Identification of oncodomains using Bayesian False Discovery Rate

Iris Ivy M. Gauran

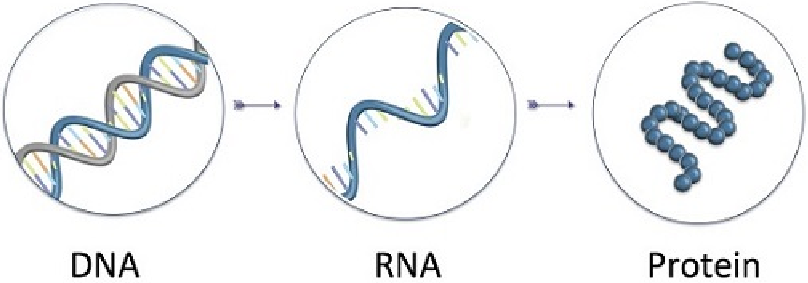
School of Statistics,
University of the Philippines

03 October 2019

Outline

Motivation: Protein Domain Analysis

Motivation: Protein Domain Analysis



Protein domains

Protein domains

- **Domain:** Unit of protein structure which evolve, function, and exist independently of the rest of the protein chain

Protein domains

- **Domain:** Unit of protein structure which evolve, function, and exist independently of the rest of the protein chain
- **Examples:**

cd00054

cd00180

Source: Domain Mapping of Disease Mutations (`http://bioinf.umbc.edu/dmdm/`)

Protein domains

- **Domain:** Unit of protein structure which evolve, function, and exist independently of the rest of the protein chain
- **Examples:**

cd00054

cd00180

Source: Domain Mapping of Disease Mutations (`http://bioinf.umbc.edu/dmdm/`)

  - Catalytic domain of protein kinases (PKs)

Protein domains

- **Domain:** Unit of protein structure which evolve, function, and exist independently of the rest of the protein chain

- **Examples:**

cd00054

cd00180

Source: Domain Mapping of Disease Mutations (`http://bioinf.umbc.edu/dmdm/`)

- ▶ Catalytic domain of protein kinases (PKs)
- ▶ Implicated in the development of various human diseases including different types of cancer

Data Description

- N positions in a domain

Data Description

- N positions in a domain
- Data: $\mathbf{a}_N = (a_1, a_2, \ldots, a_N)'$
  a_i is the number of mutations in the ith position, $i = 1, 2, \ldots N$

Data Description

- N positions in a domain
- Data: $\mathbf{a}_N = (a_1, a_2, \ldots, a_N)'$
  a_i is the number of mutations in the ith position, $i = 1, 2, \ldots N$
- Define $n_j = | \{i : a_i = j\} | =$ number of positions with j mutations $j \in \mathcal{J}$,
  $\mathcal{J} \equiv \{j : n_j > 0\}$,

Data Description

- N positions in a domain
- Data: $\mathbf{a}_N = (a_1, a_2, \ldots, a_N)'$
  a_i is the number of mutations in the ith position, $i = 1, 2, \ldots N$
- Define $n_j = | \{i : a_i = j\} | =$ number of positions with j mutations $j \in \mathcal{J}$,
  $\mathcal{J} \equiv \{j : n_j > 0\}$, $K = \max(\mathbf{a}_N)$,

Data Description

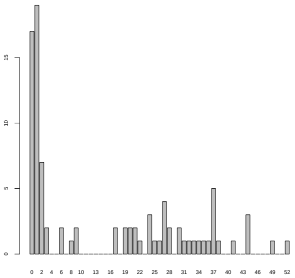
- N positions in a domain
- Data: $\mathbf{a}_N = (a_1, a_2, \ldots, a_N)'$
  a_i is the number of mutations in the ith position, $i = 1, 2, \ldots N$
- Define $n_j = \mid \{i : a_i = j\} \mid =$ number of positions with j mutations $j \in \mathcal{J}$,
  $\mathcal{J} \equiv \{j : n_j > 0\}$, $K = \max(\mathbf{a}_N)$, $\sum_{j \leq K} n_j = N$,

Data Description

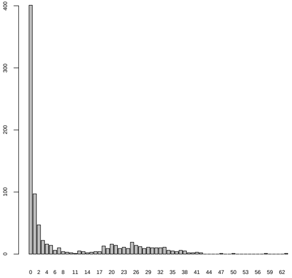
- N positions in a domain
- Data: $\mathbf{a}_N = (a_1, a_2, \ldots, a_N)'$
  a_i is the number of mutations in the ith position, $i = 1, 2, \ldots N$
- Define $n_j = \mid \{i : a_i = j\} \mid =$ number of positions with j mutations $j \in \mathcal{J}$,
  $\mathcal{J} \equiv \{j : n_j > 0\}$, $K = \max(\mathbf{a}_N)$, $\sum_{j \le K} n_j = N$, $J = |\mathcal{J}|$

Data Description

- N positions in a domain
- Data: $\mathbf{a}_N = (a_1, a_2, \ldots, a_N)'$
  a_i is the number of mutations in the ith position, $i = 1, 2, \ldots N$
- Define $n_j = \mid \{i : a_i = j\} \mid =$ number of positions with j mutations $j \in \mathcal{J}$,
  $\mathcal{J} \equiv \{j : n_j > 0\}$, $K = \max(\mathbf{a}_N)$, $\sum_{j \leq K} n_j = N$, $J = |\mathcal{J}|$



cd00054



cd00180

Overarching objective

We want to test N hypotheses of

$$\begin{aligned} H_{0i} &: \quad a_i \sim f_0 \quad \text{background mutations} \\ H_{1i} &: \quad a_i \sim f_1 \quad \text{functional (disease) mutations} \end{aligned}$$

for $i = 1, 2, \ldots, N$

while controlling a given level of Type I error such as False Discovery Rate (FDR).

False Discovery Rate

Suppose there are N hypotheses. Let

R: total number of rejections of H_{0i} (observed)

V: number of falsely rejected hypotheses among R (unobserved)

- **False Discovery Proportion (FDP):** (unobserved) proportion of false discoveries among total rejections

$$FDP = \frac{V}{R} I(R > 0)$$

- **False Discovery Rate (FDR)**

$$FDR = E(FDP) = E\left(\frac{V}{R} I(R > 0)\right)$$

FDR controlling procedures

Benjamini & Hochberg (BH) Procedure (1995, JRSS-B) For each hypotheses (H_{0i}), we have p-value, p_i.

- Order p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(N)}$

FDR controlling procedures

Benjamini & Hochberg (BH) Procedure (1995, JRSS-B) For each hypotheses (H_{0i}), we have p-value, p_i.

- Order p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(N)}$
- Reject the first $r = \max\left\{ i : p_{(i)} \leq \alpha \frac{i}{N} \right\}$ p-values.

FDR controlling procedures

Benjamini & Hochberg (BH) Procedure (1995, JRSS-B) For each hypotheses (H_{0i}), we have p-value, p_i.

- Order p-values: $p_{(1)} \le p_{(2)} \le \ldots \le p_{(N)}$
- Reject the first $r = \max\left\{i : p_{(i)} \le \alpha \frac{i}{N}\right\}$ p-values.

Some modification: Storey (2002, JRSS-B)

- Reject all hypotheses corresponding to $p_{(1)}, p_{(2)}, \ldots, p_{(\ell)}$ where

$$\ell = \max\left\{i : p_{(i)} \le \frac{\alpha i}{\widehat{\pi}_0 N}\right\}$$

FDR controlling procedures

Benjamini & Hochberg (BH) Procedure (1995, JRSS-B) For each hypotheses (H_{0i}), we have p-value, p_i.

- Order p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(N)}$
- Reject the first $r = \max\left\{i : p_{(i)} \leq \alpha\frac{i}{N}\right\}$ p-values.

Some modification: Storey (2002, JRSS-B)

- Reject all hypotheses corresponding to $p_{(1)}, p_{(2)}, \ldots, p_{(\ell)}$ where

$$\ell = \max\left\{i : p_{(i)} \leq \frac{\alpha i}{\widehat{\pi}_0 N}\right\}$$

- The BH procedure and Storey's procedure are equivalent, that is $r = \ell$, if we take $\widehat{\pi}_0 = 1$ where $\pi_0 = P(H_{0i})$.

FDR controlling procedures (cont'd)

Local FDR or Local q-value (Efron, 2004, JASA)

- Consider N gene expressions, $(z_1, \ldots, z_N)$, $\quad z_i \sim f$ where

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z)$$

FDR controlling procedures (cont'd)

Local FDR or Local q-value (Efron, 2004, JASA)

- Consider N gene expressions, $(z_1, \ldots, z_N)$, $\quad z_i \sim f$ where

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z)$$

- The local FDR (local q-value)at z_i is defined as

$$fdr(z_i) = Pr(H_{0i} \mid Z = z_i) = \frac{\pi_0 f_0(z_i)}{f(z_i)}$$

FDR controlling procedures (cont'd)

Local FDR or Local q-value (Efron, 2004, JASA)

- Consider N gene expressions, $(z_1, \ldots, z_N)$, $\quad z_i \sim f$ where

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z)$$

- The local FDR (local q-value)at z_i is defined as

$$fdr(z_i) = Pr(H_{0i} \mid Z = z_i) = \frac{\pi_0 f_0(z_i)}{f(z_i)}$$

Reject H_{0i} if $fdr(z_i) \leq \alpha$.

FDR controlling procedures (cont'd)

Local FDR or Local q-value (Efron, 2004, JASA)

- Consider N gene expressions, $(z_1, \ldots, z_N)$, $\quad z_i \sim f$ where

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z)$$

- The local FDR (local q-value)at z_i is defined as

$$fdr(z_i) = Pr(H_{0i} \mid Z = z_i) = \frac{\pi_0 f_0(z_i)}{f(z_i)}$$

Reject H_{0i} if $fdr(z_i) \leq \alpha$.
Storey's procedure : $\ell = \max\{i : Pr(H_{0i} \mid Z \geq z_i) \leq \alpha\}$.

FDR controlling procedures (cont'd)

Local FDR or Local q-value (Efron, 2004, JASA)

- Consider N gene expressions, $(z_1, \ldots, z_N)$, $\quad z_i \sim f$ where

$$f(z) = \pi_0 f_0(z) + (1 - \pi_0) f_1(z)$$

- The local FDR (local q-value)at z_i is defined as

$$fdr(z_i) = Pr(H_{0i} \mid Z = z_i) = \frac{\pi_0 f_0(z_i)}{f(z_i)}$$

  Reject H_{0i} if $fdr(z_i) \leq \alpha$.
  Storey's procedure : $\ell = \max\{i : Pr(H_{0i} \mid Z \geq z_i) \leq \alpha\}$.

- **Challenge :** Estimate of $\pi_0, f_0(z)$ (parametric form) and $f(z)$ from given data

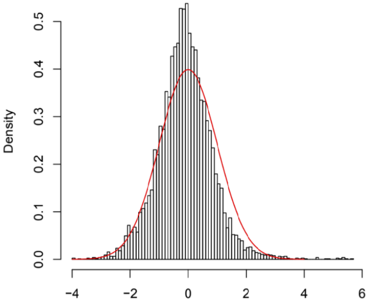
Assumption on f_0

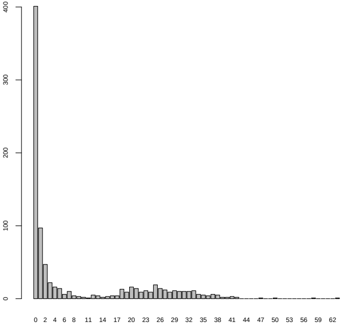
Assumption on f_0

- **Empirical Null Estimation:** f_0 is estimated based on data, rather than using some predetermined null distribution (e.g., N(0,1))

Assumption on f_0

- **Empirical Null Estimation:** f_0 is estimated based on data, rather than using some predetermined null distribution (e.g., N(0,1))
- **Zero Assumption:** most of the data ($z-$ values) near mode of f are generated from f_0
- f_0 and π_0 are estimated based on the data around the mode





Assumption on f_0 for discrete data

Zero Assumption on mutation data: Gauran et. al. (2017, Biometrics)

- The mutation count which belongs to $\mathcal{I}_0 = [0, C]$, for some unknown C, is generated from f_0, i.e., $f_1 = 0$ on $\mathcal{I}_0$.

Assumption on f_0 for discrete data

Zero Assumption on mutation data: Gauran et. al. (2017, Biometrics)

- The mutation count which belongs to $\mathcal{I}_0 = [0, C]$, for some unknown C, is generated from f_0, i.e., $f_1 = 0$ on $\mathcal{I}_0$.
- The number of mutations at position i, $a_i \sim f = \pi_0 f_0 + (1 - \pi_0) f_1$
  - If $a_i \leq C$, $f(a_i) = \pi_0 f_0(a_i)$ ($f_1(a_i) = 0$ for $a_i \leq C$) for some C.
  - If $a_i > C$, $f(a_i) = \pi_0 f_0(a_i) + (1 - \pi_0) f_1(a_i)$.

Assumption on f_0 for discrete data

Zero Assumption on mutation data: Gauran et. al. (2017, Biometrics)

- The mutation count which belongs to $\mathcal{I}_0 = [0, C]$, for some unknown C, is generated from f_0, i.e., $f_1 = 0$ on $\mathcal{I}_0$.
- The number of mutations at position i, $a_i \sim f = \pi_0 f_0 + (1 - \pi_0) f_1$
  - If $a_i \leq C$, $f(a_i) = \pi_0 f_0(a_i)$ ($f_1(a_i) = 0$ for $a_i \leq C$) for some C.
  - If $a_i > C$, $f(a_i) = \pi_0 f_0(a_i) + (1 - \pi_0) f_1(a_i)$.
- Choice of C is of paramount importance since the estimation of π_0 and f_0 depends on C.
  - Smaller values than the true value of C result in unreliable estimation of f_0.
  - Larger values result in loss of power of the testing procedure since the estimate of f_0 tends to have a heavy tail.

Choice of parametric form of f_0

Zero-inflated Generalized Poisson

- **Generalized Poisson**, $GP(\lambda, \theta)$ (Consul and Jain, 1970, Ann. Math. Stat.)

$$P(T = t) = g(t) = \frac{\lambda(\lambda + \theta t)^{t-1}}{t!} e^{-\lambda - \theta t}$$

where $|\theta| < 1$ and $\lambda > 0$ and $P(T = t) = 0$ for $t \geq m$ if $\lambda + m\theta \leq 0$.

Choice of parametric form of f_0

Zero-inflated Generalized Poisson

- **Generalized Poisson**, $GP(\lambda, \theta)$ (Consul and Jain, 1970, Ann. Math. Stat.)

$$P(T = t) = g(t) = \frac{\lambda(\lambda + \theta t)^{t-1}}{t!} e^{-\lambda - \theta t}$$

where $|\theta| < 1$ and $\lambda > 0$ and $P(T = t) = 0$ for $t \geq m$ if $\lambda + m\theta \leq 0$.

Due to large number of zero mutation counts, we use

- **Zero-inflated Generalized Poisson)**, $ZIGP(\eta, \lambda, \theta)$

$$f_0(j) = \eta\delta(0) + (1 - \eta)g(j)$$

$$f_0(j) = \begin{cases} \eta + (1 - \eta)e^{-\lambda} & j = 0 \\ (1 - \eta)g(j) & j = 1, 2, \ldots \end{cases}$$

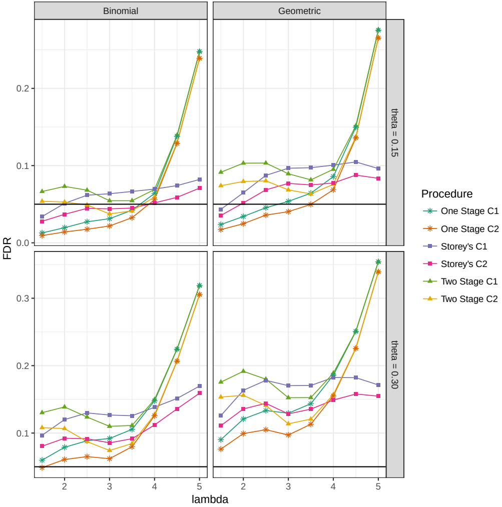
where $0 \leq \eta < 1$.

Bayesian Multiple Testing Procedures

Rationale for Bayesian Approach

- Model for f_0 is correctly specified
- As θ increases, $\widehat{\text{FDR}}$ increases
- C is underestimated in all of the scenarios
- π_0 is consequently underestimated
- violates the property

$$\pi_0 \leq \max\{\widehat{\pi}_0, E(\widehat{\pi}_0)\} < 1$$



Data

- Data: $\mathbf{a}_N = (a_1, a_2, \ldots, a_N)'$
  a_i is the number of mutations in the ith position, $i = 1, 2, \ldots N$
- Ordered data: $\mathbf{x}_N$ can be represented as a partition of the unique values of $\mathbf{a}_N$,

$$\mathbf{x}'_N = (\mathbf{x}'_0, \mathbf{x}'_1, \ldots, \mathbf{x}'_K) = (\underbrace{0, 0, \ldots, 0}_{\mathbf{x}'_0}, \underbrace{1, 1, \ldots, 1}_{\mathbf{x}'_1} \ldots, \underbrace{K, K, \ldots, K}_{\mathbf{x}'_K})$$

where $\mathbf{x}_j$ is the column vector containing n_j of js.

Model Specification

- C is integer-valued and count data are often modeled using the Poisson distribution, we consider the hierarchical model

$$C|\tau \sim \text{Poisson}(\tau) \tag{1}$$

$$\tau|\kappa_\tau, \vartheta_\tau \sim \text{Gamma}(\kappa_\tau, \vartheta_\tau) \tag{2}$$

Model Specification

- C is integer-valued and count data are often modeled using the Poisson distribution, we consider the hierarchical model

$$C|\tau \sim \text{Poisson}(\tau) \quad (1)$$

$$\tau|\kappa_\tau, \vartheta_\tau \sim \text{Gamma}(\kappa_\tau, \vartheta_\tau) \quad (2)$$

- In modeling f_0, we consider either

$$f_0 \sim \text{ZIGP}(\eta, \lambda, \theta), \qquad \text{or}$$

$$f_0 \sim \text{ZIGP}(\eta, \lambda, \theta = 0)$$

Specifications of $f(x_i \mid \phi)$

1. *Parametric Case*: $f(x_i \mid \phi) = \pi_0 f_0(x_i \mid \phi_0) + (1 - \pi_0) f_1(x_i \mid \phi_1)$, where $f_1(x_i \mid \phi_1)$ is a known parametric discrete distribution (e.g., Poisson or Generalized Poisson).

2. *Semi-parametric Case*: $f(x_i \mid \phi) = \pi_0 f_0(x_i \mid \phi_0) + (1 - \pi_0) f_1(x_i \mid \beta)$, where $f_1(x_i \mid \beta)$ is the Dirichlet distribution with concentration parameter β.

3. *Non-parametric Case*: $f(x_i \mid \phi)$ is the Dirichlet distribution with concentration parameter β.

Likelihood Function

- Split the data:

  $\mathbf{x}_n = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_C)$ for the null sample, where $n = n_0 + n_1 \ldots + n_C$ is the number of observations in the null sample

  $\mathbf{x}_{N-n} = (\mathbf{x}_{C+1}, \mathbf{x}_{C+2}, \ldots, \mathbf{x}_K)$ for the mixture of null and non-null samples

Likelihood Function

- Split the data:

  $\mathbf{x}_n = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_C)$ for the null sample, where $n = n_0 + n_1 \ldots + n_C$ is the number of observations in the null sample

  $\mathbf{x}_{N-n} = (\mathbf{x}_{C+1}, \mathbf{x}_{C+2}, \ldots, \mathbf{x}_K)$ for the mixture of null and non-null samples

- The sampling distribution for the null sample is f_0, while f is the sampling distribution of the non-null sample.

Likelihood Function

- Split the data:

  $\mathbf{x}_n = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_C)$ for the null sample, where $n = n_0 + n_1 \ldots + n_C$ is the number of observations in the null sample

  $\mathbf{x}_{N-n} = (\mathbf{x}_{C+1}, \mathbf{x}_{C+2}, \ldots, \mathbf{x}_K)$ for the mixture of null and non-null samples

- The sampling distribution for the null sample is f_0, while f is the sampling distribution of the non-null sample.

- The likelihood function for $\mathbf{x}_N$ is

$$\begin{aligned} \prod_{i \leq N} f(x_i|\phi) &= \prod_{i \leq n} \pi_0 f_0(x_i|\phi_0) \prod_{i > n} f(x_i|\phi) \\ &= \prod_{j \leq C} (\pi_0 f_0(j \mid \phi_0))^{n_j} \prod_{j > C} f(j \mid \phi)^{n_j} \qquad (3) \end{aligned}$$

Full Likelihood Function

- Define the vector of latent variables $\mathbf{z}_N = (z_1, z_2, \ldots, z_N)$ where

$$z_i = \begin{cases} 1, & x_i \sim f_0 \\ 0, & x_i \sim f_1 \end{cases}$$

Full Likelihood Function

- Define the vector of latent variables $\mathbf{z}_N = (z_1, z_2, \ldots, z_N)$ where

$$z_i = \begin{cases} 1, & x_i \sim f_0 \\ 0, & x_i \sim f_1 \end{cases}$$

- Define $n_j = n_{0j} + n_{1j}$ where n_{0j} and n_{1j} are the number of positions with $x_i = j$ mutations generated from f_0 and f_1, respectively.

$$n_{0j} = \sum_{i \leq N} z_i I(x_i = j) \quad \text{and} \quad n_{1j} = \sum_{i \leq N} (1 - z_i) I(x_i = j).$$

$$n_j = \begin{cases} n_{0j} & \text{if } j \leq C \\ n_{0j} + n_{1j} & \text{if } j > C \end{cases}$$

Full Likelihood Function

The full likelihood function for $(\mathbf{x}_N, \boldsymbol{z}_N)$ is

$$
\begin{aligned}
L(\phi \mid \mathbf{x}_N, \boldsymbol{z}_N) &= \pi_0^{\sum_{i=1}^{N} z_i} (1-\pi_0)^{N - \sum_{i=1}^{N} z_i} \prod_{i \leq N} f_0(x_i|\phi_0)^{z_i} f_1(x_i|\phi_1)^{1-z_i} \\
&= \pi_0^{\sum_{i=1}^{N} z_i} (1-\pi_0)^{N - \sum_{i=1}^{N} z_i} \prod_{j \leq C} f_0(j|\phi_0)^{n_{0j}} \prod_{j > C} f_0(j|\phi_0)^{n_{0j}} f_1(j|\phi_1)^{n_{1j}}
\end{aligned}
$$

where $\phi = (\phi_0, \phi_1, \pi_0, C, \tau)$

ϕ_0: vector of the null distribution parameters,

ϕ_1: vector of alternative distribution parameters,

π_0: proportion of observations from the null distribution,

C: cut-off for the implementation of the zero assumption, and

τ: hyperparameter of C

Choice of Prior Distributions

- Jeffrey's prior for λ, $g(\lambda) = \lambda^{-0.5}$
- Non-informative prior for π_0, η and θ

$$\begin{aligned} \pi_0 &\sim \mathcal{U}(0,1) \\ \eta &\sim \mathcal{U}(0,1) \\ \theta &\sim \mathcal{U}(0,1) \end{aligned}$$

Choice of Prior Distributions

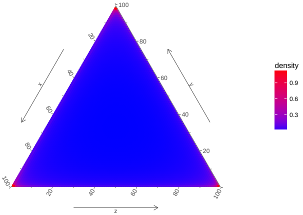
Based on specifications of $f(x_i \mid \phi)$

Non-parametric Case: $g(\boldsymbol{\beta}) \equiv \mathcal{D}(\boldsymbol{\beta})$ where

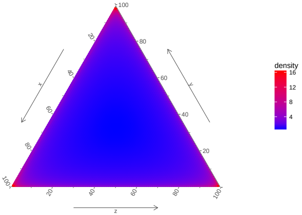
$$\boldsymbol{\beta} = (\beta, \beta \ldots, \beta),$$

P is a pre-specified value which is not data-dependent, and $P > K$.

Density Plots for Dirichlet($\beta = \beta \cdot \mathbf{1}_3$)

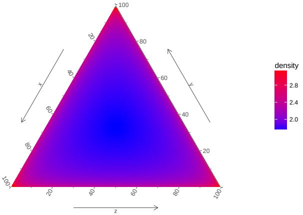


$\beta = 0.02$

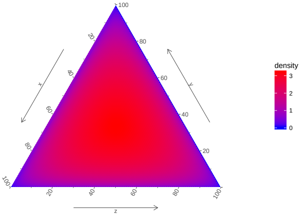


$\beta = 0.50$

Density Plots for Dirichlet($\beta = \beta \cdot \mathbf{1}_3$)



$\beta = 0.90$



$\beta = 1.50$

Adaptive MH within Gibbs sampling algorithm

1. **Initialization**:
   a. **Time instants**: Set $t = 0$ and choose the values $T_{\text{start}} < T_{\text{stop}} < T_{\text{total}}$ where T_{start} is the iteration to begin adaptation, T_{stop} is the iteration to end adaptation and T_{total} is the total number of iterations of the chain.
   b. **Proposal**: Choose the initial settings for $\phi_0^{(0)}$, $\pi_0^{(0)}$, $\mathbf{\Psi}^{(0)}$, $\tau^{(0)}$, $\mathbf{z}_N^{(0)}$ and $\mathbf{\Sigma}^{(0)}$.
2. **Gibbs step for** C: Update $C^{(t)}$ by sampling from (5).
3. **Gibbs step for** τ: Update $\tau^{(t)}$ by sampling from (6).
4. **Gibbs step for** $\mathbf{z}_N$: Update $z_i^{(t)}$ by sampling from (7), for $i = 1, 2, \ldots N$.
5. **Gibbs step for** π_0: Update $\pi_0^{(t)}$ by sampling from (8).

Adaptive MH within Gibbs sampling algorithm

6. **Metropolis-Hastings Steps**:
   a. Randomly generate $\boldsymbol{w}_t$ from ℓ_0-variate Standard Normal and let

   $$\boldsymbol{\varphi}_0{}^{(t)} = \left(\boldsymbol{\Sigma}^{(t)}\right)^{1/2} \boldsymbol{w}_t + \boldsymbol{\phi}_0{}^{(t)}.$$

   b. Accept $\boldsymbol{\phi}_0^{(t+1)} = g^{-1}(\boldsymbol{\varphi}_0^{\star})$ with probability defined in (9). Otherwise, set $\boldsymbol{\phi}_0^{(t+1)} = g^{-1}(\boldsymbol{\varphi}_0) = \boldsymbol{\phi}_0{}^{(t)}$.

Adaptive MH within Gibbs sampling algorithm

7. **Updating**: Suppose T_{thin} is the frequency with which updating occurs and T_{prop} is the proportion of previous states to include when updating. If $T_{\text{start}} < t < T_{\text{stop}}$ and and $t \equiv 0 \pmod{T_{\text{thin}}}$, identify the set of indices $\mathcal{I}$ to be used for updating.

$$\mathcal{I} = \{\lfloor t \cdot T_{\text{prop}} \rfloor, \lfloor t \cdot T_{\text{prop}} \rfloor + 1, \ldots, t\}$$

Update the parameters of the proposal covariance matrix as follows:

$$\boldsymbol{\Sigma}^{(\text{t + 1})} = \frac{1}{|\mathcal{I}|} \sum_{i=1}^{|\mathcal{I}|} \left(\phi_0^{(i)} - \overline{\phi_0}\right) \left(\phi_0^{(i)} - \overline{\phi_0}\right)^T$$

where $\overline{\phi_0} = \frac{1}{\mid \mathcal{I} \mid} \sum_{i=1}^{|\mathcal{I}|} \phi_0^{(\text{i})}$. If $t < T_{\text{total}}$, repeat from Step 6.

Adaptive MH within Gibbs sampling algorithm

8. **Gibbs step for Ψ**: Update $\Psi^{(t)}$ by sampling from (11).
9. Repeat Steps (2) to (8) for $t = 1, 2, \ldots, T$.

Local False Discovery Rate

Following the method presented by Do et al. (2005), we use the marginal posterior distribution to calculate the local false discovery rate

$$\begin{aligned} \text{fdr}(j \mid \mathbf{x}_N) &= \text{E}_{\mathbf{z}_N,\phi|\mathbf{x}_N}\left[\ \text{fdr}(j \mid \phi, \mathbf{x}_N, \boldsymbol{z}_N)\ \right] \\ &= \frac{1}{T}\sum_{t=1}^{T} \text{fdr}^{(t)}(j \mid \mathbf{x}_N, \mathbf{z}_N^{(t)}, \phi^{(t)}) \end{aligned} \tag{4}$$

for mutation counts $j = 0, 1, \ldots, K$. We reject H_{0j} if $\text{fdr}(j \mid \mathbf{x}_N) \leq \alpha = 0.05$.

False Discovery Rate and True Positive Rate

False Discovery Rate:

$$\widehat{\text{FDR}} = \frac{1}{1000}\sum_{\ell=1}^{1000} \text{FDP}_\ell = \frac{1}{1000}\sum_{\ell=1}^{1000} \frac{V_\ell}{R_\ell} I(R_\ell > 0)$$

For the ℓth generated data:

- V_ℓ: number of falsely rejected hypotheses
- R_ℓ: total number of rejected hypotheses

True Positive Rate:

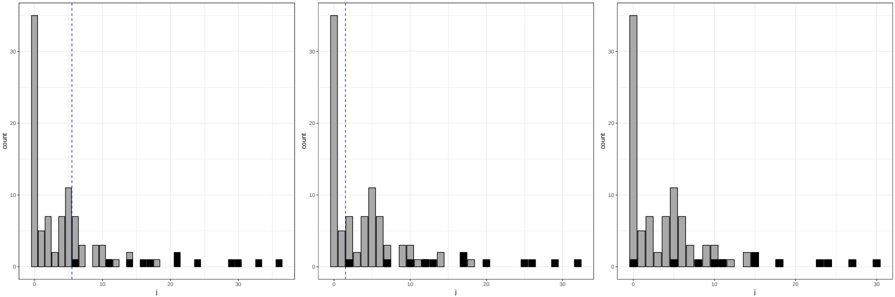
$$\widehat{\text{TPR}} = \frac{1}{1000}\sum_{\ell=1}^{1000} \left(\frac{S_\ell}{S_\ell + T_\ell} \right)$$

For the ℓth generated data:

- S_ℓ: number of correctly rejected hypotheses
- T_ℓ: number of falsely accepted hypotheses

Histograms

f_0: **ZIGP**$(\eta = 0.4, \lambda = 4, \theta = 0.3)$, f_1: **shifted Geometric,** $\pi_0 = 0.85$, $N = 100$



(a) $C = 5$ (b) $C = 1$ (c) without C

Non-parametric vs. Empirical Bayes Method

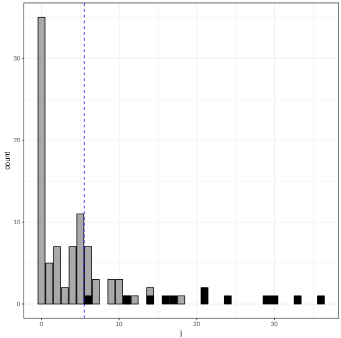
Numerical comparison when true $C = 5$ and $P = 50$

Procedure	Model for f_0: ZIGP R	$\widehat{FDR}$	$\widehat{TPR}$	Model for f_0: ZIP R	$\widehat{FDR}$	$\widehat{TPR}$
$f \sim \mathcal{D}(\boldsymbol{\beta} = 1.5 \cdot \mathbf{1}_P)$	**5.89**	**0.0286**	**0.3297**	15.17	0.1770	0.7235
	(8.51)	(0.1154)	(0.2197)	(11.65)	(0.1776)	(0.1561)
$f \sim \mathcal{D}(\boldsymbol{\beta} = 0.9 \cdot \mathbf{1}_P)$	**7.03**	**0.0413**	**0.3493**	17.49	0.2272	0.761
	(11.31)	(0.1519)	(0.2321)	(12.99)	(0.1987)	(0.159)
$f \sim \mathcal{D}(\boldsymbol{\beta} = 0.5 \cdot \mathbf{1}_P)$	7.94	0.0522	0.3844	24.18	0.3465	0.8329
	(12.06)	(0.1654)	(0.2378)	(16.45)	(0.2365)	(0.1509)
$f \sim \mathcal{D}(\boldsymbol{\beta} = 1/P \cdot \mathbf{1}_P)$	9.52	0.0760	0.4429	41.81	0.5919	0.9177
	(12.95)	(0.1927)	(0.2251)	(17.73)	(0.2256)	(0.1587)
Two-stage Procedure	12.02	0.1720	0.4678	44.31	0.6617	0.9891
	(13.03)	(0.2484)	(0.3293)	(7.12)	(0.0868)	(0.0547)
One-stage Procedure	11.94	0.1698	0.4672	30.17	0.4817	0.8987
	(12.97)	(0.2473)	(0.3285)	(13.43)	(0.2106)	(0.1541)
Storey's Procedure	12.42	0.1642	0.5929	30.16	0.4847	0.9236
	(8.98)	(0.2075)	(0.2710)	(12.17)	(0.1855)	(0.1252)

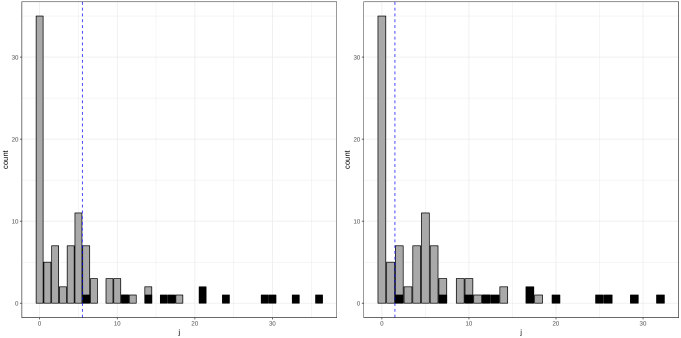
Bias of the parameter estimates

f_0: ZIGP$(\eta = 0.4, \lambda = 4, \theta = 0.3)$, $\pi_0 = 0.85$ and $C = 5$

Procedure	Model for f_0: ZIGP $\widehat{\eta}$	$\widehat{\lambda}$	$\widehat{\theta}$	$\widehat{\pi_0}$	$\widehat{C}$
$\beta = 1.5$	-0.008	1.418	-0.019	0.041	7.137
	(0.099)	(6.419)	(0.112)	(0.098)	(3.406)
$\beta = 0.9$	0.003	2.013	-0.017	0.025	6.231
	(0.119)	(8.141)	(0.114)	(0.118)	(3.545)
$\beta = 0.5$	0.014	2.300	-0.026	0.001	4.621
	(0.120)	(8.971)	(0.114)	(0.124)	(3.447)
$\beta = 1/P$	0.133	2.693	-0.021	**-0.202**	**-3.547**
	(0.113)	(10.034)	(0.098)	(0.095)	(0.794)
EB	-0.002	0.908	-0.027	**-0.045**	**-1.260**
	(0.209)	(2.359)	(0.225)	(0.198)	(1.16)



C = 5 versus C = 1



(a) $C = 5$ (b) $C = 1$

Bias	$\widehat{\eta}$	$\widehat{\lambda}$	$\widehat{\theta}$	$\widehat{\pi_0}$	$\widehat{C}$
$C = 5, \ \beta = 0.9$	0.003	2.013	-0.017	0.025	6.231
$C = 1, \ \beta = 0.9$	-0.004	2.054	-0.023	0.045	10.242

Some Remarks

Empirical Bayes method

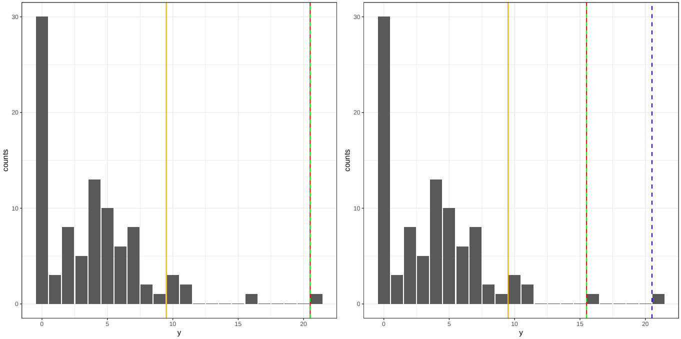
- C is underestimated using the proposed cut-off method
- π_0 is underestimated (as a consequence of the underestimation of C)
- $\widehat{\text{FDR}}$ is not controlled for any Empirical Bayes method

C = 5 versus C = 1

- $C = 1$ represents the heavily mixed case as compared to $C = 5$
- Increase in $\widehat{\text{FDR}}$ when $C = 1$
- Decrease in $\widehat{\text{TPR}}$ when $C = 1$ (for methods that control $\widehat{\text{FDR}}$)

Helix-loop-helix domain: cd00083

Data	Model for f_0: ZIGP					Model for f_0: ZIP				
	EB	NP	SP	P	GP	EB	NP	SP	P	GP
cd00083	7	1	0	0	1	7	2	1	2	2

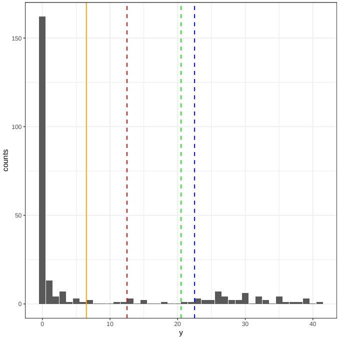


(a) Model for f_0: ZIGP (b) Model for f_0: ZIP

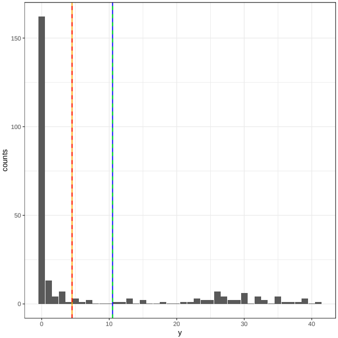
Legend: Orange: Empirical Bayes, Red: Non-parametric, Blue: Semi-parametric, Green = Parametric (Gen. Poisson)

SUSHI repeats: smart00032

Data	Model for f_0: ZIGP					Model for f_0: ZIP				
	EB	NP	SP	P	GP	EB	NP	SP	P	GP
smart00032	57	53	45	47	47	61	61	55	55	55



(a) Model for f_0: ZIGP

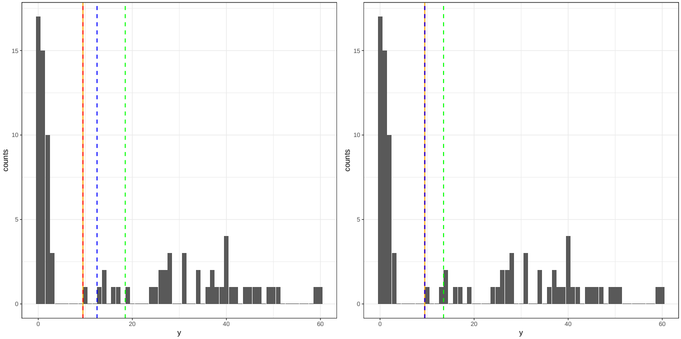


(b) Model for f_0: ZIP

Legend: Orange: Empirical Bayes, Red: Non-parametric, Blue: Semi-parametric, Green = Parametric (Gen. Poisson)

Epidermal Growth Factor domain: cd00053

Data	Model for f_0: ZIGP					Model for f_0: ZIP				
	EB	NP	SP	P	GP	EB	NP	SP	P	GP
cd00053	41	41	40	0	35	41	41	41	40	39

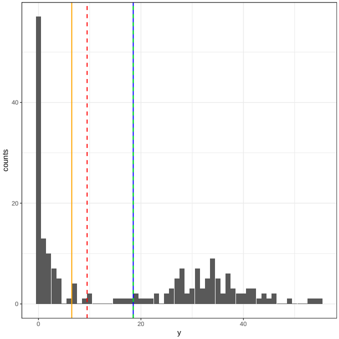


(a) Model for f_0: ZIGP (b) Model for f_0: ZIP

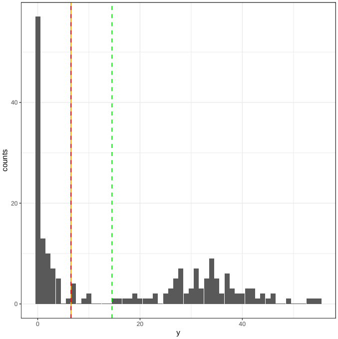
Legend: Orange: Empirical Bayes, Red: Non-parametric, Blue: Semi-parametric, Green = Parametric (Gen. Poisson)

Fibronectin Type III domain: cd00063

Data	Model for f_0: ZIGP					Model for f_0: ZIP				
	EB	NP	SP	P	GP	EB	NP	SP	P	GP
cd00063	100	95	89	90	89	100	100	99	93	93



(a) Model for f_0: ZIGP

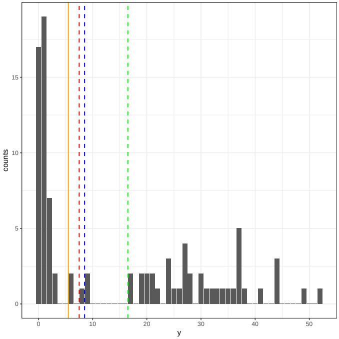


(b) Model for f_0: ZIP

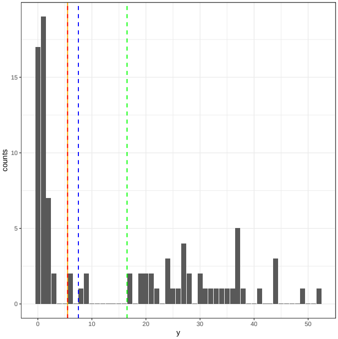
Legend: Orange: Empirical Bayes, Red: Non-parametric, Blue: Semi-parametric, Green = Parametric (Gen. Poisson)

Calcium-binding EGF-like domain: cd00054

Data	Model for f_0: ZIGP					Model for f_0: ZIP				
	EB	NP	SP	P	GP	EB	NP	SP	P	GP
cd00054	45	43	42	28	40	45	45	43	40	40



(a) Model for f_0: ZIGP



(b) Model for f_0: ZIP

Legend: Orange: Empirical Bayes, Red: Non-parametric, Blue: Semi-parametric, Green = Parametric (Gen. Poisson)

Summary and Future Work

Summary

- In general, the Empirical Bayes method leads to more rejections than any of the full Bayesian methods.
- The non-parametric method works best when f_0 is modeled using ZIGP.

Future Work

- Incorporate covariates
- Provide weights for n_j
- Explain patients not domains

References I

Do, K.-A., Müller, P., and Tang, F. (2005). A bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):627–644.

Garraway, L. A. and Lander, E. S. (2013). Lessons from the cancer genome. *Cell*, 153(1):17–37.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214.

Peterson, T. A., Nehrt, N. L., Park, D., and Kann, M. G. (2012). Incorporating molecular and functional context into the analysis and prioritization of human variants associated with cancer. *Journal of the American Medical Informatics Association*, 19(2):275–283.

Peterson, T. A., Park, D., and Kann, M. G. (2013). A protein domain-centric approach for the comparative analysis of human and yeast phenotypically relevant mutations. *BMC Genomics*, 14(3):S5.

Raim, A. M., Neerchal, N. K., and Morel, J. G. (2017). An extension of generalized linear models to finite mixture outcome distributions. *Journal of Computational and Graphical Statistics*, (just-accepted).

Stratton, M. R. (2011). Exploring the genomes of cancer cells: progress and promise. *Science*, 331(6024):1553–1558.

Conditional Posterior Distribution of C

The conditional posterior density of C given all other parameters is

$$f(C \mid \mathbf{x}_N, \boldsymbol{z}_N, \phi_0, \phi_1, \pi_0, \tau) \quad \propto \quad L(\phi \mid \mathbf{x}_N, \boldsymbol{z}_N) g(C \mid \tau) g(\tau).$$

The conditional posterior distribution of C is

$$C \mid \mathbf{x}_N, \boldsymbol{z}_N, \phi_0, \phi_1, \pi_0, \tau \sim \mathcal{M}(n = 1, \mathbf{q} = (q_0, q_1, \ldots, q_K)) \tag{5}$$

where $q_\ell, \ \ell = 0, 1, \ldots, K$ is defined as

$$q_\ell = \frac{\left\{ \prod_{j \le \ell} \{\pi_0 f_0 (j|\phi_0)\}^{n_j} \prod_{j \ge \ell+1} f(j|\phi_0, \phi_1, \pi_0)^{n_j} \right\} g(\ell \mid \tau) g(\tau)}{\sum_{\ell \le K} \left\{ \prod_{j \le \ell} \{\pi_0 f_0 (j|\phi_0)\}^{n_j} \prod_{j \ge \ell+1} f(j|\phi_0, \phi_1, \pi_0)^{n_j} \right\} g(\ell \mid \tau) g(\tau)}$$

where $\mathbf{q}^T \mathbf{1} = 1$, $g(\ell \mid \tau) = \dfrac{e^{-\tau} \tau^\ell}{\ell!}$ and $g(\tau) \equiv \mathcal{G}(\tau \mid \kappa_\tau, \vartheta_\tau)$.

Conditional Posterior Distribution of τ

The conditional posterior density of τ depends only on C, that is,

$$f(\tau \mid C) \propto g(C \mid \tau)g(\tau)$$

where $g(\tau) \equiv \mathcal{G}(\tau \mid \kappa_\tau, \vartheta_\tau)$ is the conjugate prior. The conditional posterior distribution of τ given C is then

$$\tau \mid C \sim \mathcal{G}(C + \kappa_\tau,\ \vartheta_\tau + 1). \qquad (6)$$

Conditional Posterior Distribution of z_N

The conditional posterior distribution of z_i, for any $i = 1, 2, \ldots, N$ is

$$z_i \mid \mathbf{x}_N, \phi_0, \phi_1, \pi_0, C \;\sim\; \text{Bernoulli}\,(p_i) \tag{7}$$

where

$$p_i = \max\left(I(x_i \leq C), \frac{\pi_0 f_0(x_i|\phi_0)}{f(x_i|\phi_0, \phi_1, \pi_0, C, \tau)} \right).$$

Conditional Posterior Distribution of π_0

When the information on $\mathbf{z}_N = (z_1, z_2, \dots, z_N)$ is available, we can compute

$$N_0 = \sum_{j \leq K} n_{0j} \qquad\qquad N_1 = \sum_{j \leq K} n_{1j}$$

The conditional posterior distribution of π_0 given the rest of the parameters is

$$\pi_0 \mid \mathbf{x}_N, \boldsymbol{z}_N, \phi_0, \phi_1, \pi_0, C \ \sim\ \mathcal{B}\left(N_0 + 1,\ N_1 + 1\right), \tag{8}$$

where $\mathcal{B}(a, b)$ is the Beta distribution with shape parameters a and b.

Conditional Posterior Distribution of η, λ and θ

The conditional posterior distribution of the null distribution parameters given the rest of the parameters

$$f(\phi_0 \mid \mathbf{x}_N, \boldsymbol{z}_N) \quad \propto \quad f(\mathbf{x}_N, \boldsymbol{z}_N \mid \phi_0) g(\phi_0)$$

where

$$\begin{aligned} f(\mathbf{x}_N, \boldsymbol{z}_N \mid \phi_0) \quad &\propto \quad \prod_{i \leq N} f_0(x_i|\phi_0)^{z_i} = \prod_{j \leq K} f_0(j|\phi_0)^{n_{0j}} \\ &= \quad \left[\eta + (1-\eta)e^{-\lambda}\right]^{n_{00}} \left[(1-\eta)\lambda e^{-\lambda}\right]^{\sum_{j\geq 1} n_{0j}} e^{-\theta \sum_{j \geq 1} j n_{0j}} \\ &\quad\quad \prod_{j \geq 1} \left(\frac{(\lambda + \theta j)^{j-1}}{j!}\right)^{n_{0j}} \end{aligned}$$

and $g(\phi_0) = g(\eta)g(\lambda)g(\theta) = I_{(0,1)}(\eta) \ \times \ I_{(0,1)}(\theta) \ \times \ \lambda^{-0.5} I_{(0,\infty)}(\lambda)$.

The previous expression can be reduced to the following conditional posterior densities

$$
\begin{aligned}
f(\lambda \mid \eta, \theta) &\propto \left[\eta + (1-\eta)e^{-\lambda}\right]^{n_{00}} \lambda^{-0.5+\sum_{j\geq 1} n_{0j}} e^{-\lambda \sum_{j\geq 1} n_{0j}} \prod_{j\geq 1} \left(\frac{(\lambda+\theta j)^{j-1}}{j!} \right)^{n_{0j}} \\
f(\eta \mid \lambda) &\propto \left[\eta + (1-\eta)e^{-\lambda}\right]^{n_{00}} (1-\eta)^{\sum_{j\geq 1} n_{0j}} \\
f(\theta \mid \lambda) &\propto e^{-\theta \sum_{j\geq 1} j n_{0j}} \prod_{j\geq 1} \left(\frac{(\lambda+\theta j)^{j-1}}{j!} \right)^{n_{0j}}
\end{aligned}
$$

Draws for ϕ_0

- $\phi_0 = (\eta, \lambda, \theta) \in [0,1) \times (0,\infty) \times [0,1)$

- We draw unconstrained random variables using a Metropolis-Hastings sampler and transform them to the constrained space (e.g. Raim et al. (2017)).

- Let H be a bijection from the space of ϕ_0 to the Euclidean space $\mathbb{R}^3$. The density of $\varphi_0 = H(\phi_0)$ is $f\left(H^{-1}(\varphi_0) \mid \cdot\right) \cdot \mid \det \Im(\varphi_0) \mid$ where $\Im = \partial\phi_0 / \partial\varphi_0$.

  Given $\varphi_0 = H(\phi_0)$, a proposed $\varphi_0^\star$ will be accepted with probability

$$\min\left\{ 1,\ \frac{f\left(H^{-1}(\varphi_0^\star) \mid \cdot\right) \mid \det \Im(\varphi_0^\star) \mid}{f\left(H^{-1}(\varphi_0) \mid \cdot\right) \mid \det \Im(\varphi_0) \mid} \right\} \tag{9}$$

Non-parametric Bayesian False Discovery Rate

Suppose that for a given value of C, $f(j)$ has the probability

$$\boldsymbol{\Psi} = (\psi_0, \ldots, \psi_C, \psi_{C+1}, \ldots, \psi_P).$$

The prior distribution of $\boldsymbol{\Psi}$ is $\mathcal{D}(\beta)$. The posterior distribution of $\boldsymbol{\Psi}$ is

$$\boldsymbol{\Psi} \mid (\mathbf{x}_N, \mathbf{z}_N, \beta, C) \sim \mathcal{D}(\beta_0, \beta_1, \ldots, \beta_K, \beta_{K+1}, \ldots, \beta_P) \tag{10}$$

where

$$\beta_j = \begin{cases} \beta + n_{0j} + n_{1j}, & j \le K \\ \beta, & j > K \end{cases}$$

for $j = 0, 1, 2, \ldots, P$ and $\sum_{j \le P} \psi_j = 1$.

Some Remarks

- Estimation procedure for f:

$$\widehat{\boldsymbol{f}} = \left(\frac{n_0}{N}, \frac{n_1}{N}, \ldots, \frac{n_C}{N}, \frac{n_{C+1}}{N}, \ldots, \frac{n_K}{N}\right) \qquad \text{where} \qquad \sum_{j \leq K} f(j) = 1$$

- Zero assumption on f_0:

$$\widehat{\boldsymbol{f}} = \left(\frac{n_0}{N} \approx \widehat{\pi}_0 \widehat{f}_0(0), \frac{n_1}{N} \approx \widehat{\pi}_0 \widehat{f}_0(1), \ldots, \frac{n_C}{N} \approx \widehat{\pi}_0 \widehat{f}_0(C), \frac{n_{C+1}}{N}, \ldots, \frac{n_K}{N}\right)$$

- When N is small, $\widehat{\boldsymbol{f}}$ would display sparsity wherein many cells have zero probability.
- When P is large relative to the maximum number of mutations, we allocate probabilities to cells without data points.

Non-parametric Bayesian False Discovery Rate

Implementation of Zero Assumption

- Instead, we sample from the conditional distribution $\boldsymbol{\Psi}_{(1)} \mid \boldsymbol{\Psi}_{(0)} = \boldsymbol{\psi}_{(0)}$, where $\boldsymbol{\Psi}_{(0)} = (\psi_0, \psi_1, \ldots, \psi_C)$ and $\boldsymbol{\Psi}_{(1)} = (\psi_{C+1}, \psi_{C+2}, \ldots, \psi_P)$. The (unnormalized) conditional density of $\boldsymbol{\Psi}_{(1)} \mid \boldsymbol{\Psi}_{(0)} = \boldsymbol{\psi}_{(0)}$ is given by

$$\prod_{j>C} \left[\psi_j \left(1 - \alpha_0\right)^{-1}\right]^{\beta_j - 1}$$

which indicates that $(1 - \alpha_0)^{-1} \boldsymbol{\Psi}_{(1)} \mid \boldsymbol{\Psi}_{(0)} \sim \mathcal{D}(\beta_{C+1}, \ldots, \beta_P)$, where $\alpha_0 = \sum_{k \leq C} \psi_k$ and $\psi_j = \pi_0 f_0(j)$ for $j \leq C$.

- Equivalently,

$$\boldsymbol{\Psi}_{(1)} \mid \boldsymbol{\Psi}_{(0)} \sim (1 - \alpha_0)\, \mathcal{D}(\beta_{C+1}, \ldots, \beta_P) \tag{11}$$