FAB Inference

Peter Hoff
Duke University
2021/03/24

The Big Picture

Goal:

- Infer group-specific parameters $\theta_1, \ldots, \theta_p$
- from group-specific samples $(y_{1,1}, \ldots, y_{n_1,1}), \ldots, (y_{1,p}, \ldots, y_{n_p,p})$

Direct methods: $\hat{\theta}_j = \bar{y}_j$, $\bar{y}_j \pm 2s_j/\sqrt{n_j}$

- unbiased for each group, group-specific error control
- no data sharing, high variance, poor across-group performance

Indirect methods: $\hat{\theta}_j = \frac{\hat{\tau}^2}{\hat{\tau}^2+s_j^2/n_j}\bar{y}_j + \frac{s_j^2/n_j}{\hat{\tau}^2+s_j^2/n_j}\bar{y}$, $\hat{\theta}_j \pm 2/\sqrt{n_j/s_j^2 + 1/\hat{\tau}^2}$

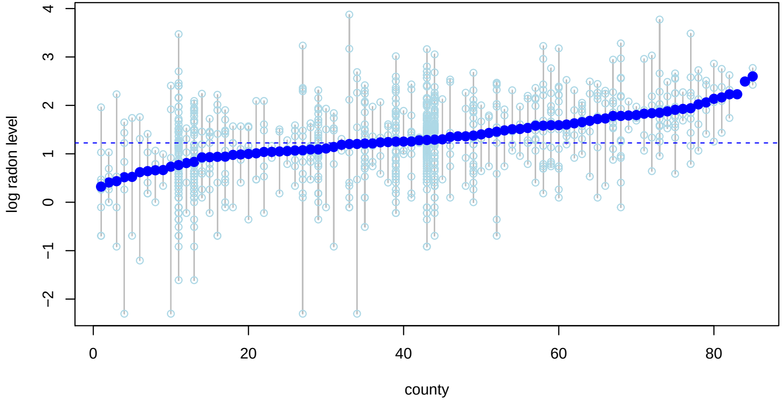
- biased for each group, no group-specific error control
- data sharing, low variance, good across-group performance

FAB methods:
data sharing, good across-group performance, group-specific error control

This talk: FAB CIs, p-values

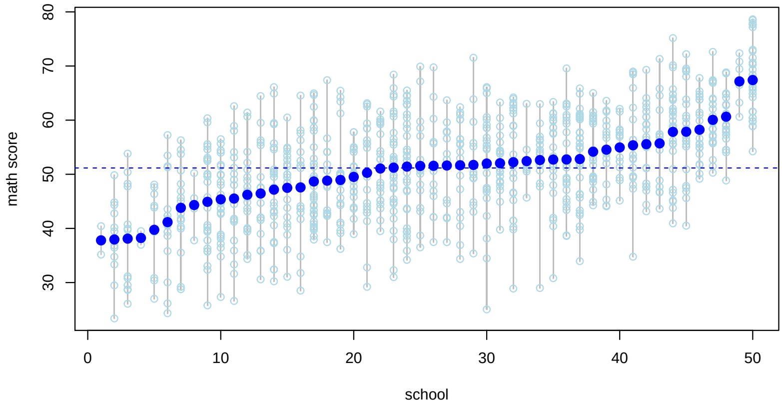
Radon data

Log radon levels of 85 Minnesota counties.



ELS data

Subset of 50 schools from 684 total.



Multilevel data

Data: $y_{i,j}$ = outcome for unit i in group j:

$$\begin{aligned} \mathbf{y}_1 &= (y_{1,1}, \ldots, y_{n_1,1}) \\ &\vdots \qquad\quad \vdots \\ \mathbf{y}_p &= (y_{1,p}, \ldots, y_{n_p,p}) \end{aligned}$$

Estimand: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, vector of group-specific means.

Group-specific inference: Obtain $\hat{\theta}_j$ and $C_j(\mathbf{y}) = [l_j(\mathbf{y}), u_j(\mathbf{y})]$ for each j so

$$\mathrm{E}[(\hat{\theta}_j - \theta_j)^2 | \boldsymbol{\theta}] \text{ is small}$$

$$\Pr\left(\theta_j \in C_j(\mathbf{y})\,|\boldsymbol{\theta}\right) = 1 - \alpha.$$

Point estimation

How should we estimate θ_j?

Unbiased estimation:

$$\hat{\theta}_j \stackrel{?}{=} \bar{y}_j = \sum_i y_{i,j}/n_j$$

- Guaranteed to be unbiased, $\mathrm{E}[\bar{y}_j] = \theta_j$.
- Minimum variance among linear unbiased estimators.
- Error depends on sample size:

$$\mathrm{E}[(\bar{y}_j - \theta_j)^2|\boldsymbol{\theta}] = \sigma^2/n_j.$$

Bayes/Shrinkage estimation:

$$\hat{\theta}_j \stackrel{?}{=} w\bar{y}_j + (1-w)\bar{y}_{-j}$$

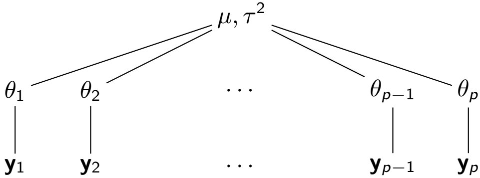
- Bias is bigger than that of $\bar{y}_j$.
- Variance is lower than that of $\bar{y}_j$.

How to pick w?

Indirect information

Obtain w from **indirect information**:

- Share information across groups with a *linking model*;
- linking model + sampling model = *hierarchical model*.



$$y_{1,j}, \ldots, y_{n_j,j} | \boldsymbol{\theta} \sim \text{i.i.d. } N(\theta_j, \sigma^2)$$
$$\theta_1, \ldots, \theta_p \sim \text{i.i.d. } N(\mu, \tau^2)$$

Optimal shrinkage

If μ, τ^2, σ^2 were known, the optimal estimator *in terms of total MSE* is

$$\check{\theta}_j = \frac{n_j/\sigma^2}{n_j/\sigma^2 + 1/\tau^2}\bar{y}_j + \frac{1/\tau^2}{n_j/\sigma^2 + 1/\tau^2}\mu.$$

(Bayes estimator, BLUP, oracle)

Since μ, τ^2, σ^2 are generally unknown, the following estimator is typically used:

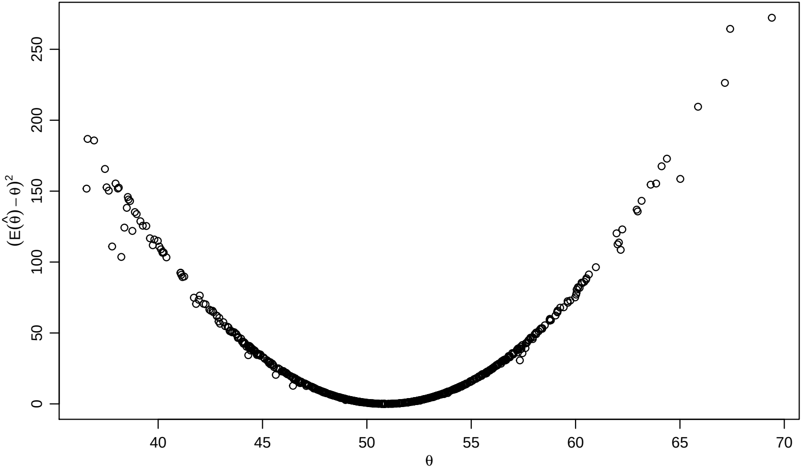
$$\hat{\theta}_j = \frac{n_j/\hat{\sigma}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\bar{y}_j + \frac{1/\hat{\tau}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\hat{\mu}.$$

(EBayes, EBLUP, JS estimator)

Performance:

- across-group : $E[||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2|\boldsymbol{\theta}] \lesssim E[||\bar{\mathbf{y}} - \boldsymbol{\theta}||^2|\boldsymbol{\theta}]$
- within-group : $\hat{\theta}_j$ better than $\bar{y}_j$ if θ_j near μ, worse if far.

Bias: ELS data



Intervals

Confidence interval:

$$C(\mathbf{y}) \stackrel{?}{=} \bar{y}_j \pm \frac{\hat{\sigma}}{\sqrt{n_j}} t_{1-\alpha/2}$$

- Exact constant coverage:

$$\Pr(\theta_j \in C(\mathbf{y})|\boldsymbol{\theta}) = 1 - \alpha \text{ for all values of } \theta_j.$$

- Narrowest interval among "unbiased" intervals.
- Doesn't use indirect information.

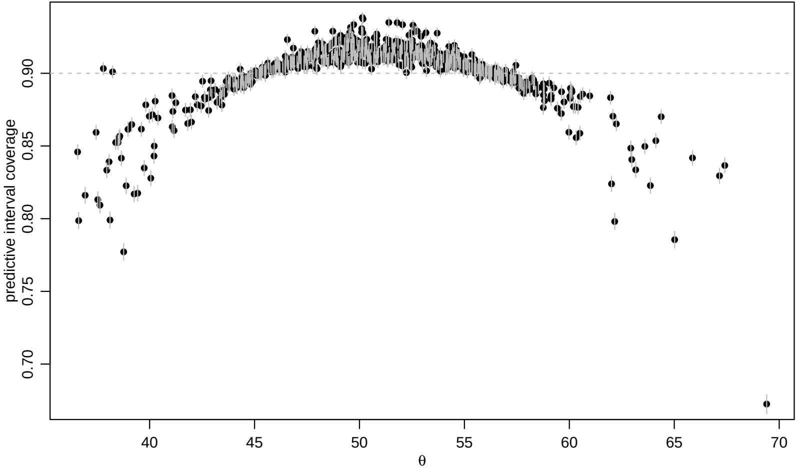
Can indirect information be incorporated, while maintaining constant coverage?

"Prediction" interval:

$$C(\hat{\theta}_j) = \hat{\theta}_j \pm t_{1-\alpha/2}/\sqrt{1/\hat{\tau}^2 + n_j/\hat{\sigma}^2}$$

- $1 - \alpha$ coverage *on average across groups*.
- Lower for some groups, higher for others, and you don't know which.

Nonconstant coverage: ELS data



$$\Pr(\theta_j \in C(\hat{\theta}_j)) \approx 1 - \alpha$$

$$\Pr(\theta_j \in C(\hat{\theta}_j)|\boldsymbol{\theta}) \quad \text{depends on } j,\ \boldsymbol{\theta}.$$

Valid confidence intervals that share information

Goal: Construct confidence intervals $C^1, \ldots, C^p$ having

- constant coverage: $\Pr(\theta_j \in C^j(\mathbf{y})|\boldsymbol{\theta}) = 1 - \alpha$ for all groups/$\boldsymbol{\theta}$'s.
- improved precision: $\mathrm{E}[|C^j(\boldsymbol{y})|] < 2t_{1-\alpha/2}$ on average across groups/$\boldsymbol{\theta}$'s.

The first criterion is group-specific/frequentist - conditional on θ_j.

The second is study-specific/Bayes - on average across $\theta_1, \ldots, \theta_p$.

All CIPs

Standard procedure:

$$C_{1/2}(y) = \{\theta : y + \sigma z_{\alpha/2} < \theta < y + \sigma z_{1-\alpha/2}\}$$

Any procedure:

$$C_w(y) = \{\theta : y + \sigma z_{\alpha(1-w)} < \theta < y + \sigma z_{1-\alpha w}\}$$

In fact, w may depend on θ: If $w : \mathbb{R} \to [0,1]$ then

$$C_w(y) = \{\theta : y + \sigma z_{\alpha(1-w(\theta))} < \theta < y + \sigma z_{1-\alpha w(\theta)}\}$$

satisfies $\Pr(\theta \in C_w(y)|\boldsymbol{\theta}) = 1 - \alpha$

- Examples in Bartholomew [1971], Stein [1962].
- Essentially complete class result in Yu and Hoff [2018].

FAB: Bayes-optimal frequentist interval

Simplified model:

- $y|\theta \sim N(\theta, \sigma^2)$, σ^2 known.
- $\pi(\theta)$ is prior information about θ.

Idea: Find the w-function that minimizes the prior expected width

$$\int \int |C_w(y)| \, p(dy|\theta)\pi(d\theta) < \int \int |C(y)| \, p(dy|\theta)\pi(d\theta)$$

Such an interval will have

- constant coverage, because C_w has constant coverage for any w-function;
- optimal precision on average with respect to π, by construction.

We call it FAB - Frequentist And Bayesian.

Optimal w-function

If $\pi(\theta)$ is the $N(\mu, \tau^2)$ density, then

$$E[|C_w|] = \int \int |C_w(y)| \, p(dy|\theta) \, \pi(d\theta)$$

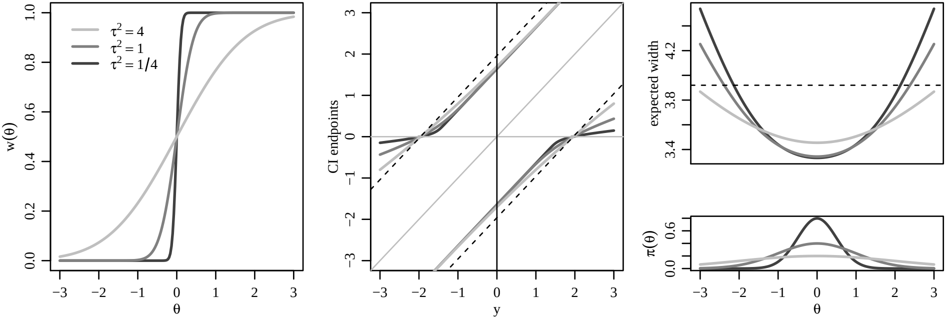
is minimized by

$$w(\theta) = g^{-1}(2\sigma(\theta - \mu)/\tau^2)$$
$$g(w) = \Phi^{-1}(\alpha w) - \Phi^{-1}(\alpha(1 - w))$$

This w-function yields Pratt's (1963) z-interval.

- Yu and Hoff (2018): Extension to t-intervals, multigroup inference.
- Hoff and Yu (2019): Linear regression coefficients.

Bayes-optimal procedure



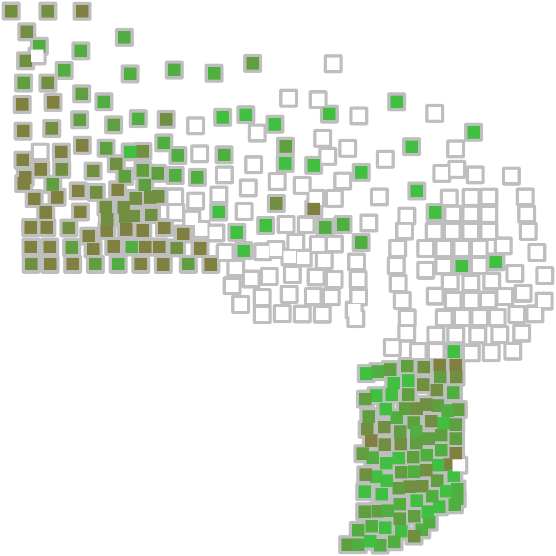
Adaptive FAB for multigroup inference

For each group $j = 1, \ldots, p$:

1. Obtain $\hat{\mu}$, $\hat{\tau}^2$, $\hat{\sigma}^2$ using data from groups other than j;
2. Obtain $\hat{w}_j(\theta) = g^{-1}(2\hat{\sigma}(\theta - \hat{\mu})/\hat{\tau}^2)$;
3. Construct $C_{\hat{w}_j}(\bar{y}_j)$.

- Exact $1 - \alpha$ coverage *for each group*, even if hierarchical model is wrong.
- Improved precision *on average across groups*.

Radon data



Small area estimation (Burris and Hoff 2019)

Sampling model: $\bar{y}_j \sim N(\theta_j, \sigma_j^2)$ independently across groups.

Linking Model: $\theta_j = \boldsymbol{\beta}^\top \mathbf{x}_j + e_j$, $\text{Cov}[\boldsymbol{\theta}] = \Sigma$ (spatial FH model).

Direct interval: $\bar{y}_j \pm \hat{\sigma}_j t_{1-\alpha/2}$

AFAB interval: For each area $j = 1, \ldots, p$

1. using areas other than j, obtain estimates of $\boldsymbol{\theta}_{-j}$, $\boldsymbol{\beta}$ and Σ;
2. obtain "prior" distribution for θ_j from estimates and working model;
3. compute optimal w-function and construct FAB interval for θ_j.

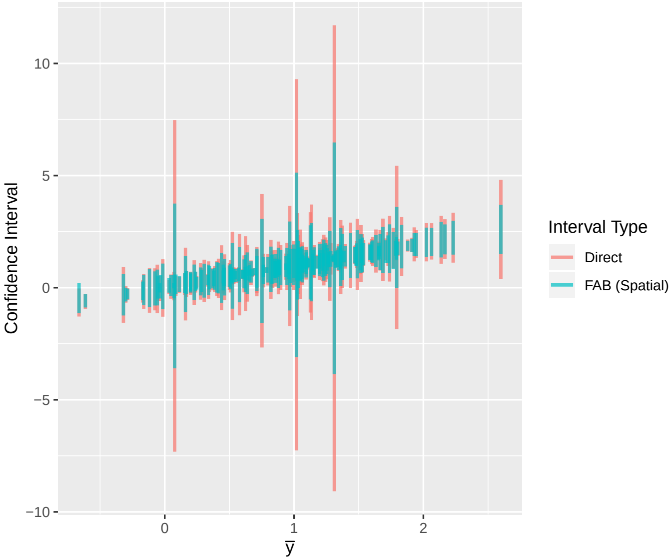
- Both intervals have $1 - \alpha$ area-specific coverage, under random sampling within each area. The linking model need not be correct.
- FAB intervals make use of information from neighboring areas and known area-level characteristics (surficial radium).

Interval comparisons

Type	Hierarchical model	relative width	fraction intervals improved
Direct	-	1.0	-
FAB	exchangeable	.77	.898
FAB	covariate	.77	.888
FAB	spatial	.74	.964
FAB	spatial, covariate	.74	.955

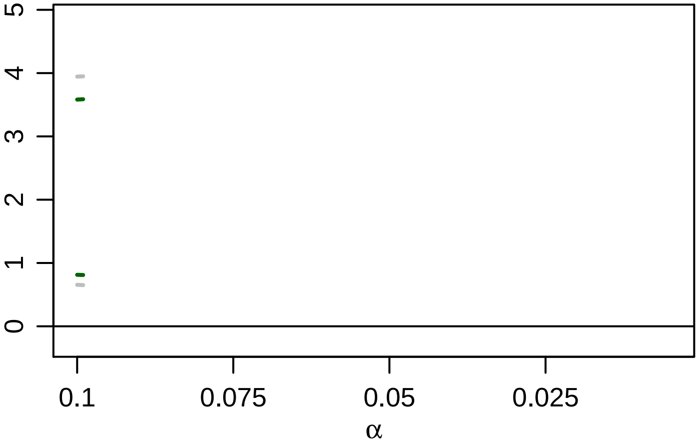
By sharing information, hierarchical models can improve across-group performance, even if the hierarchical model is wrong.

Interval comparisons



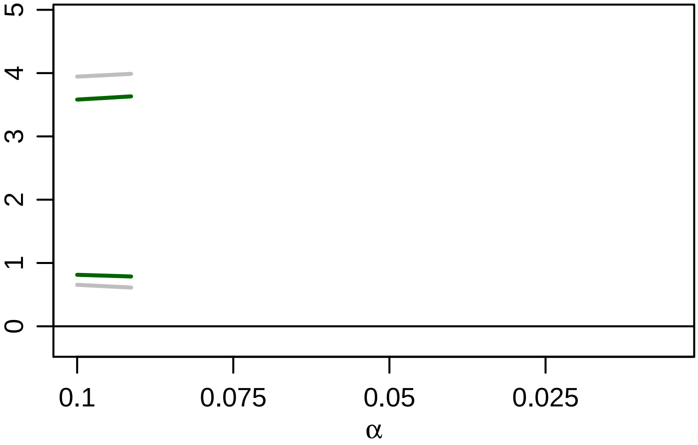
FAB p-values

$$(\mu, \tau^2, \sigma^2) = (1, 1, 1) \quad y = 2.3$$



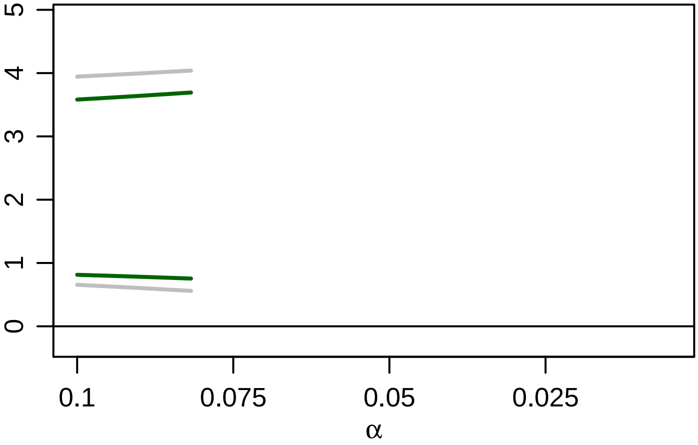
FAB p-values

$$(\mu, \tau^2, \sigma^2) = (1, 1, 1) \quad y = 2.3$$



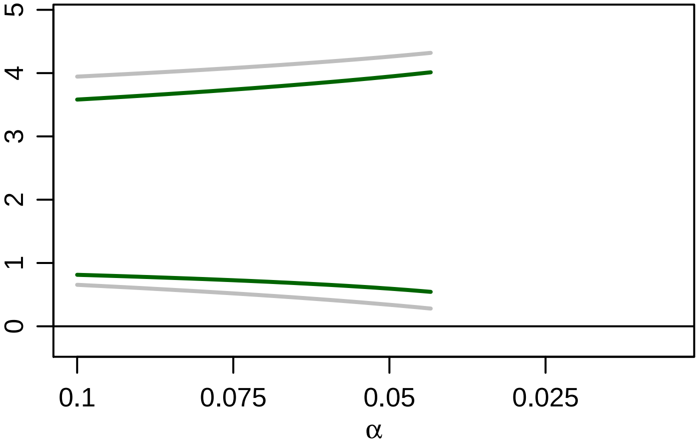
FAB *p*-values

$$(\mu, \tau^2, \sigma^2) = (1, 1, 1) \qquad y = 2.3$$



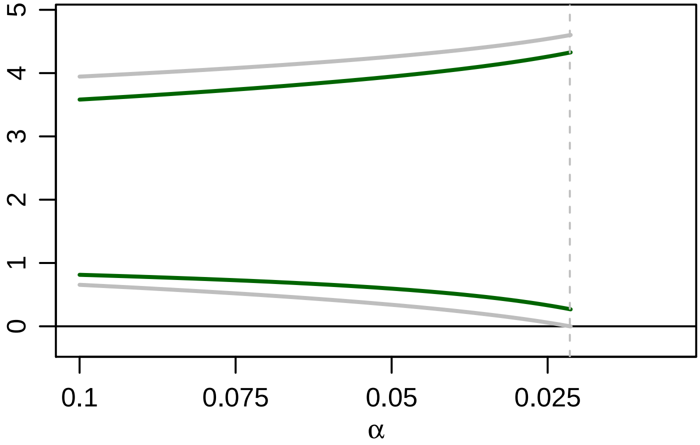
FAB p-values

$$(\mu, \tau^2, \sigma^2) = (1, 1, 1) \quad y = 2.3$$



FAB p-values

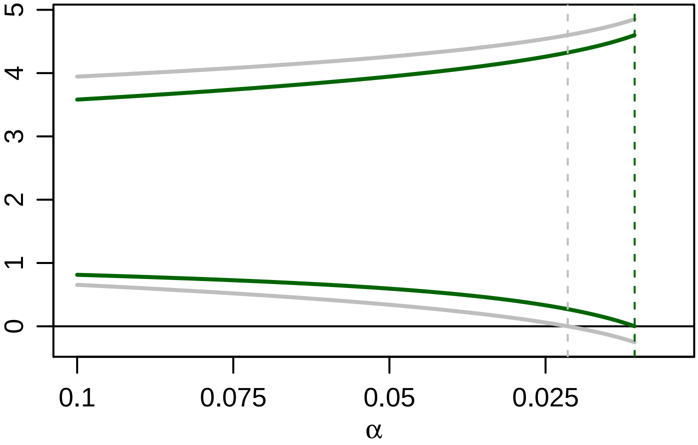
$$(\mu, \tau^2, \sigma^2) = (1, 1, 1) \qquad y = 2.3$$



$p_U = 0.021$

FAB p-values

$(\mu, \tau^2, \sigma^2) = (1, 1, 1) \quad y = 2.3$



$p_U = 0.021 \quad p_F = 0.011$

FAB p-value function

FAB CI endpoints:

$$\theta^U(\alpha) = \frac{y + \sigma\Phi^{-1}\{1 - \alpha + \Phi(\frac{y-\theta^U}{\sigma})\}}{1 + 2\sigma^2/\tau^2} + \mu\frac{2\sigma^2/\tau^2}{1 + 2\sigma^2/\tau^2},$$

$$\theta^L(\alpha) = \frac{y + \sigma\Phi^{-1}\{\alpha - \Phi(\frac{\theta^L-y}{\sigma})\}}{1 + 2\sigma^2/\tau^2} + \mu\frac{2\sigma^2/\tau^2}{1 + 2\sigma^2/\tau^2}.$$

FAB p-value:

$$p = \min\{\alpha : \theta^U(\alpha) \times \theta^U(\alpha) > 0\}$$

FAB p-value function:

$$p_F(y) = 1 - |\Phi(y/\sigma + 2\mu\tfrac{\sigma}{\tau^2}) - \Phi(-y/\sigma)|$$

Bayes optimal test and p-value

The test of $H : \theta = 0$ with best power wrt $\theta \sim N(\mu, \tau^2)$ is MP test of

$$H : Y \sim N(0, \sigma^2) \text{ versus } K : Y \sim N(\mu, \sigma^2 + \tau^2).$$

The p-value associated with this test is

$$p_F(Y) = 1 - |\Phi(Y/\sigma + 2\mu\tfrac{\sigma}{\tau^2}) - \Phi(-Y/\sigma)|.$$

In particular, $1 - |\Phi(Y/\sigma + 2\mu\frac{\sigma}{\tau^2}) - \Phi(-Y/\sigma)| \sim U[0,1]$ if $Y \sim N(0, \sigma^2)$.

More generally, $1 - |F(Z + b) - F(-Z)| \sim U[0,1]$ if

- $Z \sim F$, F symmetric about zero;
- b is constant or otherwise independent of Z.

Adaptive FAB p-values (Hoff 2021)

Sampling model: $\mathbf{Y} \sim N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$

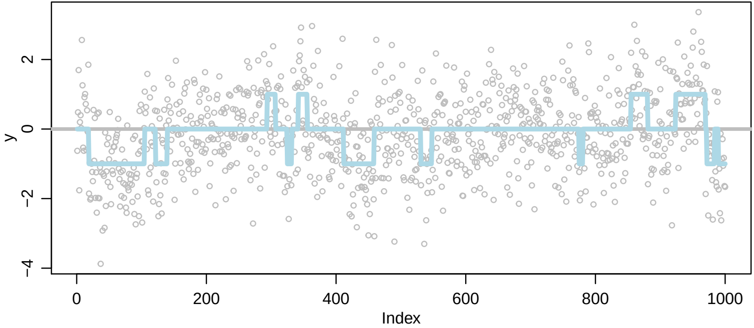
Linking model: $\boldsymbol{\theta} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Psi}) \quad \boldsymbol{\mu} = \boldsymbol{\mu}(\gamma, \mathbf{X}),\ \boldsymbol{\Psi} = \boldsymbol{\Psi}(\gamma, \mathbf{X})$.

Adaptive FAB *p*-value: For each parameter θ_j,

1. Construct *indirect information* $\mathbf{G}_j^\top \mathbf{Y}$ so that $Y_j \perp (\mathbf{G}_j^\top \mathbf{Y})$.
2. Find $(\mu_j, \tau_j^2) = (E[\theta_j | \mathbf{G}_j^\top \mathbf{Y}]\ ,\ V[\theta_j | \mathbf{G}_j^\top \mathbf{Y}])$.
3. Compute FAB p-value $p_F(Y_j) = 1 - |\Phi(Y_j/\sigma_j + 2\mu_j \frac{\sigma_j}{\tau_j^2}) - \Phi(-Y_j/\sigma_j)|$.

- If $\theta_j = 0$ then $p_F(Y_j) \sim U[0, 1]$.
- Exact p-values if $\mathrm{Cor}[Y]$ is available (independence, or linear regression).

Example: Spatial linking model



Truth: $\mathbf{Y} \sim N_p(\boldsymbol{\theta}, \mathbf{I})$, $\boldsymbol{\theta} \in \{-1, 0, +1\}^p$ follows HMM.

Assume: $\mathbf{Y} \sim N_p(\boldsymbol{\theta}, \mathbf{I})$, $\boldsymbol{\theta} \sim N_p(\mathbf{0}, \tau^2 \boldsymbol{\Psi}(\rho))$

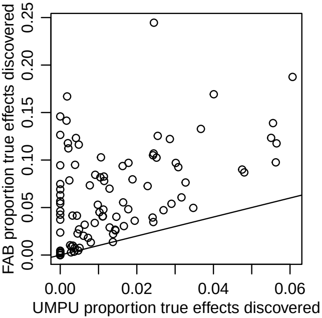
Example: Spatial linking model

For each of 100 simulations,

1. compute p-value for each θ_j, $j \in \{1 \ldots 1000\}$;
2. compute BH threshold for FDR < 0.2;
3. record number of true, false discoveries.

Results

UMPU	FDP=0.106	TDP = 0.02
FAB	FDP=0.108	TDP = 0.08

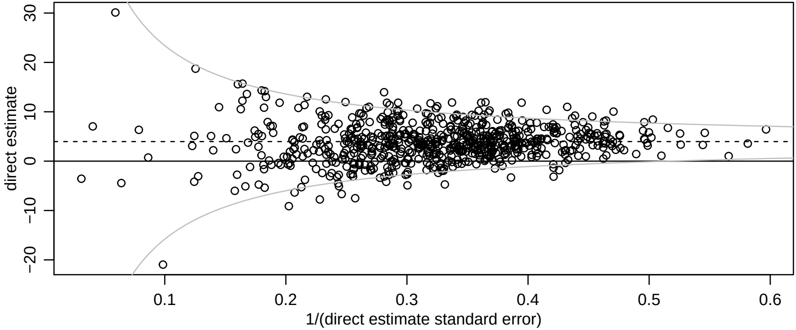


Example: Interactions in linear models

Educational Longitudinal Study:

- Student data and reading scores from $p = 684$ schools.
- Evaluate relationship between SES and reading score in each school.

Sampling model: $y_{i,j} = \mu_j + \boldsymbol{\alpha}^\top \mathbf{w}_i + \beta_j \times x_{i,j} + \epsilon_{i,j}$



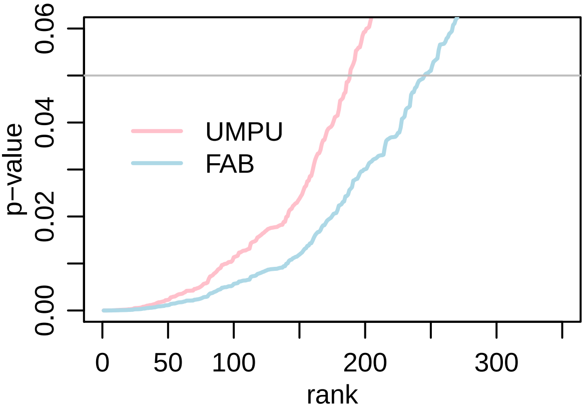
Example: Interactions in linear models

Adaptive FAB p-values: $\hat{\boldsymbol{\beta}}_{OLS} \sim N_p(\boldsymbol{\beta}, \sigma^2 \boldsymbol{V})$, $\boldsymbol{V}$ is known.

1. Construct *indirect information* $\mathbf{G}_j^\top \hat{\boldsymbol{\beta}}$ so that $\hat{\beta}_j \perp (\mathbf{G}_j^\top \hat{\boldsymbol{\beta}})$.
2. Obtain $\hat{\mu}$, $\hat{\tau}^2$ from $\mathbf{G}_j^\top \hat{\boldsymbol{\beta}}$ and linking model $\boldsymbol{\beta} \sim N_p(\mu \mathbf{1}, \tau^2 \mathbf{I})$.
3. Compute FAB p-value $p_F(\hat{\beta}_j) = 1 - |F(\hat{\beta}_j/s_j + 2\hat{\mu}\tilde{s}_j/\hat{\tau}^2) - F(-\hat{\beta}_j/s_j)|$.

Results:

245 and 188 FAB and UMPU p-values less than 0.05.



```
lmFAB               package:FABInference               R Documentation

FAB inference for linear models

Description:

     FAB p-values and confidence intervals for parameters in linear
     regression models

Usage:

     lmFAB(cformula, FABvars, lformula = NULL, alpha = 0.05,
       rssSplit = TRUE, silent = FALSE)

Arguments:

cformula: formula for the control variables

 FABvars: matrix of regressors for which to make FAB p-values and CIs

lformula: formula for the linking model (just specify right-hand side)
```

```
glmFAB              package:FABInference              R Documentation

FAB inference for generalized linear models

Description:

     asymptotic FAB p-values and confidence intervals for parameters in
     generalized linear regression models

Usage:

     glmFAB(cformula, FABvars, lformula = NULL, alpha = 0.05,
       silent = FALSE, ...)

Arguments:

cformula: formula for the control variables

 FABvars: matrix of regressors for which to make FAB p-values and CIs

lformula: formula for the linking model (just specify right-hand side)
```

```
Call:
glm(formula = y ~ . + 0, family = "binomial", data = as.data.frame(cbin
    X)))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3222  -1.0003   0.4131   1.0049   2.3274

Coefficients:
               Estimate Std. Error z value Pr(>|z+bfab|)
`(Intercept)` -0.061399   0.113821  -0.539       0.58959
fv1            0.182868   0.112903   1.620       0.05265 .
fv2            0.338750   0.111676   3.033       0.00121 **
fv3            0.103124   0.122601   0.841       0.20014
fv4            0.294321   0.118451   2.485       0.00648 **
fv5           -0.002105   0.106101  -0.020       0.50792
fv6            0.008249   0.111921   0.074       0.47062
fv7            0.224184   0.115102   1.948       0.02573 *
fv8            0.240131   0.125869   1.908       0.02821 *
fv9            0.025059   0.115629   0.217       0.41421
fv10           0.448530   0.115415   3.886      5.09e-05 ***
```

Summary

Group-level inference motivates group-level error rate guarantees.

Such error rate guarantees do not preclude inclusion of indirect information.

FAB inference

- maintains group-level error rates;
- improves performance on-average across groups;
- rates maintained even if the linking model is wrong.

Extensions and future work

- FAB F-tests and prediction regions;
- non-Guassian sampling and linking models;
- selection adjusted inference;
- FDR control.