Smaller *p*-values in genomics studies using distilled auxiliary information

Jordan Bryan joint work with Peter D. Hoff

Duke University, Department of Statistical Science

Table of contents

1. Multiple hypothesis testing in genomics
2. Frequentist, assisted by Bayes *p*-values
3. Distilled auxiliary information from genomics data
4. Validation by simulation and application to selected studies

Multiple hypothesis testing in genomics

Functional genomics

Functional genomics is the field that seeks to catalogue the function of genes and their protein products

Data generation in functional genomics has exploded in recent years, thanks to advent of RNAi and CRISPR, improvements in DNA and RNA sequencing.

The role of ‘omics’ data in functional genomics (from 2009)

> *‘Omics’ data can provide information on the size and composition of biological entities and thus determine the boundaries of the problem at hand. Biologists can then proceed to investigate function using classical hypothesis-driven experiments. It is still unclear whether even this marriage of the two methods will deliver a complete understanding of biology, but it arguably has a better chance than either method on its own.*
>
> *- Nature Methods editorial [1]*

An updated idea for 2020

‘Omics’ data should provide more than an ad-hoc mechanism for generating hypotheses. The quantitative evidence contained in the correlative structure of large omics datasets should directly aid the validation of biological hypotheses.

Specifically...

An experimental biologist wants to test the existence of a collection of biological effects

$$\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_m\}$$

A statistician has access to a massive genomics data corpus

$$\mathsf{T} = \{\mathsf{T}_1, \ldots, \mathsf{T}_K\}$$

Can the statistician use T to help the biologist resolve $\boldsymbol{\Theta}$?

Frequentist, assisted by Bayes

p-values

Hypothesis testing

1. Identify a biological effect of interest: θ_j
2. Conduct an experiment to obtain samples $Y_{1j}, \dots, Y_{n_j j} \overset{iid}{\sim} N(\theta_j, \sigma^2)$
3. Find estimator $\hat{\sigma}^2$ such that $\nu\hat{\sigma}^2/\sigma^2 \sim \chi^2_\nu$
4. Calculate $T_j = \bar{Y}_j / \sqrt{\hat{\sigma}^2/n_j}$
5. Decide whether $\theta_j = 0$ or $\theta_j \neq 0$ based on T_j

The classical p-value

Supposing $|T_j|$ is the test statistic, Hoff (2019) [2] notes that the formula for the classical p-value for the two-sided test may be written as

$$p_j = 1 - \left|F_\nu(T_j) - F_\nu(-T_j)\right| \tag{1}$$

where F_ν is the cumulative distribution function of the t_ν distribution

- Recall that $p_j \sim U(0, 1)$ under the null hypothesis $\theta_j = 0$
- Type I error can be controlled at level α by rejecting when $p_j < \alpha$

The FAB p-value

For any b_j statistically independent of T_j, Hoff (2019) [2] also notes that another quantity is uniformly distributed under $\theta_j = 0$, namely:

$$p_j^{\text{FAB}} = 1 - \left| F_\nu \left(T_j + b_j \right) - F_\nu \left(-T_j \right) \right| \tag{2}$$

Important

- Valid frequentist p-value
- Corresponding test is **more powerful** than the classical test if b_j and θ_j have the same sign
- Approaches p-value from a one-sided (oracle) test as $b_j \to \pm\infty$ for $\text{sign}(\theta_j) = \pm 1$

Bayes optimal choice of b_j

If prior knowledge about θ_j represented by

$$\theta_j \sim N(m_j, s_j^2) \tag{3}$$

then FAB p-value corresponding to Bayes-optimal level-α test has

$$b_j^{OPT} = 2m_j\sigma/s_j^2 \tag{4}$$

If σ^2 must be estimated from the data, can use plug-in estimator:

$$b_j := 2m_j\tilde{\sigma}/s_j^2 \tag{5}$$

Linking multiple hypotheses

Q: How do we get information m_j, s_j^2 that is independent of T_j?

A: Model relationships among biological effects $\{\theta_j : j = 1, \ldots, m\}$

Linking multiple hypotheses

Suppose we collect auxiliary information about the biological effects $\{\theta_j : j = 1, \ldots, m\}$ into the rows of a matrix X. Then let

$$\begin{aligned} \bar{\mathsf{Y}}|\boldsymbol{\theta} &\sim N_m\left(\boldsymbol{\theta}, \mathrm{diag}(\sigma^2/n_j)\right) \\ \boldsymbol{\theta}|\boldsymbol{\beta} &\sim N_m(\mathsf{X}\boldsymbol{\beta}, \tau^2 \mathsf{I}_m) \\ \boldsymbol{\beta} &\sim N_p(\mathbf{0}, \psi^2 \mathsf{I}_p) \end{aligned} \tag{6}$$

Under this combined sampling and *linking* model, $\bar{\mathsf{Y}}_{-j}$ gives us indirect information about T_j via estimators like

$$\tilde{m}_j = \mathrm{E}(\theta_j|\bar{\mathsf{Y}}_{-j}) \qquad \tilde{s}_j^2 = \mathrm{Var}(\theta_j|\bar{\mathsf{Y}}_{-j}) \tag{7}$$

Interpreting the marginal likelihood

The combined sampling and linking model states that $\bar{\mathsf{Y}}$ is marginally normally distributed with mean $\mathbf{0}$ and covariance

$$\psi^2 \mathsf{X}\mathsf{X}^T + \tau^2 \mathsf{I}_m + \sigma^2 \begin{bmatrix} 1/n_1 & & \\ & \ddots & \\ & & 1/n_m \end{bmatrix} \tag{8}$$

The variation in $\bar{\mathsf{Y}}$ is decomposed into

1. Variation along the principal directions of X
2. Isotropic variation (i.e. variation not explained by X)
3. Measurement error

Empirical Bayes estimation

Conditional on σ^2, τ^2, ψ^2, closed form solutions for $\tilde{m}_j, \tilde{s}_j^2$ exist

Empirical Bayes for variance components

1. $\tilde{\sigma}^2$ can be computed from replicate measurements
2. Given $\tilde{\sigma}^2$, can compute $\tilde{\tau}^2$ and $\tilde{\psi}^2$ by maximizing the marginal likelihood function $p(\bar{\mathrm{Y}}|\tau^2, \psi^2, \tilde{\sigma}^2)$

$$\arg\max_{\tau^2,\psi^2} \left\{ -\sum_{i=1}^{m} \left[\log(\psi^2\lambda_i + \tilde{\sigma}^2/\bar{n} + \tau^2) + \frac{||Q_i^T Y||_2^2}{\psi^2\lambda_i + \tilde{\sigma}^2/\bar{n} + \tau^2} \right] \right\} \qquad (9)$$

Efficient representation in terms of eigenvalues, eigenvectors of XX^T by making approximation $\tilde{\sigma}^2/n_j \approx \tilde{\sigma}^2/\bar{n}, \forall j$

Distilled auxiliary information from genomics data

The common actors

Ideally we would like to have auxiliary information X that is relevant to a wide range of genomics contexts:

- Differential expression analysis
- CRISPR modifier screens
- Drug discovery screens

Experimental techniques and technologies differ, but *genes* and *cancer cell lines* recur

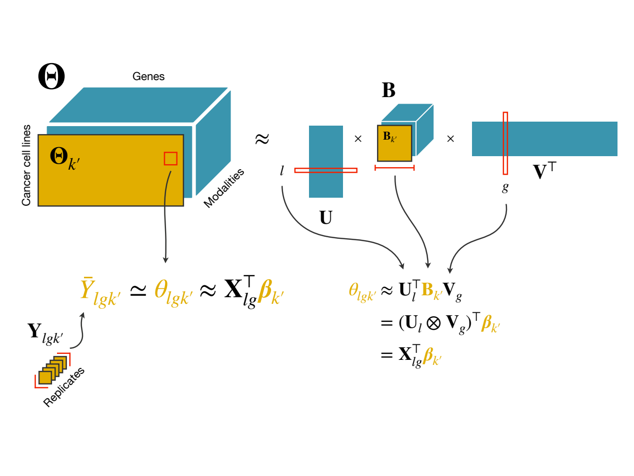
The case for tensor factorization

If we had auxiliary features for **genes** and **cancer cell lines**, we could use the FAB framework to boost power of hypothesis tests for any experiment in which these entities appear

Where to find auxiliary features?

- Recall massive genomics data $\mathsf{T} = \{\mathsf{T}_1, \ldots, \mathsf{T}_K\}$
- Let each T_k be a matrix of measurements for experimental modality k; cancer cell lines on the rows, genes on the columns
- Can decompose tensor T into gene, cancer cell line, and experimental modality constituents

The case for low-rank tensor factorization



$$\bar{Y}_{lgk'} \simeq \theta_{lgk'} \approx \mathbf{X}_{lg}^\top \boldsymbol{\beta}_{k'}$$

$$\begin{aligned}\theta_{lgk'} &\approx \mathbf{U}_l^\top \mathbf{B}_{k'} \mathbf{V}_g \\ &= (\mathbf{U}_l \otimes \mathbf{V}_g)^\top \boldsymbol{\beta}_{k'} \\ &= \mathbf{X}_{lg}^\top \boldsymbol{\beta}_{k'}\end{aligned}$$

Figure 1: Low-rank tensor factorization for covariate distillation. Model explored in different forms in [4], [3], [5]

The case for low-rank Bayesian tensor factorization

The probability model for tensor entries:

$$T_{lgk} | \mathsf{R}_k, \mathsf{U}_l, \mathsf{V}_g, \tau_k^2 \sim N\left(\mu_k + \mathsf{U}_l^T \mathsf{R}_k \mathsf{V}_g, \tau_k^2\right) \tag{10}$$

Convenient extensions

- Probit likelihood for binary data (e.g. mutations)
- “Tobit” likelihood for positive, continuous data (e.g. gene expression)
- MARginalize over missing data within sampling steps

Five step plan for FAB hypothesis testing in genomics

1. Obtain estimators of gene and cancer cell line auxiliary features via tensor probability model and Gibbs sampling: $\hat{\mathsf{U}} = \mathrm{E}(\mathsf{U}|\mathsf{T})$, $\hat{\mathsf{V}} = \mathrm{E}(\mathsf{V}|\mathsf{T})$
2. Construct auxiliary information matrix: $\mathsf{X} = \hat{\mathsf{V}} \otimes \hat{\mathsf{U}}$
3. Obtain estimators $\tilde{m}_j = \mathrm{E}\left(\theta_j|\bar{\mathsf{Y}}_{-j}\right)$, $\tilde{s}_j^2 = \mathrm{Var}(\theta_j|\bar{\mathsf{Y}}_{-j})$ for each biological effect θ_j using sampling, linking model
4. Set $\tilde{b}_j := 2\tilde{m}_j\tilde{\sigma}/\tilde{s}_j^2$
5. Calculate FAB p-value $p_j^{\mathrm{FAB}} = 1 - \left|F_\nu\left(T_j + \tilde{b}_j\right) - F_\nu\left(-T_j\right)\right|$

Validation by simulation and application to selected studies

Validation under null

Simulate 10,000 null datasets ($\boldsymbol{\theta} = \mathbf{0}$, $m = 250$, $n_j = 6\ \forall j$)

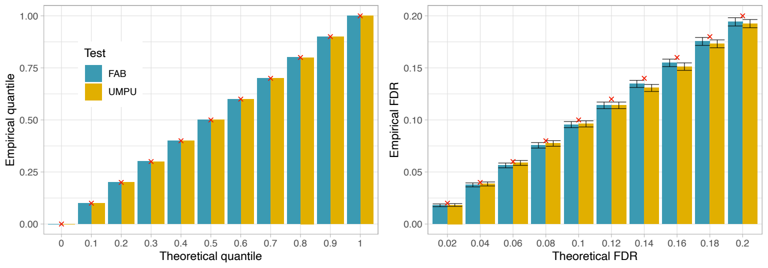


Figure 2: Empirical distribution of UMPU and FAB *p*-values under null simulation. Both achieve target FDR up to Monte Carlo error.

Validation under varying signal-to-noise

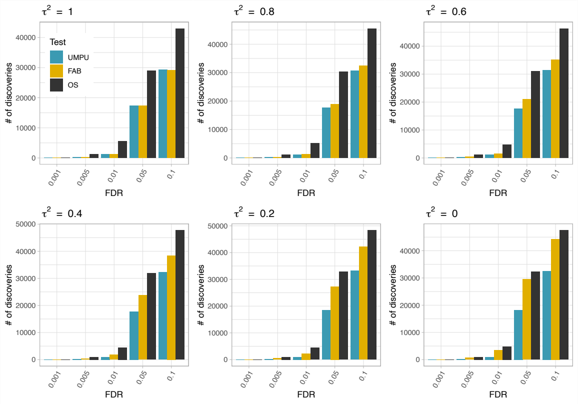


Figure 3: FAB procedure cleanly interpolates between two-sided test and one-sided oracle test as $\psi^2/\tau^2 \to \infty$.

Auxiliary information from the Cancer Dependency Map

The tensor T has a publicly available incarnation: *`depmap.org`*

1. RNAseq
2. CRISPR KO
3. RNAi KD
4. Mutation calls

500 to 1000+ cancer cell lines, 15,000 to 30,000+ genes in each dataset

Application to CRISPR knockout data

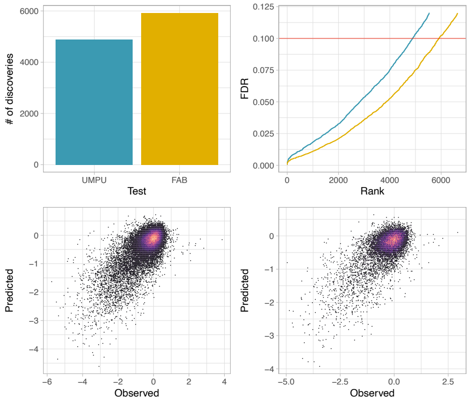


Figure 4: FAB procedure leads to more discoveries on new experiments using modalities contained in **T**

Application to drug discovery

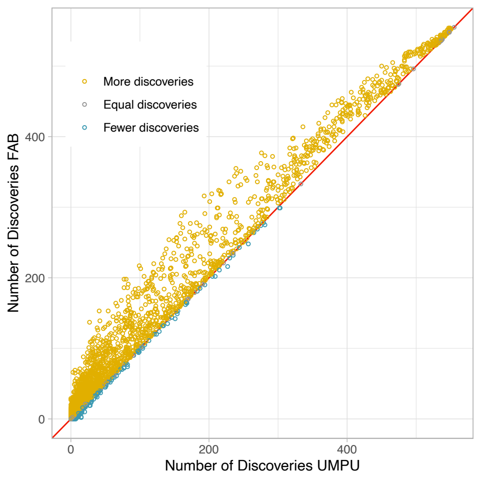


Figure 5: Relationships among cancer cell lines lead to more discoveries for 67% of tested compounds. At least as many discoveries as classical procedure for 87% of tested compounds.

Application to differential expression analysis

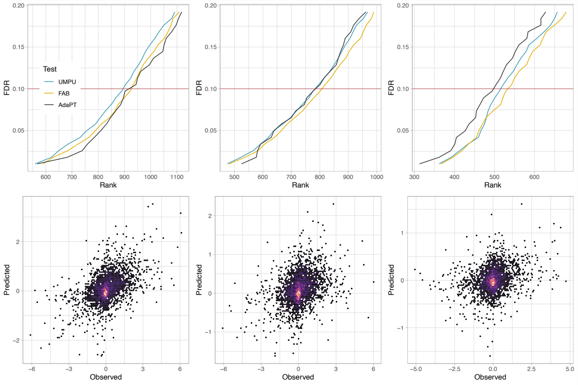


Figure 6: Gene-gene relationships distilled from cancer genomics data contain information relevant to non-cancer study.

Take home points

1. FAB procedure offers increase in statistical power, which can accumulate to hundreds of additional discoveries in the presence of many hypotheses. Strict type I error and FDR control.
2. Cell line and gene representations derived from tensor model can be used with FAB procedure to improve statistical power in many genomics contexts.
3. Little downside to proposed FAB procedure. Reverts to classical hypothesis testing when $\psi^2 \ll \tau^2$.

Code available at *`https://github.com/j-g-b/BTF`*

References i

- Defining the scientific method.
  Nature Methods, 6(4):237–237, Apr. 2009.

- P. D. Hoff.
  Smaller p-values via indirect information.
  arXiv:1907.12589 [stat], July 2019.
  arXiv: 1907.12589.

- R. Salakhutdinov and A. Mnih.
  Bayesian probabilistic matrix factorization using Markov chain Monte Carlo.
  In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 880–887, Helsinki, Finland, 2008. ACM Press.

References ii

I. Sutskever, J. B. Tenenbaum, and R. R. Salakhutdinov.
Modelling Relational Data using Bayesian Clustered Tensor Factorization.
In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1821–1828. Curran Associates, Inc., 2009.

J. Ye.
Generalized Low Rank Approximations of Matrices.
Machine Learning, 61(1-3):167–191, Nov. 2005.

Back to the linking model

Compare the vectorized form of the tensor probability model for the k^{th} modality

$$\operatorname{vec}(\mathsf{T}_k)\,|\mathsf{R}_k, \mathsf{V}, \mathsf{U}, \tau_k^2 \sim N_{n_V \times n_U}\left((\mathsf{V} \otimes \mathsf{U})\operatorname{vec}(\mathsf{R}_k), \tau_k^2 \mathsf{I}_{n_V \times n_U}\right) \tag{11}$$

to the linking model for multiple hypotheses

$$\boldsymbol{\theta}|\boldsymbol{\beta}, \mathsf{X}, \tau^2 \sim N_m(\mathsf{X}\boldsymbol{\beta}, \tau^2 \mathsf{I}_m) \tag{12}$$

Conditional on $\mathsf{X} = \mathsf{V} \otimes \mathsf{U}$, the linking model is equivalent to the tensor model applied to a *new* experimental modality k'

Application to differential dependency analysis (small m)

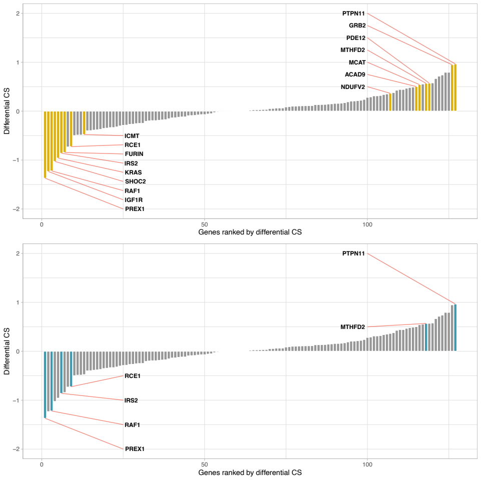


Figure 7: FAB procedure can be effective even when m is small compared to $d_V \times d_U$