CANONICAL ANALYSIS BASED ON MUTUAL INFORMATION

Allan A. Nielsen and Jacob S. Vestergaard

Technical University of Denmark
DTU Compute – Applied Mathematics and Computer Science
DK-2800 Kgs. Lyngby, Denmark

ABSTRACT

Canonical correlation analysis (CCA) is an established multivariate statistical method for finding similarities between linear combinations of (normally two) sets of multivariate observations. In this contribution we replace (linear) correlation as the measure of association between the linear combinations with the information theoretical measure mutual information (MI). We term this type of analysis canonical information analysis (CIA). MI allows for the actual joint distribution of the variables involved and not just second order statistics. While CCA is ideal for Gaussian data, CIA facilitates analysis of variables with different genesis and therefore different statistical distributions and different modalities. As a proof of concept we give a toy example. We also give an example with one (weather radar based) variable in the one set and eight spectral bands of optical satellite data in the other set.

1. INTRODUCTION

In 1936 Hotelling [1] introduced canonical correlation analysis (CCA). In CCA we find linear combinations $U = \boldsymbol{a}^T\boldsymbol{X}$ and $V = \boldsymbol{b}^T\boldsymbol{Y}$ of k variables in $\boldsymbol{X}$ and ℓ variables in $\boldsymbol{Y}$. The projections $\boldsymbol{a}$ and $\boldsymbol{b}$ are found such that U and V have maximum correlation and their variances equal one. Correlation is a linear measure of association between variables, and CCA is based on second order statistics only. CCA is therefore ideal for multivariate Gaussian data.

In this paper we replace correlation as the measure of association with the information theoretical measure mutual information (MI). In this type of analysis which we term canonical information analysis (CIA) we find $\boldsymbol{a}$ and $\boldsymbol{b}$ such that the MI between U and V is maximized, [2]. MI is an entropy based measure which allows for the actual joint distribution of U and V. It is therefore more suited for non-Gaussian data and for data with different statistical distributions and different modalities.

The idea of maximizing MI between two sets of variables is mentioned in [3]. However, the authors merely propose solutions to this problem based on independent component analysis in the individual spaces of the variables and they do not provide a truly canonical approach. In [4] and [5] the problem of maximizing MI of linear combinations of variables is solved in a manner which makes its application to small sample problems feasible. Our implementation is applicable to large sample problems including image data also.

Section 2 describes marginal and joint entropy as well as relative entropy (also known as the Kullback-Leibler divergence) and mutual information. Section 3 gives a toy example and a weather data example. Section 4 concludes the paper.

Parts of the abstract, the introduction, Section 2 and Subsection 3.1 are identical to sections in [6].

2. BASIC INFORMATION THEORY

In 1948 Shannon [7] published his now classical work on information theory. Below, we describe the information theoretical concepts entropy, relative entropy and mutual information for discrete stochastic variables, see also [8, 9, 10, 11].

2.1. Entropy

Consider a discrete stochastic variable X with probability density function (pdf) $p(X = x_i),\ i = 1, \ldots, n$, i.e, the probability of observing a particular realization x_i of stochastic variable X, where n is the number of possible outcomes or the number of bins. Let us look for a measure of information content (or surprise if you like) $h(X = x_i)$ in obtaining that particular realization. If x_i is a very probable value, i.e., $p(X = x_i)$ is high, we receive little information by observing x_i. If on the other hand x_i is a very improbable value, i.e., $p(X = x_i)$ is low, we receive much information by observing x_i. The measure of information content should be a monotonically decreasing function of p. This can be obtained by choosing for example $h \propto 1/p$.

If we observe independent realizations x_i and x_j, i.e., the two-dimensional pdf $p(X = x_i, X = x_j)$ equals the product of the one-dimensional marginal pdfs $p(X = x_i)p(X = x_j)$, we would like the joint information content to equal the sum of the marginal information contents, i.e., $h(X = x_i, X = x_j) = h(X = x_i) + h(X = x_j)$. This can be obtained by transformation by means of the logarithm.

Thus the desired characteristics of the measure of information or surprise can be obtained if we define $h(X = x_i)$

as

$$h(X = x_i) = \ln \frac{1}{p(X = x_i)} = -\ln p(X = x_i).$$

The expectation $H(X)$ of the information measure, i.e., the average amount of information obtained by observing the stochastic variable X, is termed the entropy

$$H(X) = -\sum_{i=1}^{n} p(X = x_i) \ln p(X = x_i).$$

In the limit where p tends to zero and $\ln p$ tends to minus infinity, $-p \ln p$ tends to zero. $H(X) = -\mathrm{E}\{\ln p(X)\}$ is nonnegative. A discrete variable which takes on one value only has zero entropy; a uniform discrete variable has maximum entropy (equal to $\ln n$). For the joint entropy of two discrete stochastic variables X and Y we get

$$H(X, Y) = -\sum_{i,j} p(X = x_i, Y = y_j) \ln p(X = x_i, Y = y_j).$$

Probability density functions, information content and entropy may be defined for continuous variables also (and so may relative entropy and mutual information mentioned below). In this case the entropy

$$H(X) = -\int p(x) \ln(p(x)) \mathrm{d}x \tag{1}$$

is termed differential entropy. Since $p(x)$ here may be greater than 1, $H(X)$ in the continuous case may be negative (or infinite).

2.2. Empirical Entropy

Empirical entropy $\hat{H}(X)$ is an estimator of $H(X)$ in (1). The estimator is defined as

$$\hat{H}(X) = -\frac{1}{N} \sum_{i=1}^{N} \ln p(X = x_i)$$

i.e., the average of $-\ln p$ defined over a finite sample $\{x_i\}_{i=1}^{N}$ of X, where N is the number of samples. This estimator is not based on any binning of the data.

2.3. Relative Entropy

The relative entropy also known as the Kullback-Leibler divergence [12] between two pdfs $p(X = x_i)$ and $q(X = x_i)$ defined on the same set of outcomes (or bins) is

$$D_{KL}(p, q) = \sum_i p(X = x_i) \ln \frac{p(X = x_i)}{q(X = x_i)}. \tag{2}$$

This is the expectation of the logarithmic difference between p and q. Typically p represents the "true" distribution of data or a precisely calculated theoretical distribution and q typically represents a model or an approximation of p. The relative entropy is a measure of the proximity of q and p, and it satisfies the so-called Gibbs' inequality $D_{KL} \geq 0$ with equality for $p(X = x_i) = q(X = x_i)$ only. The relative entropy is not symmetric in p and q (and therefore it is not a metric).

2.4. Mutual Information

The extent to which two discrete stochastic variables X and Y are not independent, which is a measure of their mutual information content, may be expressed as the relative entropy or the Kullback-Leibler divergence between the two-dimensional pdf $p(X = x_i, Y = y_j)$ and the product of the one-dimensional marginal pdfs $p(X = x_i)p(Y = y_j)$, i.e.,

$$D_{KL}(p(X, Y), p(X)p(Y)) = \sum_{i,j} p(X = x_i, Y = y_j) \ln \frac{p(X = x_i, Y = y_j)}{p(X = x_i)p(Y = y_j)}.$$

This sum defines the mutual information $I(X, Y) = D_{KL}(p(X, Y), p(X)p(Y))$ of the stochastic variables X and Y. Mutual information equals the sum of the two marginal entropies minus the joint entropy

$$I(X, Y) = H(X) + H(Y) - H(X, Y). \tag{3}$$

Unlike the general Kullback-Leibler divergence in (2) this measure is symmetric. Mutual information is always nonnegative, it is zero for independent stochastic variables only.

Obviously we need to estimate marginal as well as joint pdfs to obtain the mutual information estimate in (3). We employ kernel density estimation, which uses N data samples to estimate these pdfs. Mutual information is subsequently estimated using the same N data points. This is possible in practice only due to a very fast estimation of pdfs, see [2]. Note, that this is in contrast to [13] where the sample is divided into smaller portions in order to lessen the computational burden.

3. CASE STUDIES

We first give a toy example as a proof of concept. This is followed by an example with (one variate) weather radar data as one set and eight spectral bands from the SEVIRI instrument onboard the Meteosat satellite MSG-2 as the other set.

3.1. Toy Example

In a simple, illustrative example consider x and x^2. On the interval [0,1] the correlation between the two is $\sqrt{15/16}$, close to one. On the interval [–1,1] the correlation is zero, but of course the two are still functionally associated. Let us hide the parabola in noise: consider a variable x_1 sampled equidistantly on the interval [0,1]. Let another variable x_2 be random Gaussian noise with mean zero and standard deviation one.

Let y_1 be x_1^2 with random Gaussian noise with mean zero and standard deviation one tenth added. Let y_2 be random Gaussian noise with mean zero and standard deviation one. For all variables we have 1000 samples. Let the first set of variables consist of x_1 and x_2, and the second set consist of y_1 and y_2. In this case the leading canonical correlation is 0.9166 and (after sphering the input) the leading eigenvector for the first set is $[1.0000\ 0.0064]$ and for the second set $[1.0000\ 0.0143]$. So in this case canonical correlation analysis makes sense: we get a high canonical correlation and eigenvectors that isolate the signal in x_1 and y_1. Maximal mutual information is 0.7867 and the leading eigenvectors are $[1.0000\ 0.0075]$ and $[1.0000\ -0.0043]$ respectively.

Let us now redo the analysis with x_1 sampled equidistantly on the interval [–1,1]. In this case the leading canonical correlation is 0.0532 and the leading eigenvector for the first set is $[0.0391\ 0.9992]$ and for the second set $[-0.8955\ 0.4450]$. In this case canonical correlation analysis makes no sense: we get a very low canonical correlation and eigenvectors that do not isolate the signal in x_1 and y_1. Here maximal mutual information is 0.5856 and the leading eigenvectors are $[1.0000\ -0.0082]$ and $[1.0000\ -0.0086]$ respectively.

For the latter case (x_1 sampled equidistantly on the interval [–1,1]), three-dimensional contours of the estimated joint pdfs and scatter plots of the leading canonical variates are shown in Figure 1 top (correlation based) and bottom (mutual information based). The top figure reveals no structure whereas in the bottom figure we clearly recognize the noisy parabola originally in variables x_1 and y_1.

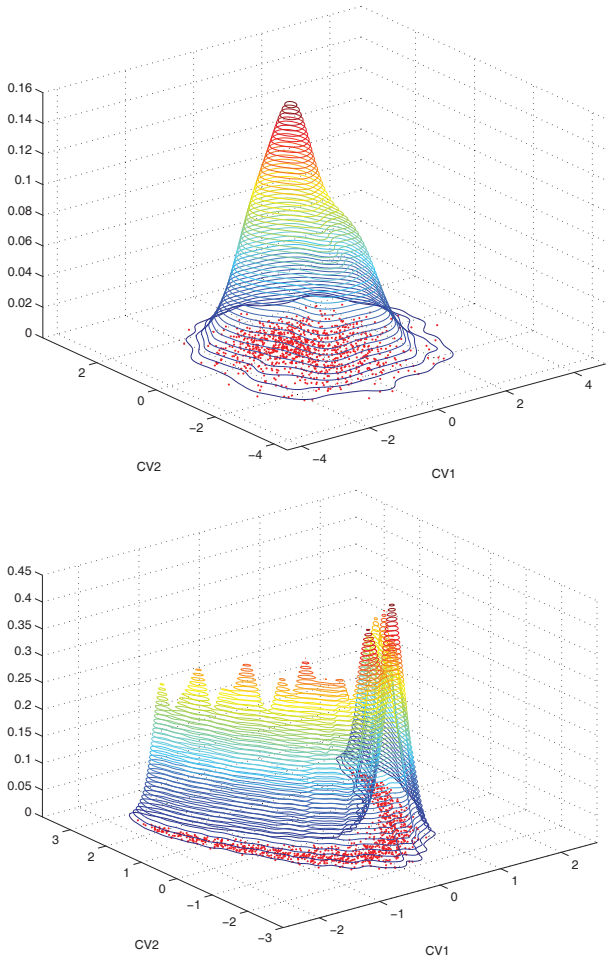

Fig. 1. 3D contours of estimated joint pdfs and scatter plots for leading canonical variates, correlation based (top) and mutual information based (bottom).

3.2. Weather Data

This data set consists of satellite and radar imagery from 20 August 2007, where extreme downpour intensities (53 millimeters in 10 minutes) were recorded in some regions of Denmark (inside the red rectangle).

The ultimate goal of this analysis (which is not dealt with in this paper) is to give a short term prediction (in the order of 30-60 minutes, the longer the better) of the extreme rain by means of the weather satellite data.

The satellite imagery is a set of $k = 8$ infrared bands from the Spinning Enhanced Visible and Infrared Imager (SEVIRI) onboard the Meteosat Second Generation (MSG-2) weather satellite. The spectral region of the infrared bands are from approximately 3.9μm to 13.4μm, and these bands monitor cloud top reflectance and emission properties. The radar data are recorded three minutes before the satellite image using the Danish Meteorological Institute (DMI) weather radars. These data consist of a single ($\ell = 1$) image of radar reflectance. The two image sources are gridded as images of 400×500 pixels with a ground sampling distance of 2 km $\times$ 2 km prior to analysis to establish pixel-to-pixel correspondence. The analysis includes the $N = 7,577$ observations in the radar imagery exhibiting reflectance from precipitation.

These data have also been treated in [2, 14]. In [14] an elaborate geometric and temporal alignment was needed to ameliorate the CCA solution. This is not needed when using the method suggested here.

Figure 2 shows the (first) canonical variate for the weather satellite data for both correlation and mutual information based solutions. The marked rectangular area is known from radar imagery to exhibit extreme rain at this particular point in time. The display range of the intensity values is within $\pm$ three standard deviations of the mean. The dashed white line marks the extent of the radar coverage. The bottom figure shows less noisy structures and a better discrimination between extreme rain areas and the remainder of the scene than the top figure.

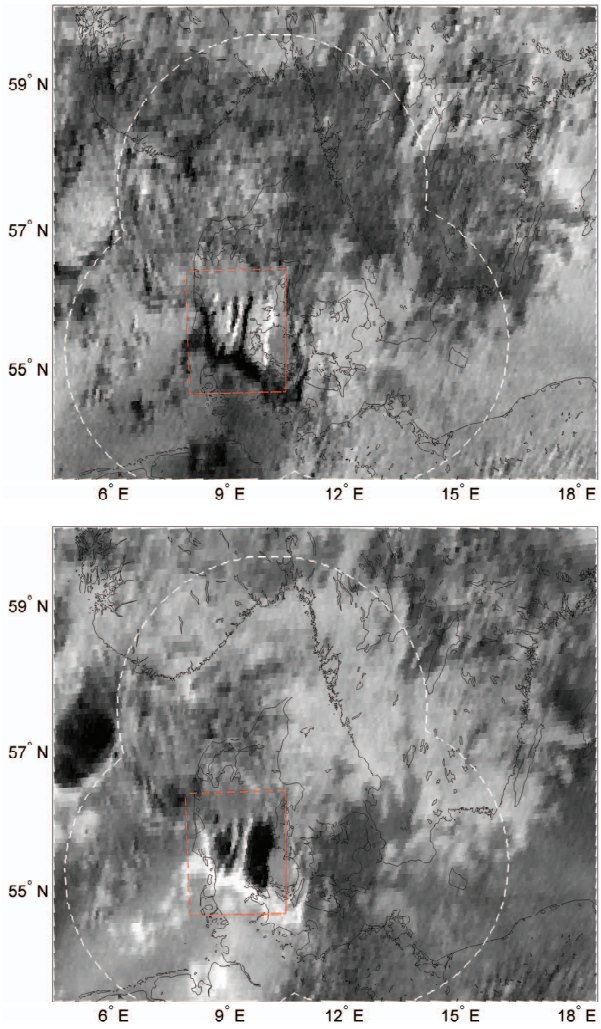

Fig. 2. Correlation (top) and mutual information based (bottom) based (first) canonical variates for the weather satellite data.

4. CONCLUSIONS

In the toy example the correlation based solution makes no sense on the interval [–1,1], whereas the mutual information based solution finds the noisy parabola in the variables analysed.

In the weather data case we see that the mutual information based canonical analysis offers less noise and a better discrimination between areas with extreme precipitation and the remainder of the area covered by the weather radar. The CIA solution provides a representation of the satellite data which carry the information most similar to that the weather radar data. This can be useful for, e.g., visualization purposes for meteorologists, or for providing pseudo-radar coverage outside the range of the radar. Also, it is hoped that a very short term prediction (in the order of 30-60 minutes) from the satellite data of very heavy rain can be obtained.

Other examples (not shown here) give a similarly better performance for the mutual information based analysis.

5. REFERENCES

[1] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. XXVIII, pp. 321–377, 1936.

[2] J. S. Vestergaard and A. A. Nielsen, "Canonical information analysis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 101, pp. 1–9, 2015, http://www.imm.dtu.dk/pubdb/p.php?6270, Matlab code at https://github.com/schackv/cia.

[3] T. de Bie and B. de Moor, "On two classes of alternatives to canonical correlation analysis, using mutual information and oblique projections," in *Proceedings of the 23rd symposium on information theory in the Benelux (ITB)*, Louvain-la-Neuve, Belgium, 2002.

[4] X. Yin, "Canonical correlation analysis based on information theory," *Journal of Multivariate Analysis*, vol. 91, pp. 161–176, 2004.

[5] M. Karasuyama and M. Sugiyama, "Canonical dependency analysis based on squared-loss mutual information," *Neural Networks*, vol. 34, pp. 46–55, 2012.

[6] A. A. Nielsen and J. S. Vestergaard, "Change detection in bi-temporal data by canonical information analysis," in *8th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, Annecy, France, 2015, http://www.imm.dtu.dk/pubdb/p.php?6888, Matlab code at https://github.com/schackv/cia.

[7] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423 and 623–656, 1948.

[8] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, J. Wiley, 2001.

[9] D. J. C. Mackay, *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.

[10] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2007.

[11] M. J. Canty, *Image Analysis, Classification and Change Detection in Remote Sensing. With Algorithms for ENVI/IDL and Python*, Taylor & Francis, CRC Press, third edition, 2014.

[12] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[13] P. Viola, "Alignment by maximization of mutual information," *International Journal of Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.

[14] J. S. Vestergaard and A. A. Nielsen, "Automated invariant alignment to improve canonical variates in image fusion of satellite and weather radar data," *Journal of Applied Meteorology and Climatology*, vol. 52, pp. 701–709, 2012.