Supplementary Materials: Exploring Non-Linear Dependencies in Atmospheric Data with Mutual Information

Petri Laarne [1,2], Emil Amnell [1], Martha Arbayani Zaidan [1,3], Santtu Mikkonen [4,5] and Tuomo Nieminen [1,6,*]

1. Derivations of Theoretical Results

1.1. Transformation Invariance

Let us recall that entropy is defined by the integral

$$H(X) = -\int_{-\infty}^{\infty} p(x) \log p(x) \, \mathrm{d}x,$$

and mutual information by

$$I(Y;X) = H(X) + H(Y) - H(X,Y).$$

We can combine these two to get the alternative definition

$$I(Y;X) = \int_{-\infty}^{\infty} p(x,y) \log \left[\frac{p(x,y)}{p(x)p(y)} \right] \mathrm{d}x \, \mathrm{d}y.$$

The probability densities satisfy $p(x,y) = p(x,f(y))J(x,y)$, where J is the Jacobian determinant of f. This means that any change of variable $y \mapsto f(y)$ cancels out. Here, f must be bijective, differentiable and have differentiable inverse. In practice, most transformations are in this category.

In fact, f only needs to be bijective, measurable and have a measurable inverse [1] (Theorem 1.6.3). The idea is to partition the space $f(Y)$ and observe that the preimages form a probability-preserving partitioning of the original space Y. This equality is preserved in the supremum over all partitionings.

Given suitable symmetry, even some periodicity can be removed. For example, the "date $\mapsto$ day of year" transformation loses no information if years are independent of each other. This is illustrated in Figure 1a of the article, where the x variable is uniformly distributed across the whole interval.

1.2. MI Correlation Coefficient

Let us consider a two-dimensional Gaussian random variable $(X,Y) \sim \mathrm{Normal}(0,\Sigma)$ with a covariance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

We may assume zero mean and unit marginal variance by the transformation invariance.

The entropy of an n-dimensional Gaussian random variable depends only on the determinant of the covariance matrix [2] (Theorem 8.4.1):

$$H(X_1, \ldots, X_d) = \frac{1}{2} \log((2\pi e)^n |\Sigma|).$$

From this it follows that

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = -\frac{1}{2} \log |\Sigma| = -\frac{1}{2} \log(1-\rho^2). \tag{S1}$$

Solving this for ρ gives

$$\rho = \pm \sqrt{1 - \exp(-2I(X;Y))},$$

from which we select the positive solution. It is not possible to extract the sign of the correlation, because the transformation $x \mapsto -x$ preserves MI, but inverts the Pearson correlation coefficient.

1.3. Partial Correlation

In this case, the normally distributed random variable is $(n+2)$-dimensional and its covariance matrix is

$$\Sigma = \begin{bmatrix} 1 & c & C_{XZ} \\ c & 1 & C_{YZ} \\ C_{XZ}^\top & C_{YZ}^\top & \Sigma_Z \end{bmatrix}.$$

This is a block matrix where the components denoted by a capital letter may be vectors. We denote by $\Sigma_{..}$ the marginal covariance matrices such as

$$\Sigma_{XZ} = \begin{bmatrix} 1 & C_{XZ} \\ C_{XZ}^\top & \Sigma_Z \end{bmatrix}.$$

The partial correlation can be expressed in terms of the precision matrix $P = \Sigma^{-1}$, the elements of which are cofactors of Σ:

$$\rho_{XY \cdot Z} = -\frac{P_{12}}{\sqrt{P_{11}P_{22}}}.$$

Conditional MI can be decomposed into a sum of (multidimensional) entropies, to which the submatrices of Σ can be plugged into. Constant terms cancel out, leaving

$$\begin{aligned} I(X;Y \mid Z) &= H(X,Z) + H(Y,Z) - H(Z) - H(X,Y,Z) \\ &= \frac{1}{2}\log\left(\frac{|\Sigma_{XZ}||\Sigma_{YZ}|}{|\Sigma_Z||\Sigma|}\right). \end{aligned}$$

Applying the normalization formula, we get

$$\begin{aligned} \rho_{I(X;Y|Z)} &= \sqrt{1 - \exp(-2I(X;Y \mid Z))} \\ &= \frac{\sqrt{|\Sigma_{XZ}||\Sigma_{YZ}| - |\Sigma_Z||\Sigma|}}{\sqrt{|\Sigma_{XZ}||\Sigma_{YZ}|}}. \end{aligned}$$

The denominator already matches that of partial correlation. It remains to show that the numerators are equal:

$$|P_{12}|^2 = |\Sigma_{XZ}||\Sigma_{YZ}| - |\Sigma_Z||\Sigma|. \tag{S2}$$

If Z is one-dimensional, the computation is easy to carry out. In the general case, we need to use block matrix results.

By a formula for block matrix determinant,

$$|P_{12}| = \begin{vmatrix} c & C_{YZ} \\ C_{XZ}^\top & \Sigma_Z \end{vmatrix} = |\Sigma_Z|(c - C_{YZ}\Sigma_Z^{-1}C_{XZ}^\top).$$

Doing the same computation for $|\Sigma_{XZ}|$, $|\Sigma_{YZ}|$ and $|\Sigma|$ (considering Σ a 2×2 block matrix, where Σ_Z is one block), we find that the $|\Sigma_Z|$ terms cancel out in (1). There remain some scalar terms and a determinant of

$$\Sigma_{XY} - \begin{bmatrix} C_{XZ} \\ C_{YZ} \end{bmatrix} \Sigma_Z^{-1} \begin{bmatrix} C_{XZ}^\top & C_{YZ}^\top \end{bmatrix},$$

which is easy to compute by block matrix multiplication and the definition of 2×2 determinant. To simplify the long expressions, it then remains to use the identities $C_{YZ}^{\top} C_{XZ} = C_{XZ}^{\top} C_{YZ}$ and $C_{XZ} \Sigma_Z^{-1} C_{YZ}^{\top} = C_{YZ} \Sigma_Z^{-1} C_{XZ}^{\top}$.

References

1. Ihara, S. *Information Theory for Continuous Systems*; World Scientific: Singapore, 1993.
2. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley-Interscience: Hoboken, NJ, USA, 2006.