# Change detection in bi-temporal data by canonical information analysis

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. Where CCA is ideal for Gaussian data, CIA facilitates analysis of variables with different genesis and therefore different statistical distributions. As a proof of concept we give a toy example. We also give an example with DLR 3K camera data from two time points covering a motor way.**

## I. 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 $\ell$ 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 II describes marginal and joint entropy as well as relative entropy (also known as the Kullback-Leibler divergence) and mutual information. Section III very briefly mentions convolution based approximate entropy estimation. Section IV sketches some aspects of mutual information maximization. Section V gives a toy example as a proof of concept and a change detection example with DLR 3K camera [6], [7] images from two time points. Section VI concludes the paper.

Parts of the abstract, the introduction, Section II and Subsection V-A are identical to sections in [8].

## II. Basic Information Theory

In 1948 Shannon [9] 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 [10], [11], [12], [13].

### A. 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

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

978-1-4673-7119-3/15/$31.00 

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

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

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

### B. Empirical Entropy

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

$$\hat{H}(X) = -\frac{1}{N}\sum_{i=1}^{N}\ln p(X = x_i) \tag{2}$$

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.

### C. Relative Entropy

The relative entropy also known as the Kullback-Leibler divergence [14] between two pdfs $p(X = x_i)$ and $q(X = x_i)$ defined on the same set of outcomes (or bins) is

$$D_{KL}(p,q) = \sum_i p(X = x_i)\ \ln\frac{p(X = x_i)}{q(X = x_i)}. \tag{3}$$

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).

### D. 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

$$I(X,Y) = H(X) + H(Y) - H(X,Y). \tag{4}$$

Unlike the general Kullback-Leibler divergence in (3) 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 (4). 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. Note, that this is in contrast to [15] where the sample is divided into smaller portions in order to lessen the computational burden.

## III. Approximate Entropy Estimation

Estimation of marginal and joint entropies is the main bottleneck in maximization of mutual information. Since it is based on pairwise distances, it has a computational complexity in the order of $\mathcal{O}(N^2)$. In [16] a fast approximate marginal (1D) entropy estimator with a complexity in the order of $\mathcal{O}(N\log N)$ is proposed. For the purpose of canonical information analysis we generalize this approximate entropy estimator to joint entropy (2D).

Approximate entropy estimation is a convolution based modification of Parzen window density estimation. Convolutions can run in the order of $O(N\log N)$ on a regular grid. The estimation procedure therefore (1) quantizes the irregular samples to a regular grid, (2) convolves with a Gaussian kernel on this grid, and (3) interpolates back onto the original positions of the samples to get an estimate of the empirical entropy in (2). See also [2].

## IV. Maximization of Mutual Information

The kernel density estimates of one- and two-dimensional pdfs by means of the method sketched above are independent of additive and multiplicative transformations of each of the original variables. Therefore the maximization of the mutual information between the two linear combinations can be carried out without constraints. This means that very many optimization schemes may be applied.

Maximization of mutual information is inherently non-convex. For problems where it is not crucial to converge to the global optimum we suggest to use a local solver, e.g., either the downhill simplex method [17] or Newton's method with the BFGS update [18] depending on whether one wishes to rely purely on function values or whether one wants to include gradient information also. For problems where convergence to the global optimum is important, we propose to use a genetic algorithm at the cost of significantly more function evaluations, see for example [19].

The choice of starting point is crucial when using local methods for global optimization. We have experimented with two different sets of starting points for each case, one being the optimum determined by canonical correlation analysis. The second set of starting points is constructed by letting the initial projections be unit vectors of length $k$ and $\ell$ respectively, with an equal weighting on all variables. It is often a good strategy to use several different starting points.

## V. Case Studies

We first give a toy example as a proof of concept. This is followed by a change detection example with bi-temporal image data from the DLR 3K camera system [6], [7].

### A. 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 left figure reveals no structure whereas in the right figure we clearly recognize the noisy parabola originally in variables $x_1$ and $y_1$.

### *B. DLR 3K Camera Data*

The images used in this example were recorded with the airborne DLR 3K camera system [6], [7] from the German Aerospace Center, DLR. This system consists of three commercially available 16 megapixel cameras arranged on a mount and a navigation unit with which it is possible to record time series of images covering large areas at frequencies up to 3 Hz. The 1000 rows by 1000 columns example images acquired 0.7 seconds apart cover a busy motorway. These data have previously been treated in [2], [20], [21]. The original RGB images can be seen in [21]. The data at the two time points were orthoprojected using global positioning system/inertial measurement unit (GPS/IMU) measurements and a digital elevation model (DEM). For flat terrain like here one pixel accuracy was obtained. In these data, the change occurring between the two time points will be dominated by the movement of the cars on the motorway. Undesired, apparent change will occur due to the movement of the aircraft and the different viewing positions at the two time points.

Using canonical information analysis as a tool for change detection, Figure 2 bottom shows the difference image between the first set of mutual information canonical variates (MICVs). Previously, a method for change detection based on canonical correlation analysis termed MAD has been proposed [22]. Comparing with the solution obtained by canonical correlation analysis in Figure 2 top it is evident that better change information is obtained by using CIA: the background is much smoother and clearly distinguishable from the areas of change (the cars) and the extreme values are present only where change has actually occurred. An iterated version of the MAD method was reported on in [23]. For space limitation reasons a comparison with results from this extension of MAD is not shown here.

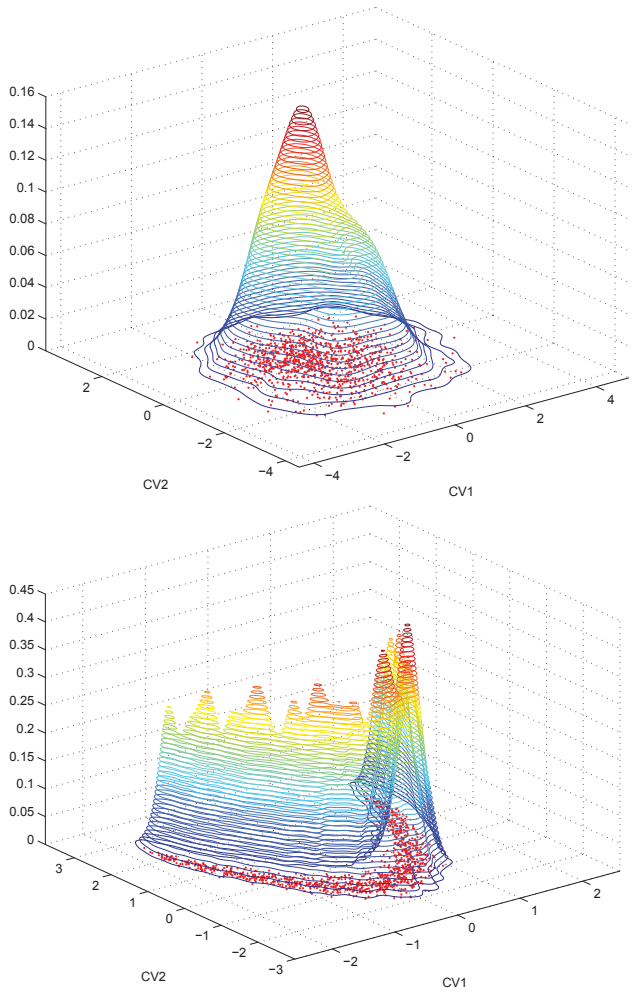


Fig. 1. 3D contours of estimated joint pdfs and scatter plots for leading canonical variates, correlation based (top) and mutual information based (bottom).

To quantify the difference between the solutions, a region marked by a red rectangle in the canonical difference image has been selected. This region is known not to have changed between the two acquisition times. The variance in this region for the solution produced by CIA is 0.265, while it is 0.878 for the correlation based solution, i.e., the ratio is 3.319. This verifies the subjective evaluation that a more homogeneous no-change background is obtained using the proposed mutual information based method. A correlation of 0.982 and 0.945 between the leading pair of canonical variates was obtained using CCA and CIA respectively, which demonstrates that a high correlation is not always the best measure for similarity. A mutual information of 1.034 and 1.335 between the leading pair of canonical variates was obtained using CCA and CIA respectively.

## VI. 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 DLR 3K camera case we see that the mutual information based canonical analysis offers less noise and a better discrimination between moving cars and the remainder of the image.

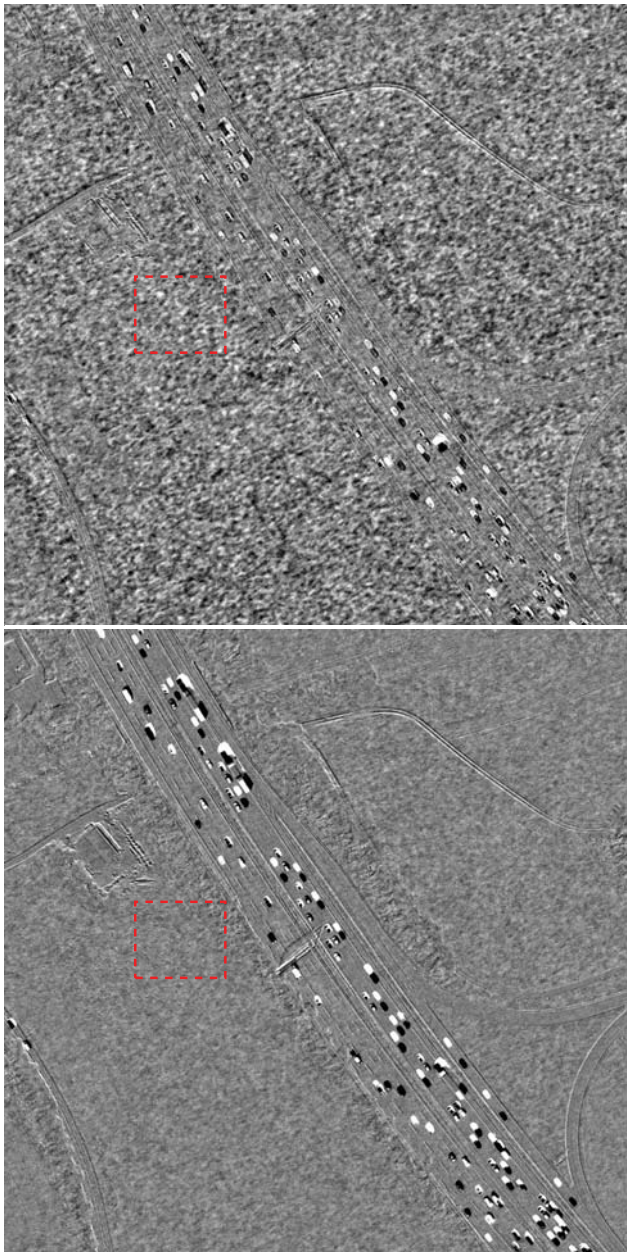

Fig. 2. Difference images of the first set of MICVs for DLR 3K data using canonical correlation analysis (top) and canonical information analysis (bottom) respectively. The display range of the intensity values is within $\pm$ three standard deviations of the mean. The marked region is used to quantify the no-change noise variance.

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

## 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] F. Kurz, B. Charmette, S. Suri, D. Rosenbaum, M. Spangler, A. Leonhardt, M. Bachleitner, R. Stätter, and P. Reinartz, "Automatic traffic monitoring with an airborne wide-angle digital camera system for estimation of travel times," in *Photogrammetric Image Analysis, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Service (PIA07), Munich, Germany*, 2007.

[7] F. Kurz, R. Müller, M. Stephani, P. Reinartz, and M. Schroeder, "Calibration of a wide-angle digital camera system for near real time scenarios," in *Proceedings of ISPRS Hannover Workshop 2007-High Resolution Earth Imaging for Geospatial Information*, 2007, pp. 1682–1777.

[8] A. A. Nielsen and J. S. Vestergaard, "Canonical analysis based on mutual information," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Milan, Italy, 2015, http://www.imm.dtu.dk/pubdb/p.php?6881, Matlab code at https://github.com/schackv/cia.

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

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

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

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

[13] 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.

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

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

[16] S. Shwartz, M. Zibulevsky, and Y. Schechner, "Fast kernel entropy estimation and optimization," *Signal Processing*, vol. 85, no. 5, pp. 1045–1058, May 2005.

[17] J. A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[18] R. Fletcher, "A new approach to variable metric algorithms," *The Computer Journal*, vol. 13, no. 3, pp. 317–322, 1970.

[19] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.

[20] A. A. Nielsen and M. J. Canty, "Kernel principal component and maximum autocorrelation factor analyses for change detection," in *SPIE Europe Remote Sensing*, 2009, vol. 7477, http://www.imm.dtu.dk/pubdb/p.php?5757.

[21] A. A. Nielsen, "Kernel Maximum Autocorrelation Factor and Minimum Noise Fraction Transformations," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 612–624, 2011, http://www.imm.dtu.dk/pubdb/p.php?5925.

[22] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sensing of Environment*, vol. 64, no. 1, pp. 1–19, 1998, http://www.imm.dtu.dk/pubdb/p.php?1220.

[23] A. A. Nielsen, "The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 463–478, Feb. 2007, http://www2.imm.dtu.dk/pubdb/p.php?4695.