Correlation analysis 1: Canonical correlation analysis

Ryan Tibshirani

Data Mining: 36-462/36-662

February 14 2013

Review: correlation

Given two random variables $X, Y \in \mathbb{R}$, the (Pearson) correlation between X and Y is defined as

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Recall that

$$\mathrm{Cov}(X,Y) = \mathrm{E}\big[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\big]$$

and

$$\mathrm{Var}(X) = \mathrm{E}\big[(X - \mathrm{E}[X])^2\big] = \mathrm{Cov}(X,X)$$

This measures a linear association between X, Y. Properties:

- $-1 \leq \mathrm{Cor}(X,Y) \leq 1$
- X, Y independent $\Rightarrow \mathrm{Cor}(X,Y) = 0$ (Homework 2)
- $\mathrm{Cor}(X,Y) = 0 \not\Rightarrow X, Y$ independent (Homework 2) ⟸

More on this later ...

Review: sample correlation

Given centered $x, y \in \mathbb{R}^n$, the sample correlation between x and y is defined as

$$\mathrm{cor}(x, y) = \frac{x^T y}{\sqrt{x^T x}\sqrt{y^T y}} .$$

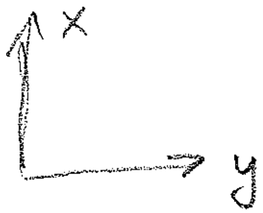
Note the analogy to the definition on the last slide—we just replace everything by its sample version. I.e., if we write cov and var for the sample covariance and variance, then

$$\mathrm{cor}(x, y) = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)}\sqrt{\mathrm{var}(y)}} .$$

Note: if $x, y \in \mathbb{R}^n$ are centered unit vectors then $\mathrm{cor}(x, y) = x^T y$

This measures a linear association between x, y. Properties:

- $-1 \le \mathrm{cor}(x, y) \le 1$
- $\mathrm{cor}(x, y) = 0 \iff x, y$ are orthogonal



Canonical correlation analysis

Principal component analysis attempts to answer the question: "which directions account for much of the observed variance in a data set?" Given a centered matrix $X \in \mathbb{R}^{n \times p}$, we first find the direction $v_1 \in \mathbb{R}^p$ to maximize the sample variance of Xv:

$$v_1 = \underset{\|v\|_2=1}{\operatorname{argmax}} \ \mathrm{var}(Xv)$$

Canonical correlation analysis is similar but instead attempts to answer: "which directions account for much of the covariance between two data sets?" Now we are given two centered matrices $X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^{n \times q}$, and we seek the two directions $\alpha_1 \in \mathbb{R}^p$, $\beta_1 \in \mathbb{R}^q$ that maximize the sample covariance of $X\alpha$ and $Y\beta$:

$$\alpha_1, \beta_1 = \underset{\|X\alpha\|_2=1,\ \|Y\beta\|_2=1}{\operatorname{argmax}} \ \mathrm{cov}(X\alpha, Y\beta) \quad \leftarrow \mathrm{cor}(X\alpha, Y\beta)$$

Subject to the constraints, this is equivalent to maximizing $\mathrm{cor}(X\alpha, Y\beta)$. (Why?)

Canonical directions and variates

The first canonical directions $\alpha_1 \in \mathbb{R}^p$, $\beta_1 \in \mathbb{R}^q$ are given by

$$\alpha_1, \beta_1 = \underset{\|X\alpha\|_2=1,\ \|Y\beta\|_2=1}{\operatorname{argmax}} \; (X\alpha)^T(Y\beta)$$

$(X\alpha_1)^T(Y\beta_1)$

$(X\cdot -\alpha_1)^T(Y\beta_1)$
$= -(X\alpha_1)^T(Y\beta_1)$

Vectors $X\alpha_1, Y\beta_1 \in \mathbb{R}^n$ are called the first canonical variates, and $\rho_1 = (X\alpha_1)^T(Y\beta_1) \in \mathbb{R}$ is called the first canonical correlation

Given the first $k-1$ directions, the kth canonical directions $\alpha_k \in \mathbb{R}^p$, $\beta_k \in \mathbb{R}^q$ are defined as

$$\alpha_k, \beta_k = \underset{\substack{\|X\alpha\|_2=1,\ \|Y\beta\|_2=1 \\ (X\alpha)^T(X\alpha_j)=0,\ j=1,\ldots k-1 \\ (Y\beta)^T(Y\beta_j)=0,\ j=1,\ldots k-1}}{\operatorname{argmax}} \; (X\alpha)^T(Y\beta)$$

PCA: $v^T v_i = 0$
$i=1,\ldots k-1$

Vectors $X\alpha_k, Y\beta_k \in \mathbb{R}^n$ are called the kth canonical variates, and $\rho_k = (X\alpha_k)^T(Y\beta_k) \in \mathbb{R}$ is called the kth canonical correlation

is ρ_k always > 0 ? yes

Example: scores data

Example: $n = 88$ students took tests in each of 5 subjects: mechanics, vectors, algebra, analysis, statistics. (From Mardia et al. (1979) “Multivariate analysis”.) Each test is out of 100 points

The tests on mechanics, vectors were closed book and those on algebra, analysis, statistics were open book. There’s clearly some correlation between these two sets of scores:

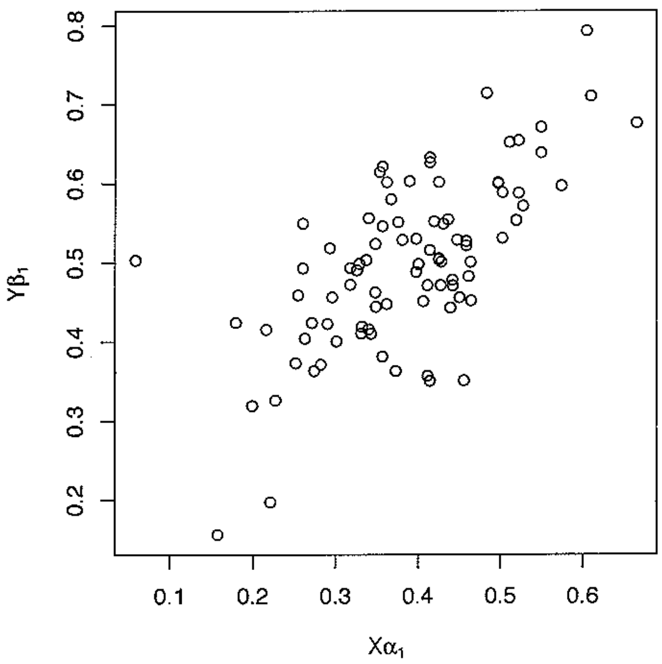
```
      alg   ana   sta
mec 0.547 0.409 0.389
vec 0.610 0.485 0.436
```

Canonical correlation analysis attempts to explain this phenomenon using the variables in each set jointly. Here X contains the closed book test scores and Y contains the open book test scores, so $X \in \mathbb{R}^{88\times 2}$ and $Y \in \mathbb{R}^{88\times 3}$

The first canonical directions (multiplied by 10^3):

$$\alpha_1 = \begin{pmatrix} 2.770 \\ 5.517 \end{pmatrix} \begin{array}{l} \text{mec} \\ \text{vec} \end{array}, \quad \beta_1 = \begin{pmatrix} 8.782 \\ 0.860 \\ 0.370 \end{pmatrix} \begin{array}{l} \text{alg} \\ \text{ana} \\ \text{sta} \end{array}$$

The first canonical correlation is $\rho_1 = 0.663$, and the variates:



The second directions are more surprising, but $\rho_2 = 0.041$

How many canonical directions are there?

We have $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$. How many pairs of canonical directions $(\alpha_1, \beta_1), (\alpha_2, \beta_2), \ldots$ are there?

We know that any n orthogonal (linearly independent) vectors in $\mathbb{R}^n$ form a basis for $\mathbb{R}^n$. Therefore there cannot be more than p orthogonal vectors of the form $X\alpha$, $\alpha \in \mathbb{R}^p$, and q orthogonal vectors of the form $Y\beta$, $\beta \in \mathbb{R}^q$. (Why?)

Hence there are exactly $\underline{r = \min\{p, q\}}$ canonical directions $(\alpha_1, \beta_1), \ldots (\alpha_r, \beta_r)$[1]

[1] This is assuming that $n \geq p$ and $n \geq q$. In general, there are actually only $r = \min\{\mathrm{rank}(X), \mathrm{rank}(Y)\}$ canonical directions

Transforming the problem

If $A \in \mathbb{R}^{p\times p}$, $B \in \mathbb{R}^{q\times q}$ are invertible, then computing

$$\tilde{\alpha}_1, \tilde{\beta}_1 = \underset{\|XA\tilde{\alpha}\|_2=1,\ \|YB\tilde{\beta}\|_2=1}{\operatorname{argmax}} (XA\tilde{\alpha})^T(YB\tilde{\beta}),$$

is equivalent to the first step of canonical correlation analysis. In particular, the first canonical directions are given by $\alpha_1 = A\tilde{\alpha}_1$ and $\beta_1 = B\tilde{\beta}_1$. The same is also true of further directions

I.e., we can transform our data matrices to be $\tilde{X} = XA$, $\tilde{Y} = YB$ for any invertible A, B, solve the canonical correlation problem with $\tilde{X}, \tilde{Y}$, and then back-transform to get our desired answers

Why would we ever do this? Because there is a transformation A, B that makes the computational problem simpler

Sphering

For any symmetric invertible matrix $A \in \mathbb{R}^{n \times n}$, there is a matrix $A^{1/2} \in \mathbb{R}^{n \times n}$, called the (symmetric) square root of A, such that $A^{1/2}A^{1/2} = A$

We write the inverse of $A^{1/2}$ as $A^{-1/2}$. Note $A^{-1/2}AA^{-1/2} = I$. (Why?)

Given centered matrices $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$,[2] we define $V_X = X^T X \in \mathbb{R}^{p \times p}$ and $V_Y = Y^T Y \in \mathbb{R}^{q \times q}$. Then

$$\tilde{X} = XV_X^{-1/2} \in \mathbb{R}^{n \times p} \text{ and } \tilde{Y} = YV_Y^{-1/2} \in \mathbb{R}^{n \times q}$$

are called the sphered versions of X and Y.[3] Note that the sample covariance of $\tilde{X}$ and $\tilde{Y}$ is

$$\mathrm{cov}(\tilde{X}) = I/n \text{ and } \mathrm{cov}(\tilde{Y}) = I/n$$

[2] Here we are assuming that $\mathrm{rank}(X) = p$ and $\mathrm{rank}(Y) = q$

[3] Alternatively, for sphering we would sometimes define $V_X = (X^T X)/n$ and $V_Y = (Y^T Y)/n$, so that the transformed sample covariances are exactly I

Transforming the problem (continued)

As suggested by the previous slide, we will take $\tilde{X} = XV_X^{-1/2}$ and $\tilde{Y} = YV_Y^{-1/2}$, and we'll solve the problem

$$\tilde{\alpha}_1, \tilde{\beta}_1 = \underset{\|\tilde{X}\tilde{\alpha}\|_2=1,\ \|\tilde{Y}\tilde{\beta}\|_2=1}{\operatorname{argmax}} (\tilde{X}\tilde{\alpha})^T(\tilde{Y}\tilde{\beta})$$

Recall that then $\alpha_1 = V_X^{-1/2}\tilde{\alpha}_1$ and $\beta_1 = V_Y^{-1/2}\tilde{\beta}_1$.

So why is this simpler? Note that the constraint says

$$1 = (\tilde{X}\tilde{\alpha})^T(\tilde{X}\tilde{\alpha}) = \tilde{\alpha}^T V_X^{-1/2} X^T X V_X^{-1/2} \tilde{\alpha} = \tilde{\alpha}^T \tilde{\alpha}$$

i.e., $\|\tilde{\alpha}\|_2 = 1$. Similarly, $\|\tilde{\beta}\|_2 = 1$. Hence our problem can be rewritten as:

$$\tilde{\alpha}_1, \tilde{\beta}_1 = \underset{\|\tilde{\alpha}\|_2=1,\ \|\tilde{\beta}\|_2=1}{\operatorname{argmax}} \tilde{\alpha}^T M \tilde{\beta}$$

where $M = \tilde{X}^T\tilde{Y} = V_X^{-1/2} X^T Y V_Y^{-1/2} \in \mathbb{R}^{p\times q}$. The same is true for further directions

Computing canonical directions and variates

Now comes the singular value decomposition to the rescue (again!). Let $r = \min\{p, q\}$. Then we can decompose

$$M = UDV^T$$

where $U \in \mathbb{R}^{p \times r}$, $V \in \mathbb{R}^{q \times r}$ have orthonormal columns, and $D = \mathrm{diag}(d_1, \ldots d_r) \in \mathbb{R}^{r \times r}$ with $d_1 \geq \ldots \geq d_r \geq 0$. Further:

- The transformed canonical directions $\tilde{\alpha}_1, \ldots \tilde{\alpha}_r \in \mathbb{R}^p$ and $\tilde{\beta}_1, \ldots \tilde{\beta}_r \in \mathbb{R}^q$ are the columns of U and V, respectively
- The canonical directions $\alpha_1, \ldots \alpha_r \in \mathbb{R}^p$ and $\beta_1, \ldots \beta_r \in \mathbb{R}^q$ are the columns of $V_X^{-1/2} U$ and $V_Y^{-1/2} V$, respectively;
- the canonical variates $X\alpha_1, \ldots X\alpha_r \in \mathbb{R}^n$ and $Y\beta_1, \ldots Y\beta_r \in \mathbb{R}^n$ are the columns of $XV_X^{-1/2}U \in \mathbb{R}^{n \times r}$ and $YV_Y^{-1/2}V \in \mathbb{R}^{n \times r}$, respectively
- The canonical correlations $\rho_1 \geq \ldots \geq \rho_r$ are equal to $d_1 \geq \cdots \geq d_r$, the diagonal entries of D

Example: olive oil data

Example: $n = 572$ olive oils, with $p = 9$ features (the `olives` data set from the R package `classifly`):

1. region ← Y 572×3
2. palmitic
3. palmitoleic
4. stearic
5. oleic
6. linoleic
7. linolenic
8. arachidic
9. eicosenoic

X 572×8

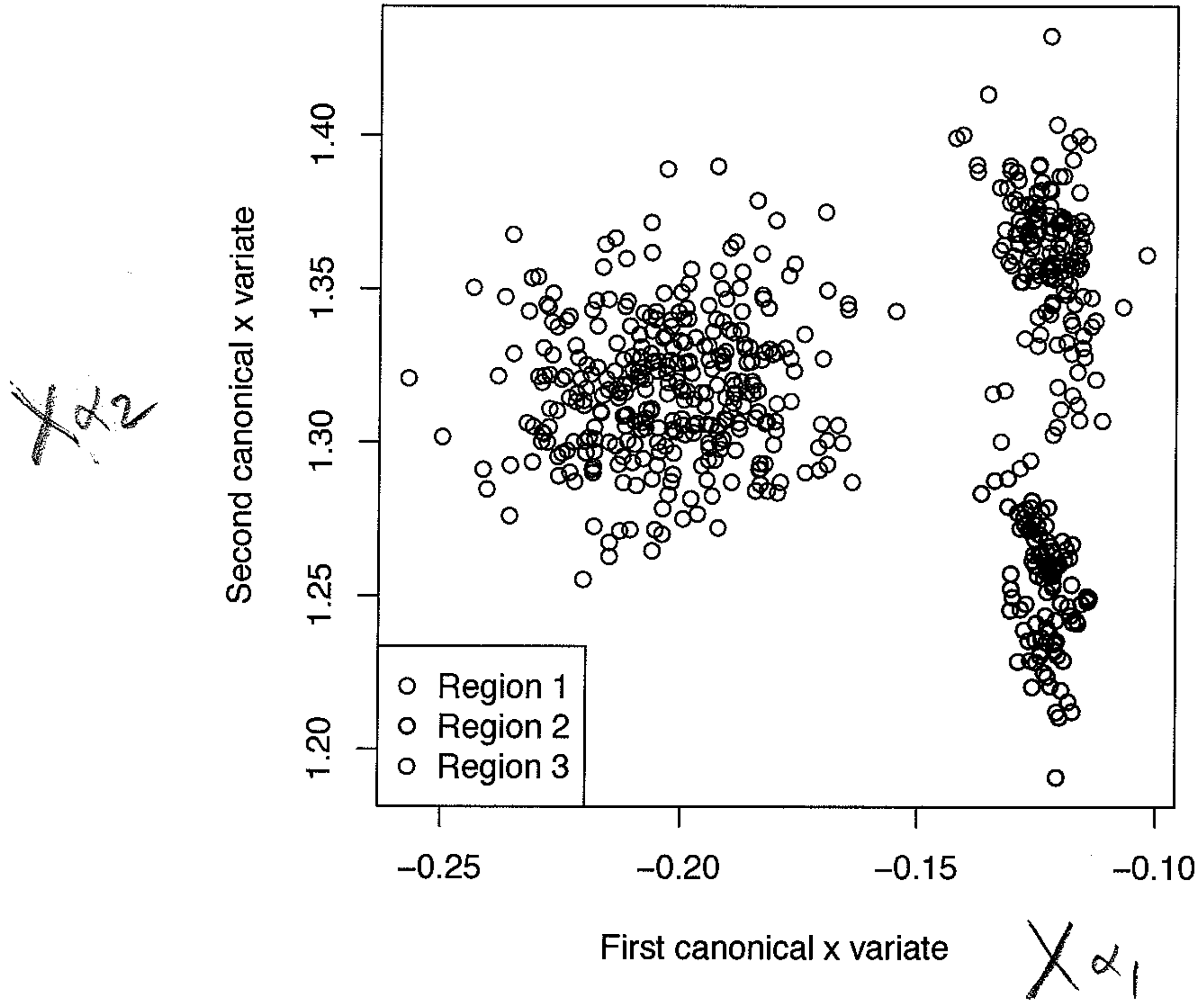
Variable 1 takes values in $\{1, 2, 3\}$, indicating the region (in Italy) of origin. Variables 2-9 are continuous valued and measure the percentage composition of 8 different fatty acids

We are interested in the correlations between the region of origin and the fatty acid measurements. Hence we take $X \in \mathbb{R}^{572\times 8}$ to contain the fatty acid measurements, and $Y \in \mathbb{R}^{572\times 3}$ to be an indicator matrix, i.e., each row of Y indicates the region with a 1 and otherwise has 0s. This might look like:

$$Y = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ & \cdots & \end{pmatrix} \qquad \begin{matrix} 1 \\ 1 \\ 3 \\ 2 \\ \vdots \end{matrix}$$

(In this case, canonical correlation analysis actually does the exact same thing as linear discriminant analysis, an important tool that we will learn later for classification)

The first two canonical X variates, with the points colored by region:



Canonical correlation analysis in R

Canonical correlation analysis is implemented by the `cancor` function in the base distribution. E.g.,

```
cc = cancor(x,y)
alpha = cc$xcoef
beta = cc$ycoef
rho = cc$cor
xvars = x %*% alpha
yvars = y %*% beta
```

Recap: canonical correlation analysis

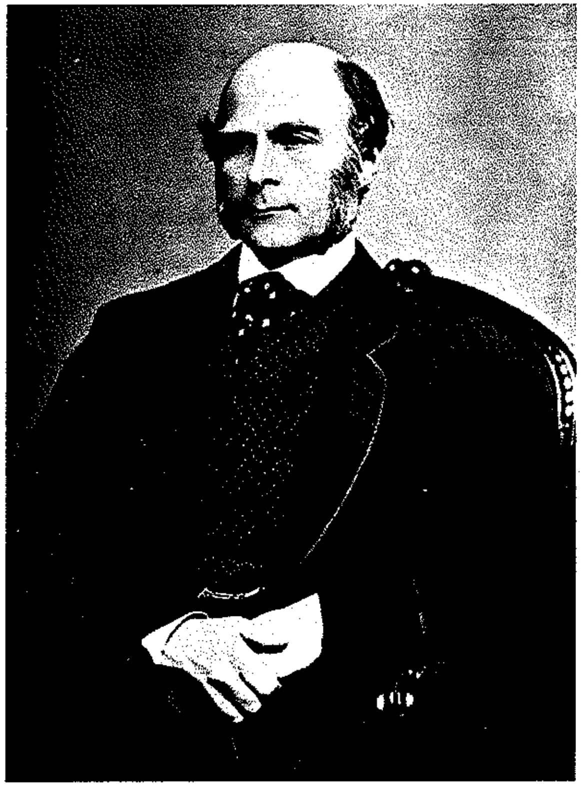
In canonical correlation analysis we are looking for pairs of directions, one in each of the feature spaces of two data sets $X \in \mathbb{R}^{n \times p}, Y \in \mathbb{R}^{n \times q}$, to maximize the covariance (or correlation)

We defined the pairs of canonical directions $(\alpha_1, \beta_1), \ldots (\alpha_r, \beta_r)$, where $r = \min\{p, q\}$, and $\alpha_j \in \mathbb{R}^p$, $\beta_j \in \mathbb{R}^q$. We also defined the pairs of canonical variates $(X\alpha_1, X\beta_1), \ldots (X\alpha_r, X\beta_r)$, where $X\alpha_j \in \mathbb{R}^n$ and $X\beta_j \in \mathbb{R}^n$. Finally, we defined the canonical correlations $\rho_1, \ldots \rho_r \in \mathbb{R}$

We saw that transforming the problem leads to a simpler form. From this simpler form we can compute the canonical directions, correlations, and variates using the singular value decomposition

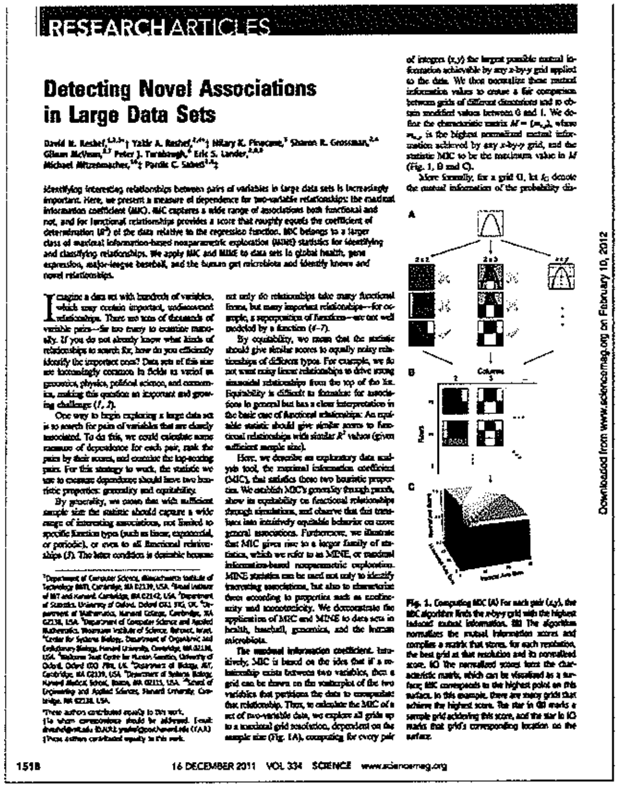
Next time: measures of correlation

A lot of work has been done, but there's still a lot of interest



1888

. . .



RESEARCH ARTICLES

Detecting Novel Associations in Large Data Sets

David N. Reshef,[1,2,3]*† Yakir A. Reshef,[2,4]*† Hilary K. Finucane,[5] Sharon R. Grossman,[2,6] Gilean McVean,[3,7] Peter J. Turnbaugh,[6] Eric S. Lander,[2,8,9] Michael Mitzenmacher,[10]‡ Pardis C. Sabeti[2,6]‡

Identifying interesting relationships between pairs of variables in large data sets is increasingly important. Here, we present a measure of dependence for two-variable relationships: the maximal information coefficient (MIC). MIC captures a wide range of associations both functional and not, and for functional relationships provides a score that roughly equals the coefficient of determination (R^2) of the data relative to the regression function. MIC belongs to a larger class of maximal information-based nonparametric exploration (MINE) statistics for identifying and classifying relationships. We apply MIC and MINE to data sets in global health, gene expression, major-league baseball, and the human gut microbiota and identify known and novel relationships.

Imagine a data set with hundreds of variables, which may contain important, undiscovered relationships. There are tens of thousands of variable pairs—far too many to examine manually. If you do not already know what kinds of relationships to search for, how do you efficiently identify the important ones? Data sets of this size are increasingly common in fields as varied as genomics, physics, political science, and economics, making this question an important and growing challenge (1, 2).

One way to begin exploring a large data set is to search for pairs of variables that are closely associated. To do this, we could calculate some measure of dependence for each pair, rank the pairs by their scores, and examine the top-scoring pairs. For this strategy to work, the statistic we use to measure dependence should have two heuristic properties: generality and equitability.

By generality, we mean that with sufficient sample size the statistic should capture a wide range of interesting associations, not limited to specific function types (such as linear, exponential, or periodic), or even to all functional relationships (3). The latter condition is desirable because not only do relationships take many functional forms, but many important relationships—for example, a superposition of functions—are not well modeled by a function (4–7).

By equitability, we mean that the statistic should give similar scores to equally noisy relationships of different types. For example, we do not want noisy linear relationships to drive strong sinusoidal relationships from the top of the list. Equitability is difficult to formalize for associations in general but has a clear interpretation in the basic case of functional relationships: An equitable statistic should give similar scores to functional relationships with similar R^2 values (given sufficient sample size).

Here, we describe an exploratory data analysis tool, the maximal information coefficient (MIC), that satisfies these two heuristic properties. We establish MIC's generality through proofs, show its equitability on functional relationships through simulations, and observe that this translates into intuitively equitable behavior on more general associations. Furthermore, we illustrate that MIC gives rise to a larger family of statistics, which we refer to as MINE, or maximal information-based nonparametric exploration. MINE statistics can be used not only to identify interesting associations, but also to characterize them according to properties such as nonlinearity and monotonicity. We demonstrate the application of MIC and MINE to data sets in health, baseball, genomics, and the human microbiota.

The maximal information coefficient. Intuitively, MIC is based on the idea that if a relationship exists between two variables, then a grid can be drawn on the scatterplot of the two variables that partitions the data to encapsulate that relationship. Thus, to calculate the MIC of a set of two-variable data, we explore all grids up to a maximal grid resolution, dependent on the sample size (Fig. 1A), computing for every pair of integers (x,y) the largest possible mutual information achievable by any x-by-y grid applied to the data. We then normalize these mutual information values to ensure a fair comparison between grids of different dimensions and to obtain modified values between 0 and 1. We define the characteristic matrix $M = (m_{x,y})$, where $m_{x,y}$ is the highest normalized mutual information achieved by any x-by-y grid, and the statistic MIC to be the maximum value in M (Fig. 1, B and C).

More formally, for a grid G, let I_G denote the mutual information of the probability dis-

Fig. 1. Computing MIC (A) For each pair (x,y), the MIC algorithm finds the x-by-y grid with the highest induced mutual information. (B) The algorithm normalizes the mutual information scores and compiles a matrix that stores, for each resolution, the best grid at that resolution and its normalized score. (C) The normalized scores form the characteristic matrix, which can be visualized as a surface; MIC corresponds to the highest point on this surface. In this example, there are many grids that achieve the highest score. The star in (B) marks a sample grid achieving this score, and the star in (C) marks that grid's corresponding location on the surface.

[1] Department of Computer Science, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. [2] Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [3] Department of Statistics, University of Oxford, Oxford OX1 3TG, UK. [4] Department of Mathematics, Harvard College, Cambridge, MA 02138, USA. [5] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. [6] Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [7] Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [8] Department of Biology, MIT, Cambridge, MA 02139, USA. [9] Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA. [10] School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA.

*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: dnreshef@mit.edu (D.N.R.); yreshef@post.harvard.edu (Y.A.R.)

‡These authors contributed equally to this work.

2012