Mixture models

Manuel A. Rivas
Department of Biomedical Data Science
BMI217
Stanford University
rivaslab.stanford.edu

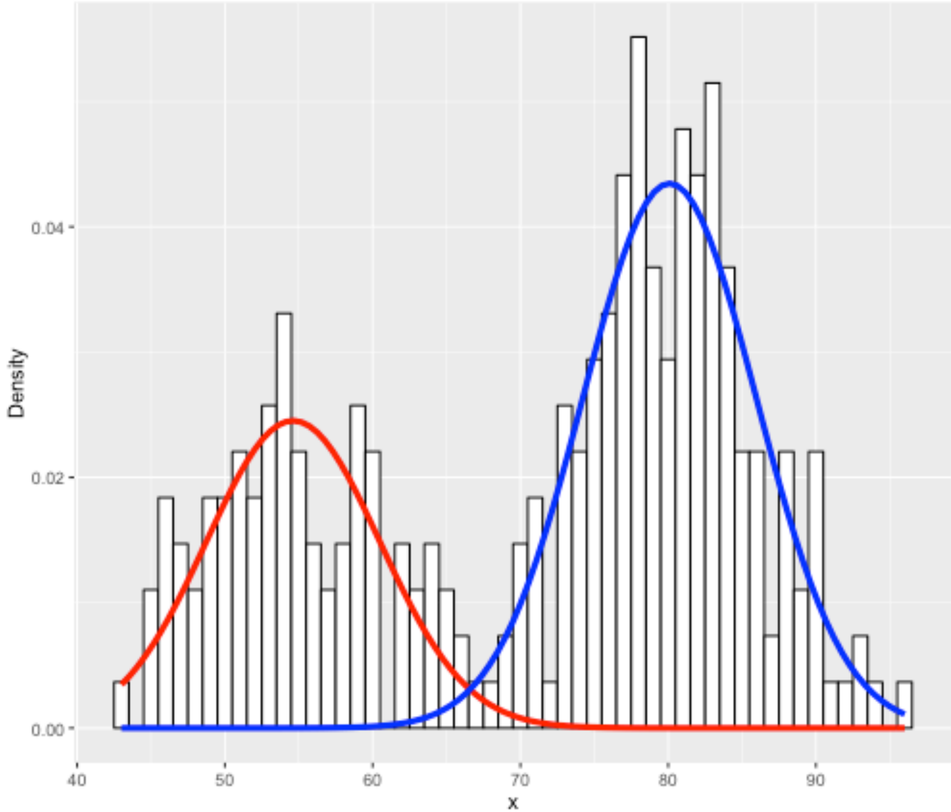
RIVASLAB

Motivating examples for today

- Rare variants and disruption on splicing
- Rare variants and enrichment across populations

Mixture models

of an outcome assume that the outcome is drawn from one of several distributions



RIVASLAB

Categorical mixing distribution

The identity of which is controlled by categorical mixing distribution

Parameterize a mixture model

One way to parameterize a mixture model with a latent categorical variable indicating which mixture component was responsible for the outcome

Example: *K* Normal Distributions

with location $\mu_k \in \mathbb{R}$

scales $\sigma_k \in (0, \infty)$

Example: K Normal Distributions

Consider mixing them in proportion π, where $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$.

Example: *K* Normal Distributions

For each outcome y_n there is a latent variable γ_n in $\{1, \ldots, K\}$

Example: K Normal Distributions

with a categorical distribution parameterized by π

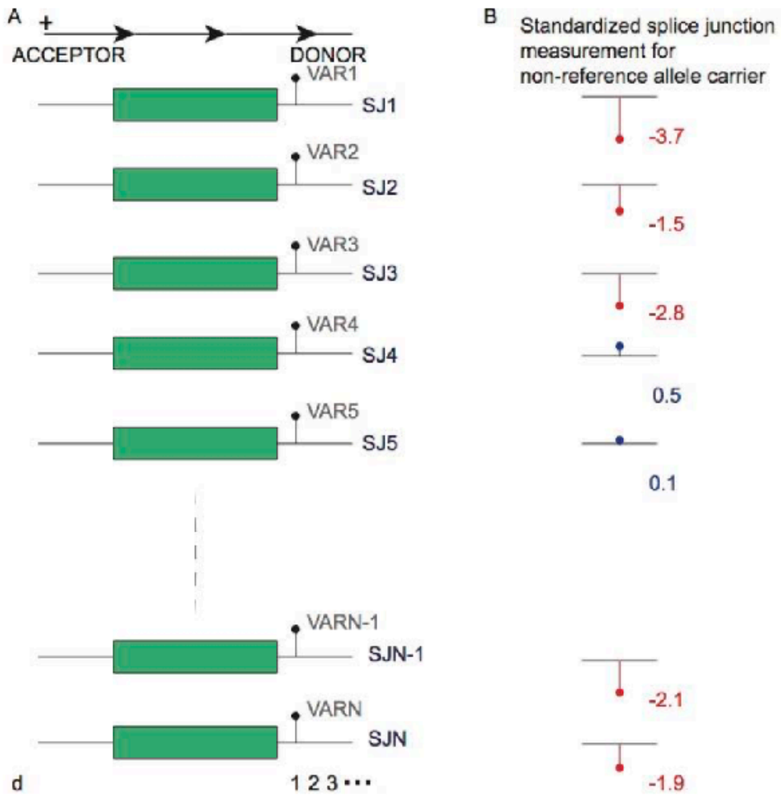
$$\gamma_n \sim \text{Categorical}\,(\pi)$$

Example: K Normal Distributions

The variable y_n is distributed according to the parameters of the mixture component

$$y_n \sim \mathcal{N}\left(\mu_{\gamma_n}, \sigma_{\gamma_n}\right)$$

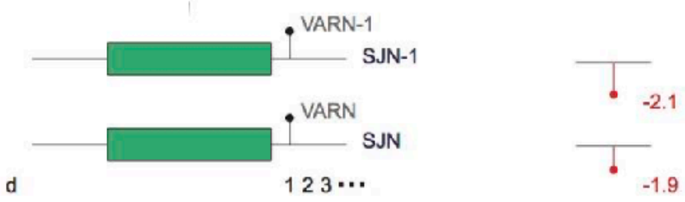
Rare variants and splicing disruption



Rare variants and splicing disruption

> We are interested in the shift of the distribution of splice junction quantification value for carriers of the alternate allele for genetic variants proximal to a splice-junction compared to noncarriers, as a function of distance from splice junction.
>
> Using the population values of reads spanning annotated splice junctions and the standardized trait value of the PTV carriers, we estimate, at each distance, the proportion of carriers belonging to

1. the no splice disruption group (0) described by the standard normal distribution, or
2. the splice disruption group (1) with alternative shift in mean μ.



Rare variants and splicing disruption: Splice Disruption Model (SDM)

$$\pi \sim \text{Beta}\,(1, 1)$$

$$\gamma_k | \pi \sim \text{Ber}\,(\pi)$$

$$\mu_0 = 0$$

$$\mu_1 \sim \mathcal{N}\,(-1, 1)$$

$$\sigma_0^2 = 1$$

$$\sigma_1^2 \sim IG\,(1, 1)$$

$$y_k | \gamma_k, \mu_0, \mu_1 \sim \mathcal{N}\left(y_k; \mu_{\gamma_k}, \sigma_{\gamma_k}^2\right)$$

π

$$\pi \sim \text{Beta}\,(1, 1)$$

Our prior for π is a uniform distribution on the interval (0, 1) (also known as Beta(1, 1) distribution)

implying that we are not favoring a priori any possible value of π over the others.

Data drawn?

The trait values have been standardized jointly and we model the trait values of the PTV carriers as either drawn from the general population distribution,

$$\mathcal{N}(\mu_0 = 0, 0 = 1),$$

or from a PTV specific normal distribution with unknown shift in mean μ_1 and unknown variance σ_1^2.

μ_1

$$\mu_1 \sim \mathcal{N}(-1, 1)$$

We model the mean μ_1 by a normal distribution with mean -1 and variance 1 to reflect our interest in those variants that decrease splicing efficiency. In principle, we could also have used another component to reflect variants with a putative increase in splicing efficiency.

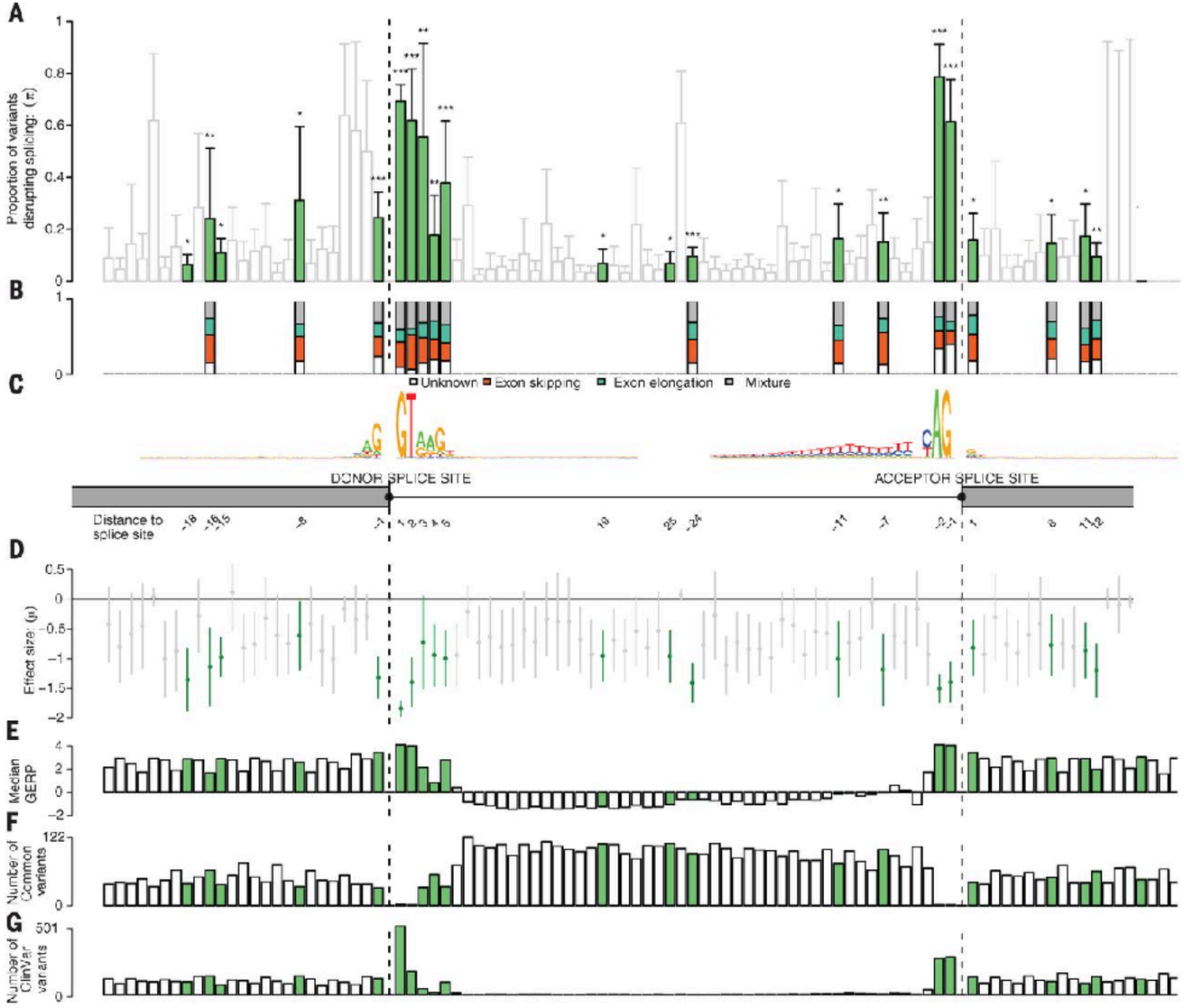
σ_1^2

$$\sigma_1^2 \sim IG(1, 1)$$

The prior for the variance parameter σ_1^2 is the inverse gamma distribution with parameters $\alpha = 1$ and $\beta = 1$.

Computation: Gibbs sampler

Inference: Rare variants and splice disruption



RIVASLAB

Rare variants and enrichment across populations

To estimate which alleles are enriched in a population isolate (e.g. Finland, Ashkenazi Jewish) compared to alleles in reference population groups in ExAC. This model estimates the proportion of observed alleles in the population isolate that are enriched or depleted as well as the shift in distribution of those alleles relative to the reference populations.

Population Isolate Enrichment Mixture Model (PIEMM)

Data input

Using the number of alternate and reference alleles observed in AJ non-IBD samples and in the population (NFE, AFR or AMR) with the highest frequency from ExAC we compute a bias corrected log odds ratio estimate, $\hat{\beta}_i$, and its standard error, $\hat{SE}_i$.

Precisely, $\hat{\beta}_i$ is the estimate of the log of the odds ratio of finding the alternate allele in AJ vs in the ExAC population with the highest allele frequency.

Model

We use these summary statistics to assign the alleles into two groups: "not enriched" and "enriched". For technical reasons described below we do this by a mixture model that has three components:

```
(0) 'similar' - AJ allele has similar frequency to reference populations,

(1) 'enriched' - AJ allele has shifted increased frequency to reference populations,

(2) 'depleted' - AJ allele has shifted decreased frequency to reference populations; and
```

Model

We estimate

i) which proportion of alleles observed in AJ belong to each of the three components, and

ii) the extent to which the enriched group is shifted in frequency from the reference populations.

PIEMM

$$\pi \sim \text{Dirichlet}(1, 1, 1)$$

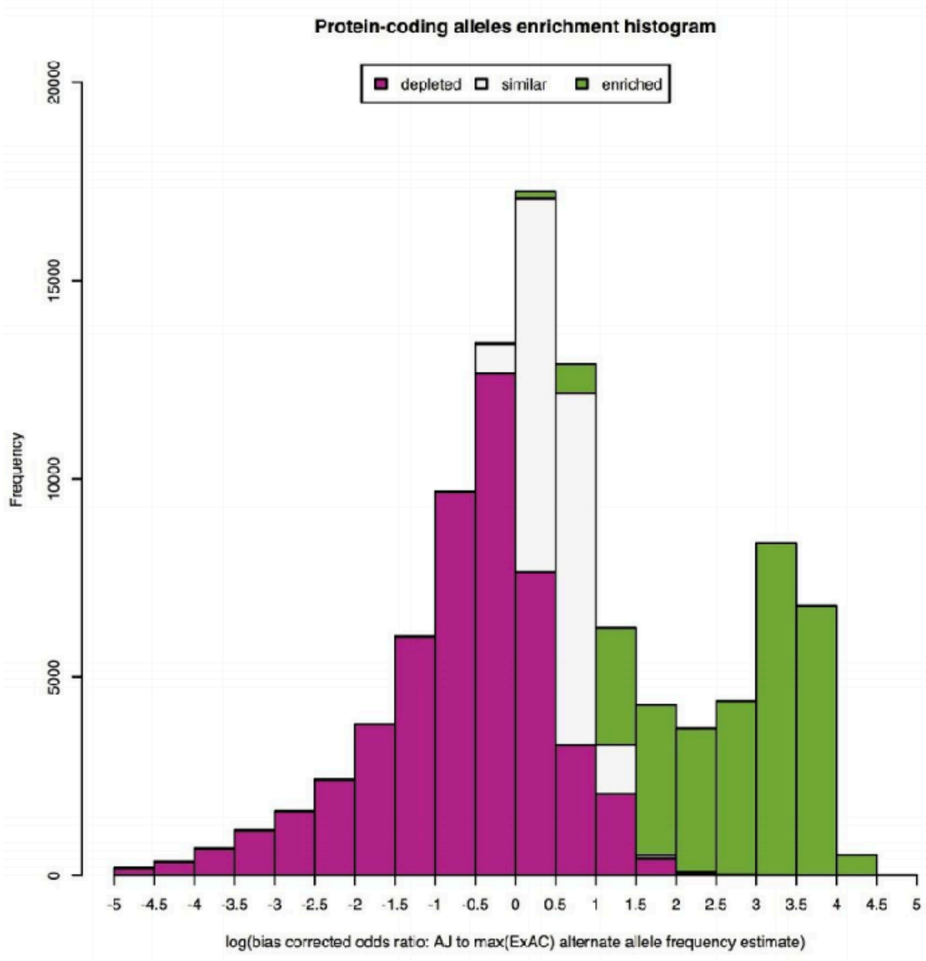
$$\gamma_i | \pi \sim \text{Multinomial}(1, \pi)$$

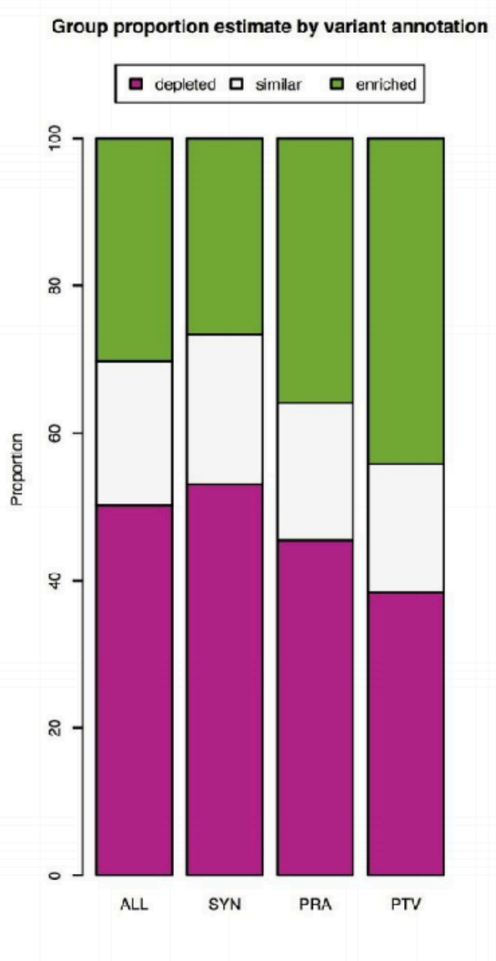
$$\mu_0 = 0$$
$$\mu_1 \sim \mathcal{N}\left(2, 0.5^2\right)$$
$$\mu_2 \sim \mathcal{N}\left(-2, 0.5^2\right)$$

$$\sigma_0^2 = 0.2^2$$
$$\sigma_1^2 = 0.5^2$$
$$\sigma_2^2 = 0.5^2$$

$$\hat{\beta}_i | \gamma_i, \mu_{\gamma_i}, \sigma_{\gamma_i}^2, \hat{\text{SE}}_i^{\,2} \sim \mathcal{N}\left(\hat{\beta}_i; \mu_{\gamma_i}, \sigma_{\gamma_i}^2 + \hat{\text{SE}}_i^{\,2}\right).$$

Enrichment of alleles discovered in AJ exome sequencing project





RIVASLAB

Enrichment of alleles discovered in AJ exome sequencing project

Variant	HGVS	Gene	Enrichment	AJ AF	max EXAC AF	Curated Traits	Inheritance
16:3293310_A/G	p.Val726Ala	MEFV	13.31	0.0416	0.0033	Familial Mediterranean fever	AR
1:155205634_T/C	p.Asn409Ser	GBA	8.40	0.0296	0.0036	Susceptibility to Lewy body dementia, Gauch	AR
4:187201412_T/C	p.Phe301Leu	F11	19.29	0.0273	0.0015	Hereditary factor XI deficiency	AR
13:20763553_CA/C	p.Leu56Argfs	GJB2	17.80	0.0199	0.0011	Autosomal recessive deafness	AR
4:187195347_G/T	p.Glu135Ter	F11	15.61	0.0195	0.0013	Factor XI deficiency	AR
12:11421038_G/A	p.Arg49Cys	PRB3	10.10	0.0189	0.0019	Salivary peroxidase	AR
9:111662096_A/G	c.2204+6T>C	IKBKAP	14.37	0.0168	0.0012	Familial dysautonomia	AR
1:215848678_C/T	p.Arg4192His	USH2A	13.06	0.0106	0.0008	Retinitis pigmentosa	AR
10:99371368_TGAG/T	p.Glu315del	HOGA1	14.05	0.0101	0.0007	Primary hyperoxaluria	AR
15:72638920_G/GGATA	p.Tyr427IlefsTer5	HEXA	11.60	0.0105	0.00106	Tay-Sachs disease	AR
7:117282620_G/A	p.Trp1282Ter	CFTR	12.67	0.0085	0.0007	Cystic fibrosis	AR

Gibbs sampler DEMO

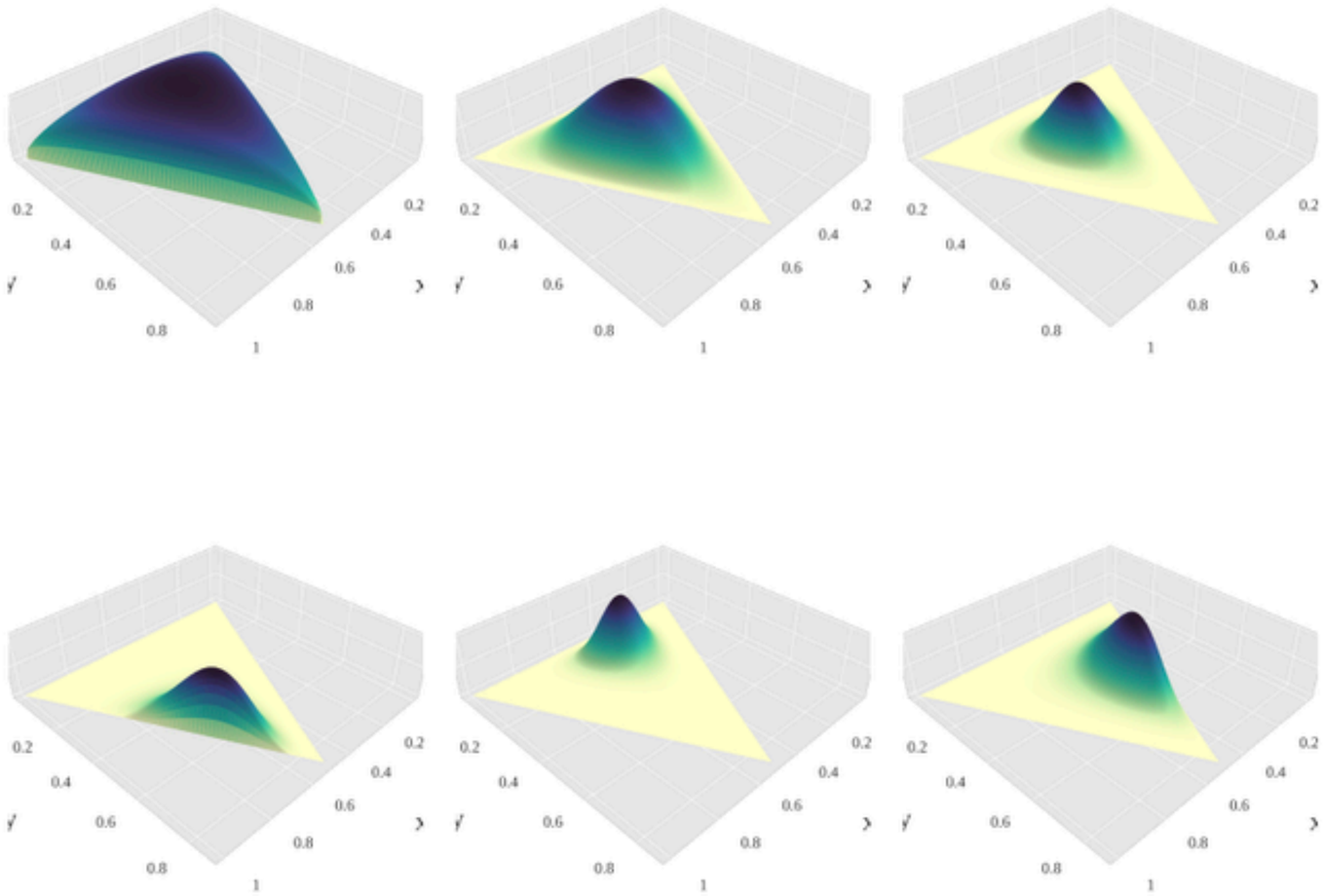
DEMO of Gibbs sampler

Gibbs sampler DEMO

Motivation of parameters and distribution: The group membership of each AJ protein-coding allele is unknown in advance. As a result the proportion π of the alleles belonging to the similar, enriched, or the depleted group (characterized by an unknown shift in mean μ and variance σ^2) needs to be estimated. Our prior for π is the uniform distribution on the 2d simplex (also known as the $\mathrm{Dirichlet}(1,1,1)$ distribution) so as to not favor *a priori* any particular value of π .

Gibbs sampler DEMO

Motivation of par
unknown in advanc
depleted group (ch
for π is the unifor
not favor *a priori* a



coding allele is
enriched, or the
be estimated. Our prior
distribution) so as to

Gibbs sampler DEMO

MCMC algorithm: We use a Gibbs sampler, an approximation algorithm, to approximate the posterior distribution of the parameters of the PIEMM. Superscripts for the variables denote their value after the corresponding iteration. Let l be the group index whenever it is not explicitly given.

Initialization

1. Initialize $\pi^{(0)}, \mu_1^{(0)}, \mu_2^{(0)}, \left(\sigma_1^2\right)^{(0)}, \left(\sigma_2^2\right)^{(0)}$, and $\gamma_i^{(0)}$ for all i.

Update

2. Repeat for $t = 1, 2, \ldots, n_{\text{burn}} + n_{\text{iter}}$

   (a) For $i = 1, 2, \ldots, n_{\text{alleles}}$, generate $\gamma_i^{(t)} \sim \text{Multinomial}\left(1, p_i^{(t)}\right)$ where

$$p_{il}^{(t)} = \frac{\pi_l^{(t-1)} \mathcal{N}\left(\hat{\beta}_i; \mu_l^{(t-1)}, \left(\sigma_l^2\right)^{(t-1)} + \hat{\text{SE}}_i^{\,2}\right)}{\pi_0^{(t-1)} \mathcal{N}\left(\hat{\beta}_i; 0, \hat{\text{SE}}_i^{\,2}\right) + \pi_1^{(t-1)} \mathcal{N}\left(\hat{\beta}_i; \mu_1^{(t-1)}, \left(\sigma_1^2\right)^{(t-1)} + \hat{\text{SE}}_i^{\,2}\right) + \pi_2^{(t-1)} \mathcal{N}\left(\hat{\beta}_i; \mu_2^{(t-1)}, \left(\sigma_2^2\right)^{(t-1)} + \hat{\text{SE}}_i^{\,2}\right)}, \text{ for } l = 0, 1, 2.$$

   (b) Generate $\pi^{(t)} \sim \text{Dirichlet}\left(1 + n_0, 1 + n_1, 1 + n_2\right)$, where $n_l = \sum_i I\left(\gamma_i^{(t)} = l\right)$, for $l = 0, 1, 2$.

   (c) Update:

$$(\sigma_l^2)^{(t)} \sim \text{IG}\left(1 + 1/2 \sum_i I\left(\gamma_i^{(t)} = l\right), 1 + 1/2 \sum_i I\left(\gamma_i^{(t)} = l\right)\left(\hat{\beta}_i - \mu_l^{(t-1)}\right)^2\right), \text{ and}$$

$\mu_l^{(t)} \sim \mathcal{N}(a, b)$, where

References

1. Rivas, Manuel A., et al. "Effect of predicted protein-truncating genetic variants on the human transcriptome." Science 348.6235 (2015): 666-669.
2. Rivas, Manuel A, et al. "Insights into the genetic epidemiology of Crohn's and rare diseases in the Ashkenazi Jewish population." bioRxiv (2016): 077180