Subsetting SNPs

Jean Morrison

November 15, 2017

Introduction

My goal is to understand two questions:

1. Given a decent estimate of the posterior distributions of b and q, can I use only a subset of SNPs to estimate the difference in ELPD between the shared and causal models.
2. Can I get a reasonable estimate of the posterior using only a subset of SNPs

In both cases, I am interested in subsetting SNPs using p-values from the first trait.

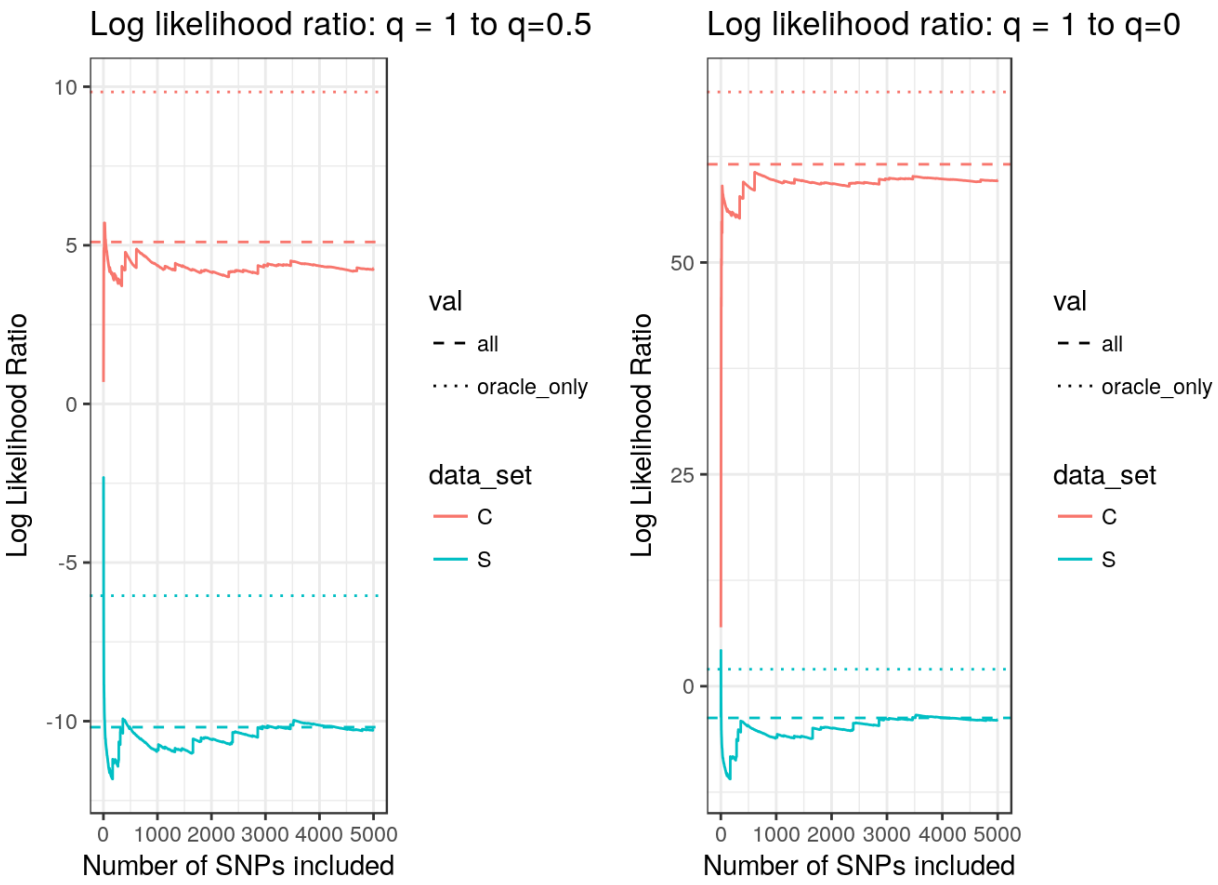
In these experiments, I will use two simluted data sets. Both have 100,000 SNPs. Data set S is generated under the shared model with $q = 0.5$ and $b = 0.4$. Data set C genrated under the causal model with $q = 1$ and $b = 0.4$.

Likelihood ratios using known parameters

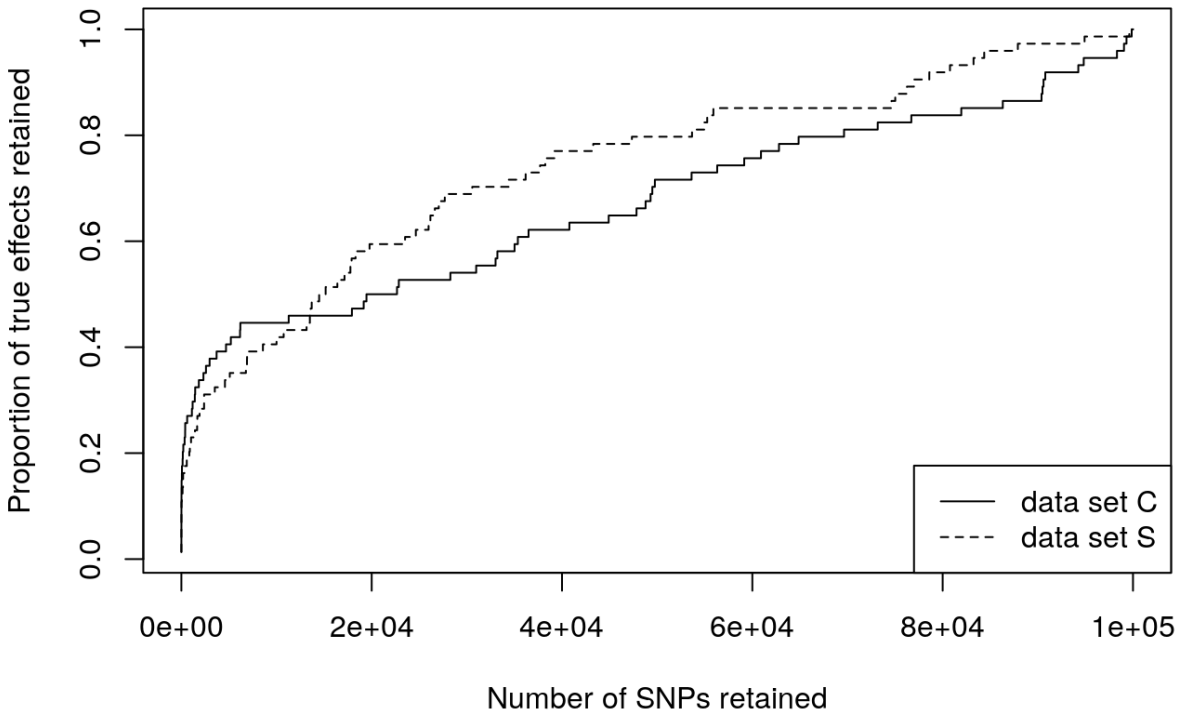
First we assume the value of b is known. We look at log likelihood ratios comparing $q = 1$ to $q = 0.5$ and $q = 1$ to $q = 0$.

```
Warning: Removed 190000 rows containing missing values (geom_path).
```

```
Warning: Removed 190000 rows containing missing values (geom_path).
```



In both data sets, with a small number of SNPs (1 or 2% of the top trait 1 SNPs) we can do a good job of approximating the log likelihood ratio we would obtain with all the SNPs. Horizontal lines show the log likelihood ratio we would obtain using only the SNPs for which the true trait 1 effect size is not zero (which we will call the "oracle" SNP set). Note that when we include the top few percent of SNPs, we retain fewer than half of the true trait 1 effect SNPs. The plot below shows the number of SNPs retained vs the proprtion of true effect SNPs retained (both data sets contain 74 trait 1 effects). SNPs in both data sets have the same trait 1 effects so the difference between these two lines is due to stochasitic differences in $\hat{\beta}_1$.



"Oracle" posteriors

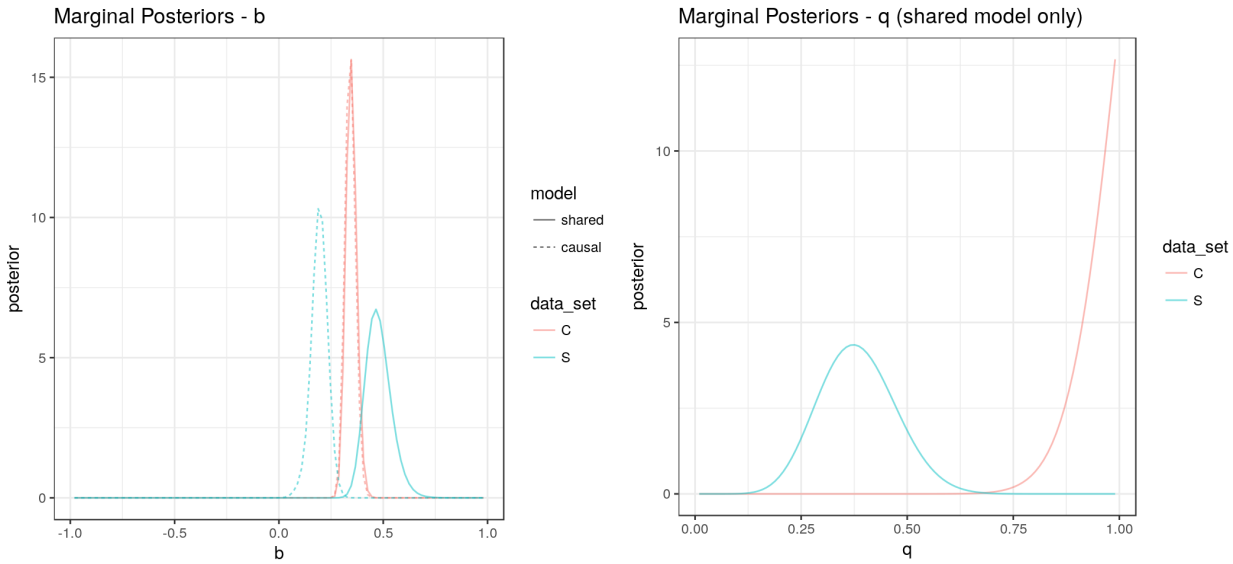
Now we consider comparing the ELPD under the causal model to the ELPD under the shared model. For the causal model, we use the data to estimate the posterior distribution of b. For the shared model we estimate the posteriors of b and q.

We start with the best possible posterior distribution we can get from these data. To estimate the "oracle" posterior, we use only SNPs with non-zero true effects on trait 1 and we assume we know the joint distribution of direct effects on trait 1 and trait 2. We also assume we know $\rho = 0$.

The plots below show the marginal posterior distributions of b and q from the shared models and the posterior of b from the causal model.

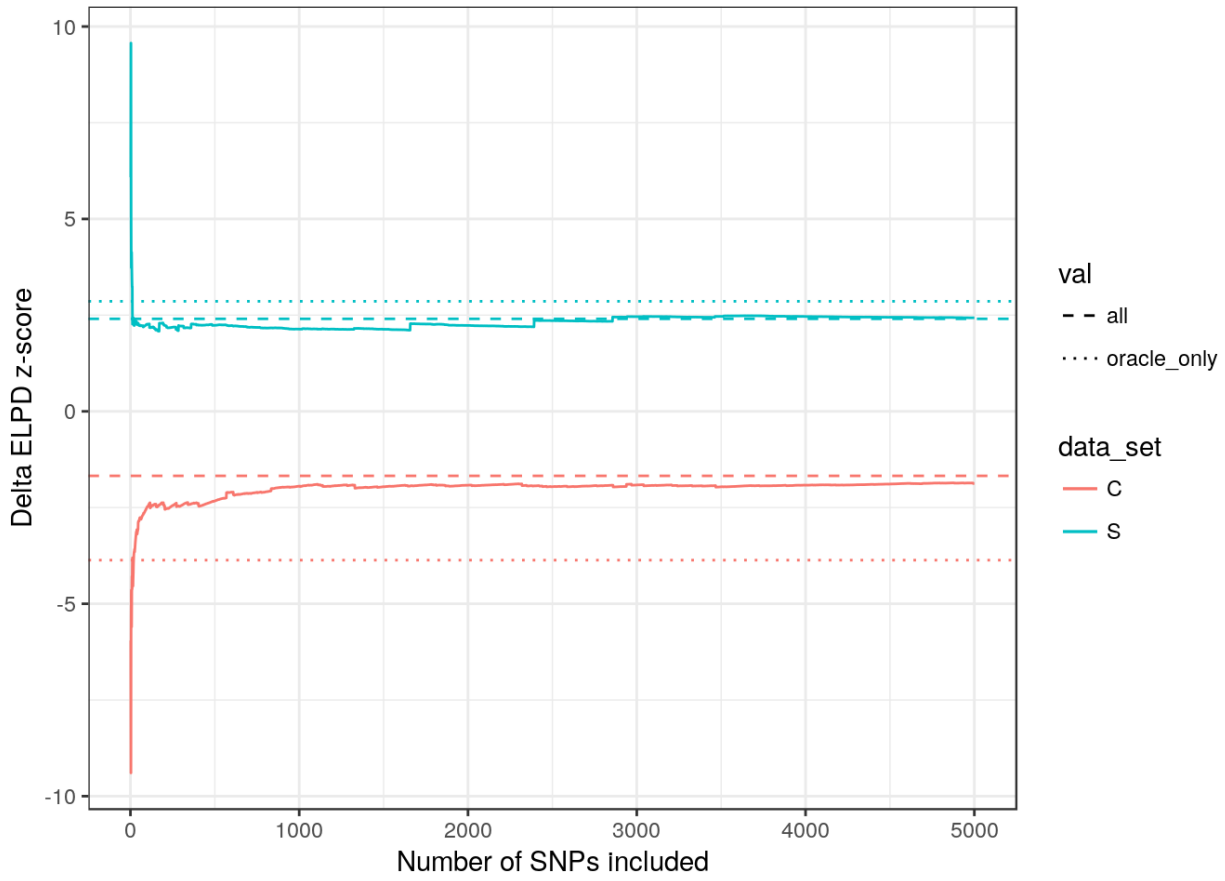
```
seed:  675987130
Model 2.
Model 3.
```

```
seed:  335984201
Model 2.
Model 3.
```



We now estimate the difference in ELPD from these two posterior distributions using subsets of the SNPs. The plots below show the z-score for the change in ELPD between the causal model and the shared model using only a subset of SNPs (number of SNPs included on horizontal axis). The dotted lines show the z-scores we would obtain using only the true trait 1 effect SNPs. The dashed line shows the z-score obtained using all SNPs. In this plot, positive z-scores are evidence for the shared model and negative z-scores are evidence for the causal model.

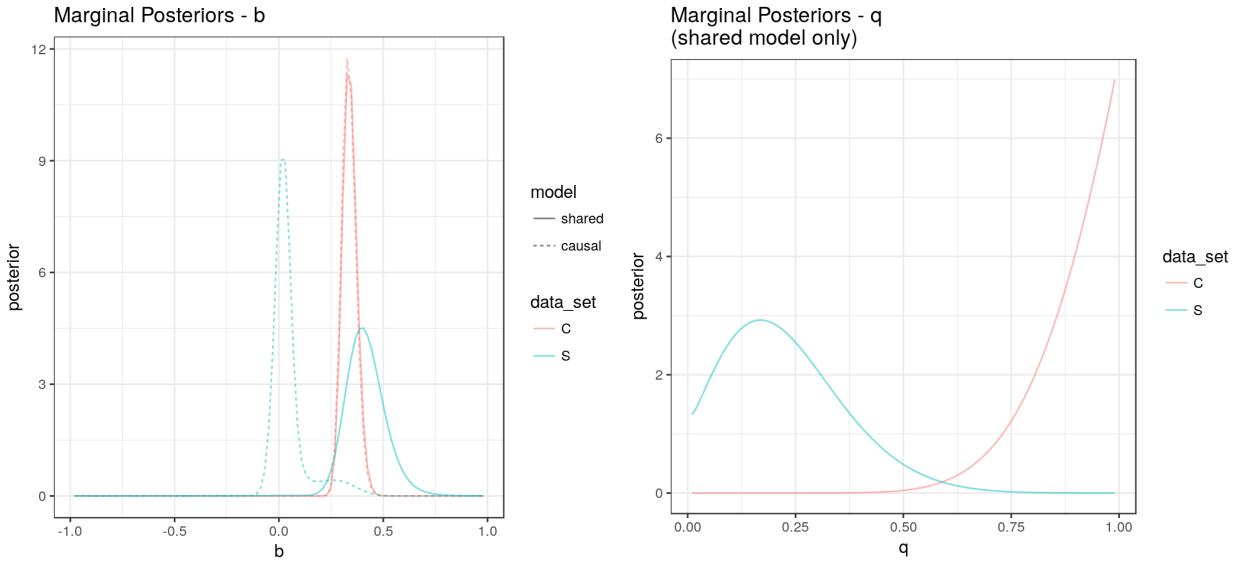
```
Warning: Removed 2 rows containing missing values (geom_path).
```



We find that we can approximate the delta ELPD z-score that we would obtain from the full set of SNPs using a few thousand of the top SNPs. Data set S shows strong evidence in favor of the shared model as we expect. Data set C shows evidence in favor of the causal model – z-score of -1.68 using all SNPs. This z-score is not larger because the posterior for q is very close to 1 and the posteriors for b are very similar under the two models. This makes the posterior distribution under the shared model for data set C look very similar to the posterior under the causal model.

Posteriors estimated with all SNPs

Now we consider the posteriors that we would get using all the SNPs and without knowing the true joint distribution of direct effects on traits 1 and 2.

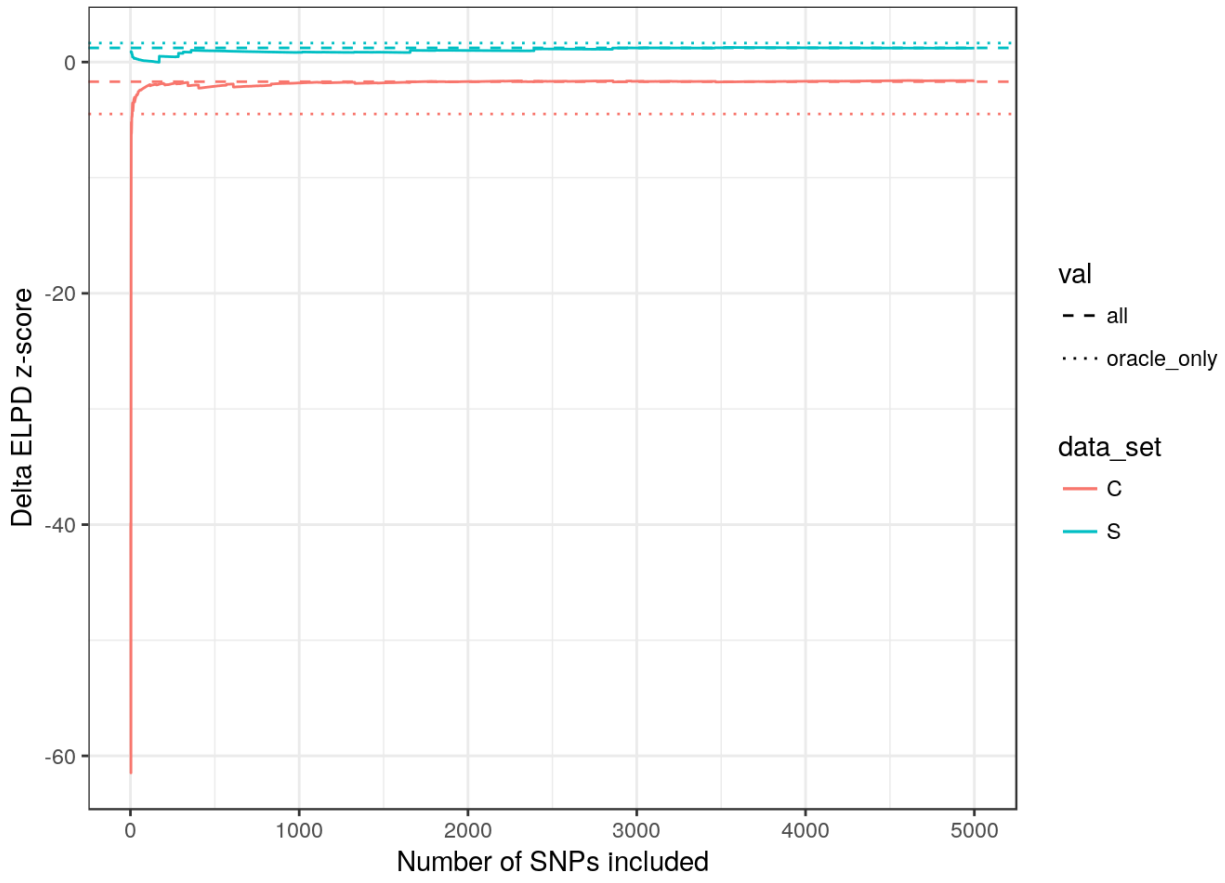


These posteriors differ from the posteriors calculated using only the true trait 1 effects in a few ways:

- The posteriors of b for data set S under the causal and shared models are lower.
- The posteriors of q for both data sets are somewhat lower. The MAP for data set S is around 0.17 using all SNPs while it was around 0.38 using only the oracle SNPs.
- The posteriors for b under both models for data set C are similar to those estimated with only the oracle SNPs.

Using these priors to calculate Δ ELPD z-scores gives the following results:

```
Warning: Removed 2 rows containing missing values (geom_path).
```

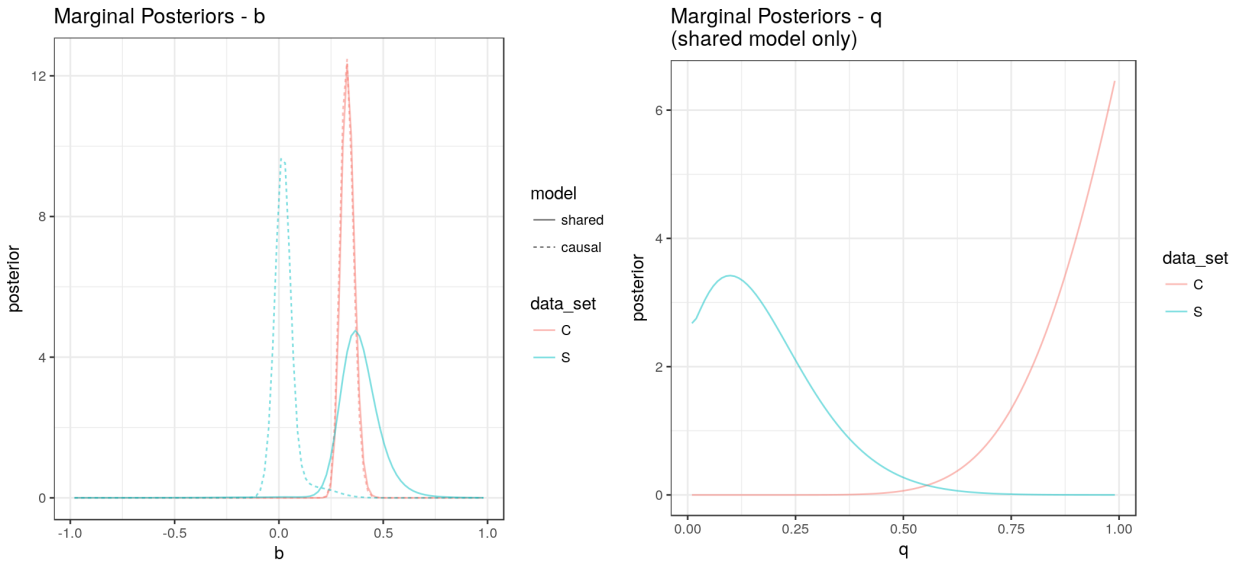


Posteriors estimated with top SNPs

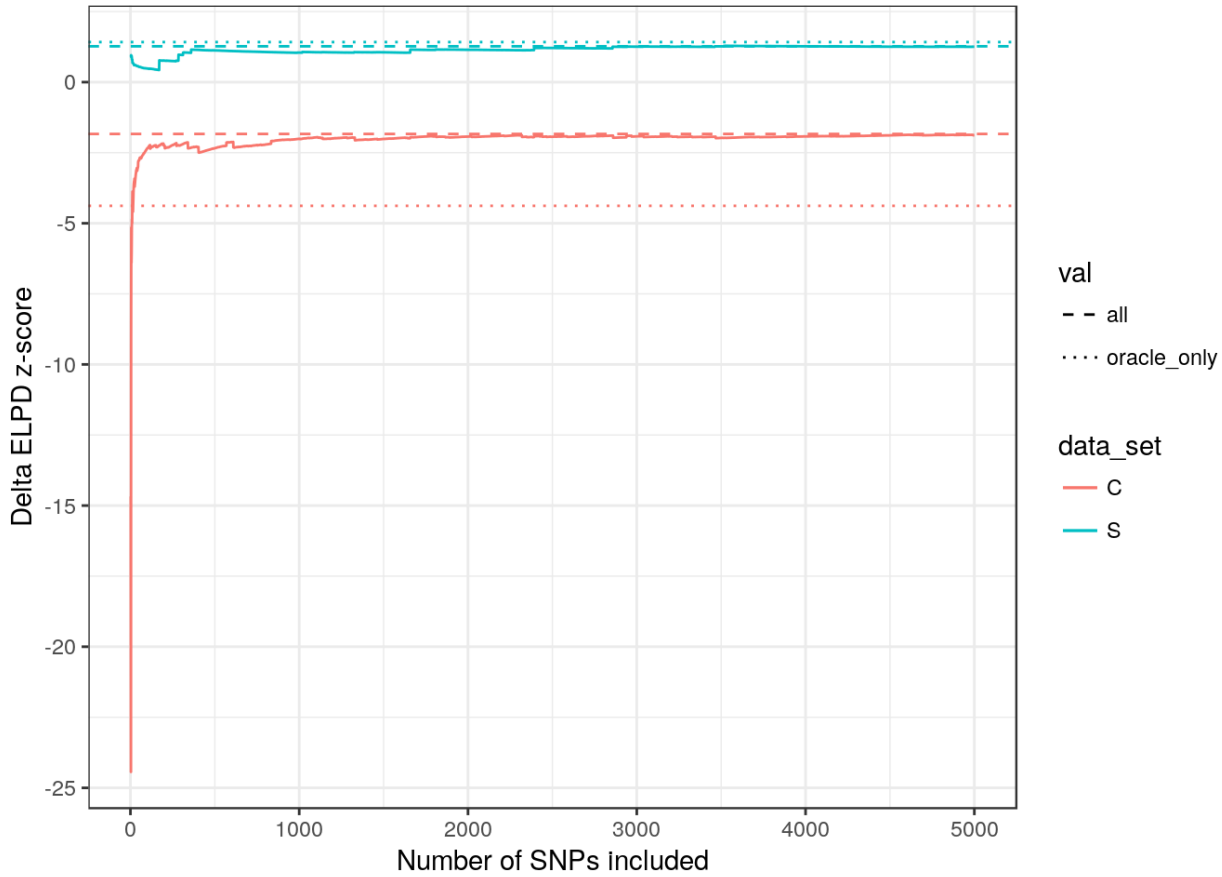
Now we estimate the posteriors using only the top 1000 SNPs. We still assume we know the joint distribution of direct effects.

```
seed:  432261163
Model 2.
Model 3.
```

```
seed:  872631807
Model 2.
Model 3.
```




```
Warning: Removed 2 rows containing missing values (geom_path).
```



Combining results for both data sets across all three estimates of the posterior gives the following results. The tables below show Δ ELPD z scores comparing the causal to the shared model. Negative z-scores are in favor of the causal model. Each column corresponds to a different estimate of the posterior distribution.

```
Data set S
```

snp_set	oracle	all	top
Oracle	2.849805	1.7246716	1.455491
All	2.395383	1.2521352	1.309378
Top 1%	2.131709	0.8419484	1.084927

Data set C

snp_set	oracle	all	top
Oracle	-3.829572	-4.382124	-4.452867
All	-1.681596	-1.780124	-1.820555
Top 1%	-1.961897	-1.949567	-1.964509

Simulation Results

I repeated the above experiments in simulations. In the results shown below, data are generated with $b = 0.4$ for all simulations. There are 20 data sets generated using each of 7 values of q: 0, 0.1, 0.2, 0.3, 0.5, 0.7 and 1. The scatter plots below compare z-scores computed using each of the following four methods:

- All/All: The posterior is calculated using all SNPs, the Δ ELPD z-score is calculated using all SNPs.
- Top/Top: The posterior is calculted using the top 1% of SNPs, the z-score is calculated using the top 1% of SNPs.
- Top/All: The posterior is calculted using the top 1% of SNPs, the z-score is calculated using all SNPs.
- All/Top: The posterior is calculted using all SNPs, the z-score is calculated using the top 1% of SNPs.

Points in the scatter plot are colored according to the value of q.

The z-scores in these plots differ from those calculated above in an important way. In the experiments we just described, we assumed that we knew the bivariate distribution of direct effects on traits 1 and 2. In the simulations results presented, this distribution is estimated from the data. In all cases, we estimate ρ and the distribution of direct effects using all SNPs.

We find that the All/All z-scores and Top/Top z-scores are highly correlated. They both do a good job of distinguishing the simulations with $q = 1$ from the other simulations. Below I show ROC curves for the same set of simulations. Each point on the curve for a particular method corresponds to a z-score cutoff. The true positive rate is the proportion of the 20 simulations with $q = 1$ that have z-scores falling below the cutoff. The false positive rate is the proportion of the 120 simulations with $q < 1$ that have z-scores below the cutoff. Colored points show the results that would be obtained with a cutoff of $\Phi(0.95)$.

Session information

```
sessionInfo()
```

```
R version 3.4.1 (2017-06-30)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 17.04

Matrix products: default
BLAS: /usr/lib/libblas/libblas.so.3.7.0
LAPACK: /usr/lib/lapack/liblapack.so.3.7.0

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
```

```
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] knitr_1.17        cumstats_1.0      gridExtra_2.2.1   sherlockAsh_0.1.0
[5] ggplot2_2.2.1     tidyr_0.6.3

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.13        highr_0.6           compiler_3.4.1
 [4] plyr_1.8.4          iterators_1.0.8     tools_3.4.1
 [7] digest_0.6.12       MHadaptive_1.1-8    evaluate_0.10.1
[10] tibble_1.3.1        gtable_0.2.0        lattice_0.20-35
[13] rlang_0.1.1         Matrix_1.2-10       foreach_1.4.3
[16] DBI_0.6-1           yaml_2.1.14         parallel_3.4.1
[19] loo_1.1.0           dplyr_0.5.0         stringr_1.2.0
[22] stats4_3.4.1        rprojroot_1.2       grid_3.4.1
[25] R6_2.2.1            rmarkdown_1.7       ashr_2.1-27
[28] magrittr_1.5        backports_1.1.0     scales_0.4.1
[31] codetools_0.2-15    htmltools_0.3.6     matrixStats_0.52.2
[34] MASS_7.3-47         assertthat_0.2.0    colorspace_1.3-2
[37] numDeriv_2016.8-1   labeling_0.3        stringi_1.1.5
[40] RcppParallel_4.3.20 lazyeval_0.2.0      munsell_0.4.3
[43] doParallel_1.0.11   pscl_1.5.1          truncnorm_1.0-7
[46] SQUAREM_2016.8-2
```