Empirical Bayes, shrinkage, and SVA

Neo Christopher Chung

Lecture 5, 1000-719bMSB

Course survey

There’s an amazing football player, who scored **1 goal per game** last year!

On average, the main strikers in this league have an average **0.4 goals per game.**

You are asked to predict what this amazing player’s **this year average goal/game**

Lewandowski has 0.56 goal/game (Poland); 0.48 goal/game (Barcelona)

0.4	0.5	0.6	0.7	0.8	0.9	1.0

Simple Linear Model

Let’s review a simple linear model:

$$y_i = b_0 + b_1 x_{i1} + e_{i1}$$

y: a dependent variable
x: a independent variable
b_0: an intercept
b_1: a coefficient
e: independently and identically distributed (i.i.d.) noise

Given many data for y and x, the method of least square provides estimates for b_0 and b_1

Multiple Linear Model

We look at multiple observations and independent variables simultaneously:

$$y_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip} + e_i$$

y_i for $i = 1, \ldots, n$ is an observed measurement (e.g., gene expression)
$\mathbf{x}_j$ for $j = 1, \ldots, p$ makes p independent variables
e.g., $\mathbf{X}_1$ may be a vector of ages
$\mathbf{X}_2$ may be a vector of susceptibility to a given disease
$\mathbf{X}_3$ may be a sequencing lane number
……

y indicates a scalar, **y** indicates a vector, **Y**/Y indicates a matrix.
A vector tends to indicate a column vector, but not always.
Notations are confusing and dependent on domains and contexts.

Multiple Linear Model

$$y_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip} + e_i$$

$$\mathbf{y} = X \mathbf{b} + \mathbf{e}$$

$$\begin{Bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{Bmatrix} = \begin{Bmatrix} x_1^T \\ x_2^T \\ . \\ . \\ . \\ x_n^T \end{Bmatrix} \begin{Bmatrix} b_0 \\ b_1 \\ . \\ b_p \end{Bmatrix} + \begin{Bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{Bmatrix}$$

Multiple Linear Model

$$\mathbf{y} = X \mathbf{b} + \mathbf{e}$$

This linear model is what we solve in **lm()** function

e.g., mod = lm(y ~ x)

Estimation and significance testing on **b** is of our interest

Least square estimator

- The method of least squares is the de facto standard method to estimate the coefficients.
- Minimizing the sum of the squares of the residuals
- It’s the maximum likelihood estimation when the noise is normally distributed with equal variances.

- Gauss–Markov theorem: it’s the best linear unbiased estimator of the coefficients

Linear Model in a Matrix Form

Finally, we consider *m* variables of *n* observations

As a convention, we are stacking vectors ($\mathbf{y_1}\ \mathbf{y_2}\ \ldots\ \mathbf{y_m}$) as rows into a *m* x *n* matrix Y:

$$\mathbf{Y} = \mathbf{B}\,\mathbf{X} + \mathbf{E}$$

$$\begin{Bmatrix} y_{11} \cdots y_{1n} \\ y_{21} \cdots y_{2n} \\ \vdots \\ y_{m1} \cdots y_{mn} \end{Bmatrix} = \begin{Bmatrix} b_{11}\, b_{12} \cdots b_{1p} \\ b_{21}\, b_{22} \cdots b_{2p} \\ \vdots \\ b_{m1}\, b_{m2} \cdots \\ b_{mp} \end{Bmatrix} \begin{Bmatrix} x_{11} x_{12} \ldots x_{1n} \\ x_{21} x_{22} \ldots x_{2n} \\ \vdots \\ x_{p1} x_{p2} \ldots x_{pn} \end{Bmatrix} + \begin{Bmatrix} e_{11} e_{12} \ldots e_{1n} \\ e_{21} e_{22} \ldots e_{2n} \\ \vdots \\ e_{m1} e_{m2} \ldots \\ e_{mn} \end{Bmatrix}$$

As we discussed gene expression

$$\mathbf{Y} = \mathbf{B}\,\mathbf{X} + \mathbf{E}$$

Y = Observed genomic data, containing m variables (rows) and n observations (cols)

X = Biological variables

E = Independently and identically distributed (i.i.d.) noise

p=1 independent variable

$$\mathbf{Y} = \mathbf{BX} + \mathbf{E}$$

$$\begin{Bmatrix} y_{11} \cdots y_{1n} \\ y_{21} \cdots y_{2n} \\ \vdots \\ y_{m1} \cdots y_{mn} \end{Bmatrix} = \begin{Bmatrix} b_{11} \\ b_{21} \\ \vdots \\ b_{m1} \end{Bmatrix} \begin{Bmatrix} x_{11} x_{12} \cdots x_{1n} \end{Bmatrix} + \begin{Bmatrix} e_{11} e_{12} \cdots e_{1n} \\ e_{21} e_{22} \cdots e_{2n} \\ \vdots \\ e_{m1} e_{m2} \cdots \\ e_{mn} \end{Bmatrix}$$

Consider there's p=1 variable labeling samples

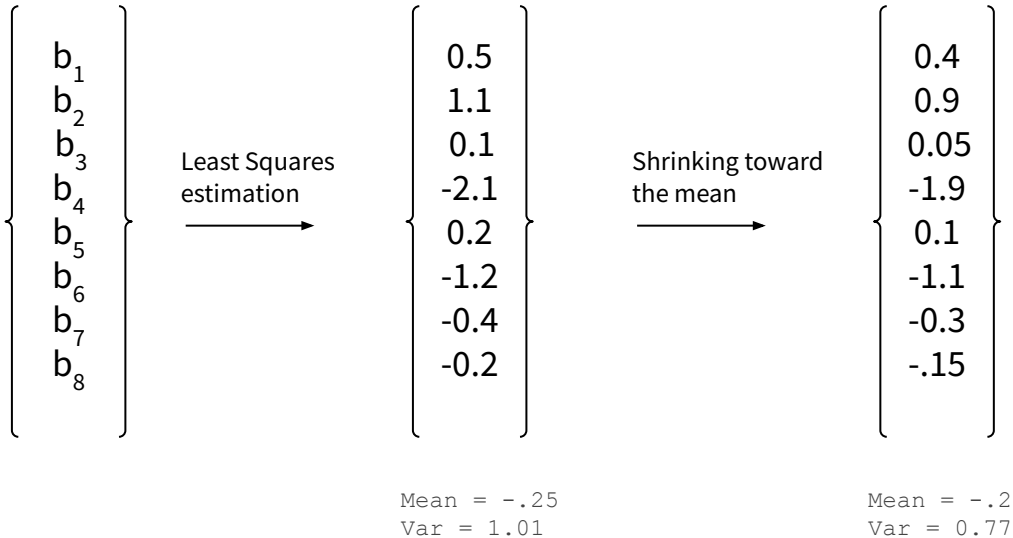
We are interested in estimating coefficients, for m genes

Least squares with a large m

- Estimate each b_{ij} independently via minimizing the sum of the squares of the residuals
- For each variable $i = 1, \ldots, m$, the errors are uncorrelated, a mean of zero, equal variances (a.k.a. optimal)

- However, when we are dealing with a large data -- many m variables, measured on a set of n observations --, consider a bias variance tradeoff
- Simply put, we may able to get “better” estimates by reducing variance and increasing bias
- Read more on James–Stein estimator

Shrinkage, simplified example



Mean = -.25
Var = 1.01

Mean = -.25
Var = 0.77

Predicting a baseball player's batting average

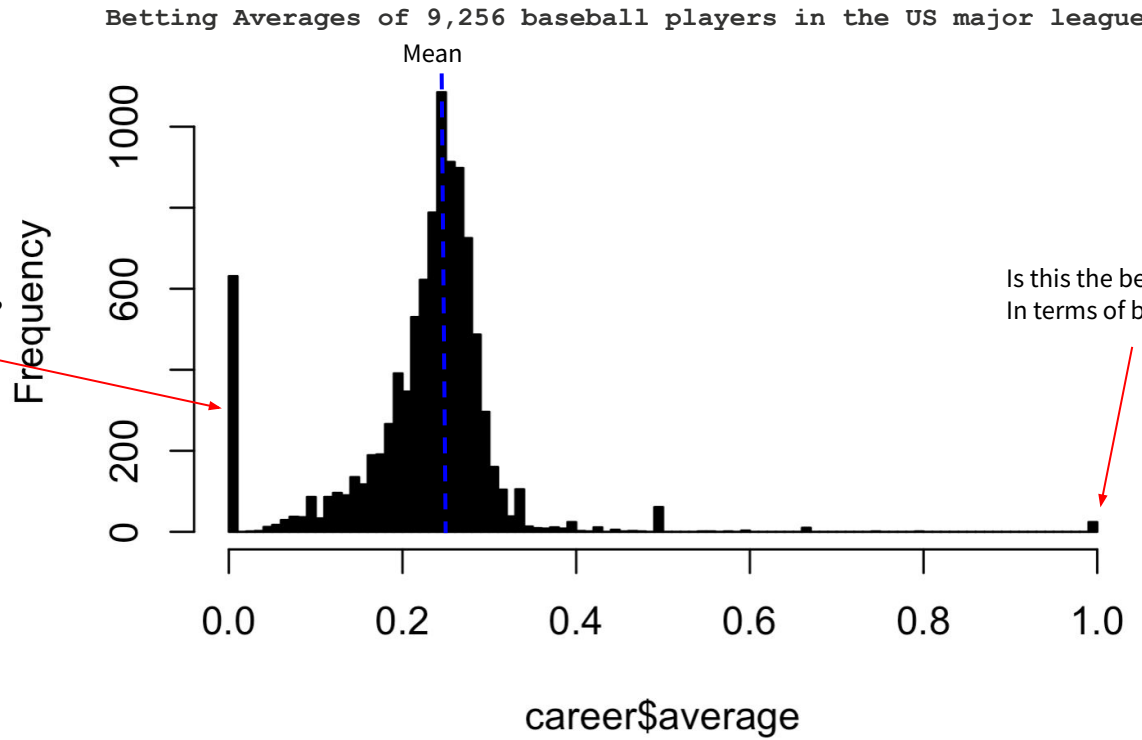
In sports analytics, we often want to predict a player's statistics in the future.
E.g., a batting average = # of a baseball player's hits divided by # of at-bats.

Given the last year's data on batting averages, you are tasked with
predicting players' batting average next year.

The unbiased linear model would suggest that you use the individual player's batting average from the last year as the predicted average for the next year.

We can do much better.

Batting averages, in the past



Will these players have an average of 0.0 again next year?

Is this the best player In terms of betting average?

Adapted from
http://varianceexplained.org/r/empirical_bayes_baseball/

Lahman (R package) has the data

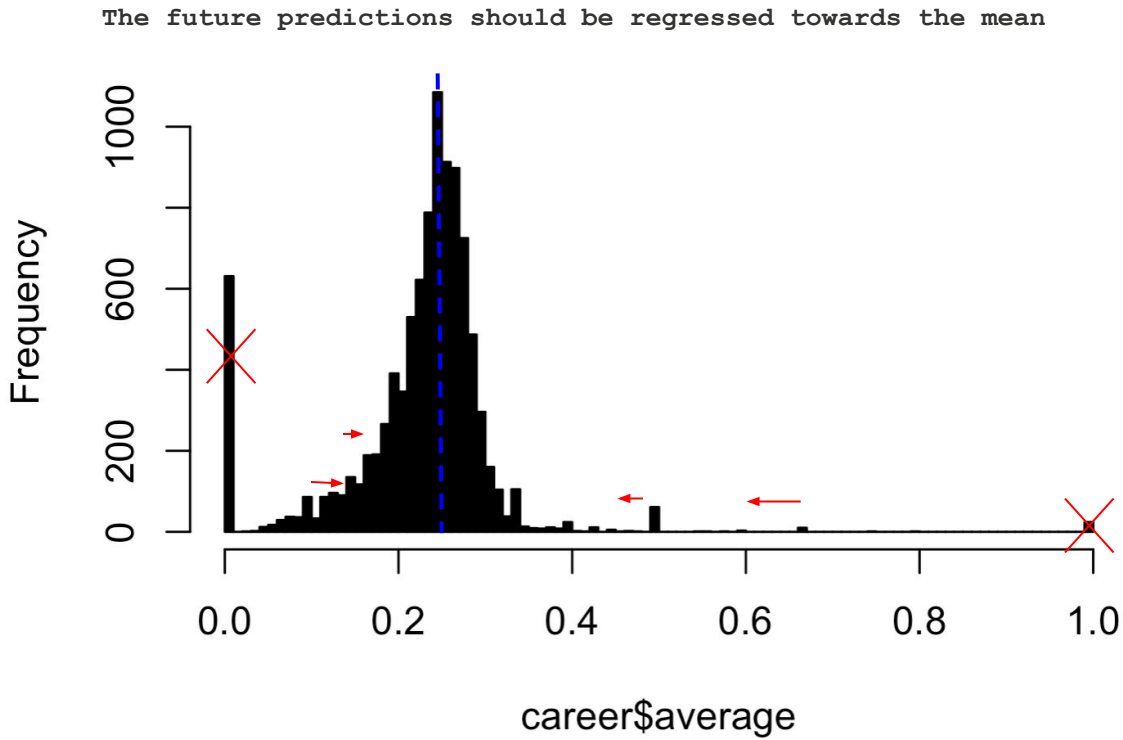
```
library(Lahman)
```

```
career
```

```
## Source: local data frame [9,256 x 4]
##
##                 name     H    AB average
##                (chr) (int) (int)   (dbl)
## 1         Hank Aaron  3771 12364  0.3050
## 2       Tommie Aaron   216   944  0.2288
## 3          Andy Abad     2    21  0.0952
## 4        John Abadie    11    49  0.2245
## 5     Ed Abbaticchio   772  3044  0.2536
## 6        Fred Abbott   107   513  0.2086
## 7        Jeff Abbott   157   596  0.2634
## 8        Kurt Abbott   523  2044  0.2559
## 9         Ody Abbott    13    70  0.1857
## 10 Frank Abercrombie     0     4  0.0000
## ..               ...   ...   ...     ...
```

H = Hit
AB = At Bats

Batting averages, shrinkage



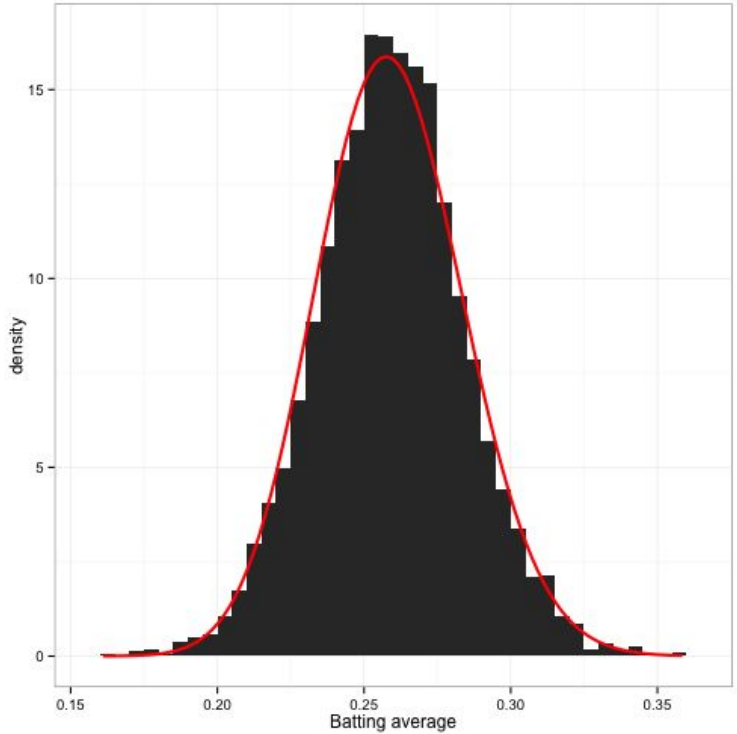
Prior distribution (empirical Bayes approach)

1. Remove pitchers
2. Filter out all players that have fewer than 500 at-bats
3. Fit a Beta distribution to the remaining data

$$X \sim \text{Beta}(\alpha_0, \beta_0)$$

```
career <- Batting %>%
  filter(AB > 0) %>%
  anti_join(Pitching, by = "playerID") %>%
  group_by(playerID) %>%
  summarize(H = sum(H), AB = sum(AB)) %>%
    mutate(average = H / AB)
career_filtered <- career %>%
    filter(AB >= 500)
m <- MASS::fitdistr(career_filtered$average, dbeta,
                    start = list(shape1 = 1, shape2 = 10))
alpha0 <- m$estimate[1]
beta0 <- m$estimate[2]
```

α_0=78.661 β_0=224.875



Estimate an individual's batting avg

We are using the estimated Beta distribution. Then update the individual's batting average accordingly.

Instead of $\text{average}_{\text{sample}} = H/AB$

$\text{average}_{EB} = (H+\alpha_0)/(AB+\alpha_0+\beta_0)$

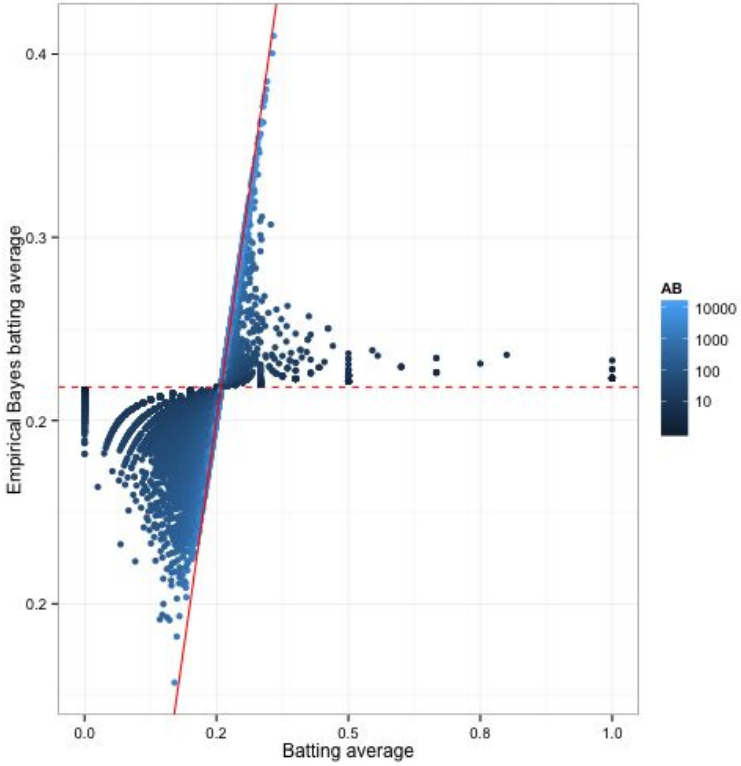
EXAMPLE: Batter A with at-bats = 1000 and hits = 300

Batter B with at-bats = 10 and hits = 4

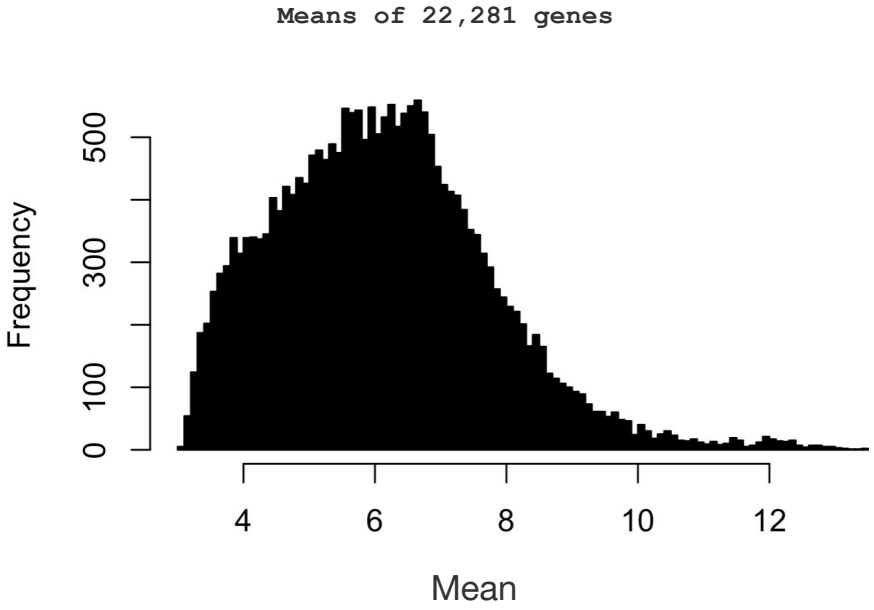
$\text{average}_A = (300+\alpha_0)/(1000+\alpha_0+\beta_0) = (300+78.7)/(1000+78.7+224.9) = 0.29$

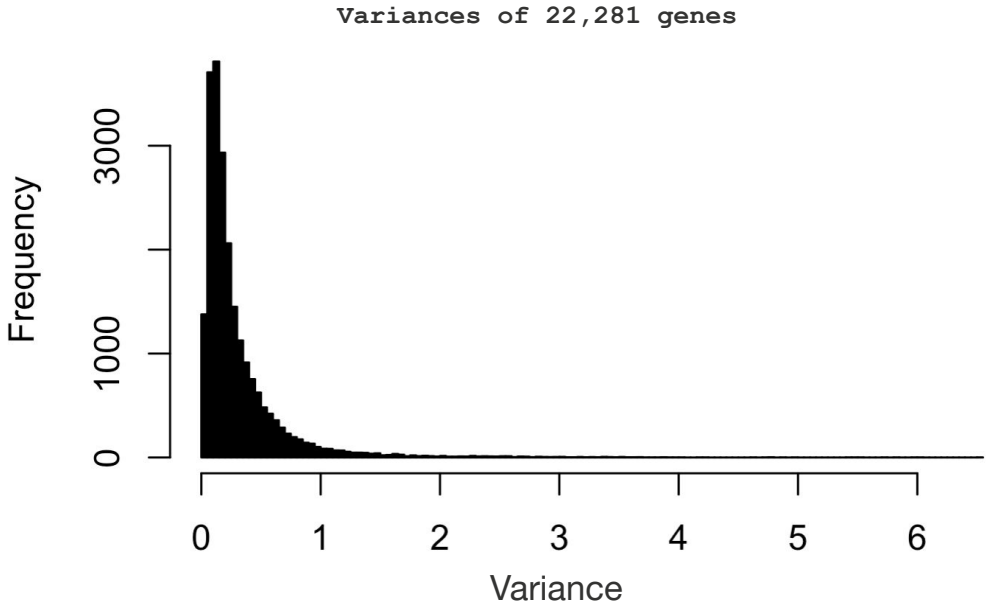
$\text{average}_B = (4+\alpha_0)/(10+\alpha_0+\beta_0) = (4+78.7)/(10+78.7+224.9) = 0.264$

Shrinkage in baseball



Bladder cancer data, Dyrskjøt et al. (2004)





Shrinkage in large scale modeling

- Regression towards mean
- Apply implicitly or explicitly.
  - Visualizing a box plot
  - Outliers are removed in a quality control step
  - Removing minimally expressed genes or zero expression values
  - In different batches, means/variances must be similar
  - Coefficients are considered simultaneously
  - Models accounts for a large m -- learning from all the data

Linear model, with technical variables

$$Y = BX + E$$

$$= BX + \Gamma G + U$$

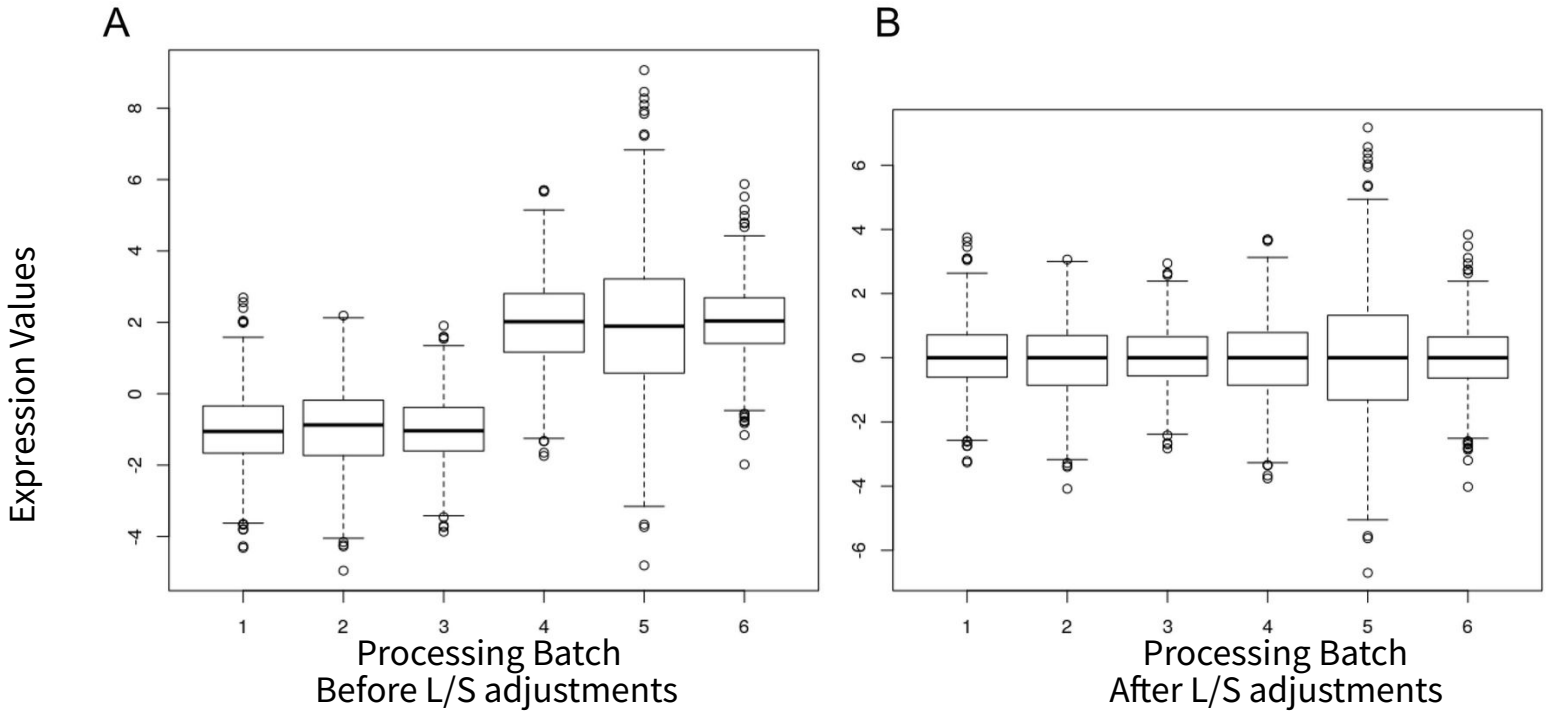
X = Biological variables

G = Technical variables

U = i.i.d. Noise

*Note that we are now dropping the bold notation.

Boxplots of data, within each batch



Čuklina et al (2019)

Location and scale (L/S) adjustments

A wide family of adjustments in which one assumes a model for the location (mean) and/or scale (variance) of the data within batches and then adjusts the batches to meet assumed model specifications.

Therefore, the batch effects can be modeled out by standardizing means and variances across batches

The simplest approach for L/S batch adjustment is to mean center and standardize the variance of each batch for each gene independently.

In more complex situations such as unbalanced designs or when incorporating numerical covariates, use a general L/S framework:

Johnson et al (2007)

Bayesian statistics

Bayesian statistics update probabilities, after obtaining new data

$P(A|B) = P(B|A)\ P(A) / P(B)$ for now assume $P(B) \neq 0$

A: a proposition, or a prior belief
B: an evidence, or observed data

→ Obtain a posterior probability, from a prior probability based on our data

Johnson et al (2007)

Frequentist and Bayesian statistics

	Frequentist	**Bayesian**
Hypothesis test	p-value	Bayes factor
Estimation	Maximum likelihood estimate with confidence interval	Posterior distribution
Probability	Frequency (Objective)	Degree of belief (Subjective)
Parameter	Fixed	Random variable

Empirical Bayes

In a standard Bayesian, **a prior = fixed** before data

In empirical Bayes, a prior distribution **is estimated from the data**

No need to impose or have a strong prior belief

Bridging two sides of statistical traditions

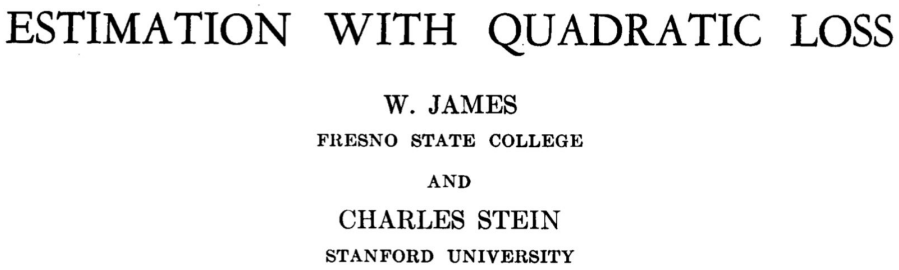
Appropriate for modeling large-scale biological data

Empirical Bayes

Stein's Paradox in Statistics by Efron & Morris (1977)

When three or more parameters are estimated simultaneously, there exist combined estimators more accurate on average (lower expected mean squared error) than any method that handles the parameters separately - Wikipedia

Named after Charles Stein, famous for James & Stein (1961)



ESTIMATION WITH QUADRATIC LOSS

W. JAMES
FRESNO STATE COLLEGE
AND
CHARLES STEIN
STANFORD UNIVERSITY

General L/S framework

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}$$

Y_{ijg} : observed data -- the expression value

for sample *j* from batch *i* containing *m* batches

n_i samples within batch *i* for *i*=1,...,m

for gene g=1,...,G

X : biological variables -- a design matrix for sample conditions

β_g : regression coefficients corresponding to *X*

γ_{ig} : additive batch effects of batch *i* for gene *g*

δ_{ig} : multiplicative batch effects of batch *i* for gene *g*

ε : noise with mean zero and variance σ^2_g

Johnson et al (2007)

Empirical bayes approach with ComBat

The most important disadvantage of many existing methods is that large batch sizes are required for implementation because such methods are not robust to outliers.

ComBat make this possible even for a smaller sample size by:

1. Estimating the L/S model parameters that represent the batch effects by pooling information across genes in each batch
2. Shrinking the batch effect parameter estimates toward the overall mean of the batch effect estimates (across genes)

Johnson et al (2007)

ComBat algorithm

1. Standardize the data

   Standardize gene-wise so that genes have similar overall mean and variance
   Standardized data, Z_{ijg}, satisfy the distributional form, $Z_{ijg} \sim N(\gamma_{ig}, \delta^2_{ig})$

2. Batch effect parameter estimates using parametric empirical priors

   Johnson et al (2007) uses the following priors $\gamma_{ig} \sim N(Y_i, \tau_i^2)$ and $\delta_{ig}^2 \sim \text{Inverse Gamma}\,(\lambda_i, \theta_i)$
   Those hyperparameters are estimated from the standardized data, Z_{ijg}
   Then, the posteriors are

$$\gamma_{ig}^* = \frac{n_i \overline{\tau}_i^2 \widehat{\gamma}_{ig} + \delta_{ig}^{2*} \overline{\gamma}_i}{n_i \overline{\tau}_i^2 + \delta_{ig}^{2*}} \quad \text{and} \quad \delta_{ig}^{2*} = \frac{\overline{\theta}_i + \frac{1}{2}\sum_j (Z_{ijg} - \gamma_{ig}^*)^2}{\frac{n_j}{2} + \overline{\lambda}_i - 1}$$

3. Adjust the data for batch effects with

$$\gamma_{ijg}^* = \frac{\widehat{\sigma}_g}{\widehat{\delta}_{ig}^*}(Z_{ijg} - \widehat{\gamma}_{ig}^*) + \widehat{\alpha}_g + X\widehat{\beta}_g$$

Johnson et al (2007)

Shrinkage in gene expression

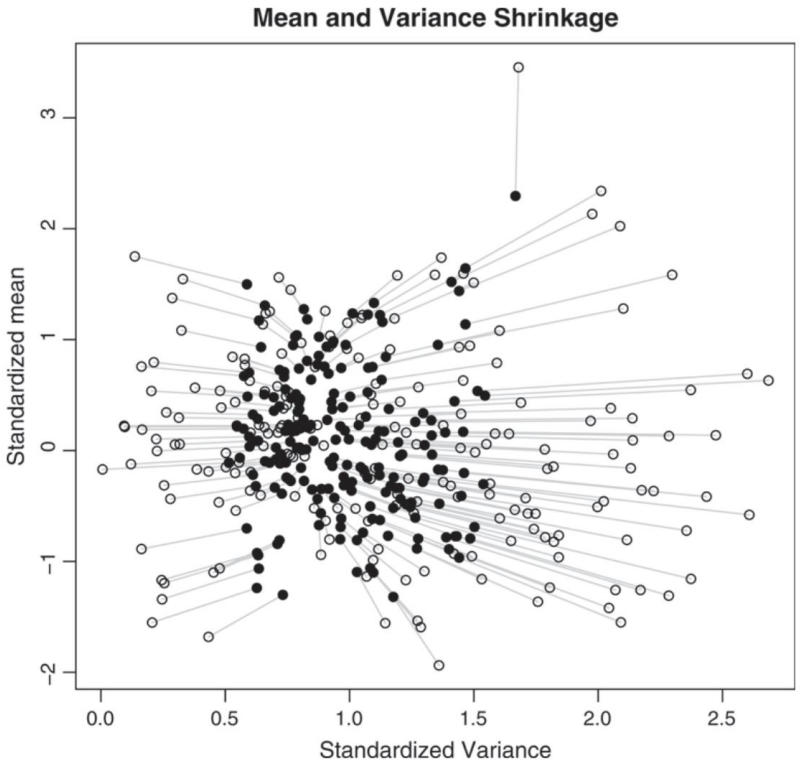
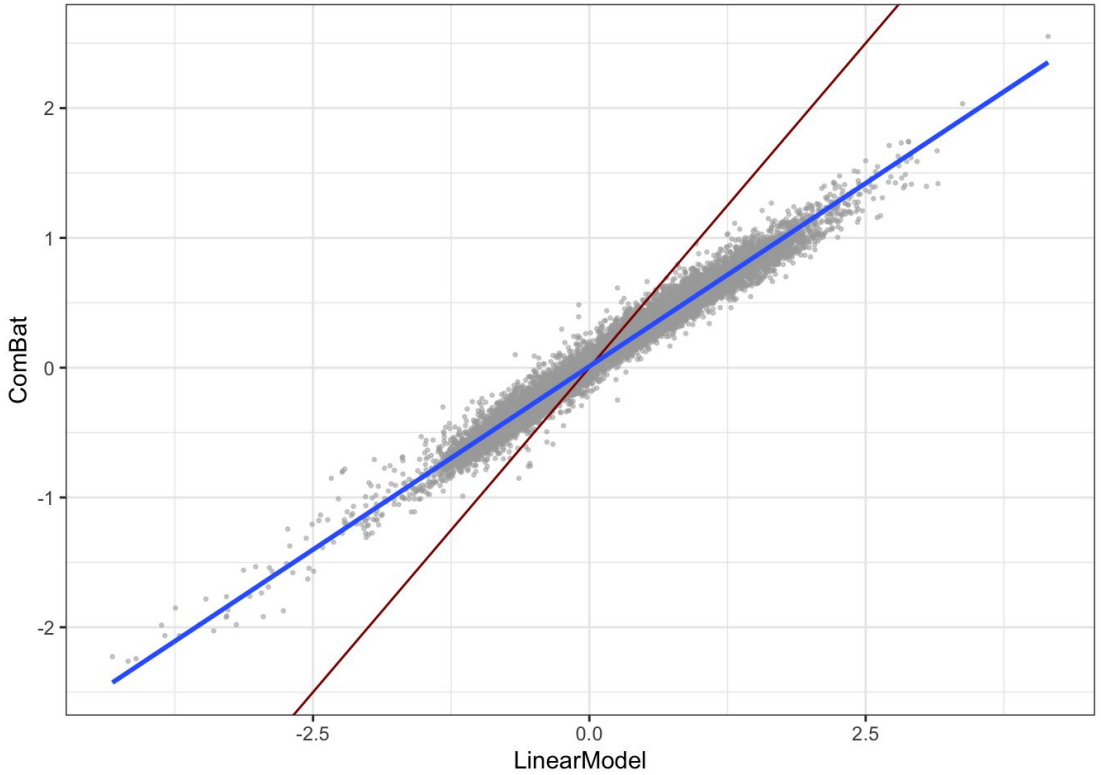
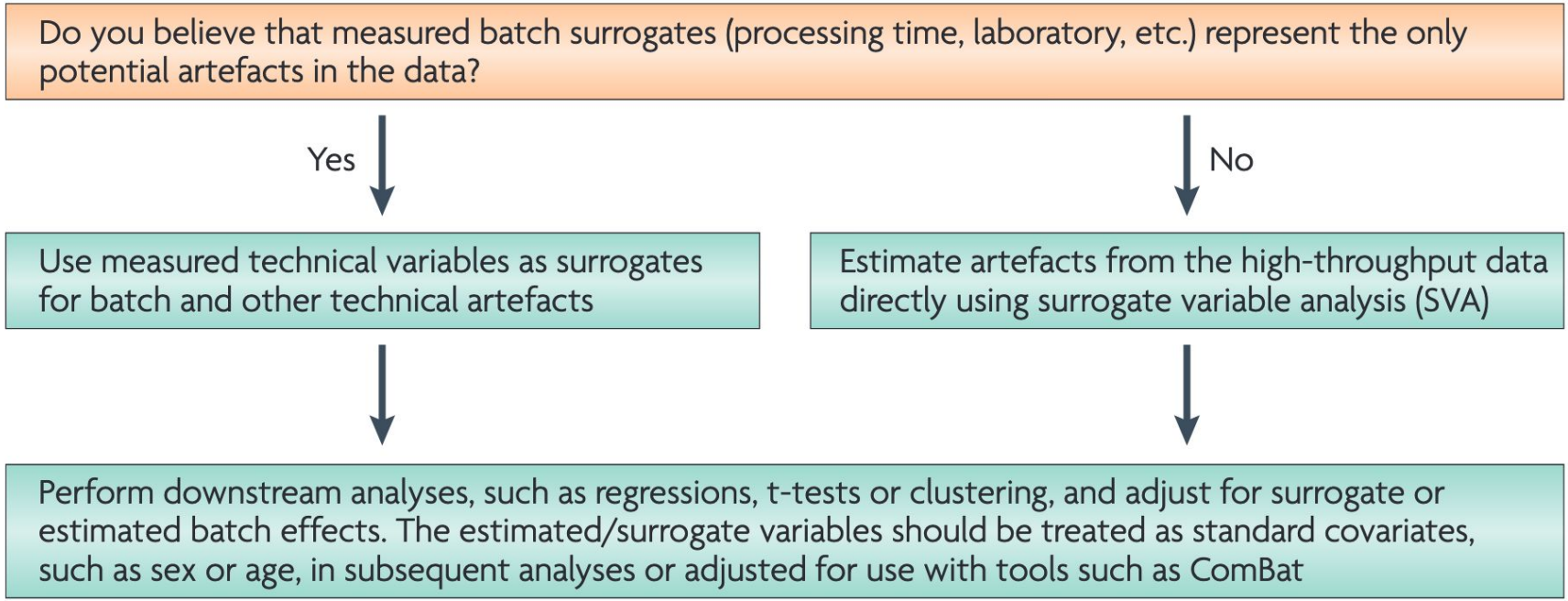


Fig. 3. Shrinkage plot for the first 200 probes from one of the batches in data set 1. The gene-wise and EB estimates of γ_{ig} and δ^2_{ig} in Section 3.1 are plotted on the Y and X axis. Open circles are the gene-wise values and the solid are after applying the EB shrinkage adjustment.

Coefficients, LM vs. ComBat (EB shrinkage)



Downstream analyses



Diagnostic analyses

Use of SVA and ComBat does not guarantee that batch effects have been addressed. After fitting models, including processing time and date or surrogate variables estimated with SVA, re-cluster the data to ensure that the clusters are not still driven by batch effects

Leek et al. 2010

Surrogate Variable Analysis

$$Y = BX + \Gamma G + U$$

Leek & Storey (2007) Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis

What if the technical variables G are not known.

We can estimate ΓG through an iterative process.

“Surrogate variables” are replacing (unknown and unmeasured) technical variables.

Surrogate Variable Analysis (SVA)

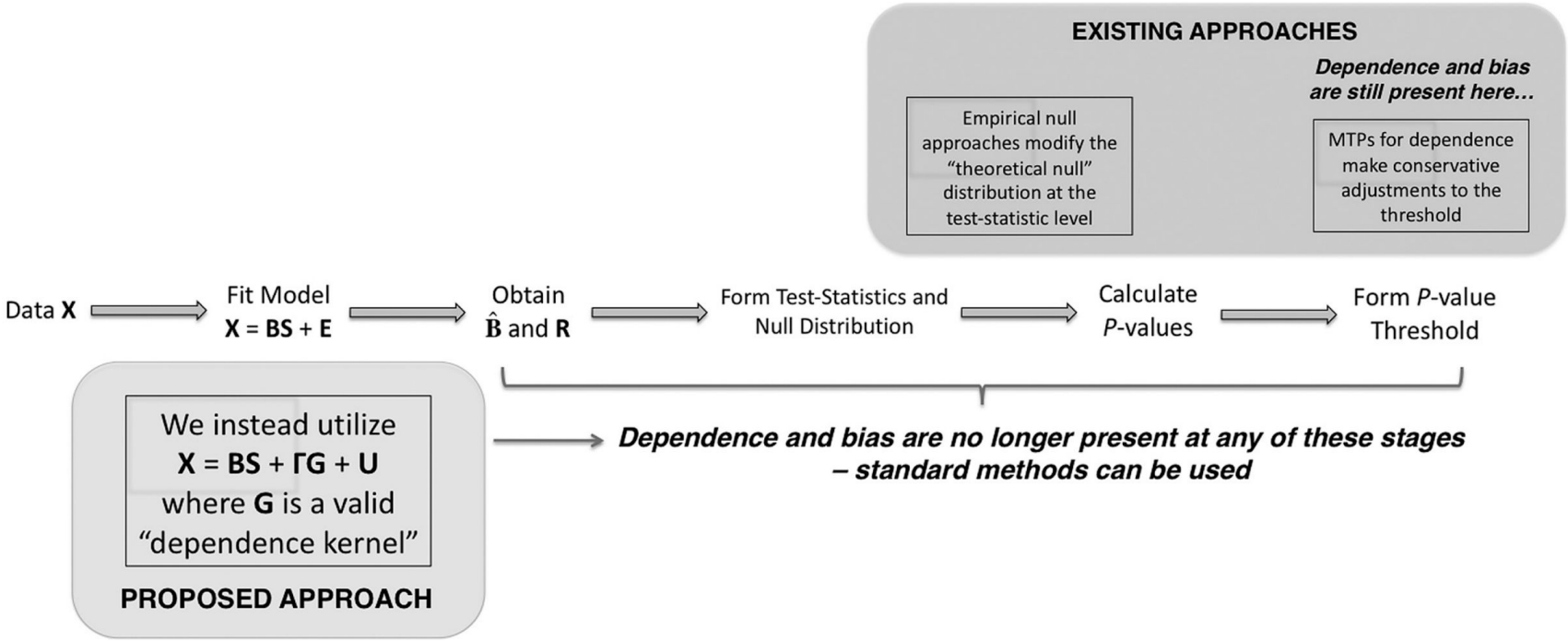


Fig. 1. A schematic of the general steps of multiple hypothesis testing. We directly account for multiple testing dependence in the model-fitting step, where all the downstream steps in the analysis are not affected by dependence and have the same operating characteristics as independent tests. Our approach differs from current methods, which address dependence indirectly by modifying the test statistics, adaptively modifying the null distribution, or altering significance cutoffs. For these downstream methods the multiple testing dependence is not directly modeled from the data, so distortions of the signal of interest and the null distribution may be present regardless of which correction is implemented.

The fundamental idea behind SVA

UNKNOWN

$$\mathbf{Y} = \mathbf{B}\,\mathbf{X} + \mathbf{\Gamma}\,\mathbf{G} + \mathbf{E}$$

$$\begin{Bmatrix} y_{11} \cdots y_{1n} \\ y_{21} \cdots y_{2n} \\ \cdot \\ \cdot \\ \cdot \\ y_{m1} \cdots y_{mn} \end{Bmatrix} = \begin{Bmatrix} b_{11} \\ b_{21} \\ \cdot \\ \cdot \\ \cdot \\ b_{m1} \end{Bmatrix} \begin{Bmatrix} x_{11} x_{12} \cdots x_{1n} \end{Bmatrix} + \begin{Bmatrix} \gamma_{11} \\ \gamma_{21} \\ \cdot \\ \cdot \\ \cdot \\ \gamma_{m1} \end{Bmatrix} \begin{Bmatrix} g_{11} g_{12} \cdots g_{1n} \end{Bmatrix} + \begin{Bmatrix} e_{11} e_{12} \cdots e_{1n} \\ e_{21} e_{22} \cdots e_{2n} \\ \cdot \\ \cdot \\ \cdot \\ e_{m1} e_{m2} \cdots \\ e_{mn} \end{Bmatrix}$$

STEP 1. Simply fitting a linear model

$$\mathbf{Y} = \mathbf{B}\,\mathbf{X} + \mathbf{E}$$

$$\begin{Bmatrix} y_{11} \cdots y_{1n} \\ y_{21} \cdots y_{2n} \\ \vdots \\ y_{m1} \cdots y_{mn} \end{Bmatrix} = \begin{Bmatrix} b_{11} \\ b_{21} \\ \vdots \\ b_{m1} \end{Bmatrix} \begin{Bmatrix} x_{11} x_{12} \cdots x_{1n} \end{Bmatrix} + \begin{Bmatrix} e_{11} e_{12} \cdots e_{1n} \\ e_{21} e_{22} \cdots e_{2n} \\ \vdots \\ e_{m1} e_{m2} \cdots e_{mn} \end{Bmatrix}$$

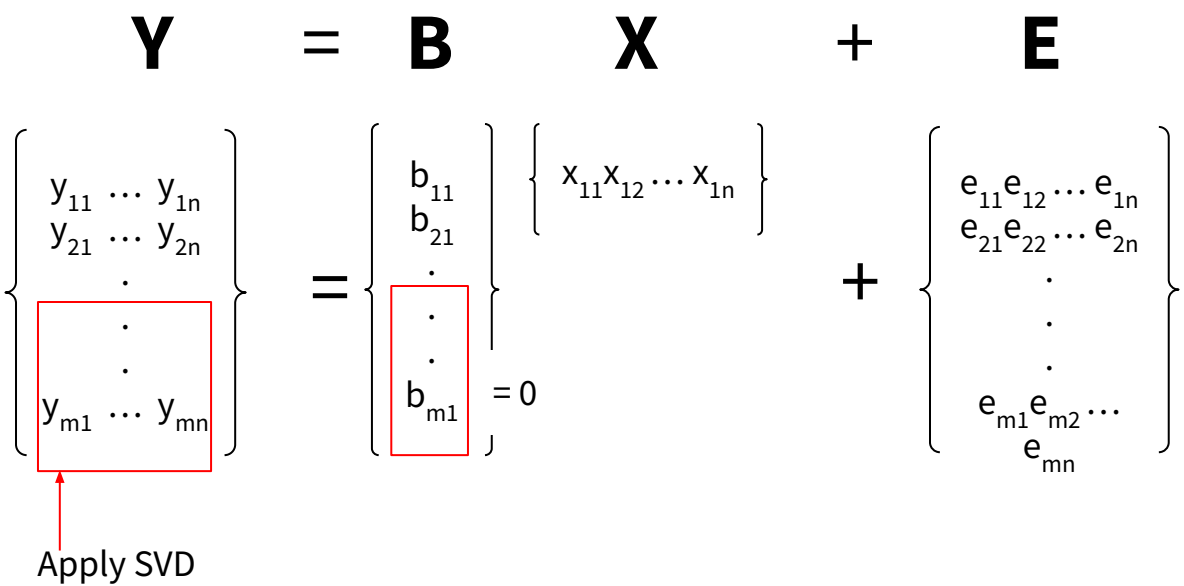
STEP 2. Find genes ($\mathbf{y}_i$) with “very small” coefficients (≈ 0)

$$\mathbf{Y} = \mathbf{B}\ \mathbf{X} + \mathbf{E}$$

$$\begin{Bmatrix} y_{11} \cdots y_{1n} \\ y_{21} \cdots y_{2n} \\ \cdot \\ \cdot \\ \cdot \\ y_{m1} \cdots y_{mn} \end{Bmatrix} = \begin{Bmatrix} b_{11} \\ b_{21} \\ \cdot \\ \cdot \\ \cdot \\ b_{m1} \end{Bmatrix} \approx 0 \quad \begin{Bmatrix} x_{11} x_{12} \cdots x_{1n} \end{Bmatrix} + \begin{Bmatrix} e_{11} e_{12} \cdots e_{1n} \\ e_{21} e_{22} \cdots e_{2n} \\ \cdot \\ \cdot \\ \cdot \\ e_{m1} e_{m2} \cdots \\ e_{mn} \end{Bmatrix}$$

These set of gene expression values are not associated with X. Therefore, any systematic variation in this subset may be associated with technical variables

STEP 3. Apply SVD on Y with ~0 coefficients



STEP 4. Taking the r Singular Vectors as r Surrogate Variables

$$\mathbf{Y} = \mathbf{B}\,\mathbf{X} + \mathbf{E}$$

$$\begin{Bmatrix} y_{11} \cdots y_{1n} \\ y_{21} \cdots y_{2n} \\ \cdot \\ \cdot \\ \cdot \\ y_{m1} \cdots y_{mn} \end{Bmatrix} = \begin{Bmatrix} b_{11} \\ b_{21} \\ \cdot \\ \cdot \\ \cdot \\ b_{m1} \end{Bmatrix} = 0 \quad \begin{Bmatrix} x_{11} x_{12} \cdots x_{1n} \end{Bmatrix} + \begin{Bmatrix} e_{11} e_{12} \cdots e_{1n} \\ e_{21} e_{22} \cdots e_{2n} \\ \cdot \\ \cdot \\ \cdot \\ e_{m1} e_{m2} \cdots \\ e_{mn} \end{Bmatrix}$$

Apply SVD

STEP n. Based on this basic procedure, Leek and others build more complex and accurate algorithms.

Now Estimated!

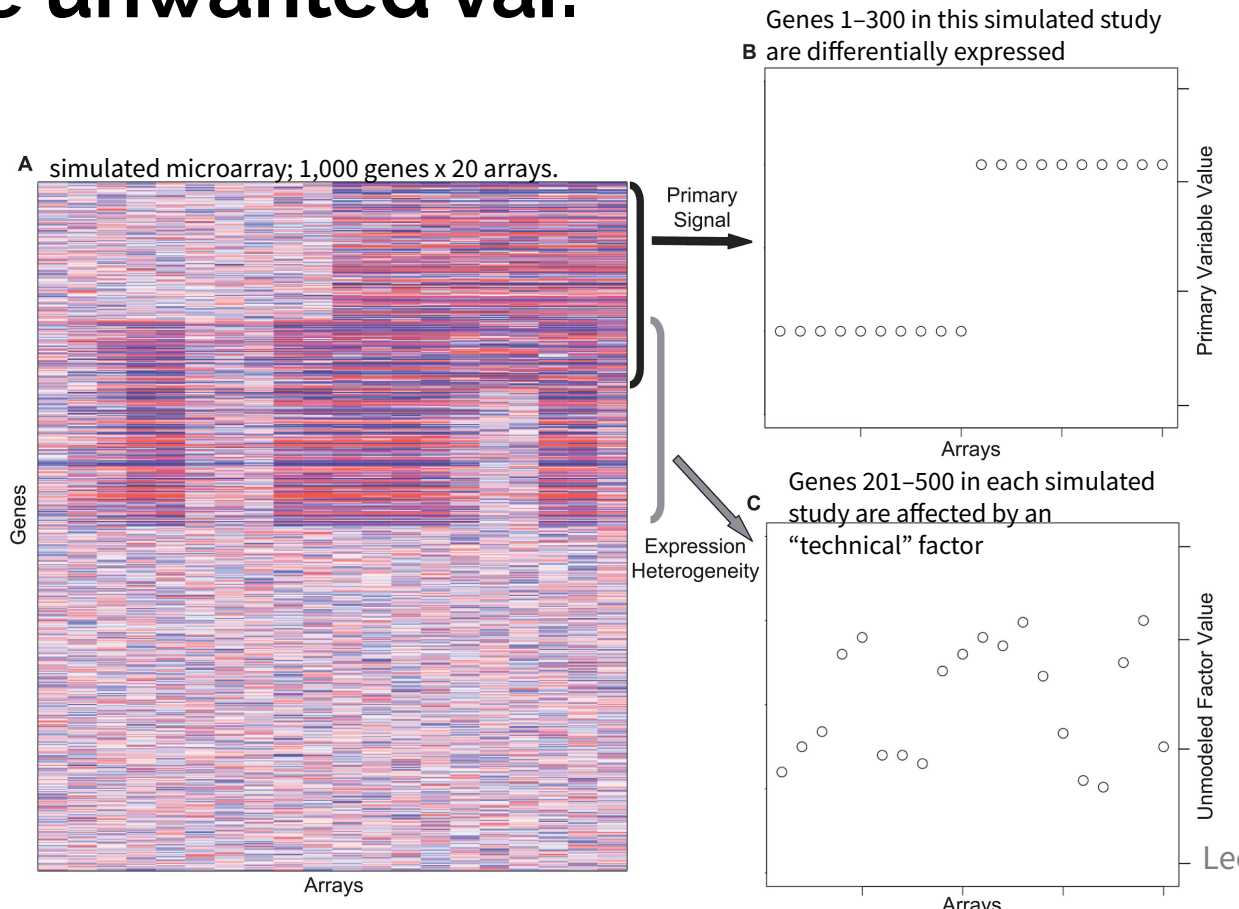
$$\mathbf{Y} = \mathbf{B}\,\mathbf{X} + \mathbf{\Gamma}\,\mathbf{G} + \mathbf{E}$$

$$\begin{Bmatrix} y_{11} \cdots y_{1n} \\ y_{21} \cdots y_{2n} \\ \vdots \\ y_{m1} \cdots y_{mn} \end{Bmatrix} = \begin{Bmatrix} b_{11} \\ b_{21} \\ \vdots \\ b_{m1} \end{Bmatrix} \begin{Bmatrix} x_{11} x_{12} \cdots x_{1n} \end{Bmatrix} + \begin{Bmatrix} \gamma_{11} \\ \gamma_{21} \\ \vdots \\ \gamma_{m1} \end{Bmatrix} \begin{Bmatrix} g_{11} g_{12} \cdots g_{1n} \end{Bmatrix} + \begin{Bmatrix} e_{11} e_{12} \cdots e_{1n} \\ e_{21} e_{22} \cdots e_{2n} \\ \vdots \\ e_{m1} e_{m2} \cdots \\ e_{mn} \end{Bmatrix}$$

Remove unwanted var.

Keep biological/relevant variation

Remove unwanted variation

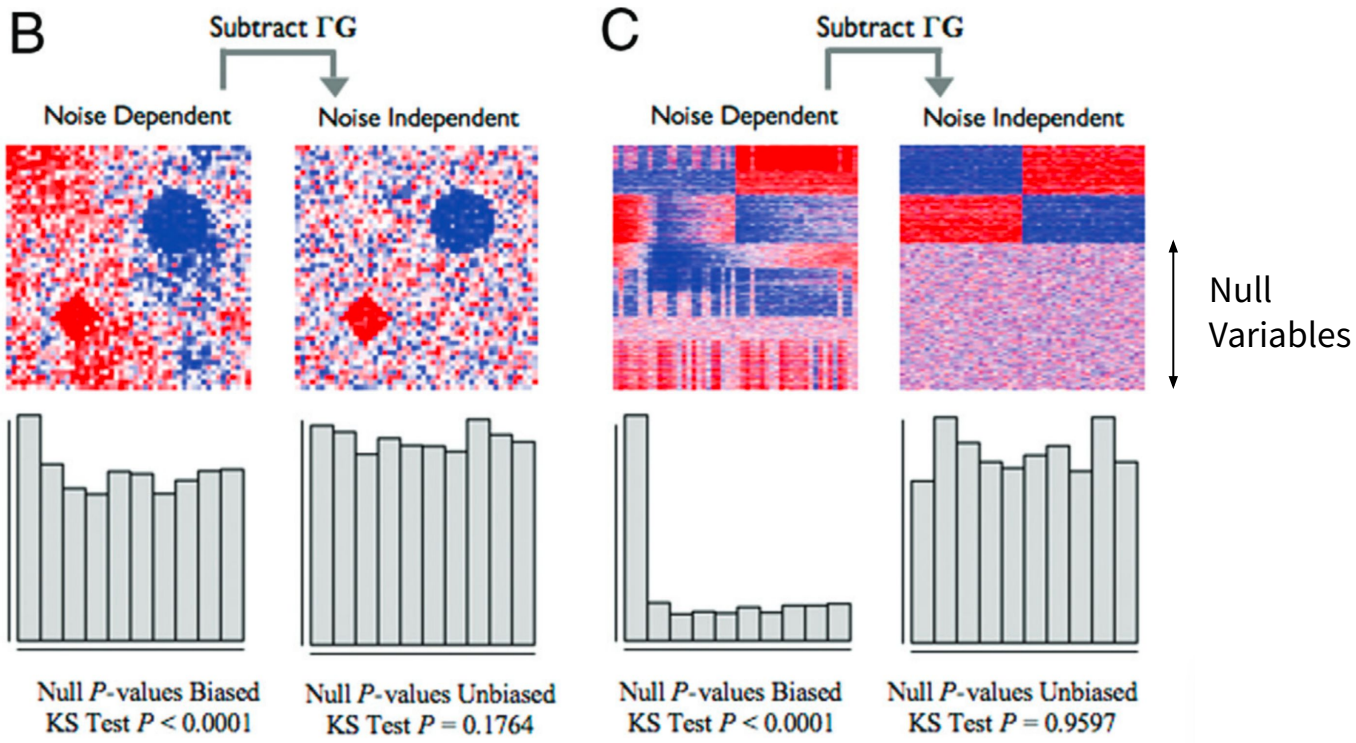


What's interesting to us! Also, what is measured.

What the dependence kernel attempts to estimate

Leek & Storey (2007)

Removing unwanted variations



Leek & Storey (2007)