using R for regression model selection with adaptive penalties procedures based on the FDR criteria”

Tal Galili

Tel Aviv University

Based on the paper by YOAV BENJAMINI and YULIA GAVRILOV

“A SIMPLE FORWARD SELECTION PROCEDURE BASED ON FALSE DISCOVERY RATE CONTROL”

(*Annals of Applied Statistics* 2009)

$X = (x_1, \ldots, x_m)$ is an $n \times m$ matrix

$\beta = (\beta_1, \ldots, \beta_m)$ (Some <u>Zeros</u>)

$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m) \sim N(0, \sigma^2 I)$

$$Y = X\beta + \varepsilon$$

<u>Task:</u> Stopping rule (finding the “Best model” on the Forward selection path)

Why forward selection ?

Motivation – Big (m) datasets:

1) Fast results
   - Simple models
   - Simple procedure
2) Good results

3) Easy to use

$$Y = X\beta + \varepsilon$$

Finding variables

Over fitting

Model
Size

Penalty

$$\overbrace{RSS_k + \sigma^2 k\lambda}^{\text{Minimize}}$$

How to choose λ ?

λ type	examples	for "big" models
constant λ	• $\lambda_a = 2$ (AIC)	•Over fitting
Non-constant (adaptive) λ	• $\lambda_n = \log(n)$ (BIC) • $\lambda_m = 2\log(m)$ (universal-threshold) • … • $\lambda_{k,m} =$?	•Better results •Faster then bootstraping.

$$\overbrace{RSS_k + \sigma^2 k \lambda}^{\text{Minimize}}$$

Multiple Step FDR (MSFDR) Adaptive Penalty

$$\boxed{\lambda_{k,m} = \frac{1}{k}\sum_{i=1}^{k} z^2_{\left[\left(\frac{q}{2}\right)\left(\frac{i}{m+1-i(1-q)}\right)\right]}}$$

$$\overbrace{RSS_k + \sigma^2 k \lambda}$$

Model selection $\Longleftrightarrow$ multiple testing

$$\underbrace{\left(\beta_1 \overset{?}{=} 0, \ldots, \beta_i \overset{?}{=} 0, \ldots, \beta_m \overset{?}{=} 0\right)}_{\left(H_{0,1}, \ldots, H_{0,i}, \ldots, H_{0,m}\right)}$$

Orthogonal X matrix => non changing, coefficients “at once”:

$$\boxed{X'X = nI \rightarrow \hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_m\right)' = n^{-1}X'y}$$

Keeping (Beta) P-values which are bellow α ⇔ forward selection

But how should we adjust for multiplicity of the many tests?

How to adjust for multiplicity?

Approach	Principle	keeping	properties
FWE (familywise error rate)	Keeping the probability of making one or more false discoveries.	α	•Conservative •Low-power
FDR (False discovery rate)	Controlling the expected proportion of incorrectly rejected null out of the rejected	q	•Not "too permissive" •high-power

Coefficient $\left(\sim t_{df}^2\right)$	P value	$\alpha=0.05$ (0.16=AIC)	Bonferroni (FWE)	BH – (FDR at $q \cdot m_0/m$) Adaptive - Step down	
		Over fitting	**Low power**	**More power**	
$\left(\frac{\hat\beta_{(1)}}{SE\left(\hat\beta_{(1)}\right)}\right)^2$ (Largest)	$P_{(1)}$ (Smallest)	α	$\frac{\alpha}{m}$	$q\frac{1}{m}$	
$\left(\frac{\hat\beta_{(2)}}{SE\left(\hat\beta_{(2)}\right)}\right)^2$	$P_{(2)}$	α	$\frac{\alpha}{m}$	$q\frac{2}{m}$	
. . .					
$\left(\frac{\hat\beta_{(i)}}{SE\left(\hat\beta_{(i)}\right)}\right)^2$	$P_{(i)}$	α	$\boxed{\frac{\alpha}{m}}$	$\boxed{q\frac{i}{m}}$	
. . .					
$\left(\frac{\hat\beta_{(m)}}{SE\left(\hat\beta_{(m)}\right)}\right)^2$ (Smallest)	$P_{(m)}$ (Largest)	α	$\frac{\alpha}{m}$	$q\frac{m}{m}=\alpha$	

Coefficient $\left(\sim t^{2}_{df}\right)$	P value	$\alpha=0.05$ (0.16=AIC)	Bonferroni (FWE)	BH – (FDR at $q\cdot{}^{m_0}\!/_{m}$) Adaptive - Step down	
		Over fitting	**Low power**	**More power**	
				$q\frac{1}{m}$	
				$q\frac{2}{m}$	
				$q\frac{i}{m}$	
				$q\frac{m}{m}=\alpha$	

$$\lambda_{k,m}=\frac{1}{k}\sum_{i=1}^{k} z^{2}_{\left(\frac{q}{2}\bullet\frac{i}{m}\right)}$$

<u>Theoretical motivation – results</u>

The minimax properties of the BH procedure were proved (in ABDJ 2006*) *asymptotically* for:

- large m, <*and* >
- orthogonal variables, <*and* >
- for **<u>sparse</u>** signals.

*ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. and JOHNSTONE, I. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*

Coefficient ($\sim t_{df}^2$)	P value	$\alpha=0.05$ (0.16=AIC)	Bonferroni (FWE)	BH – (FDR at $q\cdot {m_0}/{m}$)	Adaptive BH – (FDR at level q)
				Adaptive - Step down	
		Over fitting	**Low power**	**More power**	**More power for <u>richer</u> models**
$\left(\frac{\hat{\beta}_{(1)}}{SE\left(\hat{\beta}_{(1)}\right)}\right)^2$ (Largest)	$P_{(1)}$ (Smallest)			$q\frac{1}{m}$	$q\frac{1}{m+1-1(1-q)}$
$\left(\frac{\hat{\beta}_{(2)}}{SE\left(\hat{\beta}_{(2)}\right)}\right)^2$	$P_{(2)}$			$q\frac{2}{m}$	$q\frac{2}{m+1-2(1-q)}$
. . .					
$\left(\frac{\hat{\beta}_{(i)}}{SE\left(\hat{\beta}_{(i)}\right)}\right)^2$	$P_{(i)}$			$\boxed{q\frac{i}{m}}$	$\boxed{q\frac{i}{m+1-i(1-q)}}$
. . .					
$\left(\frac{\hat{\beta}_{(m)}}{SE\left(\hat{\beta}_{(m)}\right)}\right)^2$ (Smallest)	$P_{(m)}$ (Largest)			$q\frac{m}{m}=\alpha$	$q\frac{m}{m+1-m(1-q)}$

model size VS Penalty factor

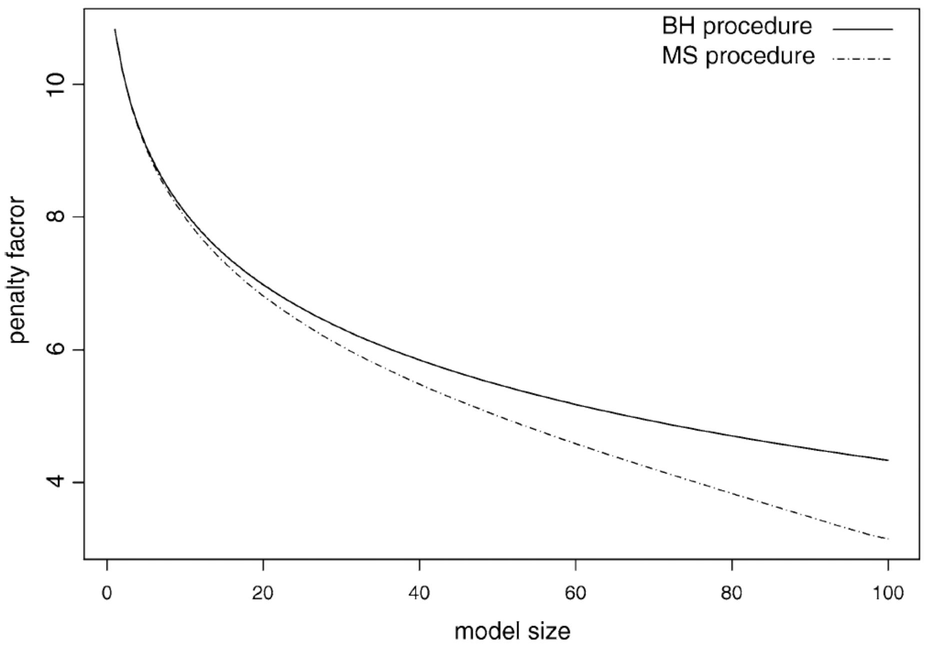

FIG. 1. *The penalty factor of BH and multiple-stage procedures at FDR level* 0.05

BH – (FDR at $q\cdot{}^{m_0}\!/\!_{m}$)	Adaptive BH – (FDR at level q)
Adaptive - Step down	
More power	**More power for <u>richer</u> models**

$$\lambda_{k,m} = \frac{1}{k}\sum_{i=1}^{k} z^2_{\left(\frac{q}{2}\bullet\frac{i}{m}\right)}$$

$$q\frac{i}{m}$$

$$q\frac{i}{m+1-i(1-q)}$$

$$\lambda_{k,m} = \frac{1}{k}\sum_{i=1}^{k} z^2_{\left(\frac{q}{2}\bullet\frac{i}{m+1-i(1-q)}\right)}$$

Forward-selection - Multiple stage FDR:

(a.k.a: MSFDR)

1. Fit Empty model
2. Find the "best" variables (x_{i*}) to enter (with the smallest P value)
3. Is this true ? $$P_i < \alpha = \frac{q}{2} \bullet \frac{i}{m+1-i(1-q)}$$
    1. Yes - Enter X*i* and repeat (step 2)
    2. No – Finish.

R implementation - stepAIC

```
MSFDR <- function(  minimal.lm, maximal.lm , FDR.q = 0.05)  { # assumes intercept

compute.Lambda <-function(k, m, Q = 0.05) { i <- c(1:k)
                        return( (1/(k + 1)) * # +1 because penalty function in stepAIC is different
                                sum(qnorm ((Q/2)  *    (  i/  (m+1-i*(1-Q))   ) )^2 ) ) }

get.model.size <- function(a.lm) { require(MASS);
                   return(extractAIC(a.lm)[1]-1)   } # without intercept

require(MASS);
the.scope <- list(lower = minimal.lm, upper = maximal.lm)
m <- get.model.size(maximal.lm)
new.model.size <- get.model.size(minimal.lm)

for(i in 1:m)
{
  old.model.size  <- new.model.size
  Lambda <- compute.Lambda(k = old.model.size + 1 , m, Q = FDR.q)

  new.model <- stepAIC(minimal.lm, direction="forward", scope=the.scope ,  k = Lambda, trace = F)
  new.model.size <- get.model.size(new.model);

  if(new.model.size <= old.model.size) break;
}
return(new.lm) }
```

```
fit1 <- MSFDR(  minimal.lm = lm.1, maximal.lm = lm.m, FDR.q = 0.05)
summary(fit1)
```

Modeling the diabetes data (Efron et al., 2004)

- n=442 diabetes patients.
- m= 64 (10 baseline variables

  with 45 paired and 9 squared interactions).
- Y - disease progression (a year after baseline)

Modeling the diabetes data (Efron et al., 2004)

Factor	t^2_{df}	P-value	P-to-enter	$\lambda_{k,m}$	$R^2_{(adj)}$
bmi	230.74	0.000000	0.000781	11.29	0.342
ltg	93.86	0.000000	0.001585	10.63	0.457
map	17.36	0.000037	0.002414	10.16	0.477
age.sex	13.56	0.000259	0.003268	9.78	0.491
bmi.map	9.60	0.002076	0.004149	9.47	0.501
hdl	9.00	0.002859	0.005059	9.20	0.510
sex	16.23	0.000066	0.005998	8.96	0.527
glu.2	5.75	0.016920	0.006969	8.75	0.531
age.2	2.58	0.109060	0.007972	8.56	0.533

Modeling the diabetes data (Efron et al., 2004)

Method	Number of variables	R^2
MS_FDR (q=.05), BIC, universal-threshold	7	0.53
AIC	9	0.54
LARS (with Cp)	16	0.55

Over fitting

Simulation - configurations

$$Y_i = X_i \beta + \varepsilon_i$$

- 7 penalty based model selection procedures
- m = 20, 40, 80, 160,   Ratio: n = 2*m
- proportion of non-zero $\beta = \sqrt{\mathrm{m}}, m/4, m/3, m/2, 3m/4, m$
- Dependencies in X: $N(0, \Sigma_{m \times m})$; $\Sigma_{m \times m} = \left[\rho^{|i-j|} \right]$

  $\rho = 0.5, 0, -0.5$

- $\beta = 1$ constant (*with* $R^2 = 0.75$), *2* rates of decrease (in one minimal β is constant)
- Computation – avg MSPE over 1000 runs
- done on 80 computers (distributed computing)

Simulation – Comparison methodology

1) Compute the ratio:

(For each model)

$$\frac{MSPE_{\text{model}}}{MSPE_{\text{random oracle}}}$$

Random Oracle = the "best" model we could find on our search path

2) For each procedure

Over all simulation configuration

find the worst ratio – and compare them

Simulation – results

Comparing the minimax between procedure

TABLE 2

The maximal relative loss (MSPE of method divided by MSPE of the random oracle). Bold figures indicate the minimax relative loss (or to within one simulation standard error). Simulation standard errors are given in parentheses

Procedure	$m = 20$	$m = 40$	$m = 80$	$m = 160$
FWD	2.80 (0.068)	2.87 (0.061)	2.84 (0.043)	3.59 (0.098)
Cp	4.77 (0.096)	4.87 (0.096)	4.88 (0.069)	6.88 (0.193)
DJ	2.34 (0.022)	2.78 (0.021)	3.05 (0.016)	2.58 (0.010)
BM	4.47 (0.064)	5.15 (0.106)	6.61 (0.137)	5.09 (0.110)
FS	3.07 (0.097)	3.62 (0.090)	3.35 (0.059)	3.35 (0.068)
TK	1.66 (0.028)	**1.71** (0.015)	**1.72** (0.010)	1.99 (0.010)
MSFDR 0.05	**1.47** (0.031)	**1.72** (0.012)	1.77 (0.010)	**1.79** (0.008)

- forward selection procedure
- Cp
- the universal threshold in Donoho and Johnstone (1994)
- Birgé and Massart (2001)
- Foster and Stine (2004)
- Tibshirani and Knight (1999)
- multiple-stage procedure in Benjamini, Krieger and Yekutieli (2006) and Gavrilov, Benjamini and Sarkar (2009)—MSFDR

R implementation – biglm + leaps

```
MSFDR.biglm <- function(biglm.obj)
{
  list.of.penalty.lambda <- function(model.size, FDR.q.level = 0.05)
  { compute.Lambda <-function(k, m, FDR.q) # k = size of model, m = maximum size of model
      { from.1.to.k <- c(1:k)
        return((1/k) * sum( qnorm ((FDR.q/2)* from.1.to.k/(m+1-from.1.to.k*(1-FDR.q))  ) ^ 2 ) )  }
    sapply(c(1:model.size),function(x) {compute.Lambda(k = x, m = model.size, FDR.q = FDR.q.level)})}

  penalized.rss <- function(obj) # get a regsubsets object
  { m <- obj$np-1
    lambda.k <- list.of.penalty.lambda(m)
    sigma2 <- obj$sserr/(obj$nn - obj$last)
            # = RSS_full_model / (number_of_obs - model_size_for_biggest_model) = sigma^2
    k <- c(1:m)
    rss <- obj$rss[-1] # RSS, without the first model with the intercept only
    return(penalized.rss.result <- rss +  sigma2 * k * lambda.k)  }

  the.FS.FDR.model.to.use <- function(obj)
  { # gives me the first model that has penaltiy RSS that is local minimum
    penalized.rss.diff <- diff(penalized.rss(aa))
    c(1:length(penalized.rss.diff))[diff(penalized.rss(aa)) > 0][1]  }

  require(leaps)
  m <- length(biglm.obj$names) - 1
  regsubsets.obj <- regsubsets(biglm.obj, method = "forward", nvmax = m+1, intercept=TRUE)

  variables.of.our.model <- the.FS.FDR.model.to.use(regsubsets.obj) # tells us which model to use
  summary(regsubsets.obj)$which[variables.of.our.model,] # vector of T/F indicators
}

MSFDR.biglm(biglm.fit)  # return the names of the variables in the final model
```

Future research

- Beyond Linear regression? (logistic and more)
- Beyond forward selection? (Mixed with Lasso and more)
- More variables then observation? (m>n)

Tel Aviv University

Based on the paper by YOAV BENJAMINI and YULIA GAVRILOV

"A SIMPLE FORWARD SELECTION PROCEDURE
BASED ON FALSE DISCOVERY RATE CONTROL"

(*Annals of Applied Statistics* 2009)

www.R-Statistics.com

Tal.Galili@gmail.com

Thank you!

Questions?

Simulation – Comparison methodology

Challenge (1): Path performance depends on simulation

(while exhaustive search over all subsets – impossible!)

What do we compare to ?

Solution (1): a "**random oracle**"

1)Find the "best" model on the forward path of nested models

Example: for the path: *X7*, *X20*, *X5*, *X9* …

The possible subsets are: {*X7*},{*X7*, *X20*}, {*X7*, *X20*, *X5*} …

2)Compare current models with random oracle

$$\boxed{\frac{MSPE_{\text{model}}}{MSPE_{\text{random oracle}}}}$$

Simulation – Comparison methodology

Challenge (2): MSPE changes per configuration, so how do we compare algorithms?

Solution (2): search for "**empirical minimax performance**" – find the minimum across "maximum relative MSPE over the configurations"

Simulation – conclusions

TABLE 1

The preferable values of q for the FDR procedures studied

		BH	TSFDR	MSFDR
$m = 20$ and 40	$\rho = -0.5$	0.05	0.05	0.05
	$\rho = 0$	0.05	0.05	0.05
	$\rho = 0.5$	0.05	0.05	0.05
	any ρ	0.05	0.05	0.05
$m = 80$	$\rho = -0.5$	0.1	0.1	0.1
	$\rho = 0$	0.1	0.05	0.05
	$\rho = 0.5$	0.1	0.05	0.05
	any ρ	0.1	0.1	0.05
$m = 160$	$\rho = -0.5$	0.25	0.25	0.25/0.1
	$\rho = 0$	0.1	0.1	0.05
	$\rho = 0.5$	0.1	0.1	0.1
	any ρ	0.1	0.1	0.05

Simulation – results (extended)

TABLE 3
Mean relative MSPE values for the k least-favorable configurations, $k = 2, 3$, ALL. The case for $k = 1$ is in Table 2

	$m = 20$			$m = 40$			$m = 80$			$m = 160$		
k	**2**	**3**	**ALL**	**2**	**3**	**ALL**	**2**	**3**	**ALL**	**2**	**3**	**ALL**
Procedure												
FWD	2.80	2.62	1.48	2.69	2.62	1.46	2.77	2.52	1.37	3.55	3.54	1.54
Cp	4.71	4.28	1.82	4.58	4.42	1.80	4.75	4.32	1.69	6.73	6.58	2.07
DJ	2.30	2.25	1.36	2.60	2.53	1.44	2.91	2.82	1.47	2.51	2.47	1.46
BM	4.39	4.20	2.03	4.77	4.63	2.41	5.55	5.06	2.35	4.81	4.44	2.30
FS	3.07	3.06	1.77	3.49	3.42	1.57	3.26	2.97	1.33	2.79	2.57	1.24
TK	1.64	1.63	1.30	1.66	1.60	1.23	1.67	1.61	1.19	1.86	1.81	1.23
MSFDR 0.05	1.46	1.45	1.27	1.63	1.60	1.25	1.72	1.67	1.21	1.79	1.78	1.21

Earlier studies limitations:

1)Constant coefficients (mostly)

2)Largest m = 50

3)NOT Compared to other non-constant adaptive penalties

The selection procedures and configurations studied by various authors.

Paper	Studied estimators	Correlation Structure $Corr(X_i, X_j)$	Number of observations	Number of potential explanatory variables	Coefficients
Yuan, Ekici,Lu and Monteiro (2007)	FES, OLS, CW, RRR, PLS, PCR, RR, CAIC	$\rho^{\|i-j\|}$, $\rho = 0.5$	20 and 50	8, 20	(3,1.5,0,0,2,0,0,0), (5,0,0,0,0,0,0,0), 10 ones and 10 zeros
Bunea, Niu and Wegkamp (2003)	AIC, BIC, FDR	Independent covariates	200 and 500	5	(10,0.1,0,0.25,0,0.05)
Tibshirani and Knight (1999)	AIC, CIC(TK) and conditional bootstrap	$\rho^{\|i-j\|}$ $\rho = 0.7$	50 and 150	21	2, 6 and 10 non-zero coefficients such that theoretical R^2=0.75
Brieman (1992)	Cp and bootstrap	$\rho^{\|i-j\|}$ $\rho = 0.7$	60, 160 and 600	40	3, 9, 15 and 21 non-zero constant coefficients such that theoretical R^2=0.75
Ye (2002)	AIC, BIC, RIC and Ye's adaptive procedure	$\rho^{\|i-j\|}$ ρ = -0.5, ρ =0, $\rho = 0.5$	200	50	0,5,10,50 non-zero constant coefficients such that theoretical R^2=0.75
Zou and Hastie (2005)	Lasso,Ridge regression, Elastic net	$\rho^{\|i-j\|}$ $\rho = 0.5$	20 and 200	8	(3,1.5,0,0,2,0,0,0), constant coefficients equal to 0.85
		0.5	100 and 400	40	20 non-zero constant coefficients
Fan and Li (2001)	Lasso, Ridge regression,Best subset SCAD	$\rho^{\|i-j\|}$ $\rho = 0.5$	40 and 60	3	(3,1.5,0,0,2,0,0,0)
Tibshiani (1996)	Least squares, Lasso, Ridge regression, Best subset	$\rho^{\|i-j\|}$ $\rho = 0.5$	20	8	(3,1.5,0,0,2,0,0,0) constant coefficients equal to 0.85
		0.5	100	40	20 non-zero constant coefficients
George and Foster (1997)	AIC, BIC, RIC, GF	$\rho^{\|i-j\|}$ ρ = -0.5,0 and 0.5	200	50	0,5,10,50 non-zero constant coefficients such that theoretical

For orthogonal X matrix: $$X'X = nI \rightarrow \hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_m\right)' = n^{-1}X'y$$

The difference of each step is of the standardized coefficient, since:

$$\frac{RSS_{k-1} - RSS_k}{\sigma^2} = \frac{RSS_{k-1} - \left(RSS_{k-1} - \hat{\beta}_k^2 \sum_{i=1}^{n} x_{ki}^2\right)}{\sigma^2} = \frac{n\hat{\beta}_k^2}{\sigma^2} = \left(\frac{\hat{\beta}_k}{SE\left(\hat{\beta}_k\right)}\right)^2$$

Forward selection is **<u>like sorting the P-values</u>** and then keeping only who ever is smaller then α:

$$RSS_k + \sigma^2 \underbrace{\sum_{i=1}^{k} z_{\alpha/2}^2}_{z_{\alpha/2}^2 + \sum_{i=1}^{k-1} z_{\alpha/2}^2} \le RSS_{k-1} + \sigma^2 \sum_{i=1}^{k-1} z_{\alpha/2}^2$$

So on which P should we stop ? $\alpha = 0.05$?

An **adaptive** penalty procedure

$$RSS_k + \sigma^2 \underbrace{\sum_{i=1}^{k} z^2_{\alpha/2}}_{z^2_{\alpha/2} + \sum_{i=1}^{k-1} z^2_{\alpha/2}} \le RSS_{k-1} + \sigma^2 \sum_{i=1}^{k-1} z^2_{\alpha/2}$$

With this λ

$$RSS_k + \sigma^2 \sum_{i=1}^{k} z^2_{\alpha/2} = RSS_k + \sigma^2 k \underbrace{\left(\frac{1}{k} \sum_{i=1}^{k} z^2_{\alpha/2} \right)}_{\lambda}$$

BH – (FDR at $q \cdot m_0/m$)	Adaptive BH – (FDR at level q)
Adaptive - Step down	
More power	**More power for <u>richer</u> models**
$q\dfrac{i}{m}$	$q\dfrac{i}{m+1-i(1-q)}$

What this is NOT:

1. Fit full model – then check the P values
2. Fit m “small” models – then check the P values

Correction for multiple testing

	# declared non-significant	# declared significant	Total
# true null hypotheses	U	V	m_0
# non-true null hypotheses	T	S	$m - m_0$
Total	$m - R$	R	m

$$FWE : P(V \geq 1) < \alpha$$

Very conservative,
Low power

$$FDR : E\left[\frac{V}{R}\right] < q$$

Different objective,
More power