StatsM254 Statistical Methods in Computational Biology Lecture 9 - 04/29/2014

Measure of Correlation & Dependence

Lecturer: Jingyi Jessica Li *Scribe: Chelsea Ju, Alden Huang*

1 Concepts

- Dependence vs. Independence
  Independence: X and Y r.v.s

$$p(x, y) = p(x)p(y)$$
$$x \in supp(X)$$
$$y \in supp(Y)$$

- Correlation:
  Broad class of relationships including dependence.
- Common Misunderstandings:
  Correlated $\neq$ dependence.
  e.g. Pearson correlation $= 0$ does not imply X and Y are independent.
  (X, Y) is bivariate Gaussian $\Leftrightarrow$ X and Y are independent.
- Correlation/dependence $\neq$ causality.

2 Methods

2.1 Pearson Correlation

- Describe linear relationships.
- Invariant to linear transformations of X and Y, that is:

$$\begin{aligned}\rho(X, Y) &= \rho(aX + b, Y) \\ &= \rho(aX + b, cY + d)\end{aligned}$$

- Population version:

$$\rho(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}$$

- Sample version (statistic):

$$\text{data } \{(X_i, Y_i)\}_{i=1}^n \text{ from the joint distribution of } (X, Y)$$

$$\hat{\rho}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left(\sum_{i=1}^n (X_i - \bar{X})^2\right)\left(\sum_{i=1}^n (Y_i - \bar{Y})^2\right)}}$$

$$\hat{\rho}(X, Y) \in [-1, 1]$$

- Coefficient of determination R^2: describes goodness of fit.
  In the simple linear model w/ univariates:

$$\begin{aligned}
Y &= a + bX + \epsilon, \epsilon \sim N(0, \sigma^2) \\
R^2 &= 1 - \frac{sum\ of\ squares\ residuals\ (SSR)}{sum\ of\ squares\ total\ (SST)} \\
&= 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \\
&= \frac{sum\ of\ squares\ explained\ (SSE)}{SST} \\
&= \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \\
&= \hat{\rho}^2 \\
R^2 &\in [0, 1]
\end{aligned}$$

2.2 Rank Correlation

- Monotonic relationships.
- Invariant to monotonic transformations.
- Robust to outliers.

1. Spearman Rank Correlation

   Pearson correlation between ranked X_i and Y_i.
   Example:

X_i	$rank$
0.8	1
1.2	2.5
1.2	2.5
2.3	4
18	5

2. Kendall's tau (Kendall, 1938) [1]

   Procedure:

$$
\begin{aligned}
& (X_i, Y_i)\ vs\ (X_j, Y_j) \\
concordant\ pairs\ & X_i > X_j\ and\ Y_i > Y_j \\
& or \\
& X_i < X_j\ and\ Y_i < Y_j \\
discordant\ pairs\ & X_i > X_j\ and\ Y_i < Y_j \\
& or \\
& X_i < X_j\ and\ Y_i > Y_j \\
neither\ concordant\ or\ discordant\ & X_i = X_j\ or\ Y_i = Y_j
\end{aligned}
$$

$$
\tau = \frac{(number\ of\ concordant\ pairs) - (number\ of\ discordant\ pairs)}{\frac{1}{2}n(n-1)} \in [-1, 1]
$$

R Function: *corr(x, y, method=c("pearson", "spearman", "kendall"))*

2.3 Maximal Correlation (Renyi Correlation)

(Gebelein, 1941; Renyi, 1959)[2, 3]

$$
R(X, Y) = \sup_{g,h}\ \rho(g(X), h(Y))
$$

$$
0 < var(g(X)) < \infty
$$

$$
0 < var(h(Y)) < \infty
$$

$$
R(X, Y) \in [0, 1]
$$

$$
R(X, Y) = 0\ iff\ X\ and\ Y\ are\ independent.
$$

Procedure to find R:

Alternating Condition Expectation (ACE) (Breiman and Friedman, 1985) [4]

Goal:

$$
\min_{g,h} \mathrm{E}[(g(X) - h(Y))^2] \tag{1}
$$

Basic ACE algorithm:

1. $$
\begin{aligned}
\text{set } h(Y) &= \frac{Y}{||Y||} \\
\text{e.g. } h(Y_i) &= \frac{Y_i}{\sqrt{\sum_{i=1}^{n} Y_i^2}}
\end{aligned}
$$

2. Iterate until equation (1) fails to converge.

$$
g_1(X) = \mathrm{E}[h(Y)|X]
$$

Replace $g(X)$ with $g_1(X)$.

$$h_1(Y) = \frac{\mathrm{E}[g(X)|Y]}{||\mathrm{E}[g(X)|Y]||}$$

Replace $h(Y)$ by $h_1(Y)$.

3. The resulting values h and g are the transformations of Y and X

$$\Rightarrow R(X,Y) = \rho(g(X), h(Y))$$

R package: "*acepack*".

2.4 Distance Correlation / Brownian Correlation

(Szekely, 2005)[5]

- $=0$ iff X and Y are independent (same as maximal correlation)
- $X \in \mathbb{R}^p, Y \in \mathbb{R}^q$ as arbitrary dimensions
- Sample distance covariance

$$a_{ij} = ||X_i - X_j|| \overset{L2}{=} \sqrt{\sum_{k=1}^{p}(X_{ik} - X_{jk})^2}$$

$$b_{ij} = ||Y_i - Y_j|| \text{ doubly centered distances}$$

$$A_{ij} = a_{ij} - \bar{a}_{i.} - \bar{a}_{.j} + \bar{a}_{..}$$

$$B_{ij} = b_{ij} - \bar{b}_{i.} - \bar{b}_{.j} + \bar{b}_{..}$$

$$dCov_n^2(X,Y) := \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}B_{ij}$$

- Sample distance variance

$$dVar^2(X) = dCov^2(X,X) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}^2$$

- Sample distance correlation

$$dCor_n(X,Y) = \frac{dCov_n(X,Y)}{\sqrt{dVar_n(X) \bullet dVar_n(Y)}}$$

- R package "*energy*", function "*dcor*".

2.5 Hoeffding's Independent Test

(Hoeffding, 1948)[6].

- Non-parametric testing statistic

$$H_o : P(X,Y) = P(X)P(Y)$$

- R-package: *"Hmisc"*

2.6 CorGC

(Delicado & Smrekar, 2009) [7]

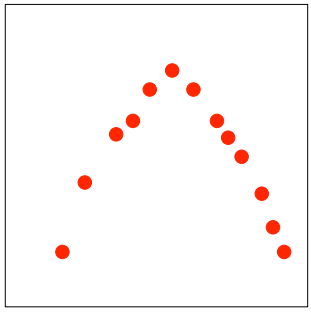
- Principle component based.
- Can be downloaded from web.

Both Hoeffding's test and CorGC can capture non-linear relationship.

2.7 Mutual Information

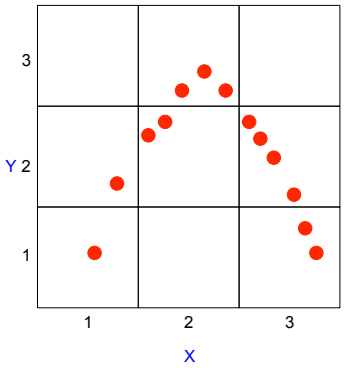
$$I(X,Y) = \int dxdy P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$
$$= E_{(X,Y)} \left[\log_2 \frac{P(X,Y)}{P(X)P(Y)} \right]$$

- $I(X,Y) = 0$ if and only if X and Y are independent
- $I(X,Y) \in [0,1]$
- Given $\{(X_i, Y_i)\}_{i=1}^n$, how to estimate MI?
  1. KDE = kernel density estimation
     smoothing (into a continuous distribution)
  2. KNN = k-th nearest neighbors
     non-parametric
- Example:



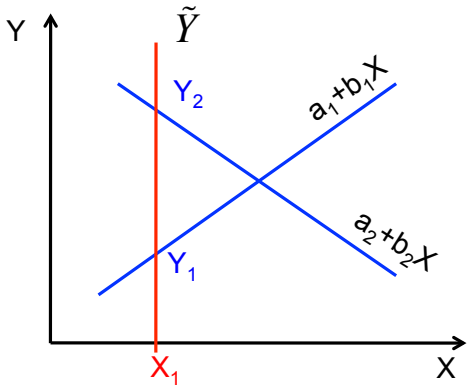
Approaches:

1. Draw grid
2. Estimate $P(X)$ and $P(Y)$



$$\hat{I}(X,Y) = \sum_{x=1}^{3} \sum_{y=1}^{3} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)}$$
$$P(1,1) = \frac{1}{13}$$
$$P(X=1) = \frac{2}{13}$$
$$P(Y=3) = \frac{3}{13}$$

2.8 New Method (by Dr. Li)



Problem:

- for a given X, how to predict 2 Ys
- mixture of regression
- recall liquid association method from lecture 5: use Z as a hidden indicator

Proposed Solution - ignore hidden indicator Z

1. Prediction
2. Measure of association

- Objective Function:
  $\sum_{k=1}^{2} \sum_{i \in c_k} (Y_i - a_k - b_k X_i)^2$

  p: Proportion of class 1

  $1-p$: Proportion of class 2

$$\hat{Y_i} = \begin{cases} a_1 + b_1 X_i & \text{with} \quad \hat{p} \\ a_2 + b_2 X_i & \text{with} \quad 1-\hat{p} \end{cases}$$

$\hat{Y_i}$: a two values discrete variable

$$Y = \begin{cases} a_1 + b_1 X + \epsilon_1 & \text{with} \quad p \\ a_2 + b_2 X + \epsilon_2 & \text{with} \quad 1-p \end{cases}$$
$$\epsilon_1 \perp \epsilon_2 \sim N(0, \sigma^2)$$

Y_i as incomplete sample point

- Example: c vs d constant
  $|c-d|$

  C vs D, 2 values discrete variables $\begin{cases} c_1 & p \\ c_2 & 1-p \end{cases}$ $\begin{cases} d_1 & q \\ d_2 & 1-q \end{cases}$

$$\begin{aligned} Dist(C,D) = & \, min(p,q)|c_1 - d_1| + \\ & (1 - max(p,q))|c_2 - d_2| + \\ & (p - min(p,q))|c_1 - d_2| + \\ & (q - min(p,q))|c_2 - d_1| \end{aligned}$$

References

[1] M. Kendall, "A New Measure of Rank Correlation",*Biometrika*, vol.30, no. 1-2, pp. 81–89.

[2] H. Gebelein, "Das statistische Problem der Korrelation als Variations- und Eigenwert-problem und sein Zusammenhang mit der Ausgleichungsrechnun", *ZAMMJournal of Applied Mathematics and Mechanics/Zeitschrift fr Angewandte Mathematik und Mechanik*, vol. 21, pp. 364–379, 1941.

[3] A. Renyi, "On measures of dependence", *Acta mathematica hungarica*, vol. 10, pp. 441-451, 1959.

[4] L. Breiman, and J. H. Friedman,"Estimating optimal transformations for multiple regression and correlation", *Journal of the American Statistical Association*, vol. 80, no. 391, pp. 580–598, 1985.

[5] G. Szkely, and M. Rizzo, "A new test for multivariate normality", *Journal of Multivariate Analysis*, vol. 93, no. 1, pp. 58–80, 2005.

[6] W. Hoeffding, "A non-parametric test of independence", *The Annals of Mathematical Statistics*, vol. 19, no. 4, pp. 546–557, 1948.

[7] P. Delicado, and M. Smrekar, "Measuring non-linear dependence for two random variables distributed along a curve", *Statistics and Computing,*, vol. 19, no.3, pp. 255–269, 2009.

[8] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Médard, M. and J. Crowcroft, "XORs in the air: practical wireless network coding", *IEEE/ACM Transactions on Networking,*, vol. 16, no. 3, pp. 497–510, 2008.

[9] H. Rahul, N. Kushman, D. Katabi, C. Sodini, and F. Edalat, "Learning to Share: Narrowband-Friendly Wideband Wireless Networks", *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 4, pp. 147–158, 2008.