SVA Simulation

Christopher Lo

9/6/2022

Simulation studies inspired from *"A general framework for multiple testing dependence"* (Leek et al. 2008)

Simulation Set-Up, one single experiment

We generate X from the following model:

$$X = BS + \Gamma G + U$$

We have $m = 1000$ genes (tests), $n = 20$ samples, and $r = 2$ latent variables.

Sampling noise: $U_{m,n} \sim N(0, 1)$.

The design matrix S is 10 cases and 10 controls: $S_{1,n} = 1$ for $n = 1 : 20$. Then, $S_{2,n} = 0$ for $n = 1 : 10$, $S_{2,n} = 1$ for $n = 11 : 20$.

Control effect for all genes: $b_{m,1} \sim N(0, 1), m = 1 : 1000$

Case effect for DE genes $m = 1 : 300$: $b_{m,2} \sim N(3, 1)$

Case effect for Non-DE genes $m = 301 : 1000$: $b_{m,2} = 0$

Latent design matrix G: $G_{r,n} \sim Bernoulli(.4), n = 1 : 10$.

$G_{r,n} \sim Bernoulli(.6), n = 11 : 20$, where $r = 1, 2$.

Latent effect 1: $\Gamma_{m,1} = 0, m = 1 : 300$,

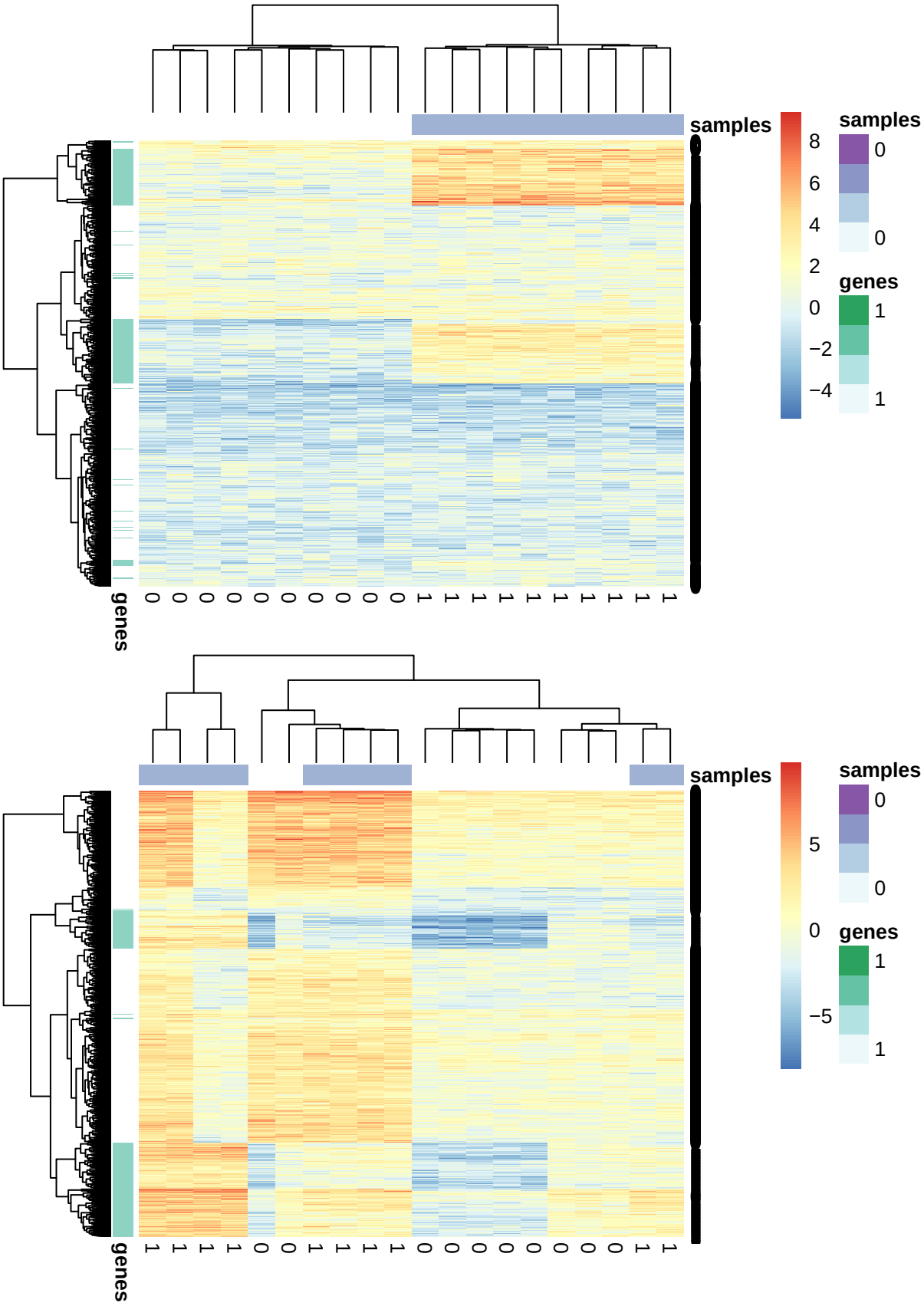
$\Gamma_{m,1} \sim N(1, 1), m = 301 : 1000$. (Positive signal overlaps with Non-DE genes, will lead to FPs if not corrected)

Latent effect 2: $\Gamma_{m,2} \sim N(-1, 1), m = 1 : 300$, $\Gamma_{m,2} = 0, m = 301 : 1000$. (Negative signal overlaps with DE genes, will lead to FNs if not corrected)

Therefore, for every gene, whether it is DE or not, it will be affected by one of the two latent variables

To ask/consider:

- Currently we have negative expression due to Latent effect 2.
- Should we normalize before running analysis?


samples
genes
0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1
8
6
4
2
0
−2
−4
samples
0
0
genes
1
1
samples
genes
1 1 1 1 0 0 1 1 1 1 0 0 0 0 0 0 0 0 1 1
5
0
−5
samples
0
0
genes
1
1

- What is the correlation between true latent variables and primary variables?

Primary case/control vs. latent 1: 0.4082483

Primary case/control vs. latent 2: 0

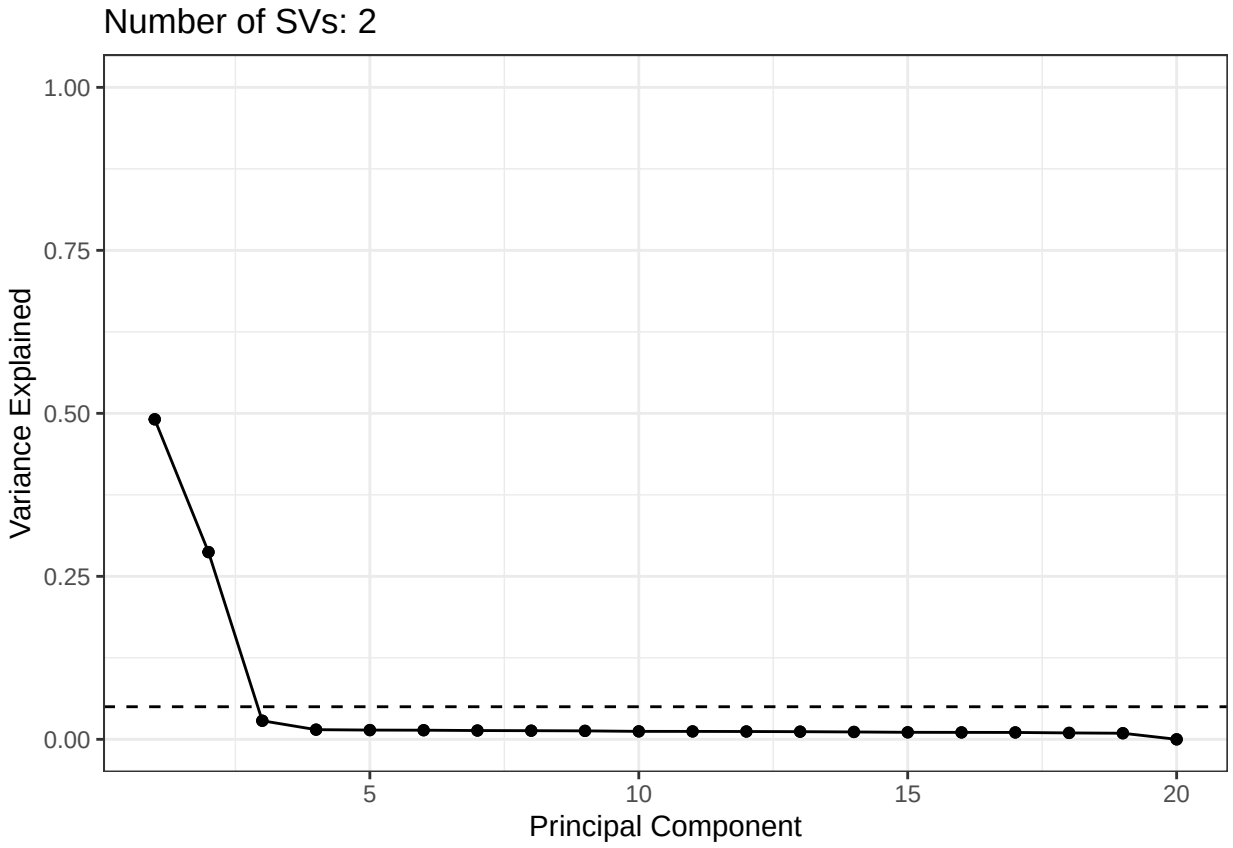
Primary design matrix vs. latent design matrix span residual (not sure): 0.3464704

Estimate the number of SVs:

```
n.sv = num.sv(X, t(S), method = "be")
cat("Number of SVs: ", n.sv, "\n")
```

```
## Number of SVs:  2
```

```
pca = prcomp(t(X))
variance = pca$sdev^2 / sum(pca$sdev^2)
qplot(c(1:length(variance)), variance) + geom_line() + geom_point() +
  geom_hline(yintercept=1/ncol(X), linetype = "dashed") +
  xlab("Principal Component") + ylab("Variance Explained") + ggtitle(paste0("Number of SVs: ", n.sv)) +
```



Estimate SVs, primary variable coefficients, and SV coefficients

- Are latent variables are spanned by the estimated SVs?
- Are the estimated coefficients similar to true coefficients?
- Is the null p-value distribution uniform?
- Do the ranks of top genes match?

```
nullMod = t(S)[, 1]
svobj = sva(X, t(S), nullMod, n.sv = n.sv)
```

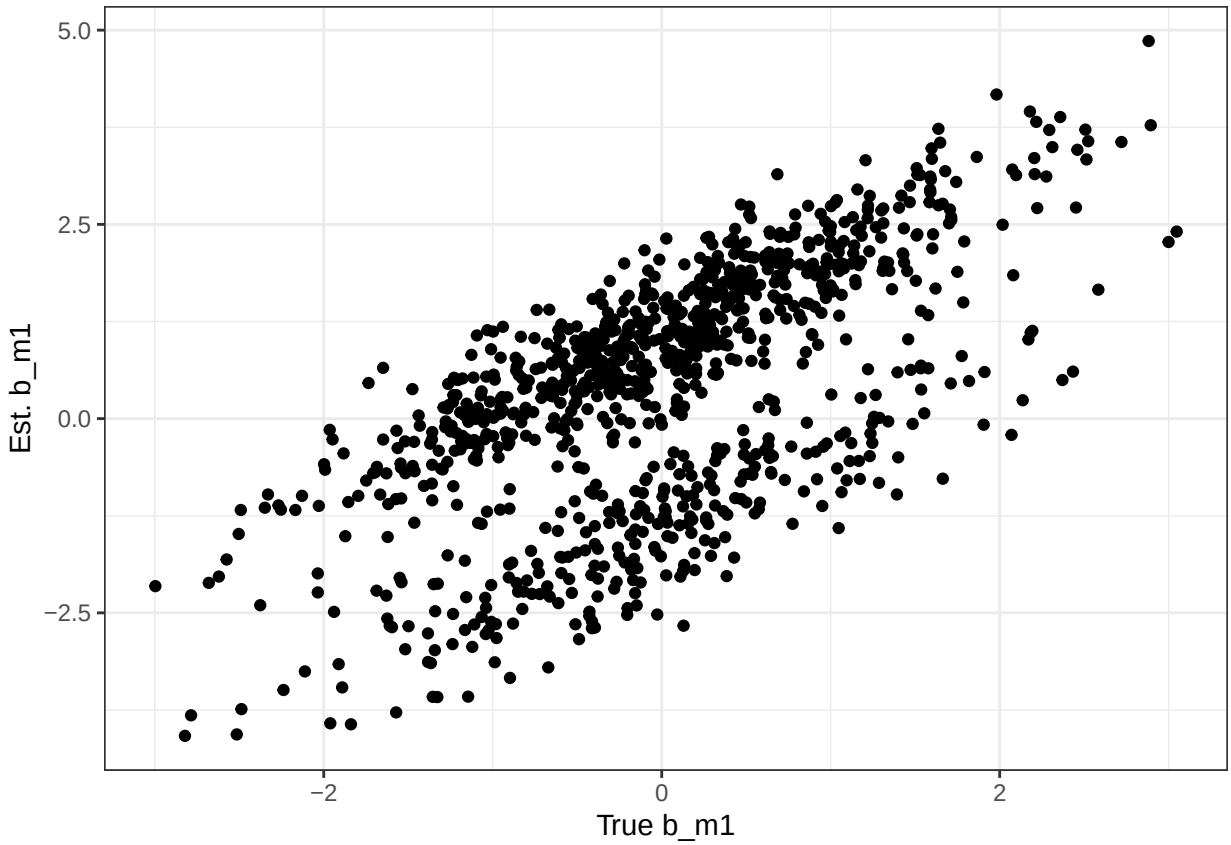
```
## Number of significant surrogate variables is:  2
## Iteration (out of 5 ):1  2  3  4  5
```

Inferred SV vs. latent design matrix span residual (not sure): 0.3381575.

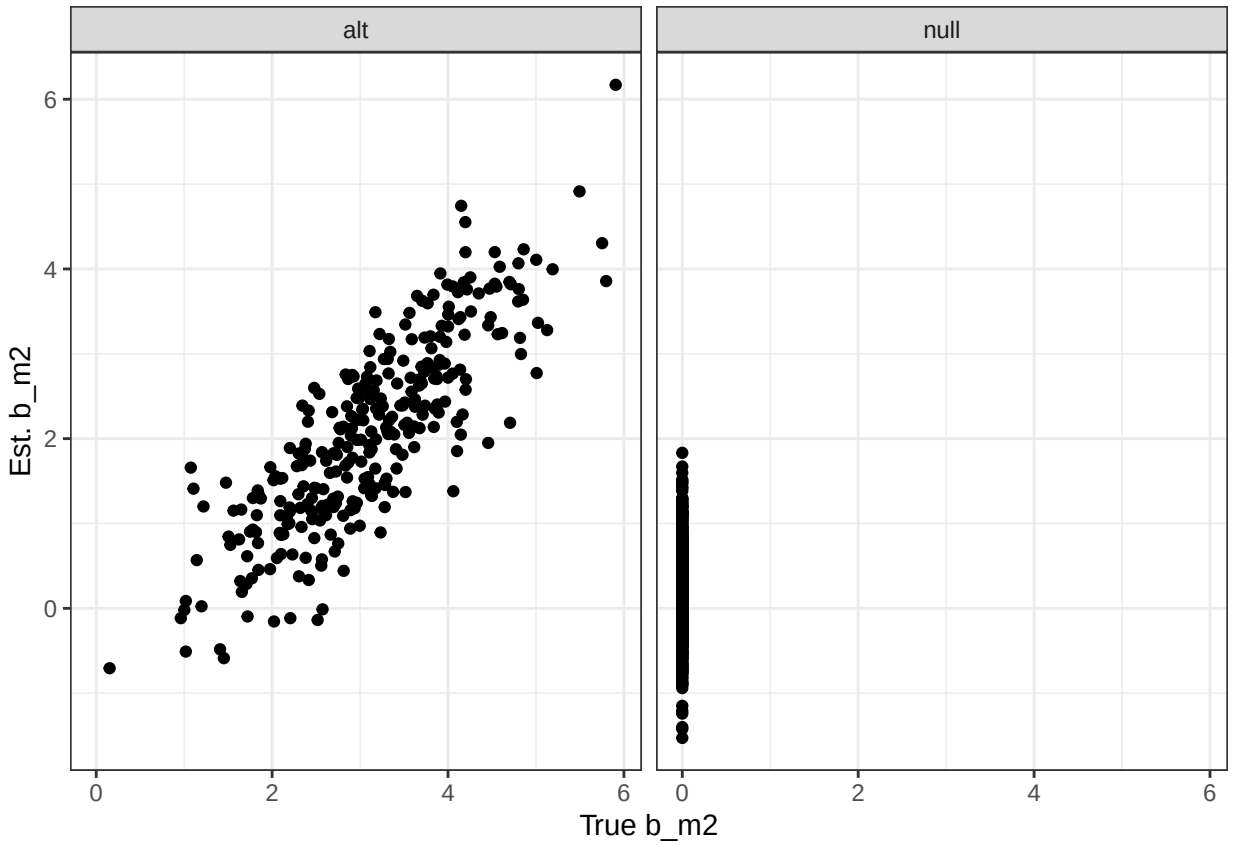
Latent 1 vs. SV 1: -0.9980922

Latent 2 vs. SV 2: -0.944612

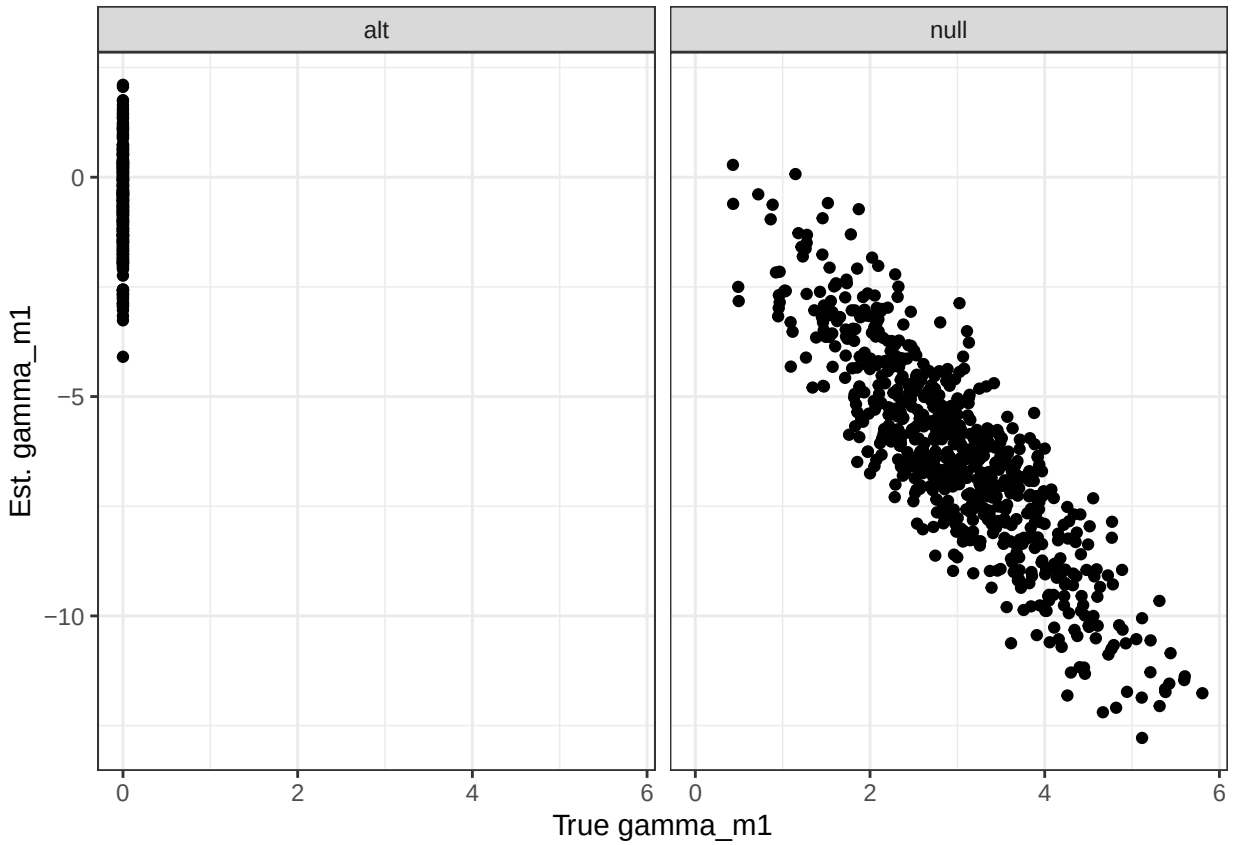
```
nullmodsv = cbind(nullMod, svobj$sv)
modsv = cbind(t(S), svobj$sv)
fitsv = lm.fit(modsv, t(X))

#visually look at predicted coefficients
plot_df = data.frame(b1 = B[, 1],
                     b2 = B[, 2],
                     b1_hat = fitsv$coefficients[1 ,],
                     b2_hat = fitsv$coefficients[2 ,],
                     b2_labels = c(rep("alt", 300), rep("null", m - 300)),
                     gamma1 = Gamma[, 1],
                     gamma1_hat = fitsv$coefficients[3 ,],
                     gamma1_labels = c(rep("alt", 300), rep("null", m - 300)),
                     gamma2 = Gamma[, 2],
                     gamma2_hat = fitsv$coefficients[4 ,],
                     gamma2_labels = c(rep("alt", 300), rep("null", m - 300)))

ggplot(plot_df, aes(b1, b1_hat)) + geom_point() + labs(x = "True b_m1", y = "Est. b_m1")
```



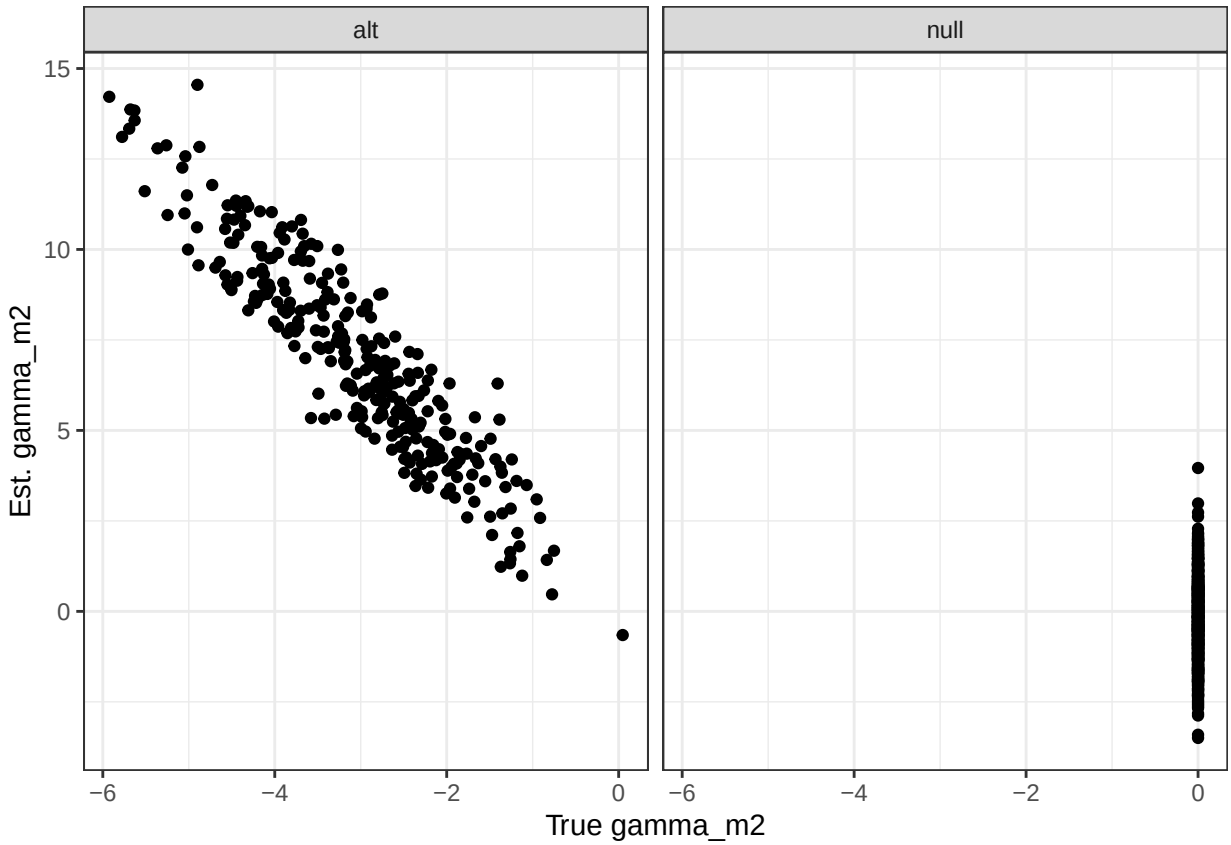

```
ggplot(plot_df, aes(b2, b2_hat)) + geom_point() + facet_wrap(~b2_labels) + labs(x = "True b_m2", y = "Es
```



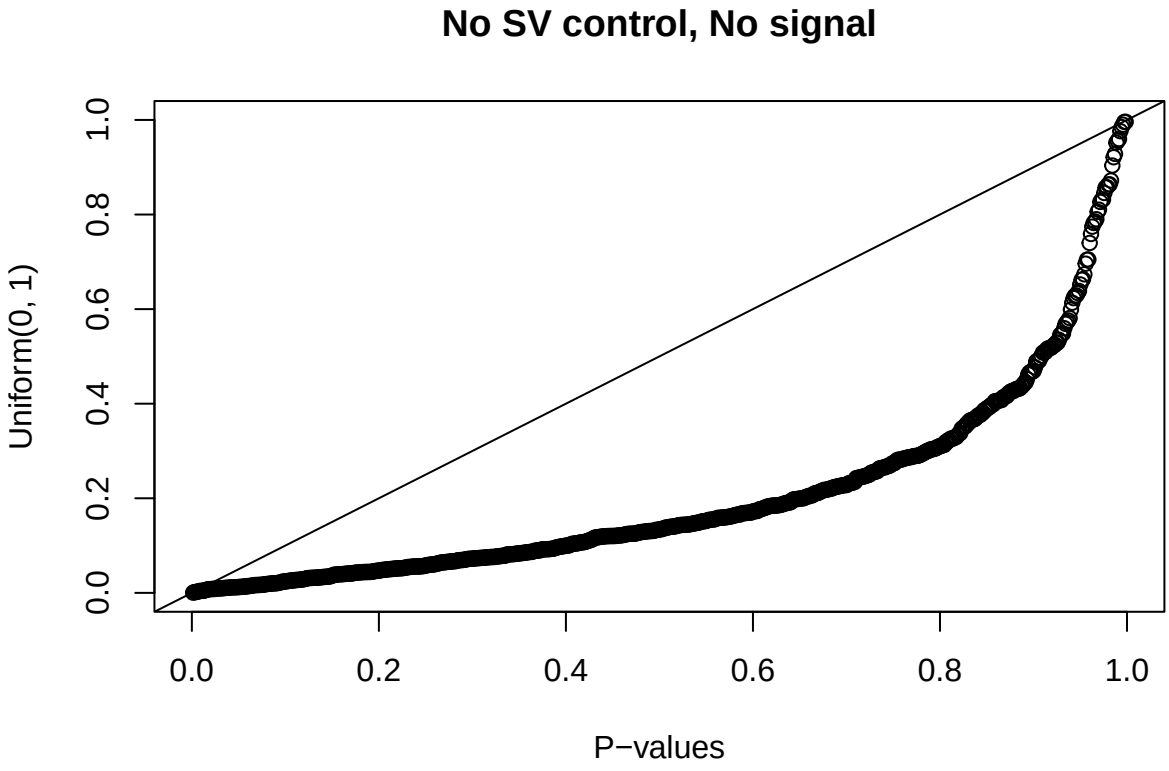

```
ggplot(plot_df, aes(gamma1, gamma1_hat)) + geom_point() + facet_wrap(~gamma1_labels) + labs(x = "True g
```

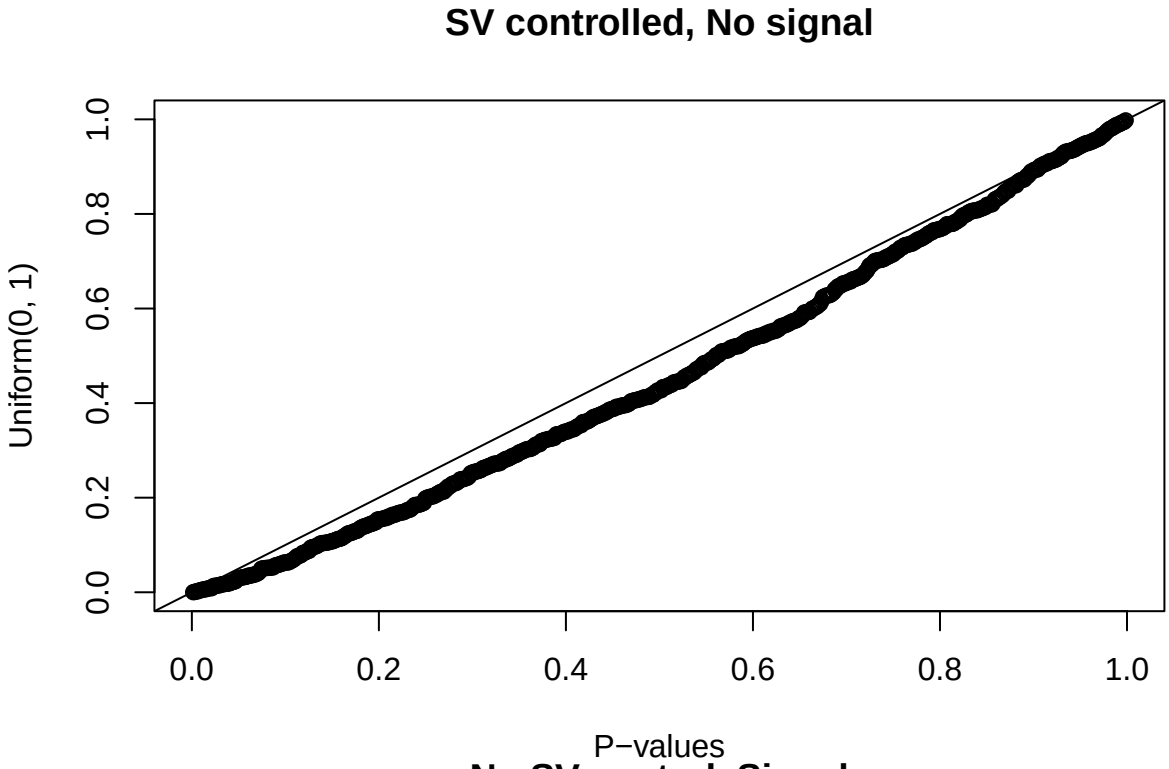


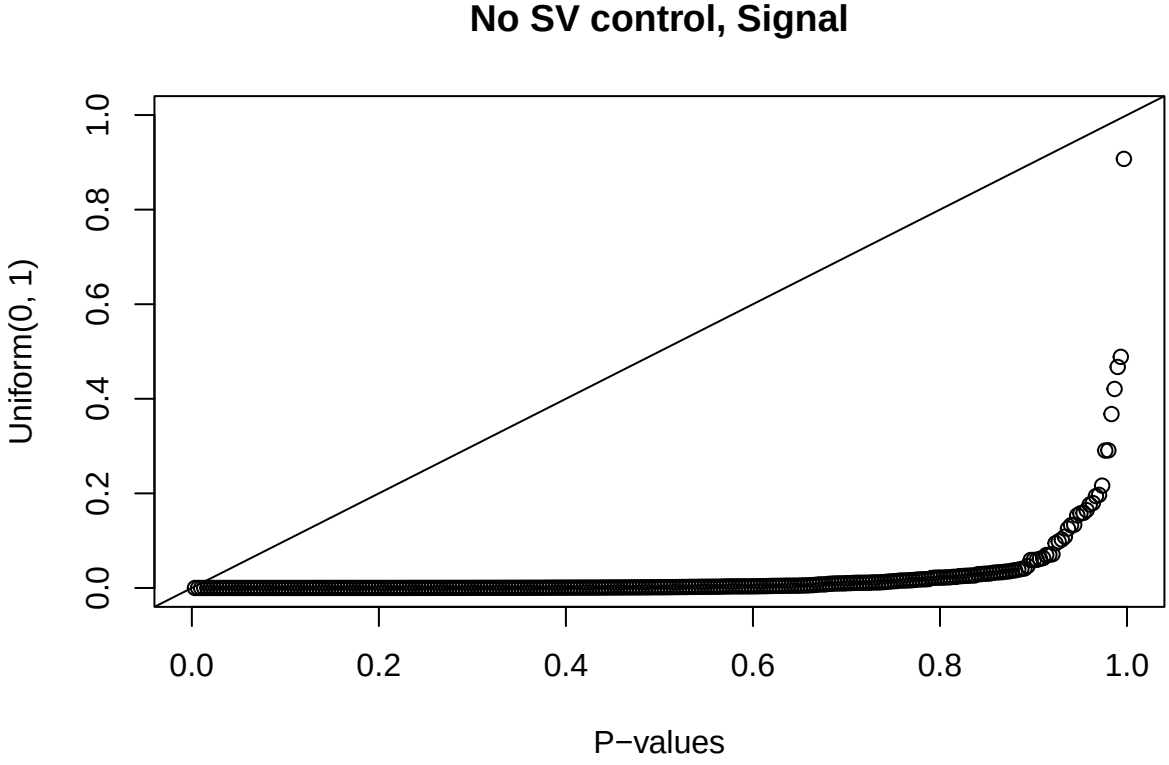

```
ggplot(plot_df, aes(gamma2, gamma2_hat)) + geom_point() + facet_wrap(~gamma2_labels) + labs(x = "True g
```

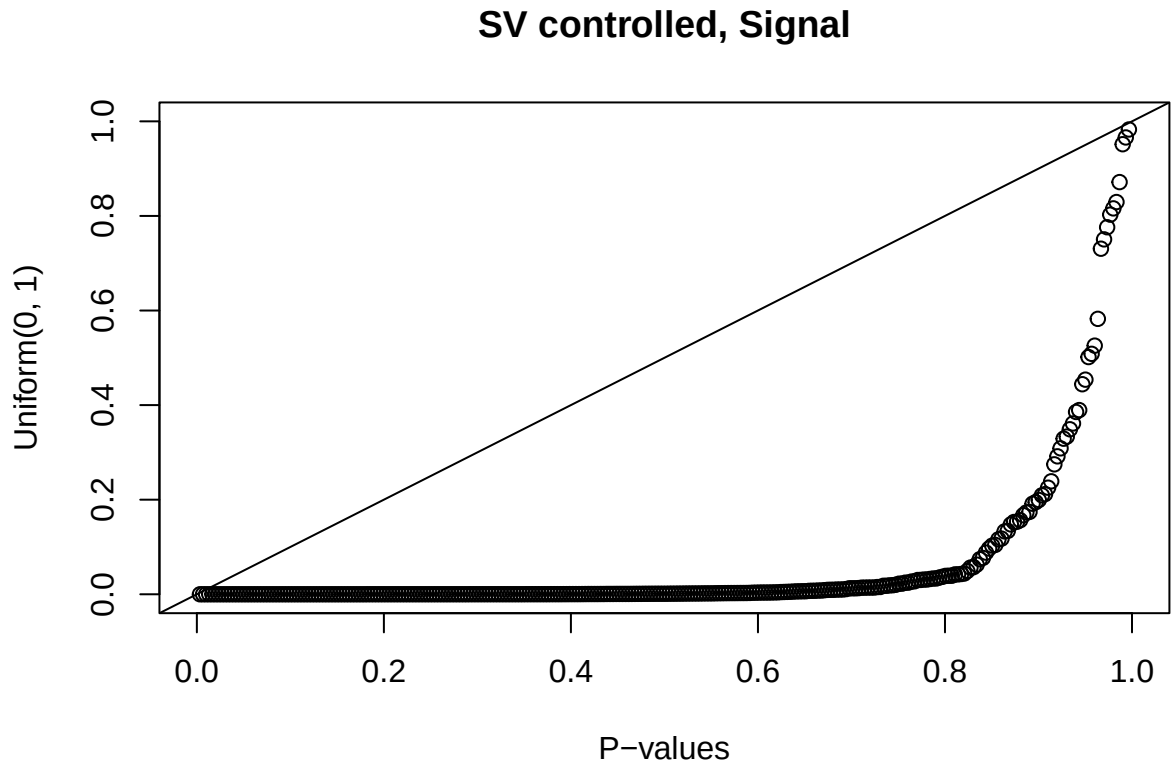


Not sure what's going on here yet regarding p-values and ranking.

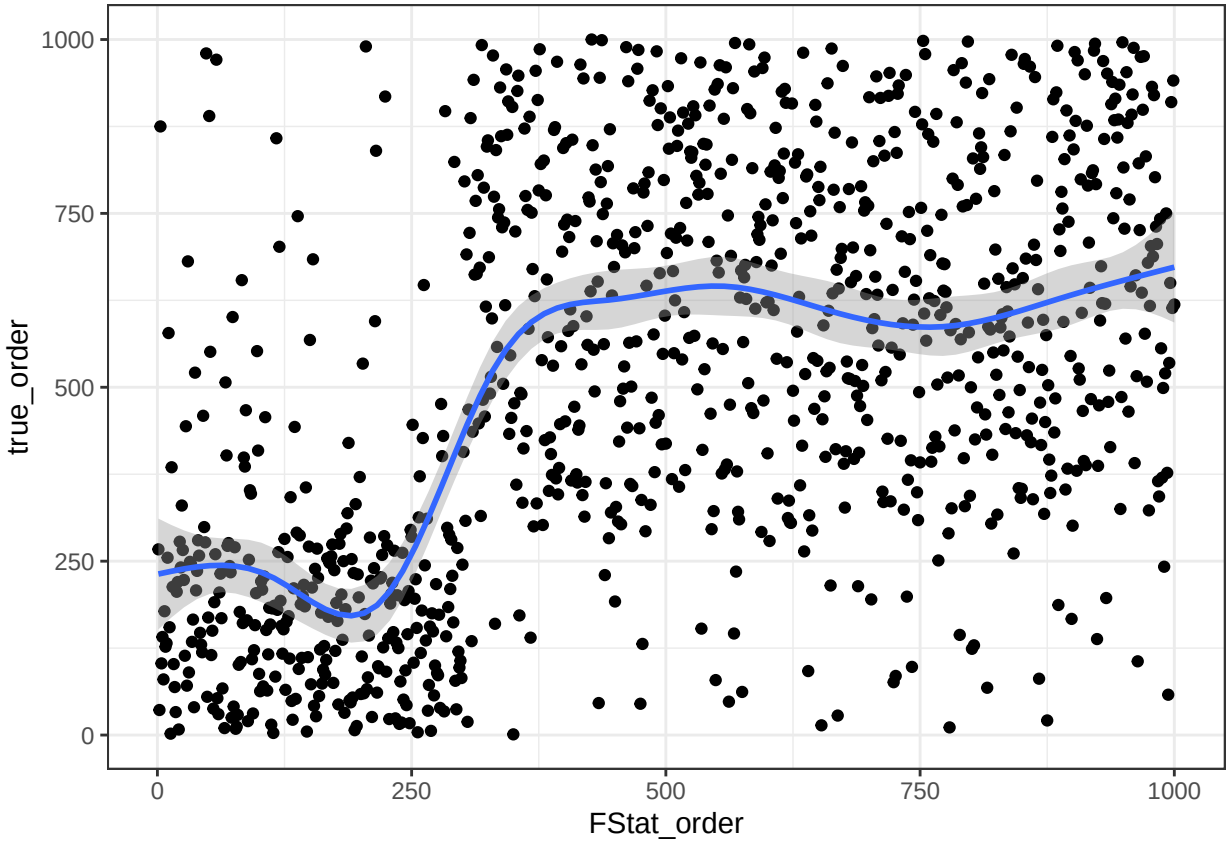



SV controlled, No signal
Uniform(0, 1)
0.0
0.2
0.4
0.6
0.8
1.0
P−values


No SV control, Signal
Uniform(0, 1)
0.0
0.2
0.4
0.6
0.8
1.0
P−values




```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



"Knobs to turn" in estimating the number of SVs

$\Gamma_{m,1}$: If strong effect relative to $b_{m,1}$ (fixed), then this will generate noise on control samples, leading to false positives.

$\Gamma_{m,2}$: If strong effect relative to $b_{m,2}$ (fixed), then this will generate noise on case samples, leading to false negatives.

Our certainty of Γ to effect case or control samples depends on "the percentage of row space of S explained by G". We appropx that by looking at $cor(G_r, S_2), r = 1, 2$. We probably can fix this value for now.

Knob Speculation, within one experiment

$\Gamma_{m,1}$	$\Gamma_{m,2}$	$cor(G_r, S_2)$	**DE**	**Scree plot**
strong	weak	strong	more FPs	more even PCA
weak	strong	strong	more FNs	more even PCA
weak	weak	strong	neutral	more dominated PCA
strong	strong	strong	more FPs and FNs	more even PCA

Simulation with multiple experiments

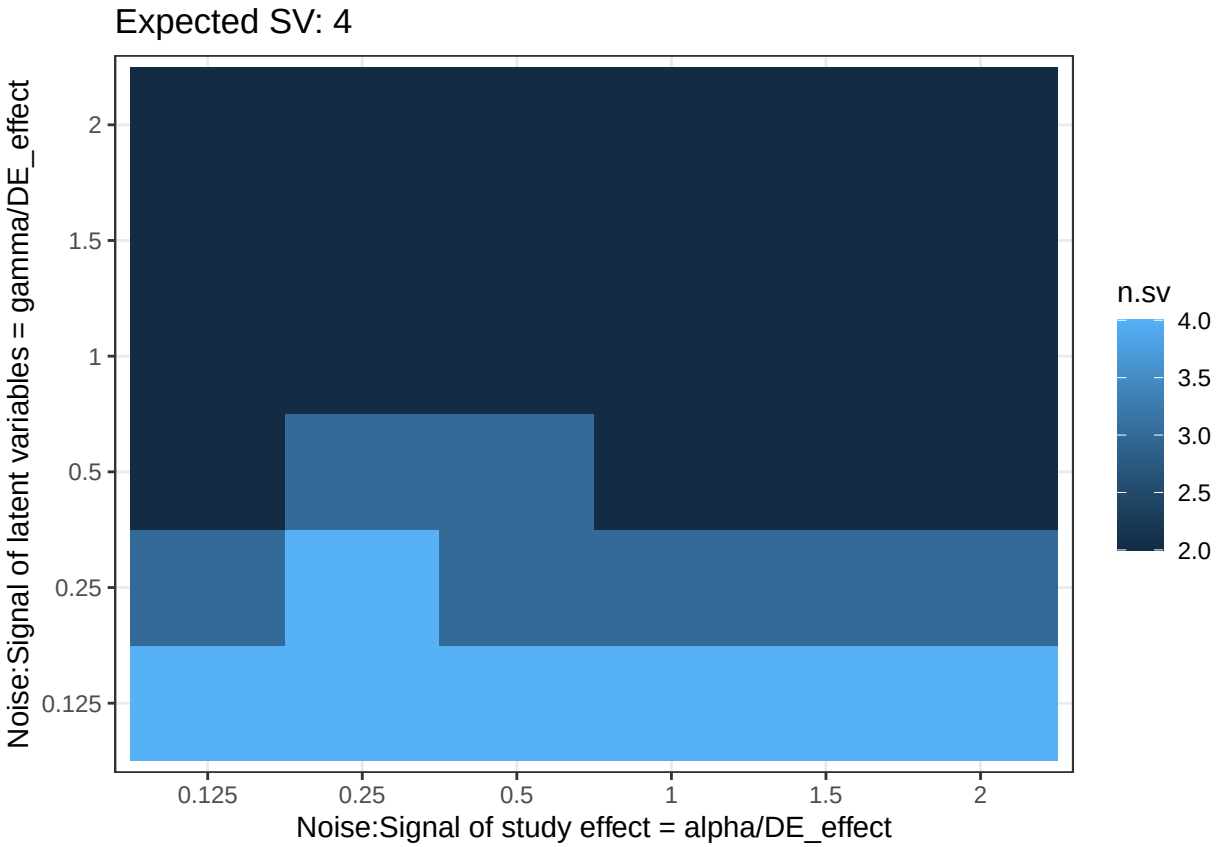
$$X = \left[\begin{array}{c|c} B_1 & B_2 \end{array}\right] \left[\begin{array}{c|c} S_1 & 0 \\ \hline 0 & S_2 \end{array}\right] + \left[\begin{array}{c|c} \Gamma_1 & \Gamma_2 \end{array}\right] \left[\begin{array}{c|c} G_1 & 0 \\ \hline 0 & G_2 \end{array}\right] + U$$

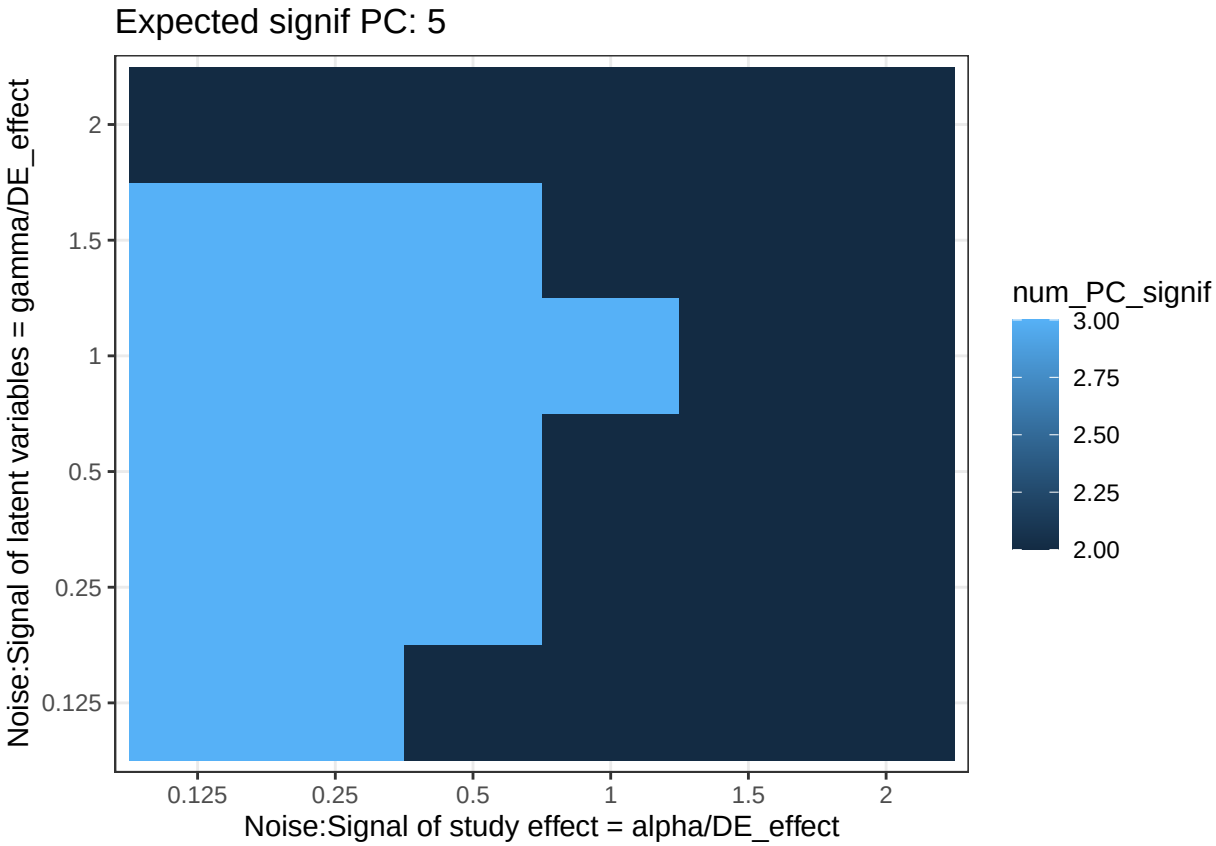
$$S_1 = \begin{bmatrix} 1 & ... & 1 & ... & 1 \\ 0 & ... & 1 & ... & 1 \end{bmatrix}$$

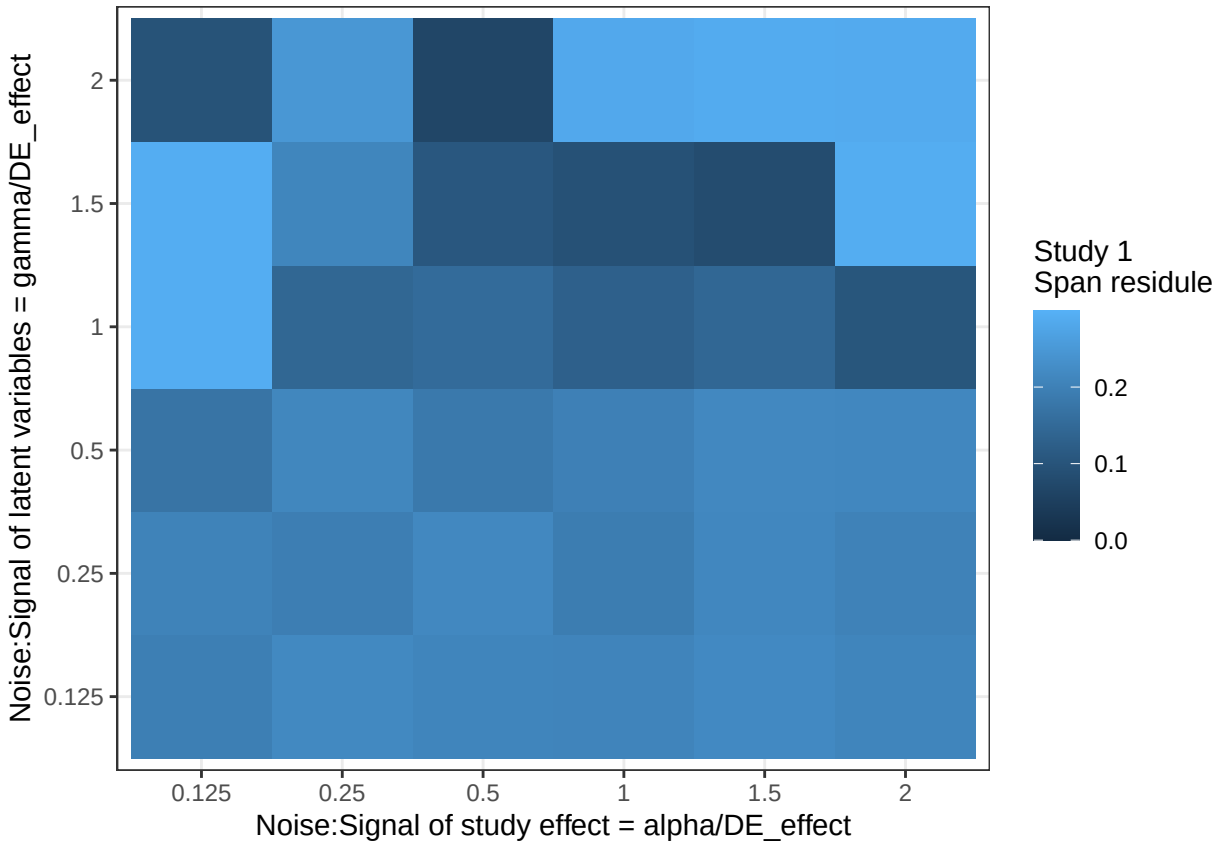
$$S_2 = \begin{bmatrix} 1 & ... & 1 & ... & 1 \\ 0 & ... & 1 & ... & 1 \end{bmatrix}$$

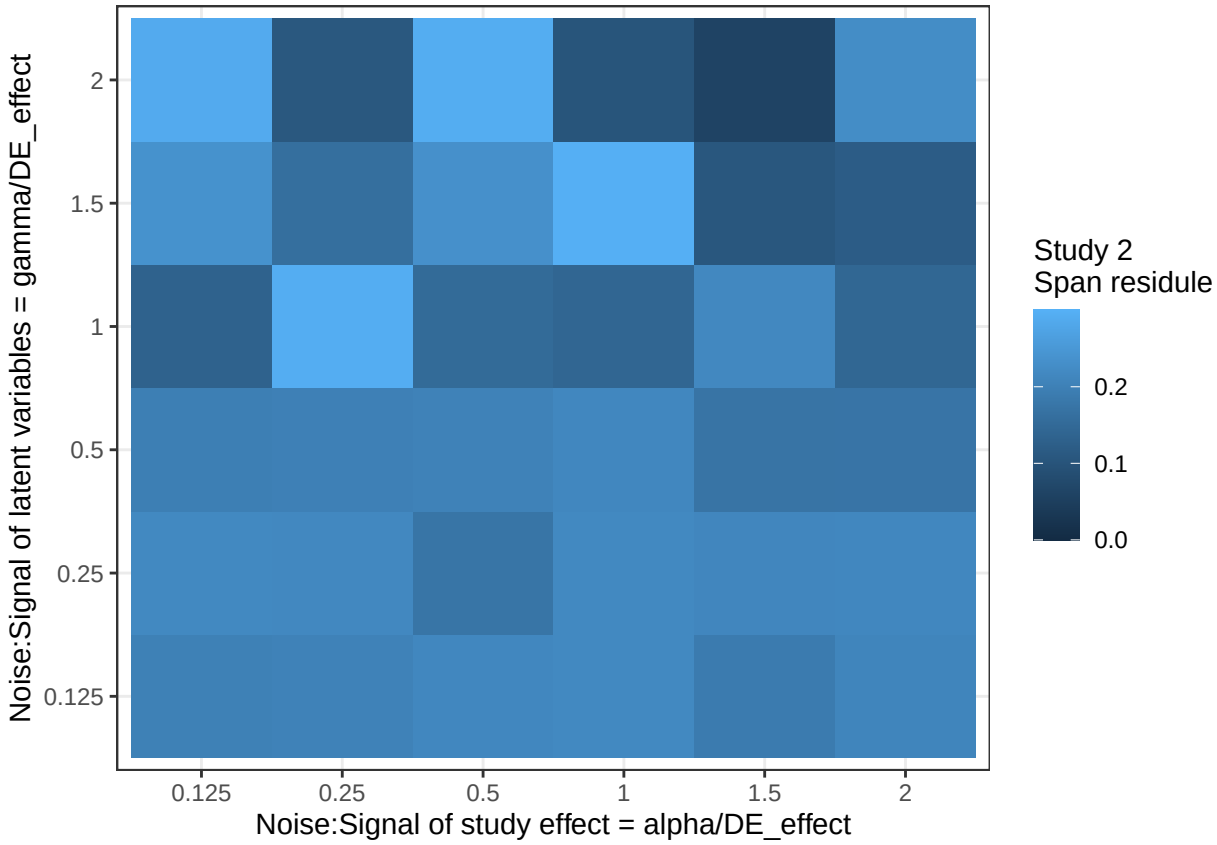
where study 2's effect relative to study 1 is modeled as $B_{21} \sim N(\alpha, 1)$ where B_1 and B_2 have the same set of DE genes with equal effects.

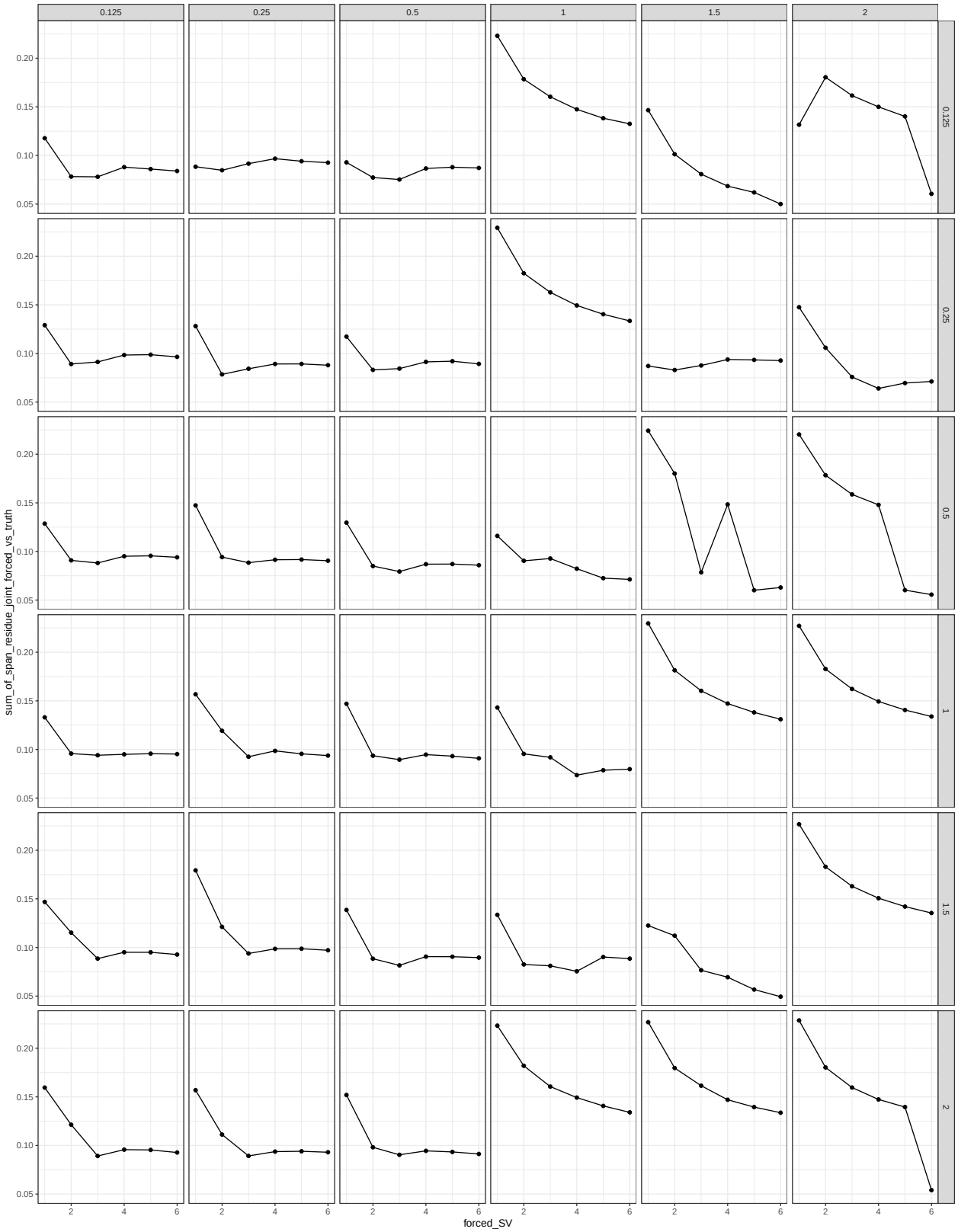
We vary SV effects γ via $\Gamma_{1:2,m,1} = 0, m = 1:300$, $\Gamma_{1:2,m,1} \sim N(\gamma, 1), m = 301:1000$.


Expected SV: 4
Noise:Signal of latent variables = gamma/DE_effect
2
1.5
1
0.5
0.25
0.125
0.125
0.25
0.5
1
1.5
2
Noise:Signal of study effect = alpha/DE_effect
n.sv
4.0
3.5
3.0
2.5
2.0


Expected signif PC: 5
Noise:Signal of latent variables = gamma/DE_effect
2
1.5
1
0.5
0.25
0.125
0.125
0.25
0.5
1
1.5
2
Noise:Signal of study effect = alpha/DE_effect
num_PC_signif
3.00
2.75
2.50
2.25
2.00


Noise:Signal of latent variables = gamma/DE_effect
2
1.5
1
0.5
0.25
0.125
0.125
0.25
0.5
1
1.5
2
Noise:Signal of study effect = alpha/DE_effect
Study 1
Span residule
0.2
0.1
0.0


Noise:Signal of latent variables = gamma/DE_effect
2
1.5
1
0.5
0.25
0.125
0.125
0.25
0.5
1
1.5
2
Noise:Signal of study effect = alpha/DE_effect
Study 2
Span residule
0.2
0.1
0.0


0.125
0.25
0.5
1
1.5
2
sum_of_span_residue_joint_forced_vs_truth
forced_SV
0.05
0.10
0.15
0.20
2
4
6