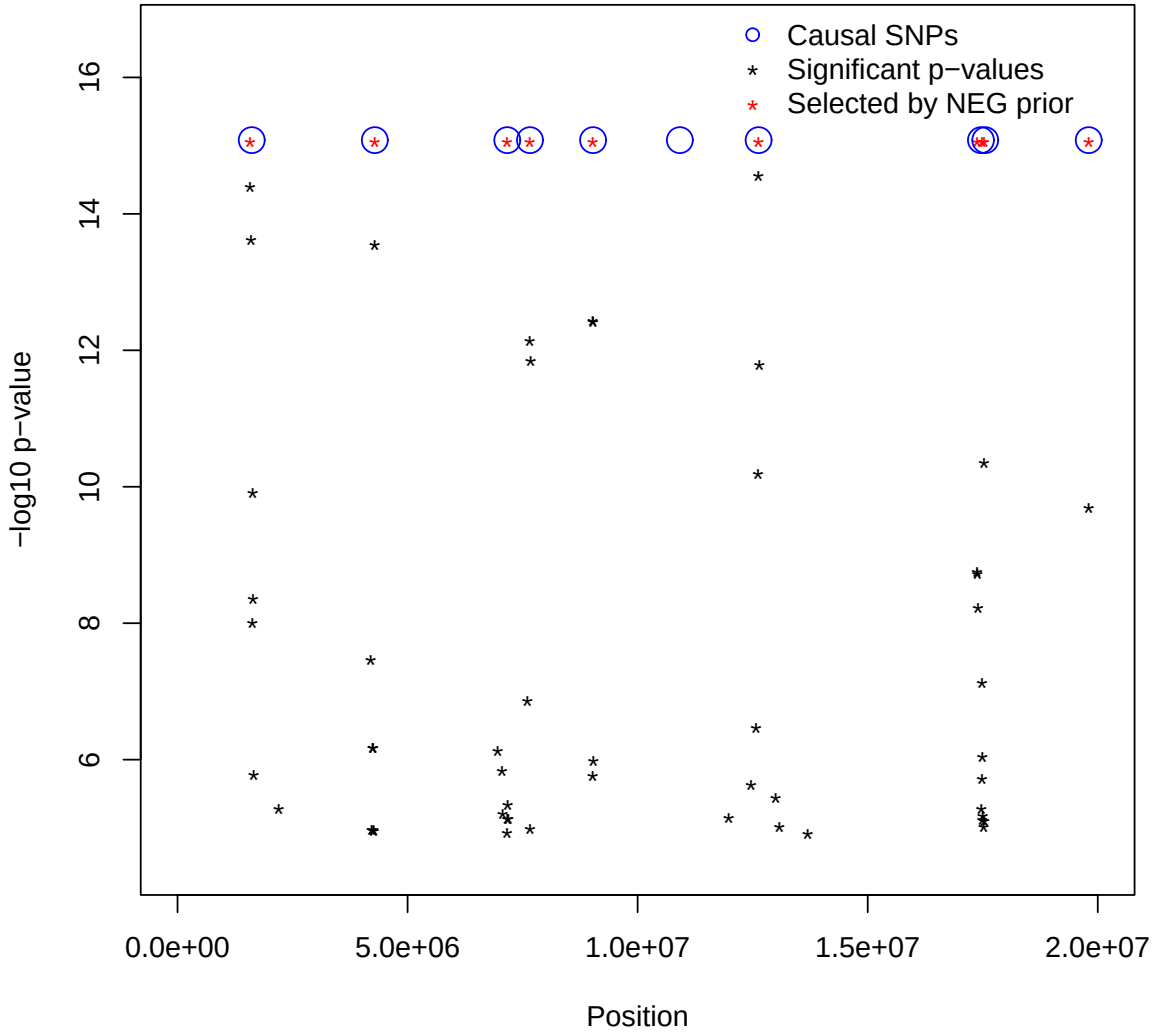
From permutations of case control status.

Error rates in simulated case-control data sets

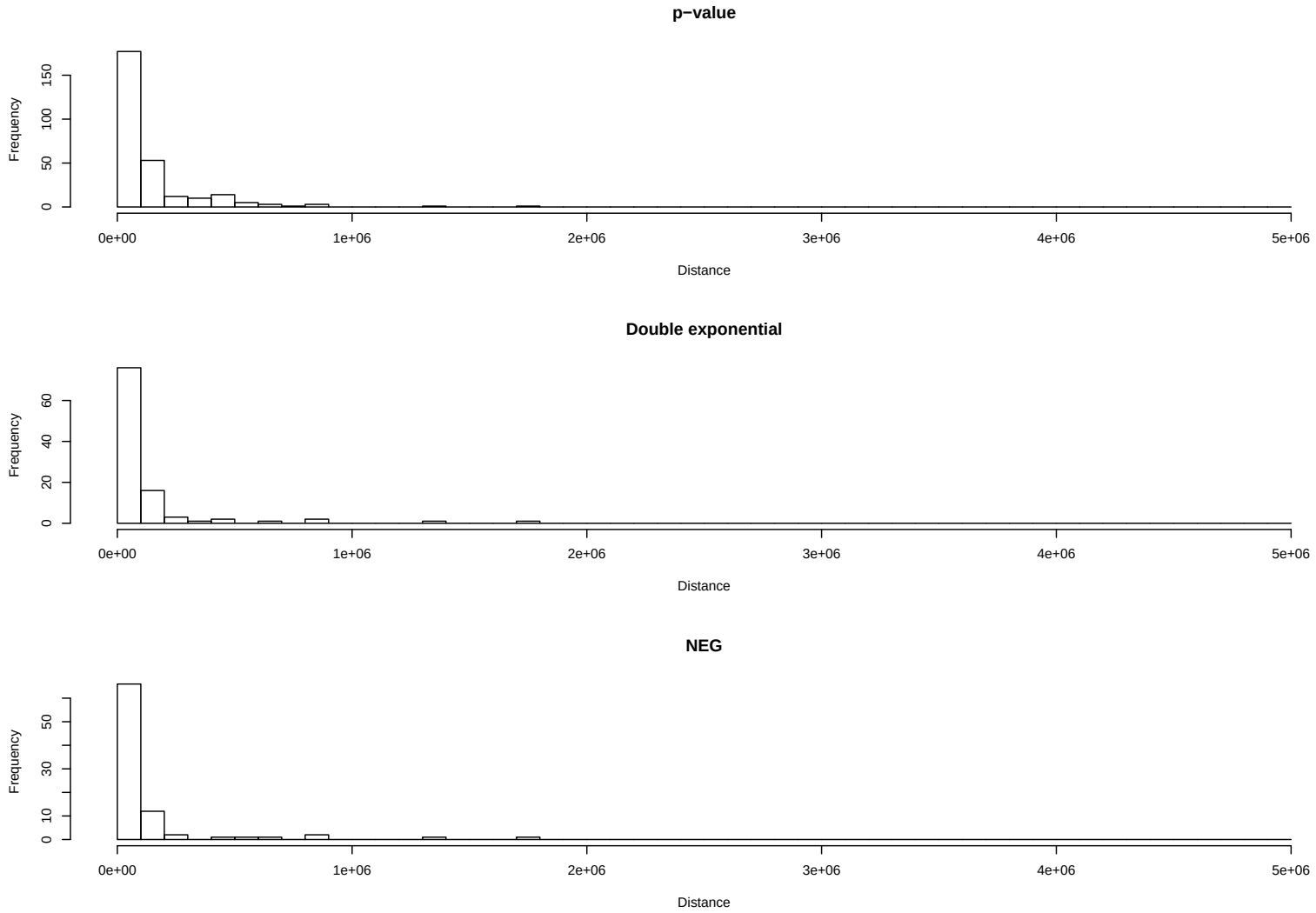
	Causal allele frequency					
	0.15		0.05		0.01	
Selection method	False +ve	False -ve	False +ve	False -ve	False +ve	False -ve
BVS	8	13	50	72	52	79
p-value	277	11	159	35	227	58
Shrinkage prior MAP:						
NEG $(\lambda = 1)$	19	18	33	46	65	69
NEG $(\lambda = 5)$	9	21	29	54	46	72
Double exponential	29	16	43	46	68	70

Numbers of false positives and false negatives in the 30 data sets analysed

MAP position of causal and detected variants



Distance of selected variant to nearest causal variant



Conclusions

- Shrinkage methods improve on simple per SNP p-value analyses by removing many false positives
- Shrinkage methods with CLG optimization algorithm is computationally feasible for analysis of dense genotype data
  - Analysis of 44,000 SNPs took approximately 15 minutes
  - Limit is size of computer memory
- Using the NEG prior instead of the DE prior does not affect computation but does improve selection of causal variants.
- Deterministic search is computationally more expensive and does not improve performance.