Statistical Divergence

And application to discrete probability distributions of mutation in cancer

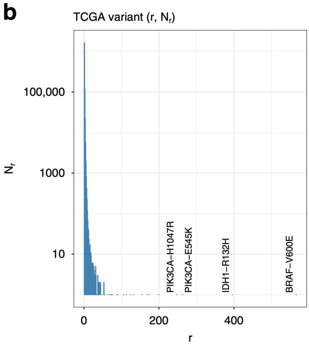
Ethan Ashby

Pomona College

September 25, 2020

Goal: use mutations to classify tumor primary site

- Cancer is a malady caused and characterized by mutation.
- Mutation frequency and spectra varies dramatically across tissue type.
  - In 27 different cancer types, median mutation frequency varied 1000 fold.[1]
- The preponderance of mutation in human cancer is **rare**[2]



[1] Lawrence et al., 2013

[2] Chakraborty et al., 2019

Statistical divergence measures the "distance" between probability distributions

Intuition: *divergence* measures the average differences between probability functions, P and Q, weighted by a function f of the odds ratio between P and Q. More formally:

Statistical divergence measures the "distance" between probability distributions

Theorem 1

Let C be a thrice differentiable, convex function with positive support s.t. $C(0) = 0$. Define the Pearson residual at x to be:

$$\delta(x) = \frac{d_n(x)}{f_\theta(x)} - 1 \tag{1}$$

*Then the **disparity, ϕ-divergence, or f-divergence** between d and f_θ is given by:*

$$\rho_C(d_n, f_\theta) = \sum_{x=0}^{\infty} C(\delta(x)) f_\theta(x) \tag{2}$$

Statistical divergence is a weak measure of distance

Theorem 2

*C, or the **disparity generating function**, meets the requirements of a statistical distance.*

- *The disparity defined in Theorem 1 is nonnegative ($\rho_C(d_n, f_\theta) \geq 0$)*
- *The disparity is only 0 iff $d_n = f_\theta$*

However, C need not satisfy symmetry ($\rho_C(d_n, f_\theta) = \rho_C(f_\theta, d_n)$), nor must C satisfy the triangle inequality

The disparity generating function yields many classical divergences

Supplying different convex, thrice differentiable functions C in Theorem 1 give different divergences.

$C(\delta)$	Formula	Divergence
$(\delta+1)\log(\delta+1)-\delta$	$\sum d_n \log(d_n/f_\theta)$	Likelihood Disparity
$\delta-\log(\delta+1)$	$\sum f_\theta \log(f_\theta/d_n)$	Kullback-Liebler
$2((\delta+1)^{1/2}-1)^2$	$2\sum[d_n^{1/2}-f_\theta^{1/2}]^2$	(twice) squared Hellinger Distance
$\delta^2/2$	$\sum \frac{(d_n-f_\theta)^2}{2f_\theta}$	(half) Pearson's chi-square
$\frac{\delta^2}{2(\delta+1)}$	$\sum \frac{(d_n-f_\theta)^2}{2d_n}$	(half) Neyman's chi-square

Subfamilies of divergences generate classic divergences

The **Cressie-Read** family of power divergences is indexed by a real parameter $\lambda \in (-\infty, \infty)$:

$$\mathrm{PD}_\lambda(d_n, f_\theta) = \frac{1}{\lambda(\lambda+1)} \sum d_n \left[\left(\frac{d_n}{f_\theta} \right)^\lambda - 1 \right] \tag{3}$$

λ	Divergence
1	PCS
0	LD
-1/2	HD
-1	KLD
-2	NCS

Rényi divergence: another intriguing family

We define the **Rényi divergence**, or alpha divergence, or the information of order α obtained if the distribution Q is replaced by P as:

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log_2 \left(\sum_{k=1}^{n} \frac{p_k^\alpha}{q_k^{\alpha-1}} \right) \tag{4}$$

Note that as $\alpha \to 1$, we obtain the KLD. As $\alpha \to 1/2$, we obtain double the Bhattacharyya distance (related to the Bhattaacharyya coefficient, the approximate overlap between two distributions). As $\alpha \to 0$, the probabilities (regardless of their value) are weighted equally. As $\alpha \to \infty$, the Rényi entropy (and therefore the divergence) is determined by the higher probabilities.

Smoothed Good-Turing frequency estimation generates mutation probabilities

Good-Turing frequency estimation allows assignment of probabilities to events (mutations) we've never seen before:

$$\hat{q_r}^{GT} = \frac{r+1}{m+1} \frac{S(N_{r+1})}{S(N_r)} \tag{5}$$

where $\hat{q_r}$ is the estimated probability (in a new tumor) of occurrence of a variant that has been observed r times in m previous tumors.

$$1 - e^{-\frac{N_1}{m+1}} \tag{6}$$

yields an exponential approximation of the probability of encountering *at least one previously unseen variant* in a new tumor.

The challenge: how proximal is Gene A to Gene B?

Gene A:

Tiss. Type	A	B	...	K
P(mut|type)	0.1	0.15	...	0.01

Gene B:

Tiss. Type	A	B	...	K
P(mut|type)	0.6	0.4	...	0.05

Approach 1: Describe each gene as a bivariate joint distribution

Tiss. Type	A	...	K
$\mathsf{P}(x_j = 1)$	$\mathsf{P}(x_j = 1 \vert A)\mathsf{P}(A)$	...	$\mathsf{P}(x_j = 1 \vert A)\mathsf{P}(K)$
$\mathsf{P}(x_j = 0)$	$\mathsf{P}(x_j = 0 \vert A)\mathsf{P}(A)$	...	$\mathsf{P}(x_j = 0 \vert A)\mathsf{P}(K)$

Approach 1: Describe each gene as a bivariate joint distribution

Tiss. Type	A	...	K
P($x_j = 1$)	P($x_j = 1 \| A$)P(A)	...	P($x_j = 1 \| A$)P(K)
P($x_j = 0$)	P($x_j = 0 \| A$)P(A)	...	P($x_j = 0 \| A$)P(K)

Benefits:

- Interpretable probabilities
- Defines prob dist that we can apply divergences to

Approach 1: Describe each gene as a bivariate joint distribution

Tiss. Type	A	...	K
P$(x_j = 1)$	P$(x_j = 1\|A)$P(A)	...	P$(x_j = 1\|A)$P(K)
P$(x_j = 0)$	P$(x_j = 0\|A)$P(A)	...	P$(x_j = 0\|A)$P(K)

Benefits:

- Interpretable probabilities
- Defines prob dist that we can apply divergences to

Drawbacks:

- In general, second row values $>$ first row values. We want to focus on the ”middle values” (i.e. largest values in first row).
- A good divergence scheme would assign the second row no weight, and only define distance based on first row signal. So why include the second row?

Approach 2: Flip conditional probabilities using Bayes Rule

Bayes Rule:

$$P(C|x_j) = \frac{P(x_j|C)P(C)}{P(x_j)} \tag{7}$$

Gene A:

Tiss. Type	A	...	K
$x_j = 1$	P($A|x_j = 1$)	...	P($K|x_j = 1$)

Approach 2: Flip conditional probabilities using Bayes Rule

Bayes Rule:

$$P(C|x_j) = \frac{P(x_j|C)P(C)}{P(x_j)} \tag{7}$$

Gene A:

Tiss. Type	A	...	K
$x_j = 1$	P$(A\|x_j = 1)$	...	P$(K\|x_j = 1)$

Benefits:

- Defines very simple discrete prob dist
- Probabilities relate to classification
- Don't have to deal with second row!

Approach 2: Flip conditional probabilities using Bayes Rule

Bayes Rule:

$$P(C|x_j) = \frac{P(x_j|C)P(C)}{P(x_j)} \tag{7}$$

Gene A:

Tiss. Type	A	...	K
$x_j = 1$	P$(A\|x_j = 1)$	...	P$(K\|x_j = 1)$

Benefits:

- Defines very simple discrete prob dist
- Probabilities relate to classification
- Don't have to deal with second row!

Drawbacks:

- Bias small genes and common cancer types?
- How do we measure tissue specificity?

Approach 3: Use softmax function to transform GT probs to pdf

The **softmax function** is a generalization of the logistic function and is often used in the last layer of a neural network to normalize the output to a probability distribution.

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{8}$$

A larger base of the exponent creates a probability distribution more concentrated around the larger input values.

Approach 3: Use softmax function to transform GT probs to pdf

The **softmax function** is a generalization of the logistic function and is often used in the last layer of a neural network to normalize the output to a probability distribution.

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{8}$$

A larger base of the exponent creates a probability distribution more concentrated around the larger input values.

Benefits:

- Generates prob dist directly from GT probs
- Don't have to deal with second row!

Approach 3: Use softmax function to transform GT probs to pdf

The **softmax function** is a generalization of the logistic function and is often used in the last layer of a neural network to normalize the output to a probability distribution.

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{8}$$

A larger base of the exponent creates a probability distribution more concentrated around the larger input values.

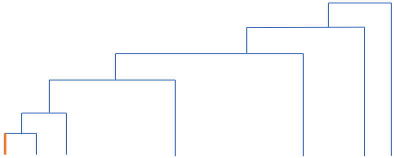
Benefits:

- Generates prob dist directly from GT probs
- Don't have to deal with second row!

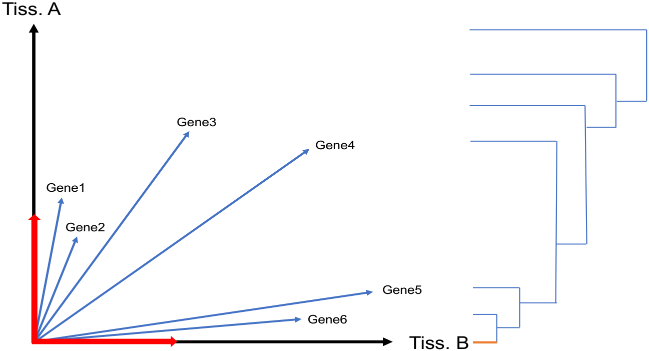
Drawbacks:

- Loss of interpretability of probs
- Distorts the tissue specific geometry

Semi-supervised hierarchical clustering to identify metagenes



	Dummy A	Dummy B
Gene 1	1	100
Gene 2	3	90
Gene 3	10	60
Gene 4	50	40
Gene 5	95	2
Gene 6	99	1



Some divergence measures that we should consider

- Jensen-Rényi
  - Tuneable
- Jensen-Shannon
  - Information Theoretic Interpretation
- Cosine
  - preserves tissue specific geometry
- Skew divergence/Jensen-Shannon-α divergence
  - tuneability and information theoretic interpretation