High-dimensional data

" many regressions model "

A single design matrix $X_{d \times n}$

Many, m, response variables: $\underline{y}_i$ length n

$i = 1, 2, \dots, m$

$$\left. Y_{m \times n} = B_{m \times d} X_{d \times n} + E_{m \times n} \right]$$

$\underline{\beta}_i$ is $i^{\underline{th}}$ row of B (length d)

$\beta_i = (\beta_{i1}, \beta_{i2}, \dots, \beta_{id})$

m regressions

$m \gg n \gg d$

Two topics

① Jackstraw

② Surrogate variable analysis

Estimating latent variables in our setting

Suppose $Y_{m \times n} = \Phi_{m \times r} Z_{r \times n} + \underline{E_{m \times n}} \leftarrow$

Z are unobserved latent variables

Φ are unknown parameters

Assume that all e_{ij} ($\in E$) are independent and

$$e_{i1}, \ldots, e_{in} \overset{iid}{\sim} (\underset{\substack{\uparrow \\ \text{mean}}}{0}, \underset{\substack{\uparrow \\ \text{variance}}}{\sigma_i^2})$$

$$\begin{aligned}
\underset{\substack{\uparrow \\ n \times n}}{\frac{1}{m} Y^T Y} = & \frac{1}{m} (Y - \Phi Z)^T (Y - \Phi Z) && (1) \\
& + \frac{1}{m} (Y - \Phi Z)^T \Phi Z && (2) \\
& + \frac{1}{m} (\Phi Z)^T (Y - \Phi Z) && (3) \\
& + \frac{1}{m} \boxed{(\Phi Z)^T \Phi Z} && (4)
\end{aligned}$$

$$\underline{Y = (Y - \Phi Z) + (\Phi Z)}$$

$$\lim_{m\to\infty} \frac{1}{m} Y^T Y \overset{\text{with prob. 1}}{=} \underset{(1)}{D} + \underset{(2)}{0} + \underset{(3)}{0} + \underset{(4)}{\Pi_Z}$$

Π_Z is the row space of Z

D is $n \times n$ diagonal matrix where each element (i,i) is $\lim_{m\to\infty} \frac{1}{m} \sum_{k=1}^{m} \sigma_k^2$

$D \propto I$ (identity)

This implies that the first r eigenvectors of $\frac{1}{m} Y^T Y$ converge to Π_Z with prob. 1.

(Leek 2011 Biometrics) $\hat{Z} = $ top r eigenvectors of $\frac{1}{m} Y^T Y$

Jackstraw

A method to perform inference on $\Phi_{m\times r}$.

Let's suppose we to test :

$$H_0: \phi_i = 0 \quad \text{vs.} \quad H_1: \phi_i \neq 0$$

① Estimat $\hat{Z}$ and obtain an association statistic t_i for each response variable $\underline{y}_i$.

② Take a subset of rows of Y of size s. Permute indepently the s rows to obtain Y^*.

$m-s$ rows are intact
s rows are permuted

③ Obtain $\hat{Z}^*$ from Y^* and obtain s statistics t_i^* from my s permuted response variables. This yield s null statistics t_i^*.

④ Repeat steps ② and ③ B times. Yield Bs null statistics.

$$p_i = \frac{1}{Bs} \sum_{b=1}^{B} \sum_{k=1}^{s} \mathbb{1}(t^*_{bk} \geq t_i)$$

Trade-off:

s small $\Rightarrow$ more accurate but slow

s large $\Rightarrow$ less accurate (conservatively) but faster

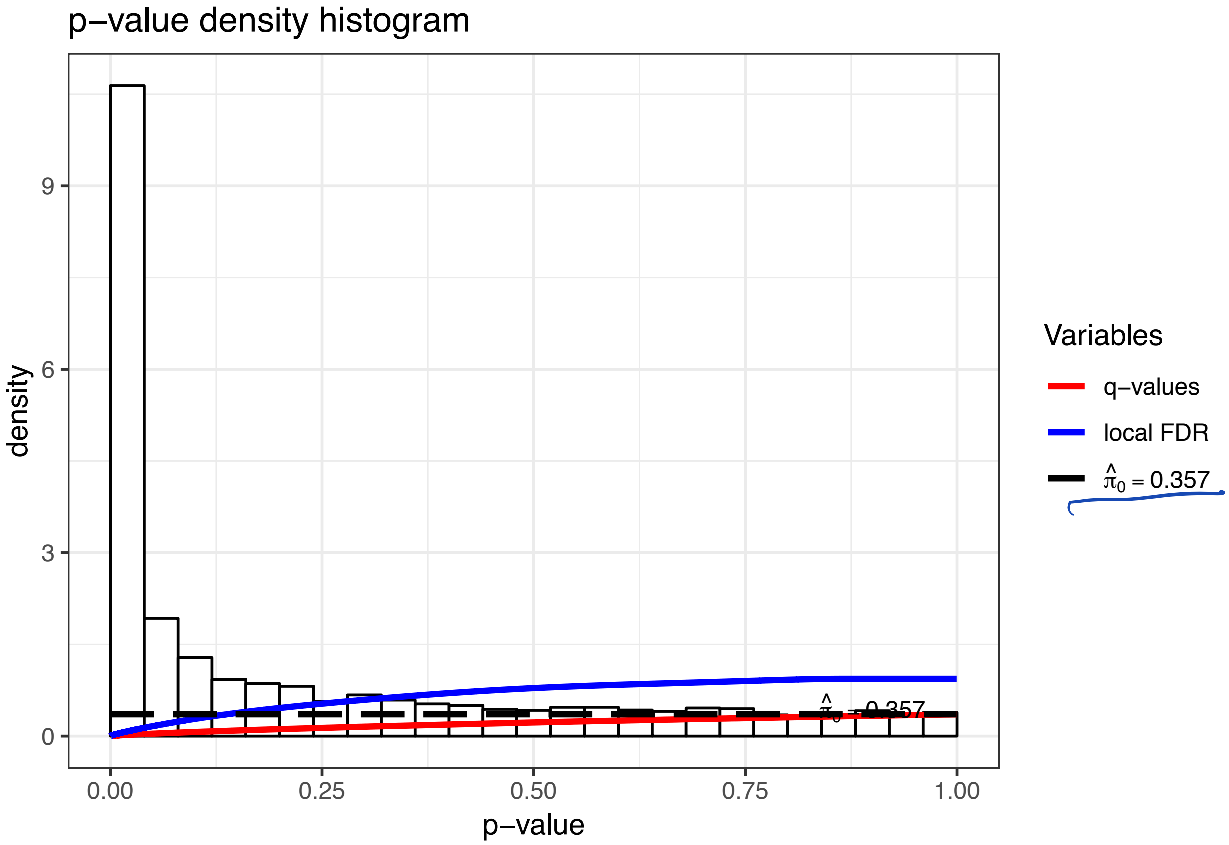
Jackstraw Example: Yeast Cell Cycle

Recall the yeast cell cycle data from earlier. We will test which genes have expression significantly associated with PC1 and PC2 since these both capture cell cycle regulation.

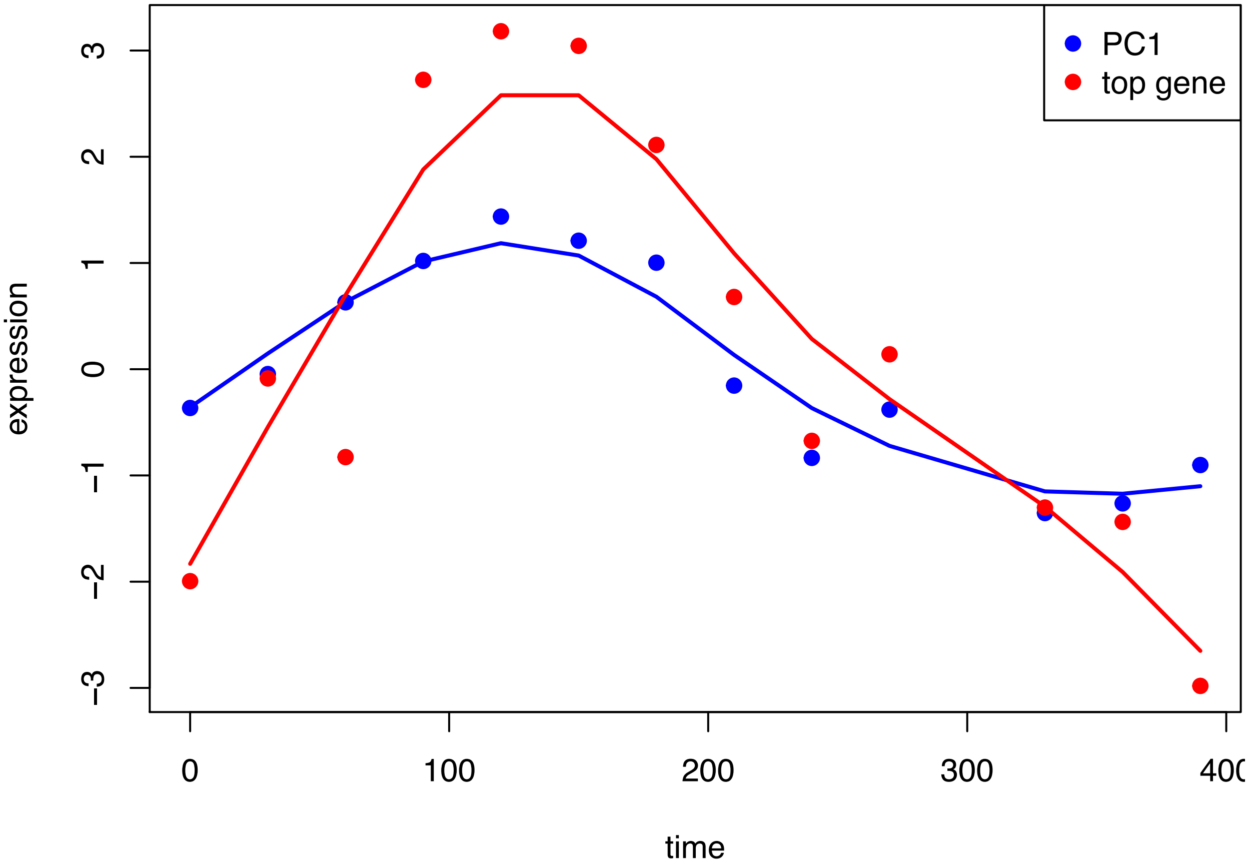
```
> load("./data/spellman.RData")
> time
 [1]   0  30  60  90 120 150 180 210 240 270 330 360 390
> dim(gene_expression)
[1] 5981   13
> dat <- t(scale(t(gene_expression), center=TRUE, scale=FALSE))
```

Test for associations between PC1 and each gene, conditioning on PC1 and PC2 being relevant sources of systematic variation.

```
> jsobj <- jackstraw_pca(dat, r1=1, r=2, B=500, s=50, verbose=FALSE)
> jsobj$p.value %>% qvalue() %>% hist()
```

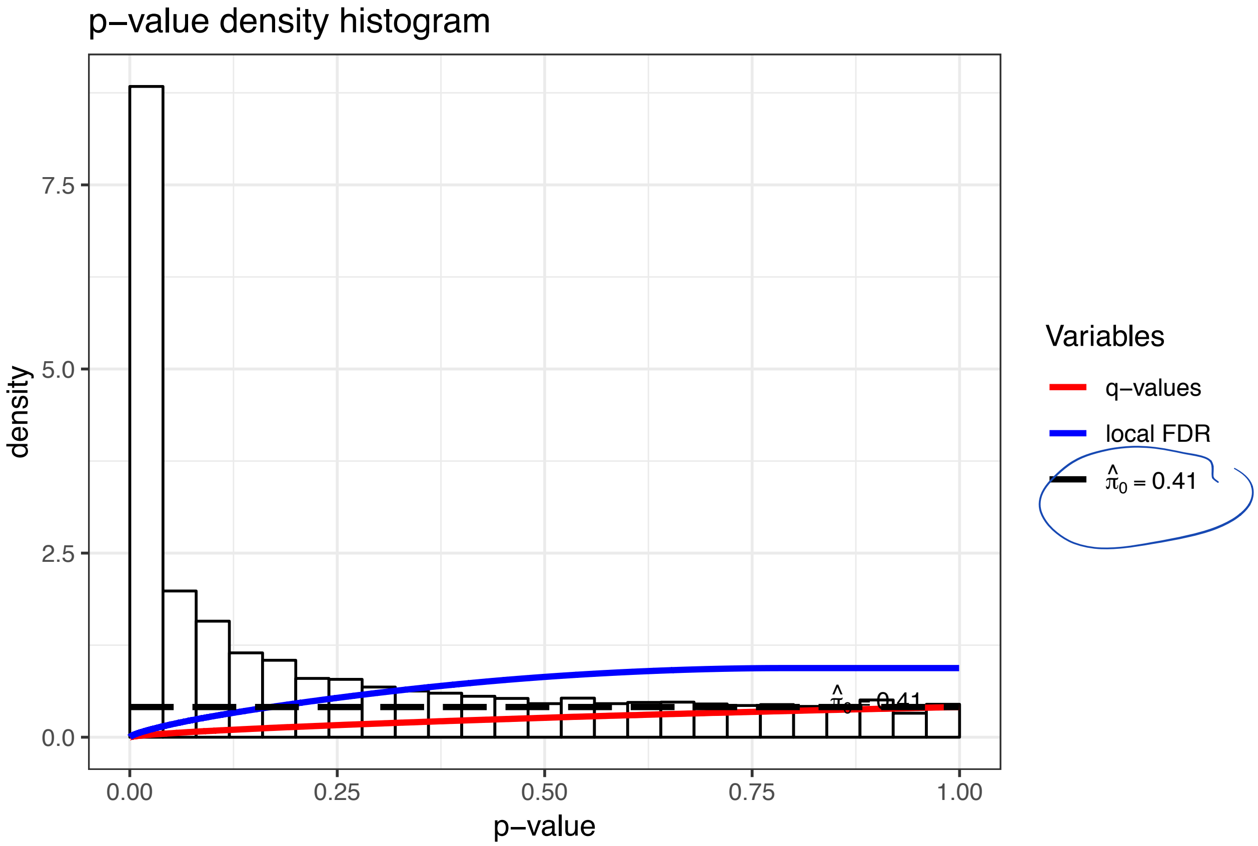


This is the most significant gene plotted with PC1.

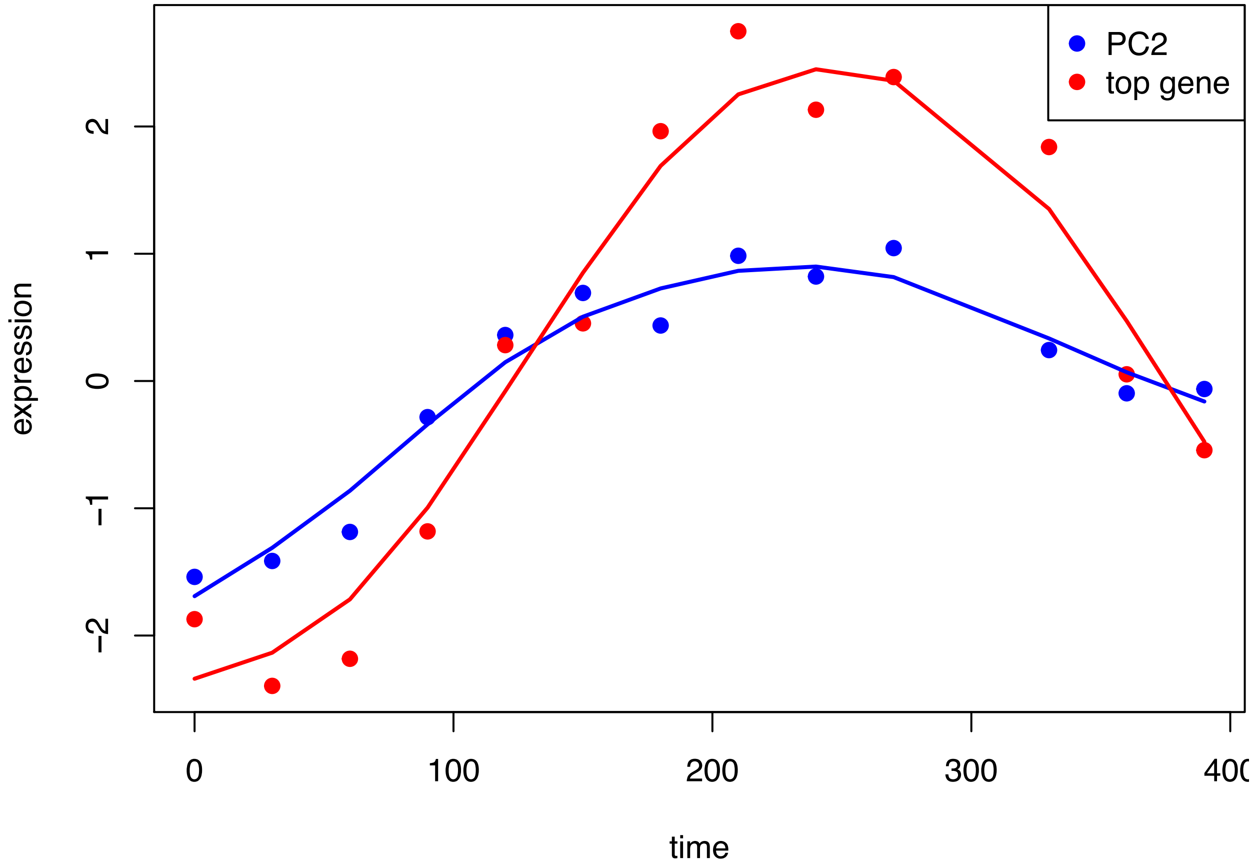


Test for associations between PC2 and each gene, conditioning on PC1 and PC2 being relevant sources of systematic variation.

```
> jsobj <- jackstraw_pca(dat, r1=2, r=2, B=500, s=50, verbose=FALSE)
> jsobj$p.value %>% qvalue() %>% hist()
```



This is the most significant gene plotted with PC2.



Surrogate Variable Analysis

$$Y_{m\times n} = B_{m\times d} X_{d\times n} + \underline{\Phi_{m\times r} Z_{r\times n}} + E_{m\times n}$$

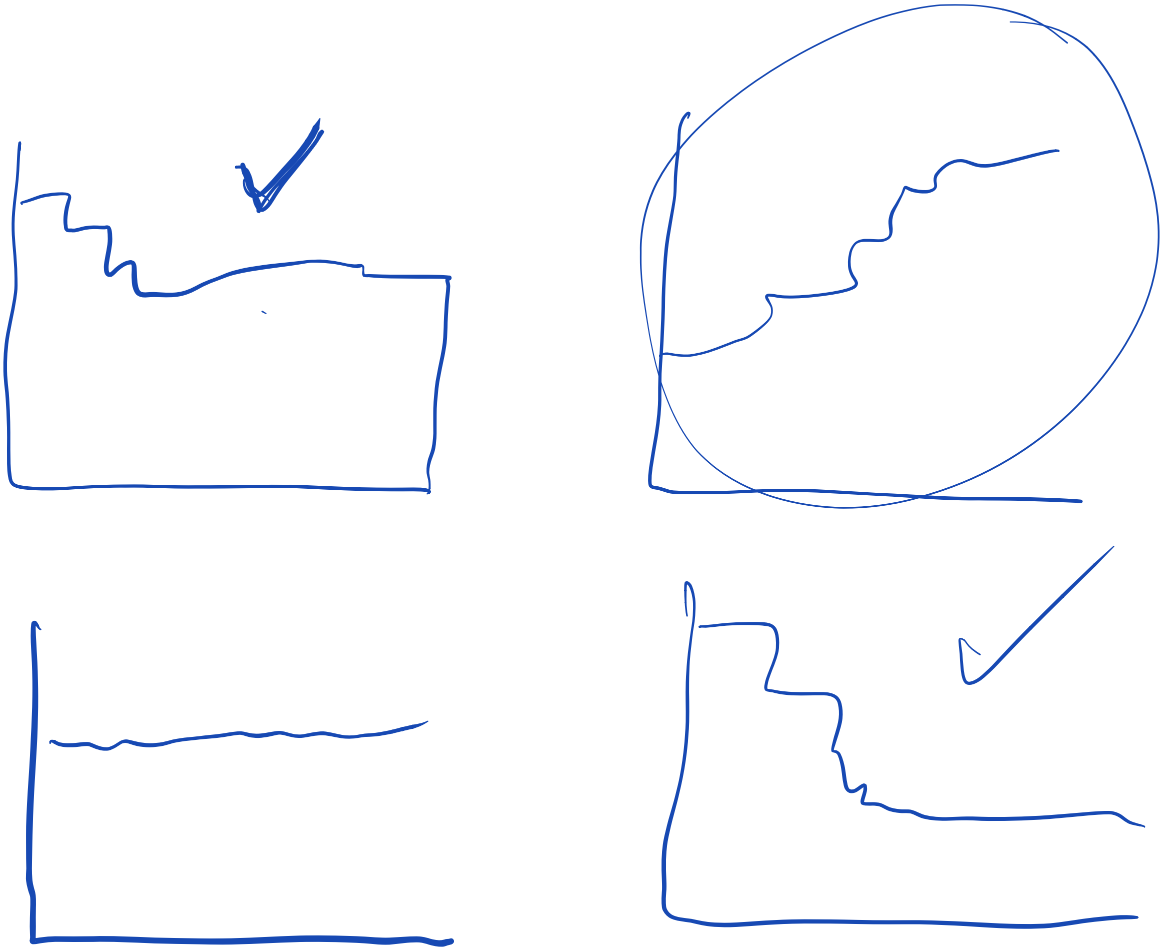
$$e_{i1}, e_{i2}, \ldots, e_{in} \overset{iid}{\sim} (0, \sigma_i^2)$$

$$m \gg n \gg d, r$$

X and Y are observed

Want to do inference on B

Need to deal with ΦZ



Basic Idea

Iterate:

1. Estimating Z from

$$Y - \tilde{B} X$$

2. Estimating B from

$$Y - \hat{\Phi} \hat{Z}$$

We showed:

- $\tilde{B}$ needs to be regularized

If $\tilde{B} = \hat{B}_{OLS}$ then

$$Y - \hat{B}_{OLS} X$$

only captures the part of ΦZ that is orthogonal to X.

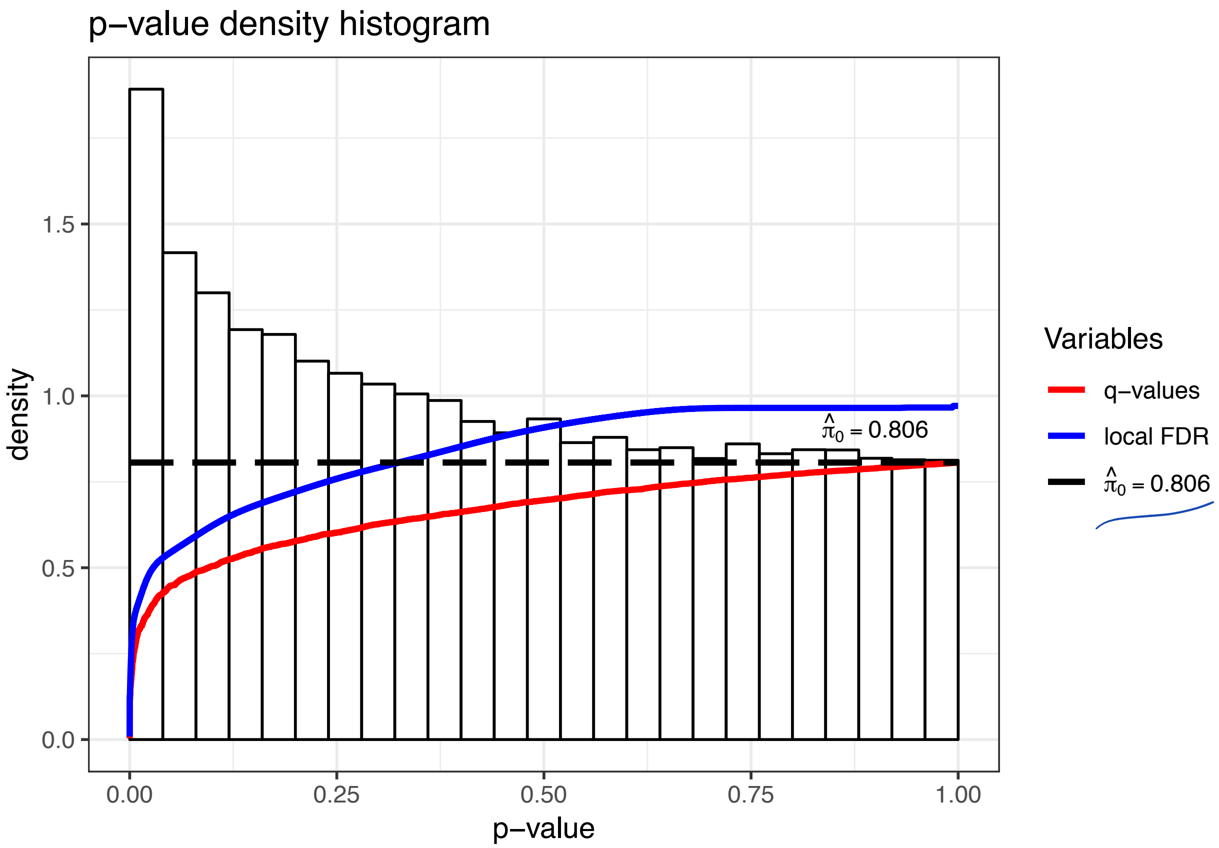
- The EM algorithm to estimate Z takes the above form

Surrogate Variable Analysis Example: Kidney Expression by Age

In Storey et al. (2005), we considered a study where kidney samples were obtained on individuals across a range of ages. The goal was to identify genes with expression associated with age.

```
> library(edge)
> library(splines)
> load("./data/kidney.RData")
> age <- kidcov$age
> sex <- kidcov$sex
> dim(kidexpr)
[1] 34061    72
> cov <- data.frame(sex = sex, age = age)
> null_model <- ~sex
> full_model <- ~sex + ns(age, df = 3)
```

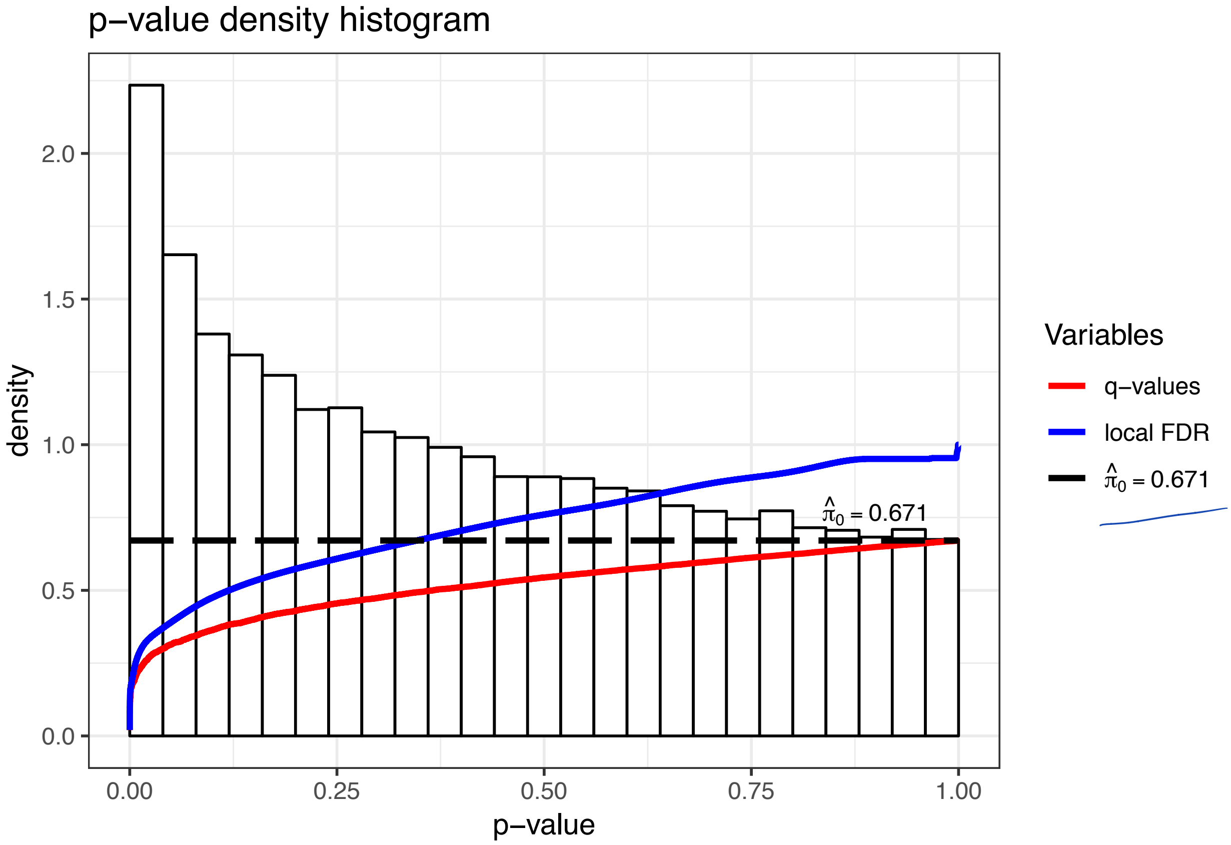
```
> de_obj <- build_models(data = kidexpr, cov = cov,
+                         null.model = null_model,
+                         full.model = full_model)
> de_lrt <- lrt(de_obj, nullDistn = "bootstrap", bs.its = 100, verbose=FALSE
> qobj1 <- qvalueObj(de_lrt)
> hist(qobj1)
```



Now that we have completed a standard generalized LRT, let's estimate $\boldsymbol{Z}$ (the surrogate variables) using the `sva` package as accessed via the `edge` package.

```
> dim(nullMatrix(de_obj))
[1] 72  2
> de_sva <- apply_sva(de_obj, n.sv=4, method="irw", B=10)
Number of significant surrogate variables is:  4
Iteration (out of 10 ):1  2  3  4  5  6  7  8  9  10
> dim(nullMatrix(de_sva))
[1] 72  6
> de_svalrt <- lrt(de_sva, nullDistn = "bootstrap", bs.its = 100, verbose=F
```

```
> qobj2 <- qvalueObj(de_svalrt)
> hist(qobj2)
```



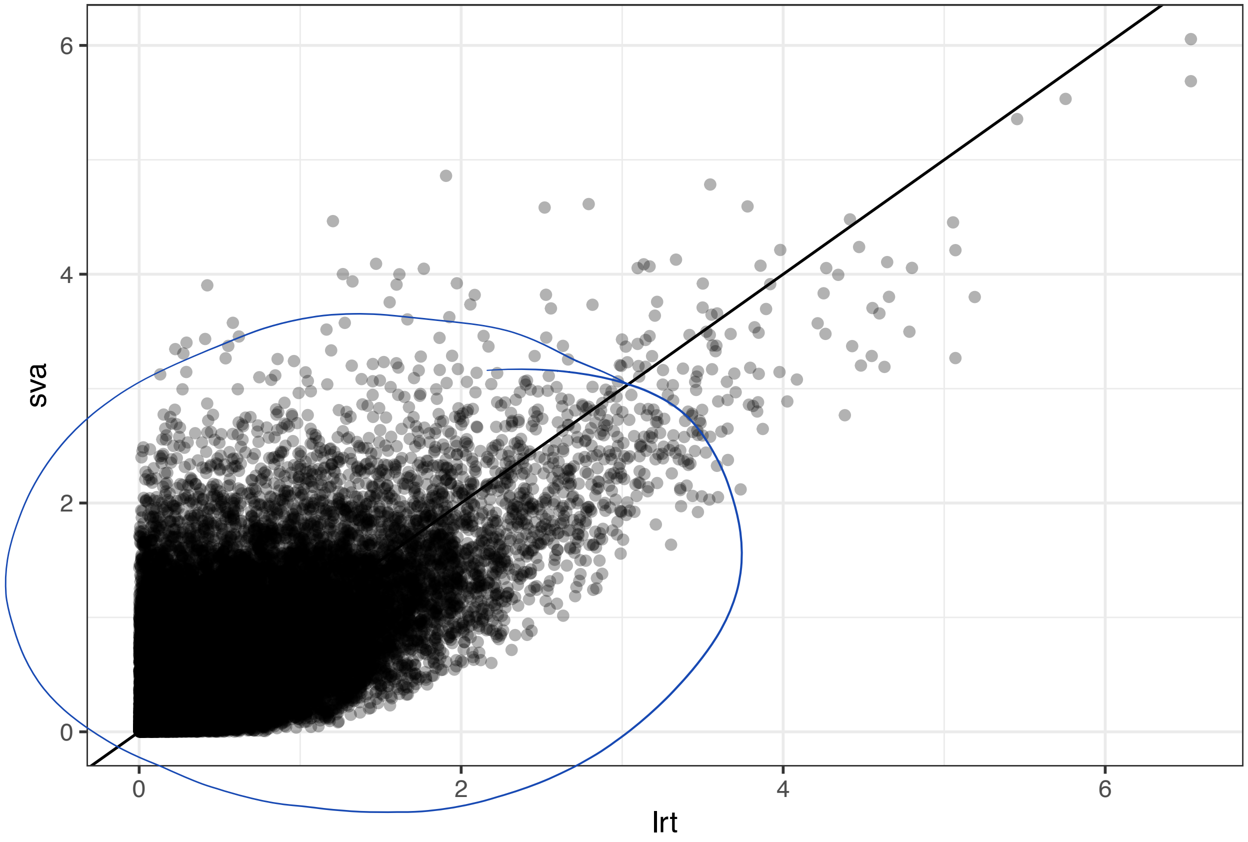
```
> summary(qobj1)

Call:
qvalue(p = pval)

pi0:    0.8059662

Cumulative number of significant calls:
```

```
           <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1    <1
p-value        28    175   879   1802  3064 5431 34061
q-value         0      0     2      4    16   30 34061
local FDR       0      0     2      2     8   21 34061
```

```
> summary(qobj2)

Call:
qvalue(p = pval)

pi0:    0.6708454

Cumulative number of significant calls:

           <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1    <1
p-value        26    151  1022   2081  3635 6279 34061
q-value         0      0     0      3     4   47 34061
local FDR       0      0     0      1     3   28 34049
```

P-values from two analyses are fairly different.

```
> data.frame(lrt=-log10(qobj1$pval), sva=-log10(qobj2$pval)) %>%
+   ggplot() + geom_point(aes(x=lrt, y=sva), alpha=0.3) + geom_abline()
```


sva
lrt