

Discussion: Hinshaw, Szapudi

Christopher R. Genovese

Department of Statistics

Carnegie Mellon University

<http://www.stat.cmu.edu/~genovese/>

12 June 2006

Part I: Evidence, Occam's Razor, and Statistical Models

Evidence and Model Selection

- Questions about unknown parameters (e.g., Is $n_s < 1$?) can often be framed as a comparison between models. This leads to techniques for *model selection*.
- The literature on model selection is vast and rich but one common theme, even today, focuses on the relative merits of AIC versus BIC.

AIC: Akaike Information Criterion (Akaike 1973)

BIC: Bayesian Information Criterion (Schwarz 1978)

Evidence and Model Selection (cont'd)

- Suppose

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, \dots, n,$$

where x is p -dimensional, ϵ are random errors, and f is unknown.

Suppose we have parametric models $\mathcal{M}_1, \dots, \mathcal{M}_m$ for f :

$$\mathcal{M}_k = \{f_k(x, \theta_k) : \theta_k \in \Theta_k\} \quad k = 1, \dots, m$$

where Θ_k has dimension d_k .

Let $\hat{\ell}_k$ be the maximized log likelihood over \mathcal{M}_k .

Comparing maximized likelihoods $\hat{\ell}_k$ by themselves can be misleading.

Evidence and Model Selection (cont'd)

- Instead we add a complexity penalty to the measure of misfit. Select a model by minimizing the resulting criterion.

Key examples:

- AIC (Akaike Information Criterion):

$$\text{AIC} = -\hat{\ell}_k + d_k$$

- BIC (Bayesian Information Criterion):

$$\text{BIC} = -\hat{\ell}_k + d_k \log \sqrt{n}$$

- Many alternatives have been proposed.
- But how do we decide which penalty to use?

Two Related Goals

- Goal 1: Model Identification

Suppose the “true” model, \mathcal{M}_{k^*} , is among $\mathcal{M}_1, \dots, \mathcal{M}_m$.

We would like a *consistent* model selection scheme:

$$\hat{k} \rightarrow k^* \text{ as } n \rightarrow \infty.$$

- Goal 2: Optimal Estimation/Prediction

Will select a model and compute an estimate \hat{f} of the unknown f .

We want the best worst-case (minimax) performance over the assumed set of functions.

- These two goals are in conflict.
- Loosely speaking, BIC gives consistent model selection (Goal 1) and AIC gives optimal minimax estimators (Goal 2).
- A host of alternative criteria have been proposed to balance the competing strengths of AIC and BIC.

The Bayesian in BIC

Besides its consistency properties, BIC has two other (though related) substantive justifications.

1. BIC can be derived from information theoretic arguments in terms of the Minimum Description Length principle.

Two-stage MDL: $-\hat{\ell}_k$ to describe the data given the MLE, and $\frac{k}{2} \log n$ to describe the MLE.

Other of forms of description length give different criteria.

2. BIC offers an approximation to the log Bayes Factors and thus to approximate model posterior probabilities.

Bayes Factors and Model Probabilities

Suppose we have two models \mathcal{M}_0 and \mathcal{M}_1 . The Bayes Factor B_{10} of model 0 relative to model 1 is

$$B_{10} = \frac{p(Y | \mathcal{M}_1)}{p(Y | \mathcal{M}_0)} = \frac{\frac{p(\mathcal{M}_1|Y)}{p(\mathcal{M}_0|Y)}}{\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}},$$

where $p(Y | \mathcal{M})$ is the marginal distribution of the data under model \mathcal{M} and so forth. In words,

$$\text{Bayes Factor} = \frac{\text{posterior odds}}{\text{prior odds}}.$$

If $p(\mathcal{M}_0) = p(\mathcal{M}_1) = 0.5$, then $p(\mathcal{M}_1 | Y) = \frac{B_{10}}{1+B_{10}}$.

Posterior model probabilities can likewise be computed in more general situations.

Bayes Factors and Model Probabilities (cont'd)

- The Bayes factor has the form of a likelihood ratio and may or may not depend on the priors used. In general,

$$p(Y | \mathcal{M}_k) = \int_{\Theta_k} d\theta p(Y | \theta, \mathcal{M}_k) p(\theta | \mathcal{M}_k).$$

- Jeffrey's (1961) and Kass and Raftery (1994) suggest benchmarks for B_{10} that are commonly used to weigh evidence against \mathcal{M}_0 :
 - Barely worth mentioning ($0 \leq \log B_{10} \leq 1$)
 - Positive ($1 \leq \log B_{10} \leq 3$)
 - Strong ($3 \leq \log B_{10} \leq 5$)
 - Decisive ($\log B_{10} > 5$)

But there are no uniformly correct criteria.

Bayes Factors and BIC (cont'd)

- BIC enables an approximation of the log Bayes Factors without the need to introduce the prior distributions on the model. Let

$$S = \text{BIC}_0 - \text{BIC}_1 = \hat{\ell}_1 - \hat{\ell}_0 + (d_0 - d_1) \log \sqrt{n}.$$

Then, as $n \rightarrow \infty$,

$$\frac{S - \log B_{10}}{\log B_{10}} \rightarrow 0.$$

The relative error of e^S is only $O(1)$, but under certain conditions can be $O(n^{-1/2})$.

- Thus when the priors on the models are equal, we have approximately that

$$p(\mathcal{M}_k | Y) \approx \frac{e^{-\text{BIC}_k}}{\sum_j e^{-\text{BIC}_j}}.$$

Choosing the model with the smallest BIC is approximately the same as maximizing the posterior probability of the model.

Alternative Approaches: Confidence Sets

Generate a uniformly valid confidence set for f : a data-determined set \mathcal{C} such that $P\{\mathcal{C} \ni f\}$ is at least the target confidence level *for all functions in the assumed class*.

Then answer question of interest by examining properties of the functions in \mathcal{C} .

For example, look at the range of n_s among functions in \mathcal{C} .

Genovese et al. (2004) constructs such confidence sets for the CMB temperature power spectrum.

Bryan et al. (2005) develops algorithms to searching to confidence sets to, for instance, compute a confidence interval for n_s .

Alternative Approaches: Model Averaging

Account for the uncertainty about which model is best by making the model itself a part of the parameter specification.

Introduce a parameter M to index the model, and let the parameter space Θ to be

$$\Theta = \bigsqcup_{k=1}^m \Theta_k \times \{k\}.$$

For example, can include a model where n_s is fixed at 1 and a model where $n_s < 1$ is allowed to vary.

Then do a full Bayesian analysis on this parameter space; posterior probabilities are computed in the regular way.

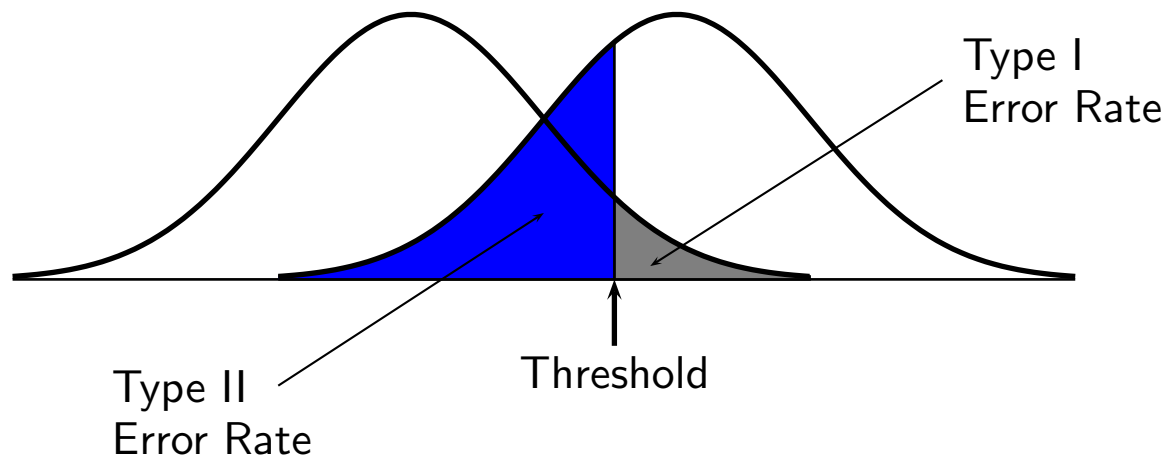
For instance, the model probabilities become $P\{M = k \mid Y\}$.

Reversible-jump MCMC (Green 1995) is very useful and practical in such problems.

Part II: False Discovery Control for Multiple Testing

One Test, One Threshold

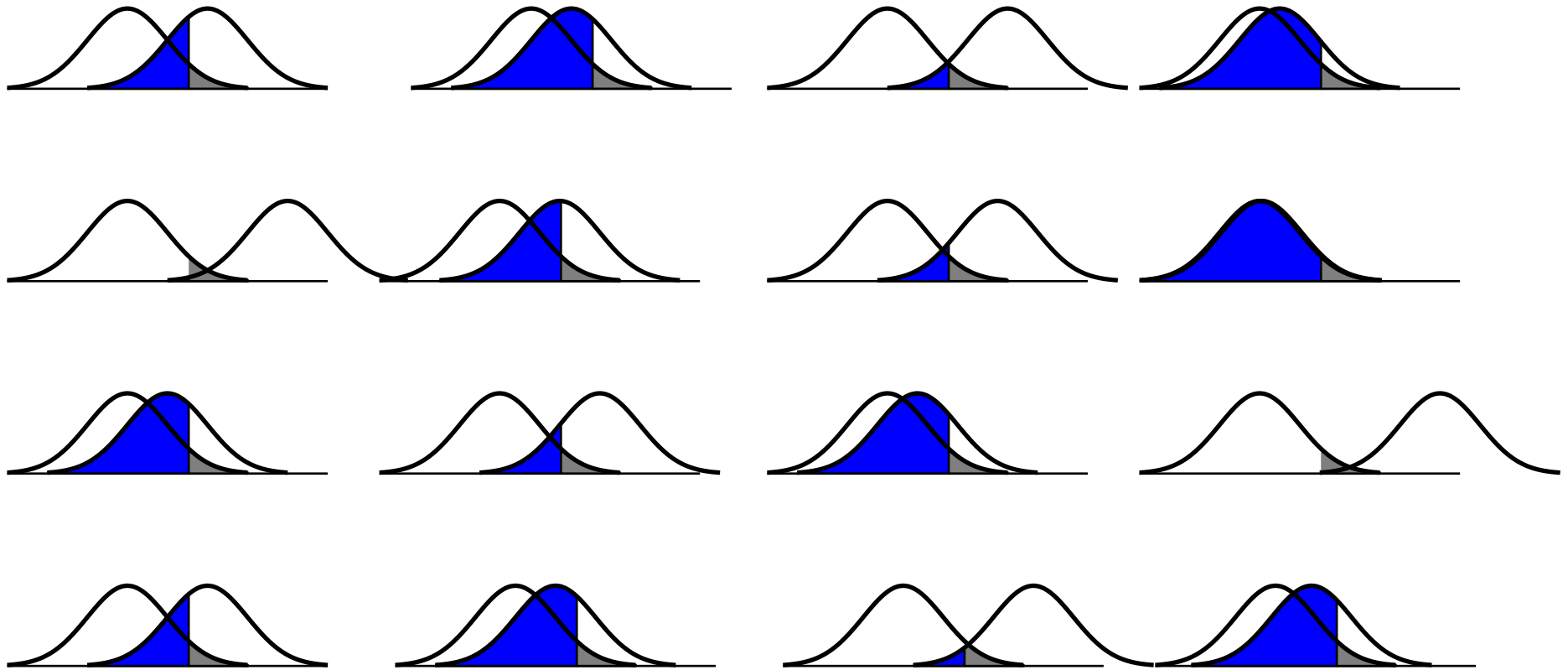
With a single hypothesis test, we choose a rejection threshold to control the Type I error rate,



while achieving a desirable Type II error rate for relevant alternatives.

Many Tests, One Threshold

With multiple tests, the problem is more complicated



Each test has possible Type I and Type II errors, and there are many possible ways to combine them. The probability of a Type I error grows with the number of tests.

The Multiple Testing Problem

- Perform m simultaneous hypothesis tests with a common procedure.
- For any given procedure, classify the results as follows:

	H_0 Retained	H_0 Rejected	Total
H_0 True	TN	FD	T_0
H_0 False	FN	TD	T_1
Total	N	D	m

Mnemonics: T/F = True/False, D/N = Discovery/Nondiscovery

All quantities except m , D , and N are *unobserved*.

- The problem is to choose a procedure that balances the competing demands of sensitivity and specificity.

How to Choose a Threshold?

- Control Per-Comparison Type I Error (PCER)
 - a.k.a. “uncorrected testing,” many type I errors
 - Gives $P\{FD_i > 0\} \leq \alpha$ marginally for all $1 \leq i \leq m$
- Control Familywise Type I Error (FWER)
 - e.g.: Bonferroni: use per-comparison significance level α/m
 - Guarantees $P\{FD > 0\} \leq \alpha$
- Control False Discovery Rate (FDR)
 - first defined by Benjamini & Hochberg (BH, 1995, 2000)
 - Guarantees $FDR \equiv E\left(\frac{FD}{D}\right) \leq \alpha$
- ...

A Practical Problem

- While FWER-control is appealing, it tends to produce low power.
- FDR control offers a way to increase power while maintaining some principled bound on error.

It is based on the assessment that

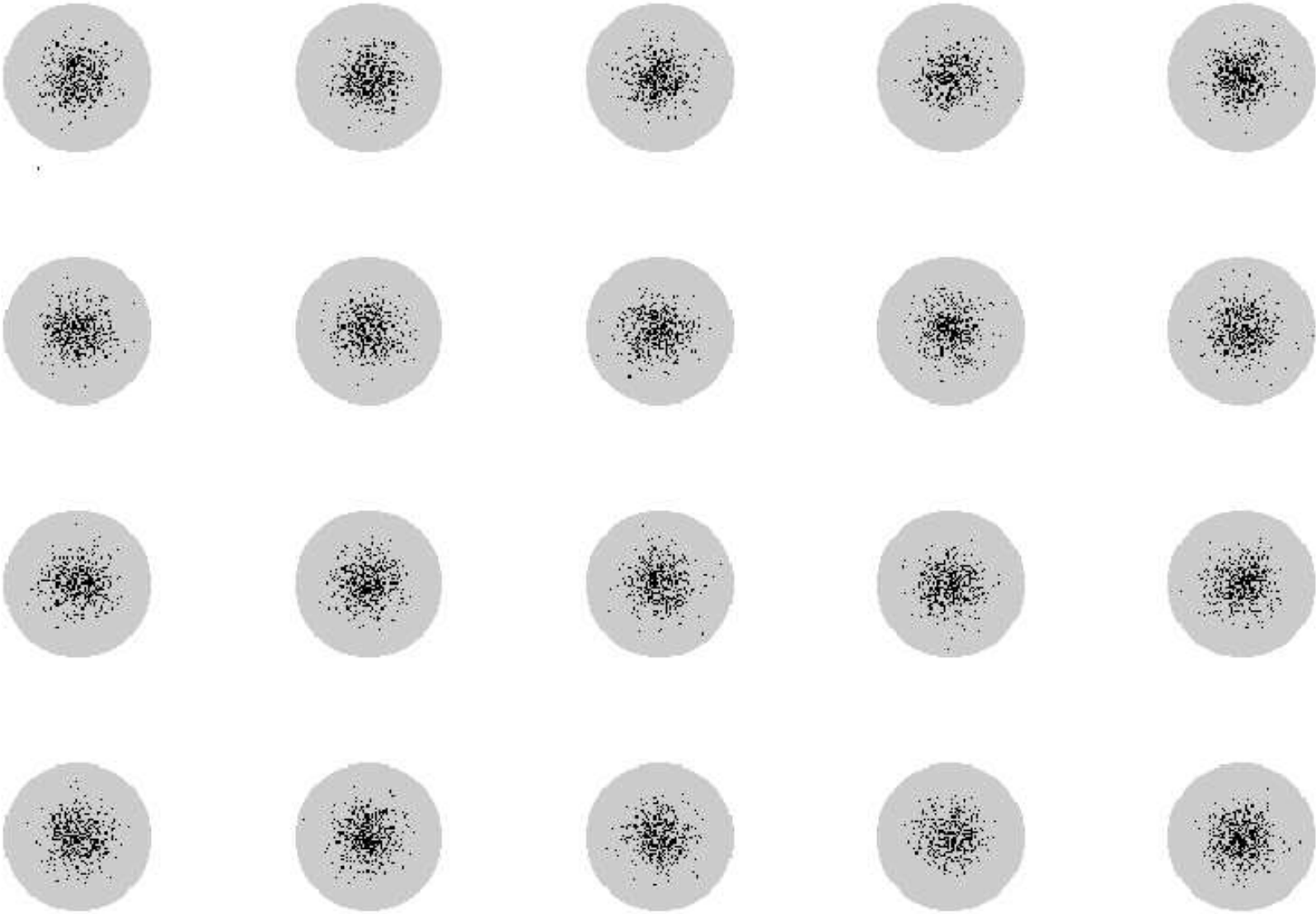
4 false discoveries out of 10 rejected null hypotheses

is a more serious error than

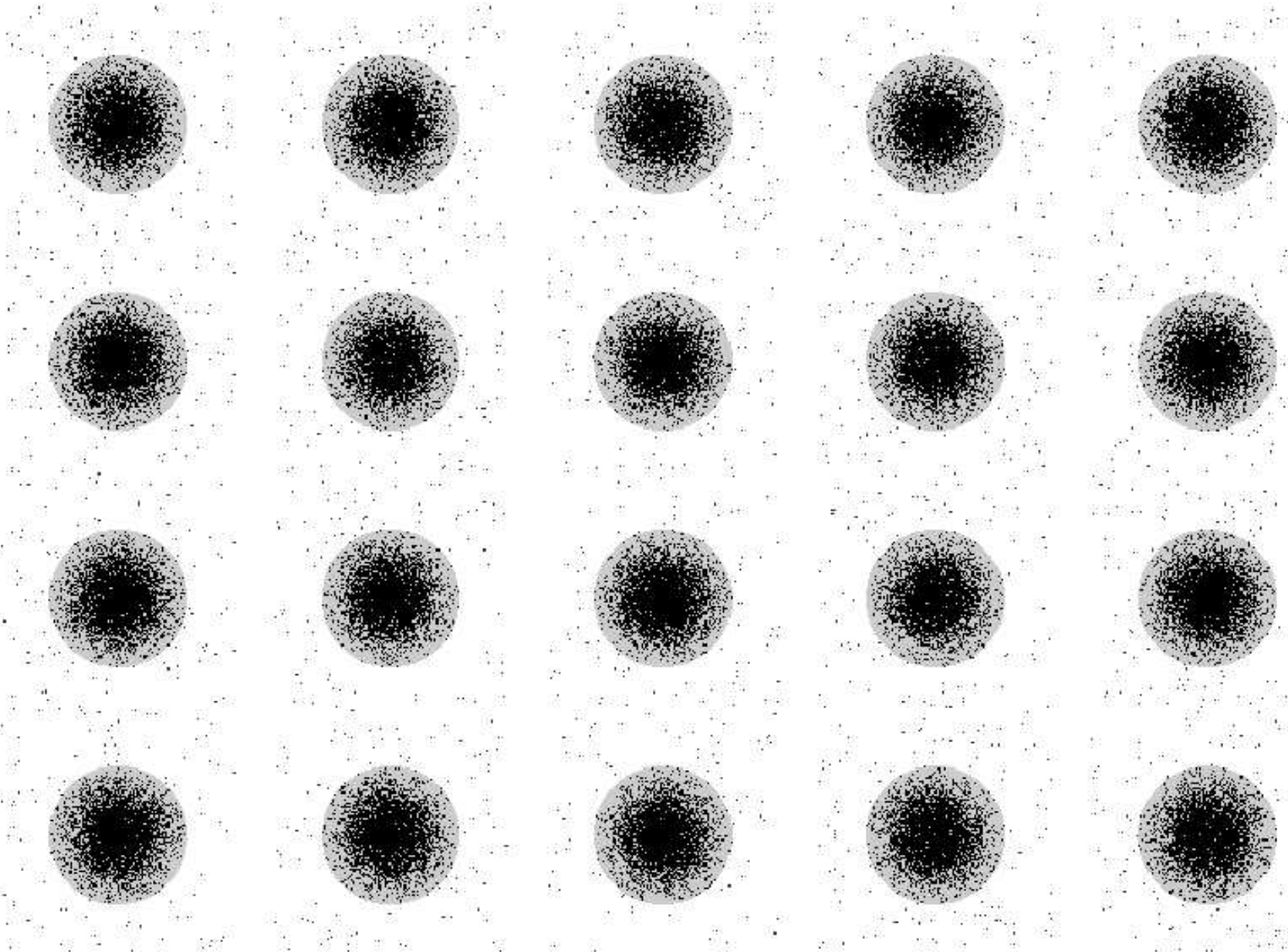
20 false discoveries out of 100 rejected null hypotheses.

- A simple illustration ...

FWER Control



FDR Control



The False Discovery Rate

For any threshold T , the False Discovery Proportion is given by

$$\text{FDP}(T) = \frac{\text{\#False Discoveries}}{\text{\#Discoveries}}$$

and the False Discovery Rate is given by

$$\text{FDR}(T) = \langle \text{FDP}(T) \rangle .$$

The Benjamini-Hochberg (BH) procedure is fast and broadly applicable method for selecting a multiple testing threshold T_{BH} .

For any selected $0 < \alpha < 1$,

$$\text{FDR}(T_{\text{BH}}) \leq \alpha \frac{T_0}{m} .$$

Issue 1: Improving Power

BH is derived with a conservative bound, so it works best in the sparse case where $T_0 \approx m$.

To improve power, we need a better estimate of T_0 . Many such adaptive methods have been developed.

Benjamini, Krieger, and Yekutieli (BKY, 2004) offer a simple and effective improvement to BH:

- Use BH at level β_1 . Let r_1 be the number of rejected null hypotheses.
- If $r_1 = 0$, stop.
- Otherwise, let $\hat{T}_0 = m - r_1$.
- Use BH at level $\alpha' = \beta_2 m / \hat{T}_0$.

Default case: $\beta_1 = \beta_2 = \alpha / (1 + \alpha)$.

Issue 2: Dependence

- Benjamini and Yekutieli (2001) show that the original BH method still controls FDR at the nominal level even for some kinds of dependent tests.
- Under general dependence, BH controls FDR at level

$$\alpha \frac{T_0}{m} \sum_{i=1}^m \frac{1}{i}.$$

Thus, a distribution-free procedure for FDR control is to apply BH at level $\alpha / \sum_{i=1}^m \frac{1}{i}$. Unfortunately, this is typically very conservative, sometimes even more so than Bonferroni.

- Practically speaking, BH is quite hard to break even beyond what has been proven.
- Better control under dependence can be obtained with good estimates of the dependence structure.

Other Variations

- Exceedance Control (Genovese and Wasserman 2002):
Control $P\{\text{FDP} > \gamma\}$ rather than $\langle \text{FDP} \rangle$
- Higher Criticism (Jin and Donoho 2004):
Optimal detection of sparse mixtures
- Estimating the Proportion of False Nulls (Meinshausen and Rice 2004; Meinshausen and Bühlmann 2005; Cai, Jin, and Low 2006)
- False Discovery Control for Regions (Perone Pacifico et al. 2004, 2006)
- ...