

# Factor regression for dimensionality reduction and data integration techniques with applications to cancer data

Alejandra Avalos Pacheco

 @AleAviP

 avalos@hms.harvard.edu



Harvard-MIT Center  
for Regulatory Science



**Dana-Farber**  
Cancer Institute

Wirtschaftsuniversität Wien

October 16th, 2020



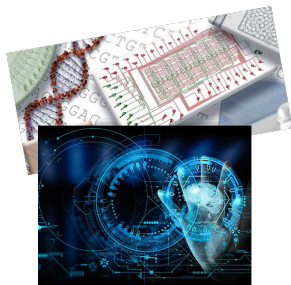
Dr. David Rossell  
UPF



Dr. Richard Savage  
Pinpoint



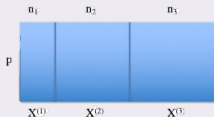
# Motivation



- New technologies enable the gathering of large datasets.
- Two main statistical challenges:
  - Volume: High dim. data  $\implies$  hard to handle and interpret.
  - Variety: Data are often not collected all at once  $\implies$  systematic biases.

## GOAL

Combine multiple studies into a single analysis.



# Solution



## Solution

A sparse latent factor regression model to integrate heterogeneous data  
Factor analysis + factor regression + sparsity + batch effect correction

### Contributions:

- 1 Showing that these issues are practically-relevant in cancer genomics.
- 2 A flexible Bayesian factor regression model to integrate large datasets, jointly learning batch and covariate effects and sparse low-rank covariances.
- 3 A novel and scalable non-local prior based formulation to induce sparsity and learn the number of factors. The first adaptation of non-local priors to factor models.
- 4 A scalable EM algorithm with closed-form updates to obtain Mode a Posteriori (MAP) estimates and an R implementation publicly available <https://github.com/AleAviP/BFR.BE>.

Cancer statistics <sup>1</sup>

## CASES

18 million

New cases of cancer,  
worldwide, 2018.

## DEATHS

9.6 million

Deaths from cancer,  
worldwide, 2018  
1 in 6 deaths

## SURVIVAL

50%

Survive cancer for 10 or  
more years  
UK, 2010-2011

## PREVENTION

38%

Preventable cases  
2015, UK

## Large scale projects:

- The Cancer Genome Atlas (TCGA),
- Cancer Genome Project (CGP)
- International Cancer Genome Consortium (ICGC)

<sup>1</sup><https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type>  
<https://www.who.int/news-room/fact-sheets/detail/cancer>

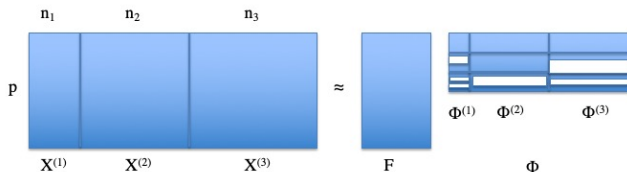


# Naïve approach

- Edefonti et al (2012) stack all the studies in **one** data-set:

$$\mathbf{x}_i^\top = \left( (\mathbf{x}_i^{(1)})^\top, (\mathbf{x}_i^{(2)})^\top, \dots, (\mathbf{x}_i^{(S)})^\top \right)$$

- Perform factor analysis





# Factor analysis

- Goal: Given  $X \in \mathbb{R}^{n \times p}$  obtain  $F \in \mathbb{R}^{n \times q}$ ,  $q \ll p$

- Model:  $\mathbf{x}_i = \phi \mathbf{f}_i + \mathbf{e}_i$

$$\mathbf{f}_i \sim N(0, I)$$

$$\mathbf{e}_i | \mathcal{T} \sim N(0, \mathcal{T}^{-1})$$

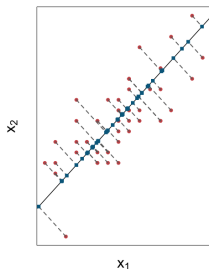
$$\mathbf{x}_i | \mathbf{f}_i, \phi, \mathcal{T} \sim N(\phi \mathbf{f}_i, \mathcal{T}^{-1})$$

$$\mathbf{x}_i | \phi, \mathcal{T} \sim N(0, \phi \phi^\top + \mathcal{T}^{-1})$$

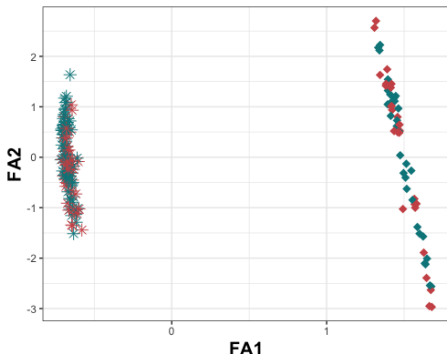
- MLE, optimise:  $\log p(X | \phi, \mathcal{T})$

$\phi$  and  $\mathcal{T}$  do not have a closed-form.

- When  $\mathcal{T}^{-1} = \sigma_\epsilon^2 I$ , we recover PPCA and PCA when  $\mathcal{T}^{-1} = 0$ .



# Ovarian cancer dataset



## Problem

Provides limited flexibility to account for systematic biases or sources of variation that are not of interest

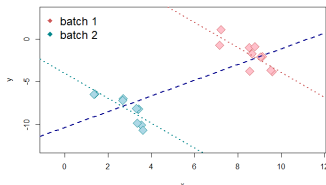


# Batch effects



Batch effects are non-biological experimental variation

- Arise when data are generated under different experimental conditions.
- Are inevitable and can lead to incorrect conclusions when combining data without adjusting for it.
- Account for a large part of the covariance and thus have a strong effect on the solution, limiting our ability to see biological patterns of interest.



BFR with batch effect correction <sup>2</sup>

Bayesian factor regression with batch effect correction:

- Model:  $\mathbf{x}_i = \phi \mathbf{f}_i + \theta \mathbf{v}_i + \beta \mathbf{b}_i + \mathbf{e}_i$ 
  - $\theta \in \mathbb{R}^{p \times p_v}$ : regression coefficients
  - $\beta \in \mathbb{R}^{p \times p_b}$ : additive batch effects
  - $\mathbf{v}_i \in \mathbb{R}^{p_v}$ : observed covariates
  - $\mathbf{b}_i \in \{0, 1\}^{p_b}$ : batch indicators
  - $\mathbf{e}_{ij} \sim N(0, \tau_{js}^{-1})$ ,  $\tau_{js}$ : the  $j^{\text{th}}$  idiosyncratic precision element in batch  $s$ .
- Priors
  - Idiosyncratic precisions:  $\tau_{jl} \mid \eta, \xi \sim \text{Gamma}(\eta/2, \eta\xi/2)$
  - Regression parameters:  $(\theta_j, \beta_j) \sim N(0, \psi I)$

<sup>2</sup>Avalos-Pacheco A. , Rossell D. , Savage R. S. , (2020+) arXiv



# Flat prior on the loadings

## Bayesian factor regression with batch effect correction model

$$\mathbf{x}_i = \phi \mathbf{f}_i + \theta \mathbf{v}_i + \beta \mathbf{b}_i + \mathbf{e}_i$$

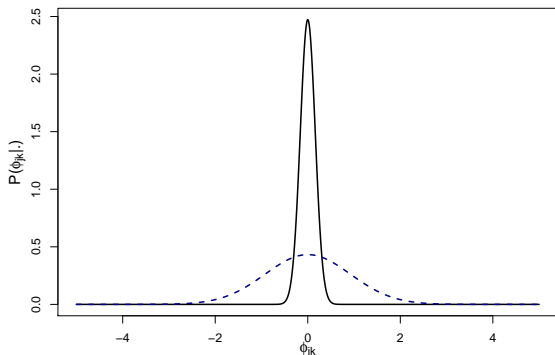
- ✓ Enables a more complete understanding of multi-study data.
- ✓ Corrects mean and variance batch effects.
- ✓ EM algorithm is able to effectively estimate and remove such biases.
- ✗ Dimension of latent factors needs to be specified.

## Spike-and-slab prior on the loadings



$$p(\phi_{jk} | \gamma, \lambda_0, \lambda_1) = (1 - \gamma_{jk})p(\phi_{jk} | \lambda_0, \gamma_{jk} = 0) + \gamma_{jk}p(\phi_{jk} | \lambda_1, \gamma_{jk} = 1)$$

★ Normal-spike-and-slab<sup>3</sup>

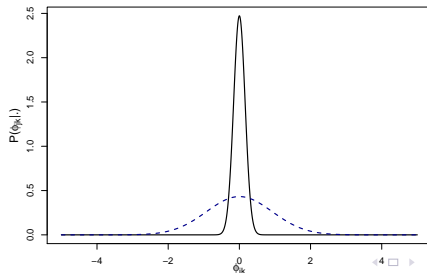


<sup>3</sup>George and McCulloch (1993) Journal of the American Statistical Association



# Normal-spike-and-slab prior model

- ✓ Enables a more complete understanding of multi-study data.
- ✓ Corrects mean and variance batch effects.
- ✓ EM algorithm is able to effectively estimate and remove such biases.
- ✓ Dimension of the latent factors is learned
- ✓ Discriminates the important (slab), from the ignorable factors (spike).
- ✗ Slab prior assigns non-negligible positive probability to regions consistent with null hypotheses.



## Spike-and-slab prior on the loadings



$$p(\phi_{jk} \mid \gamma, \lambda_0, \lambda_1) = (1 - \gamma_{jk})p(\phi_{jk} \mid \lambda_0, \gamma_{jk} = 0) + \gamma_{jk}p(\phi_{jk} \mid \lambda_1, \gamma_{jk} = 1)$$

- ★ Normal-spike-and-slab
- ★ Normal-spike-and-MOM-slab <sup>4</sup>

## Non-local priors

An absolutely continuous measure with density  $p(\phi_{jk} \mid \gamma_{jk} = 1)$  is a non-local prior if  $\lim_{\phi_{jk} \rightarrow 0} p(\phi_{jk} \mid \gamma_{jk} = 1) = 0$ .

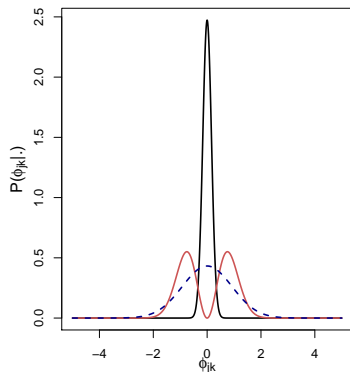
$$p(\phi_{jk} \mid \lambda_1, \gamma_{jk} = 1) = \frac{\phi_{jk}^2}{\lambda_1} N(\phi_{jk}; 0, \lambda_1)$$

<sup>4</sup>Johnson V. E., Rossell, D., (2010) Journal of the Royal Statistical Society Series B

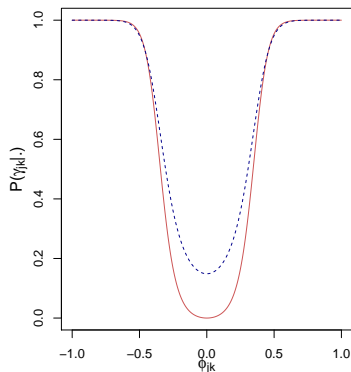
## Novel non-local spike-and-slab priors



prior densities



inclusion probabilities





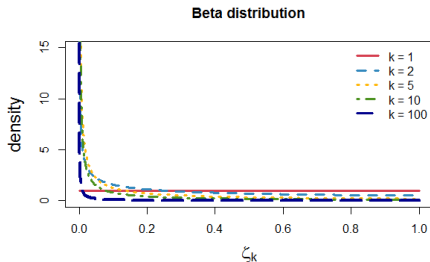
## Hyper prior on the latent indicators

Indian buffet process (IBP) prior <sup>5</sup>

$$\Gamma = \{\gamma_{jk}\}_{i,k=1}^{P,\infty}$$

$$\gamma_{jk} | \zeta_k \sim \text{Bernoulli}(\zeta_k)$$

$$\zeta_k | \alpha, 1 \sim \text{Beta}(\alpha/k, 1)$$

Inference is done via EM algorithm<sup>6</sup>, providing closed-form expressions.

<sup>5</sup>Griffiths and Ghahramani (2005) Technical report, Gatsby Computational Neuroscience Unit

<sup>6</sup>Ročková, V., George, E. I., (2016, 2014) Journal of the American Statistical Association





## Algorithm

**initialise**  $\hat{\phi} = \phi^{(0)}$ ,  $\hat{\theta} = \theta^{(0)}$ ,  $\hat{\beta} = \beta^{(0)}$ ,  $\widehat{\mathcal{T}}_{b_i} = \mathcal{T}_{b_i}^{(0)}$ ,  $\hat{\zeta} = \zeta^{(0)}$

**while**  $\epsilon > \epsilon^*$ ,  $\epsilon_{\phi} > \epsilon_{\phi}^*$  and  $t < T$

E-step:

*Latent factors:*  $\mathbb{E}[f_i | \hat{\Delta}, \mathbf{X}] = (\mathbf{I}_q + \hat{\phi}^\top \widehat{\mathcal{T}}_{b_i} \hat{\phi})^{-1} \hat{\phi}^\top \widehat{\mathcal{T}}_{b_i} (\mathbf{x}_i - \hat{\theta} \mathbf{v}_i - \hat{\beta} \mathbf{b}_i)$

*Latent indicators<sup>+</sup>:*  $\mathbb{E}[\gamma_{jk} | \hat{\Delta}] = \hat{\rho}_{jk}$

M-step:

*Loadings<sup>+</sup>:*  $\hat{\phi}_{jk} = \arg \max_{\phi_{jk}} Q_1(\hat{\Delta})$

*Variances:*  $\hat{\tau}_l^{-1} = \frac{1}{n_l + \eta - 2} \text{diag} \left\{ \sum_i: b_{il}=1 \left( \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top - 2 \tilde{\mathbf{x}}_i \mathbb{E}[f_i | \cdot]^\top \hat{\phi}^\top + \hat{\phi} \mathbb{E}[f_i f_i^\top | \cdot] \hat{\phi}^\top \right) + \eta \xi \mathbf{I}_p \right\}$

*Coefficients:*  $(\hat{\theta}_j^\top, \hat{\beta}_j^\top) = \sum_i \left[ \hat{\tau}_j^\top b_i (\mathbf{x}_{ij} - \hat{\phi}_j^\top \mathbb{E}[f_i | \cdot]) (\mathbf{v}_i, \mathbf{b}_i)^\top \right] \left[ \sum_i \left[ \hat{\tau}_j^\top b_i (\mathbf{v}_i, \mathbf{b}_i) (\mathbf{v}_i, \mathbf{b}_i)^\top \right] + \frac{1}{\psi} \mathbf{I} \right]^{-1}$

*Weights:*  $\hat{\zeta}_k = \frac{\sum_{j=1}^p \hat{\rho}_{jk} + \frac{a_{\zeta}}{k} - 1}{\frac{a_{\zeta}}{k} + b_{\zeta} + p - 1}$

set  $\Delta^{(t+1)} = \hat{\Delta}$  and  $\phi^{(t+1)} = \hat{\phi}$

compute  $\epsilon = Q(\Delta^{t+1}) - Q(\Delta^t)$ ,  $\epsilon_{\phi} = \max \|\phi_{jk}^{(t+1)} - \phi_{jk}^{(t)}\|$  and

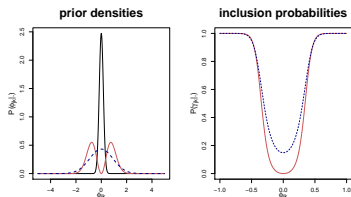
$t = t + 1$

**end**



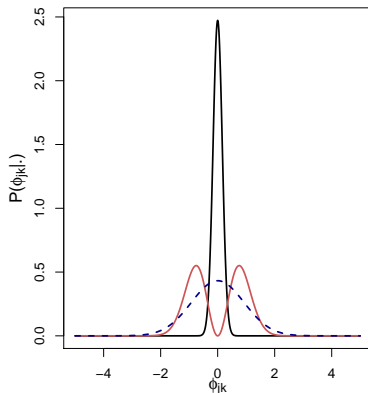
# Normal-spike-and-MoM-slab prior

- ✓ Enables a more complete understanding of multi-study data.
- ✓ Corrects mean and variance batch effects.
- ✓ EM algorithm is able to effectively estimate and remove such biases.
- ✓ Dimension of the latent factors is learned
- ✓ Discriminates the important (slab), from the ignorable factors (spike).
- ✓ Provides guidelines for the choice of  $\lambda_0$  and  $\lambda_1$
- ✓ NLPs facilitate interpretation: well-separated hypotheses.
- ✓ NLPs balance parsimony and sensitivity
- ✓ Closed-form expressions of EM available (also approximations)

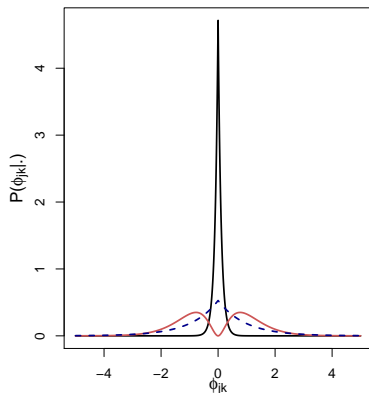


Laplace-spike-and-MoM-slab prior <sup>7</sup>

prior Normal-spike



prior Laplace-spike

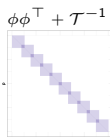
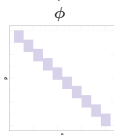


<sup>7</sup>Avalos-Pacheco A. , Rossell D. , Savage R. S. , (2020+) arXiv

## Simulation studies



Synthetic data **without** batch effects for  $n = 100$ ,  $q^* = 10$ ,  $p = 1,000$  or 1,500 parameters, truly sparse loadings  $\phi^*$ .

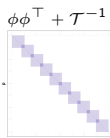


Model	$p = 1,000$				$p = 1,500$			
	$\hat{q}$	$\ \hat{\phi}\ _0$	$\ \mathbb{E}[X] - \hat{\mathbb{E}}[X]\ _F$	$\ \text{Cov}[x_i] - \widehat{\text{Cov}}[x_i]\ $	$\hat{q}$	$\ \hat{\phi}\ _0$	$\ \mathbb{E}[X] - \hat{\mathbb{E}}[X]\ _F$	$\ \text{Cov}[x_i] - \widehat{\text{Cov}}[x_i]\ $
$q = 10$								
Flat	10.0	10000.0	73.5	125.3	10.0	10000.0	89.4	203.7
Normal-SS	10.0	1298.6	43.9	89.1	10.0	1931.4	54.2	180.7
MOM-SS	10.0	1296.6	43.5	80.7	10.0	1919.3	56.2	169.4
FastBFA	9.9	778.1	60.3	165.0	9.9	1157.8	72.8	247.7
LASSO-BIC	10.0	5288.7	54.9	270.2	10.0	8414.6	67.2	408.4
$q = 100$								
Flat	100.0	100000.0	209.5	185.7	100.0	100000.0	259.2	280.2
Normal-SS	31.0	1228.6	109.0	144.6	56.4	1568.2	181.3	231.9
MOM-SS	9.7	856.8	79.4	143.3	9.2	745.4	105.0	245.6
FastBFA	83.6	1389.9	198.1	141.9	87.2	1763.9	208.2	211.3
LASSO-BIC	10.0	4787.3	54.1	271.4	10.0	7976.6	66.1	409.3

## Simulation studies

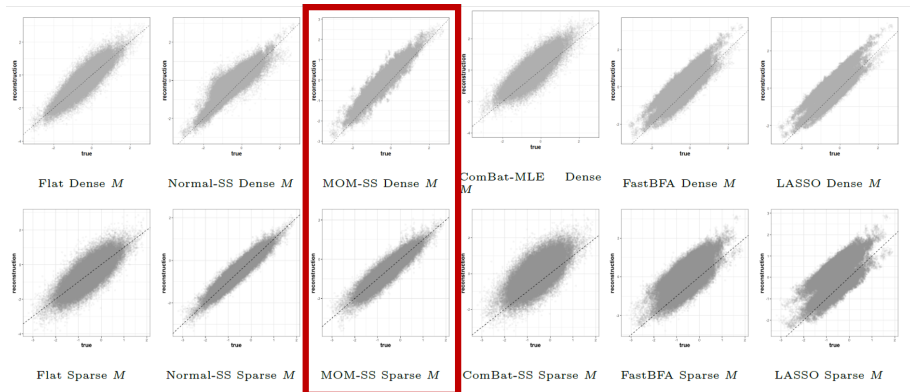


Synthetic data **with** batch effects for  $n = 200$ ,  $q^* = 10$ ,  $p = 250$  or  $500$  parameters, truly sparse loadings  $\phi^*$ .

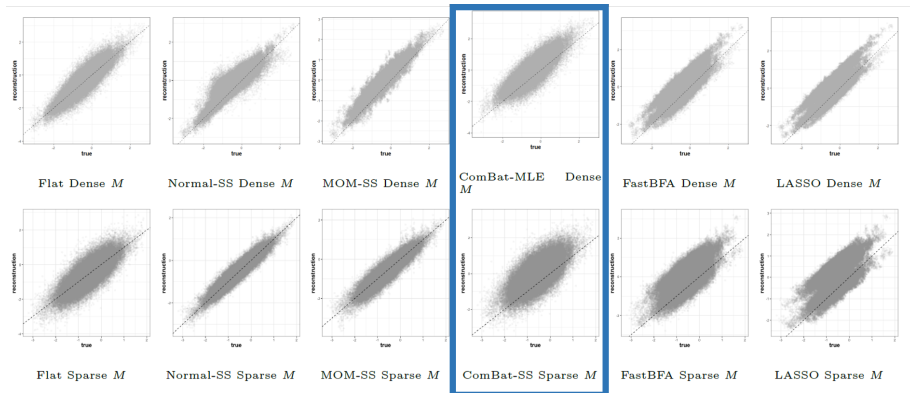


Model	$p = 250$				$p = 500$			
	$\hat{q}$	$\ \hat{\phi}\ _0$	$\ \mathbb{E}[X] - \hat{\mathbb{E}}[X]\ _F$	$\ Z\phi^T - \mathbb{E}[Z] \cdot \hat{\phi}^T\ _F$	$\hat{q}$	$\ \hat{\phi}\ _0$	$\ \mathbb{E}[X] - \hat{\mathbb{E}}[X]\ _F$	$\ Z\phi^T - \mathbb{E}[Z] \cdot \hat{\phi}^T\ _F$
$q = 10$								
Flat	10.0	2500.0	42.7	52.0	10.0	2500.0	54.8	68.2
Normal-SS	10.0	330.0	39.7	53.7	10.0	650.0	51.2	68.1
MOM-SS	10.0	330.0	39.2	61.3	10.0	650.0	49.6	86.1
ComBat-MLE	10.0	2500.0	127.2	143.3	10.0	2500.0	177.9	200.8
FastBFA	10.0	173.1	53.7	166.8	10.0	376.0	71.3	235.4
LASSO-BIC	10.0	1441.3	39.9	179.4	10.0	3159.1	50.0	254.2
$q = 100$								
Flat	100.0	25000.0	96.8	100.6	100.0	25000.0	147.8	152.5
Normal-SS	10.0	765.8	45.7	54.8	10.6	1146.3	60.0	72.6
MOM-SS	10.0	740.4	63.8	72.4	10.0	1158.7	85.7	108.3
ComBat-MLE	100.0	25000.0	169.0	182.9	100.0	25000.0	232.7	252.4
FastBFA	10.0	337.0	51.9	168.3	11.3	681.8	75.8	247.9
LASSO-BIC	10.3	1374.0	39.6	178.9	10.3	2613.9	49.8	252.1

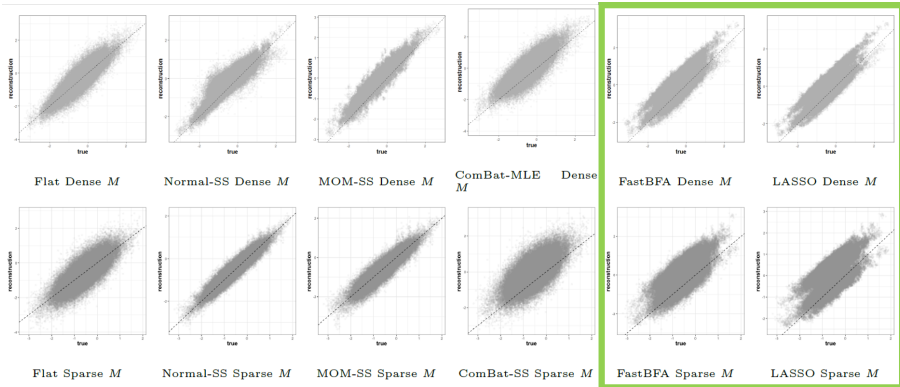
## Simulation studies



## Simulation studies



## Simulation studies



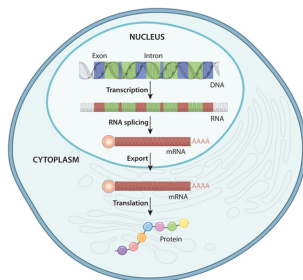


# Gene expression



## Gene expression

- It has been used as a drug discovery tool
- Key to understanding biological process such as cancer
- Useful for classifying cancer tumours into subtypes



# Cancer datasets



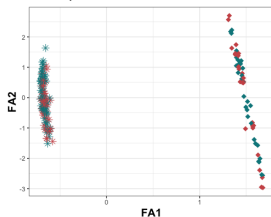
- ① Ovarian cancer: curatedOvarianData 1.16.0,  $p = 1,007$  genes
    - ① Illumina Human microRNA array E.MTAB.386,  $n_1 = 129$  patients.
    - ② GSE30161,  $n_2 = 58$  patients.
  - ② Lung cancer: TCGA2STAT 1.2,  $p = 1,198$  genes
    - ① Affymetrix Human Genome U133A 2.0 Array,  $n_1 = 133$  patients.
    - ② Affymetrix Human Exon 1.0 ST Array,  $n_2 = 112$  patients.
  - ③ Colon cancer: Gene Expression Omnibus,  $p = 172$  genes in the f-TBRS signature.
    - ① GSE17538,  $n_1 = 238$  patients.
    - ② GSE14333,  $n_2 = 101$  patients.
- Age at initial pathologic diagnosis has been used as covariate.

# Ovarian Unsupervised

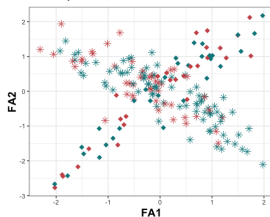


Harvard-MIT Center  
for Regulatory Science

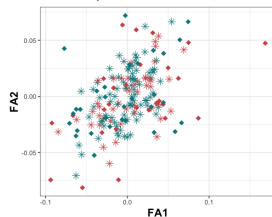
Ov, No correction



Ov, ComBat-MLE



Ov, MOM-SS

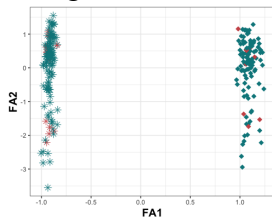


# Lung Unsupervised

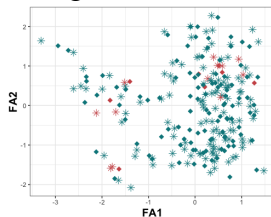


Harvard-MIT Center  
for Regulatory Science

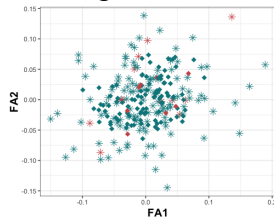
## Lung, No correction



## Lung ComBat-MLE



## Lung MOM-SS

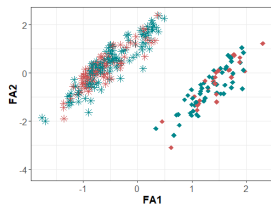


# Colon Unsupervised

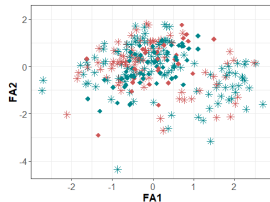


Harvard-MIT Center  
for Regulatory Science

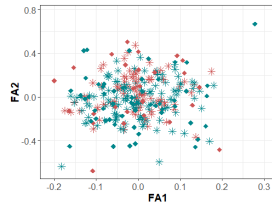
### Colon, No correction



### Colon ComBat-MLE



### Colon MOM-SS











## Supervised



Expression data of cancer datasets. Supervised analysis for ovarian ( $p = 1,007$  genes), lung ( $p = 1,198$  genes) and colon ( $p = 172$  genes) data sets.

	Ovarian			Lung			Colon		
	$\hat{q}$	$ \hat{M} _0$	Concordance index	$\hat{q}$	$ \hat{M} _0$	Concordance index	$\hat{q}$	$ \hat{M} _0$	Concordance index
Batch 1-MLE 90%	67.1	67569.7	0.618	52.1	62415.8	0.461	52.9	9081.6	0.736
Batch 1-MLE 70%	27.0	27088.3	0.632	35.2	42169.6	0.471	17.0	2924.0	0.721
Batch 2-MLE 90%	40.4	40481.4	0.522	36.6	43607.2	0.522	48.1	8256.0	0.479
Batch 2-MLE 70%	23.4	23362.4	0.524	23.2	27913.4	0.419	23.3	4007.6	0.495
Flat	100.0	100700.0	0.634	100.0	119800.0	0.669	100.0	17200.0	0.594
Normal-SS	7.8	7854.6	0.568	11.0	13178.0	0.489	7.0	1204.0	0.621
MOM-SS	4.0	4028.0	0.588	74.0	88652.0	0.665	53.4	9184.8	0.764
ComBat-MLE 90%	101.0	101707.0	0.589	79.0	94642.0	0.688	67.0	11524.0	0.738
ComBat-MLE 70%	41.0	41287.0	0.588	30.0	35940.0	0.568	24.0	4128.0	0.734
ComBat-FastBFA	100.0	100700.0	0.527	100.0	119800.0	0.707	100.0	17200.0	0.582



-  Avalos-Pacheco A., Rossell D., Savage R. S., (2020) *Heterogeneous large datasets integration using Bayesian factor regression*, Bayesian Analysis. [projecteuclid.org/euclid.ba/1600135439](https://projecteuclid.org/euclid.ba/1600135439).
-  Avalos-Pacheco A., (2019) *Factor regression for dimensionality reduction and data integration techniques with applications to cancer data*.
-  Johnson V. E., Rossell, D., (2010) *On the use of non-local prior densities in Bayesian hypothesis tests*. Journal of the Royal Statistical Society Series B 72(2),143-170.
-  Ganzfried, B. F., et. al., (2013) *Curated ovarian data: clinically annotated data for the ovarian cancer transcriptome*. Database 2013.
-  Wan, Y.-W., Allen, G. I., and Liu, Z., (2016) *TCGA2STAT: simple TCGA data access for integrated statistical analysis in R*. Bioinformatics 32(6), 952-954.
-  Calon, A., et. al., (2012) *Dependency of colorectal cancer on a TGF-beta-driven program in stromal cells for metastasis initiation*. Cancer Cell 22(5), 571-584..
-  R Packages
-  1 BMFR: <https://github.com/AleAviP/BFR.BE>