ISPRS Journal of Photogrammetry and Remote Sensing 101 (2015) 1–9

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Canonical information analysis

Jacob Schack Vestergaard *, Allan Aasbjerg Nielsen

Department of Applied Mathematics and Computer Science, Technical University of Denmark, Asmussens Alle, Building 324, DK-2800 Lyngby, Denmark

ARTICLE INFO

Article history:
Received 30 April 2014
Received in revised form 3 November 2014
Accepted 11 November 2014

Keywords:
Information theory
Probability density function estimation
Parzen windows
Entropy
Mutual information maximization
Canonical mutual information analysis
CIA
Approximate entropy

ABSTRACT

Canonical correlation analysis is an established multivariate statistical method in which correlation between linear combinations of multivariate sets of variables is maximized. In canonical information analysis introduced here, linear correlation as a measure of association between variables is replaced by the information theoretical, entropy based measure mutual information, which is a much more general measure of association. We make canonical information analysis feasible for large sample problems, including for example multispectral images, due to the use of a fast kernel density estimator for entropy estimation. Canonical information analysis is applied successfully to (1) simple simulated data to illustrate the basic idea and evaluate performance, (2) fusion of weather radar and optical geostationary satellite data in a situation with heavy precipitation, and (3) change detection in optical airborne data. The simulation study shows that canonical information analysis is as accurate as and much faster than algorithms presented in previous work, especially for large sample sizes. URL: http://www.imm.dtu.dk/pubdb/p.php?6270

1. Introduction

In canonical correlation analysis (CCA) first published by Hotelling in 1936 (Hotelling, 1936) linear combinations $U = \mathbf{a}^T\mathbf{X}$ and $V = \mathbf{b}^T\mathbf{Y}$ of two sets of stochastic variables, k-dimensional $\mathbf{X}$ and ℓ-dimensional $\mathbf{Y}$, which maximize correlation between U and V are found. Correlation considers second order statistics of the involved variables only and as such it is ideal for Gaussian data. In this paper we investigate replacement of correlation with mutual information (Hyvärinen et al., 2004; Mackay, 2003; Bishop, 2007; Canty, 2010) which is a more general, information theoretical, entropy based measure of association between variables. Entropy and mutual information (MI) depend on the actual probability density functions of the involved variables and thus on higher order statistics. The resulting method is termed canonical mutual information analysis, or in short canonical information analysis (CIA).

Since multi-source data, which is typically of different genesis, often follow very different (non-Gaussian) distributions, the application of MI facilitates analysis of such data. In one of our examples we apply the method to a joint analysis of radar and optical data (which follow very different distributions thus rendering CCA non-optimal). Other areas where the method could potentially be very useful include data of different modalities, for example SAR, LiDAR, optical and medical data. In general, this type of analysis has a strong potential for application in data fusion and other fields of data integration, see also (Ehlers, 1991; Pohl and Van Genderen, 1998; Conese and Maselli, 1993).

Mutual information as a measure of association has previously proven useful in the context of image registration. Studholme et al. (1999) proposed a normalized variant of MI for registration of medical images, which Suri and Reinartz (2010) employ for automatic registration of SAR and optical images. For the purpose of change detection, Erten et al. (2012) derive an analytical expression for the mutual information between temporal multichannel SAR images.

Other dependence measures have been considered in the literature, such as kernel canonical correlation analysis (kCCA) (Lai and Fyfe, 2000; Bach and Jordan, 2002). However, while kernel methods do indeed provide an implicit nonlinear transformation of the data maximizing some dependence measure, they do not possess the same qualities as linear methods in terms of interpretation. Specifically, a linear method, such as CIA, finds the actual functional relation between the original variables, where a kernel method, such as kCCA, would find a hidden/intrinsic transformation which makes the relation between CVs linear. This property of the linear solution immediately eases interpretation of the result.

* Corresponding author.
E-mail addresses: jsve@dtu.dk (J.S. Vestergaard), alan@dtu.dk (A.A. Nielsen).
URL: http://www.compute.dtu.dk/~jsve (J.S. Vestergaard).

http://dx.doi.org/10.1016/j.isprsjprs.2014.11.002
0924-2716/

The idea of maximizing MI between two sets of variables is mentioned by Bie and Moor (2002). However, the authors only 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. Yin (2004) and Karasuyama and Sugiyama (2012) solve the problem of maximizing MI of linear combinations of variables in a manner which makes its application to small sample problems feasible. In practical terms the solutions offered are not applicable to large sample problems including for example image data. Our fast grid-based entropy estimator (Section 5) facilitates the use of CIA to large sample problems. Both Yin (2004) and Karasuyama and Sugiyama (2012) request orthogonality between solutions (as in CCA), whereas we allow for oblique solutions (Section 2) via a structure removal procedure inspired by Friedman's projection pursuit (Friedman, 1987). The well known difficulties in estimating and optimizing entropy measures, will be addressed in Sections 4–6.

Below, Section 2 describes the concept of canonical information analysis and motivates the following sections. Section 3 describes the information theoretical concepts entropy of a univariate stochastic variable, joint entropy of two stochastic variables, relative entropy, and mutual information. Section 4 briefly describes the estimation of one- and two-dimensional probability density functions, Section 5 describes approximate entropy estimation, and Section 6 describes the maximization of mutual information of two linear combinations of stochastic variables. Section 7 gives (1) a simple, illustrative toy example, (2) a case study with weather radar data and optical data from a meteorological satellite, and (3) a case with change detection in optical airborne data. Section 8 concludes. An appendix is included, motivating some of the implementation choices made. Supplementary material is provided with additional simulation studies and results from the two case studies plus an extra application of CIA for change detection.

2. Canonical information analysis

Inspired by canonical correlation analysis (Hotelling, 1936) we propose a method for maximizing mutual information between the linear combinations $U = \mathbf{a}^T\mathbf{X}$ and $V = \mathbf{b}^T\mathbf{Y}$ of two sets of stochastic variables, k-dimensional $\mathbf{X}$ and ℓ-dimensional $\mathbf{Y}$.

The goal of CIA can be stated as

$$\mathbf{a}^\star,\ \mathbf{b}^\star = \arg\max_{\mathbf{a},\mathbf{b}} I(U,V) \tag{1}$$

where $I(U,V)$ is the mutual information between the two linear combinations U and V which can be defined as

$$I(U,V) = h(U) + h(V) - h(U,V) \tag{2}$$

where $h(U)$ and $h(V)$ are the marginal entropies and $h(U,V)$ the joint entropy. This will be detailed further in Sections 3–5.

Maximization of mutual information is known to be a non-convex optimization problem (Modersitzki, 2004; Haber and Modersitzki, 2007) wherefore we have conducted experiments with local as well as global optimization methods, see Section 6. The inherent lack of certainty of finding a global optimum will be elucidated by application of the method to different real world multispectral decomposition problems, see Section 7.

In canonical correlation analysis k and ℓ linear combinations (components) are determined with the criterion that the i'th component maximizes correlation between U and V while being orthogonal to the first $i-1$ components. Friedman (1987) introduced in projection pursuit 'structure removal' as the solution to avoid re-finding a previously found direction in space. Structure removal works by histogram equalization of the projected data to a Gaussian distribution and transforming back to the original space. In CIA we choose to adopt this principle of structure removal with the modification that the projected data U and V are substituted with uniformly distributed white noise. This modification is necessary since, in contrast to projection pursuit, CIA does not maximize non-Gaussianity of one projection, but rather it maximizes statistical dependence between two projections. This structure removal replaces the orthogonality requested by Yin (2004) and Karasuyama and Sugiyama (2012).

3. Basic information theory

In 1948 Shannon (Shannon, 1948) published his now classical work on information theory. Below, we describe the information theoretical concepts entropy and mutual information for discrete and continuous stochastic variables, see also (Hyvärinen et al., 2004; Mackay, 2003; Bishop, 2007; Canty, 2010).

3.1. Discrete variables

Consider a discrete stochastic variable X with probability density function (pdf) $p(X = x_i),\ i = 1,\ldots,N$. The information content is defined as $-\ln(p(X = x_i))$. The expectation $H(X)$ of the information content is termed the entropy of the stochastic variable X

$$H(X) = -\sum_{i=1}^{N} p(X = x_i)\ln(p(X = x_i)). \tag{3}$$

For the joint entropy of two discrete stochastic variables X and Y we get

$$H(X,Y) = -\sum_{i,j} p(X = x_i, Y = y_j)\ln(p(X = x_i, Y = y_j)). \tag{4}$$

3.2. Continuous variables

Probability density functions, information content and entropy may be defined for continuous variables also. This is necessary to represent linear combinations of sampled data. In this case the entropy

$$h(X) = -\int p(x)\ln(p(x))\mathrm{d}x \tag{5}$$

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

Empirical entropy $\hat{h}(X)$ is an estimator of $h(X)$ in (5). The estimator is defined as

$$\hat{h}(X) = -\frac{1}{N}\sum_{i=1}^{N}\ln(p(X = x_i)) \tag{6}$$

and as such it is defined over a finite sample $\{x_i\}_{i=1}^N$ of X, where N is the number of samples. As opposed to (3) and (4) this estimator is not based on any binning of the data.

Empirical entropy has previously proven useful for manipulating entropy measures (Viola, 1995). We have experienced this experimentally (not shown here) and find this estimator useful for canonical information analysis.

The extent to which two continuous 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 $p(x,y)$ and the product of the one-dimensional marginal pdfs $p(x)p(y)$, i.e.,

$$D_{KL}(p(x,y), p(x)p(y)) = \int\int p(x,y)\ln\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\,\mathrm{d}y. \tag{7}$$

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

$$I(X,Y) = h(x) + h(y) - h(x,y). \tag{8}$$

Unlike the general Kullback–Leibler divergence this measure is symmetric. Mutual information is always nonnegative, it is zero for independent stochastic variables only.

We need to estimate marginal as well as joint pdfs to obtain the mutual information estimate in (8). Karasuyama and Sugiyama (2012) estimate the ratio in (7) directly. 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 our very fast estimation of pdfs which will be described in Section 5. Note, that this is in contrast to Viola (1997) where the sample is divided into smaller portions in order to lessen the computational burden and to Yin (2004) where an explicit estimation is used that does not scale well to image analysis problems and other large sample problems.

4. Density estimation

The histogram is a simple non-parametric density estimator. However, the estimated histogram is not smooth and it depends on the end points of bins and the width of bins. By using kernel density estimators (Rosenblatt, 1956; Parzen, 1962; Silverman, 1986) where we center a kernel on each observation, we may obtain smoother histograms that do not depend on bin end points. The kernel density estimator (Parzen windows estimator) for the pdf of X at value t is

$$\hat{p}(X = t|\mathbf{x}) = \frac{1}{N\sigma}\sum_{i=1}^{N}\varphi\left(\frac{t - x_i}{\sigma}\right) \tag{9}$$

where $\mathbf{x} = \{x_i\}_1^N$ is a vector of realizations of X, $\varphi(z)$ is the kernel and σ a smoothing parameter referred to as the bandwidth. Often we choose the Gaussian kernel

$$\varphi(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}z^2\right). \tag{10}$$

The width of the Gaussian, i.e., the standard deviation is thus equivalent to the bandwidth σ.

The kernel density estimator assumes continuous distributions, thus we estimate continuous variants of the information theoretic measures mentioned in Section 1. Since only two one-dimensional projections of the data are considered, the known problems with kernel density estimators in higher dimensions (Beirlant et al., 1997; Kraskov et al., 2004) are found to be negligible for canonical information analysis.

In two dimensions the bivariate Gaussian is often chosen to have a diagonal covariance matrix leaving two parameters to be estimated, namely the bandwidth in each direction. Estimation of the bandwidth is an example of the bias-variance trade-off: a too narrow kernel causes too large variation in the density estimate and a too wide kernel oversmooths the estimated distribution (Jones and Marron, 1996).

Here we use a data-driven bandwidth selection method based on the maximal smoothing principle (Terrell, 1990). This method is known to be conservative (oversmoothing) by nature (Jones and Marron, 1996; Terrell, 1990), but this is outweighed by fulfilling two – in this context – more important properties: the bandwidth estimate is stable, i.e., it varies smoothly for small changes in projection direction of the data. Experiments (see Appendix A) have shown that this is not the case for, e.g., neither the linear diffusion process based method by Botev et al. (2010) nor for Sheather–Jones (Sheather and Jones, 1991). The second property is computational speed, where it outperforms the commonly preferred "solve-the-equation plug-in" method (Sheather and Jones, 1991). Speed is of practical importance as the density estimation will be part of calculating the objective value for a non-convex optimization problem, wherefore the bandwidth will be estimated repeatedly. This is especially true for large problems, e.g., image processing.

5. Approximate entropy estimation

Estimation of marginal and joint entropies is the main bottleneck in maximization of mutual information. Parzen window density estimation, in the explicit form presented above, has previously been used for this purpose, see e.g. Yin (2004). However, since it is based on pairwise distances, it has a computational complexity in the order of $\mathcal{O}(N^2)$. Shwartz et al. (2005) proposed a fast approximate marginal (1D) entropy estimator with a complexity in the order of $\mathcal{O}(N \log N)$. For the purpose of canonical information analysis we generalize this approximate entropy estimator to joint entropy (2D). This is described below and illustrated in Fig. 1.

Approximate entropy estimation is a convolution based modification of Parzen window density estimation. Convolution of the samples with the kernel in (10) is equivalent to the density estimation in (9). Convolutions can run in the order of $\mathcal{O}(N \log N)$ 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 samples' original positions to get an estimate of the empirical entropy in (6).

Quantization requires choosing a discretization, i.e., a number of bins B^2 and a domain $[x_a, x_b] \times [y_a, y_b]$ over which to discretize. The (m,n)'th bin in this regular grid is positioned at $(x,y)_{m,n} = (x_a + m\Delta x, x_b + n\Delta y)$ where $m,\ n \in \{0,\ldots,B-1\}$, $\Delta x = \frac{x_b - x_a}{B-1}$ and $\Delta x = \frac{y_b - y_a}{B-1}$. The i'th sample point falls into a cell spanned by the four bin centers with indices (m_i, n_i), $(m_i + 1, n_i)$, $(m_i, n_i + 1)$, $(m_i + 1, n_i + 1)$ where $m_i = \mathrm{fl}\left[\frac{x_i - x_a}{\Delta x}\right]$ and $n_i = \mathrm{fl}\left[\frac{y_i - y_a}{\Delta y}\right]$ and $\mathrm{fl}[\cdot]$ is the floor operation. The weights for each of these four bin centers are given by a bilinear interpolation scheme:

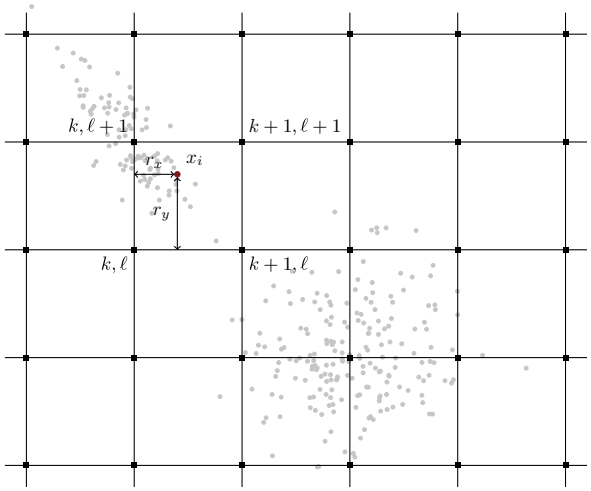


Fig. 1. Quantization. Illustration of bilinear quantization of samples to a regular grid to enable fast approximate joint entropy estimation. The gray dots are examples of irregular samples and the red dot is used to exemplify the bilinear weights. The black rectangles indicate the bins and the indices of the four bins influenced by the red dot are shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$w_i(m_i, n_i) = (1 - r_x)(1 - r_y)$$
$$w_i(m_i + 1, n_i) = r_x(1 - r_y)$$
$$w_i(m_i, n_i + 1) = (1 - r_x)r_y$$
$$w_i(m_i + 1, n_i + 1) = r_x r_y$$

where $r_x = \frac{x_i - x_a}{\Delta x} - m_i$ and $r_y = \frac{y_i - y_a}{\Delta y} - n_i$, i.e., the fraction removed by the floor operation. The quantized value $Q_{m,n}$ at a given bin is thus a weighted count of samples in the proximity of the bin. The quantization is collected in a $B \times B$ image-like matrix $\mathbf{Q}$. This bilinear weighting is the 2D analogue of the linear weighting suggested by Shwartz et al. (2005).

Convolution of the quantized signal on the regular grid with the kernel φ from (10)

$$\hat{\mathbf{Q}} = \varphi * \mathbf{Q}$$

can be performed in the order of $\mathcal{O}(B^2 \log B^2)$, i.e., dependent on the number of bins rather than the number of samples. The resulting (m, n)'th element of $\hat{\mathbf{Q}}$ is an estimate of the density at the (m, n)'th bin. Distributing this estimate back onto the original sample positions is done using the same weights as earlier, such that

$$\hat{p}(x_i) = Q_{m_i,n_i} w_i(m_i, n_i) + Q_{m_i+1,n_i} w_i(m_i + 1, n_i) + Q_{m_i,n_i+1} w_i(m_i, n_i + 1) + Q_{m_i+1,n_i+1} w_i(m_i + 1, n_i + 1).$$

This is an approximation of (9) and can be plugged directly into (6). The complexity of the quantization is linear in the number of samples, thus the complexity of the estimation is $\mathcal{O}(N + B^2 \log B^2)$. Unlike estimates of discrete entropy, the estimate of empirical entropy is not dependent on the choice of B^2, since the summation over probabilities is carried out over the sample positions, rather than the bins. The choice does, however, influence the accuracy of the approximation.

Shwartz et al. (2005) also provides a gradient of the marginal entropy estimate, which we have generalized to joint entropy. The marginal entropy gradient is given with respect to the samples $\frac{\partial H}{\partial (\mathbf{a}^T \mathbf{X})}$. For the purpose of canonical information analysis the gradient with respect to the linear weighting $\mathbf{a}$ is needed. The chain rule yields

$$\frac{\partial h_x}{\partial \mathbf{a}} = \frac{\partial h_x}{\partial (\mathbf{a}^T \mathbf{X})} \frac{\partial (\mathbf{a}^T \mathbf{X})}{\partial \mathbf{a}} = \frac{\partial h_x}{\partial (\mathbf{a}^T \mathbf{X})} \mathbf{X}^T.$$

This is completely analogous for joint entropy estimation and the reader is referred to Shwartz et al. (2005) for further details.

The computational complexity of the approximate gradient estimation is of the order $\mathcal{O}(B^2 \log B^2 + N N_{\text{dim}})$ where N_{dim} is the dimensionality of the linear weighting, i.e., either k or ℓ. In comparison, explicit calculation of the entropy gradient is of complexity $\mathcal{O}(N_{\text{dim}} N^2 + N)$ (Shwartz et al., 2005).

6. Maximization of mutual information

The kernel density estimates of one- and two-dimensional pdfs by means of the method sketched in Section 4 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 (Nelder and Mead, 1965) or Newton's method with the BFGS update (Fletcher, 1970), depending on whether one wishes to rely purely on function values or leverage the gradient introduced above. 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. Results shown below are obtained using the genetic algorithm implemented in MATLAB with a population size of $5(k + \ell)^2$.

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 $\mathbf{a}_0$ and $\mathbf{b}_0$ be unit vectors of length k and ℓ respectively, with an equal weighting on all variables, such that

$$\mathbf{a}_0 = \frac{1}{\sqrt{k}} \mathbf{1}_k, \quad \mathbf{b}_0 = \frac{1}{\sqrt{\ell}} \mathbf{1}_\ell \tag{11}$$

where $\mathbf{1}_n$ is an n-vector of ones. For some problems, several candidate starting points may exist in which case we suggest to employ an optimization strategy where multiple local solvers start from individual starting points.

7. Case studies

Here we give an illustrative toy example, an example which fuses weather radar and optical geostationary satellite data for a situation with heavy precipitation, and an example of using canonical information analysis for change detection in optical airborne data. These examples will be referred to as *toy*, *weather* and *cars* respectively for brevity.

The results are summarized in Table 2. Higher order components for these data sets were found to be trivial, wherefore only the leading component is shown.

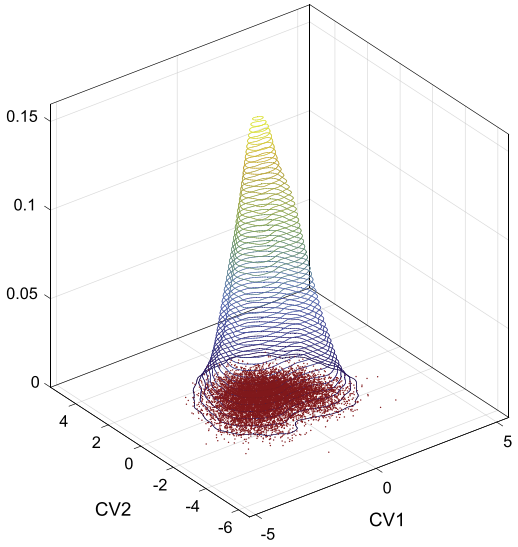
7.1. Toy example

In a simple, illustrative example consider the functions $f(x) = x$ and $g(x) = x^2$. The correlation between the functions over the interval [0, 1] is $\sqrt{15/16} = 0.9682$, close to one. The correlation between the two over the interval $[-1, 1]$ is zero and yet of course the two variables are still closely associated.

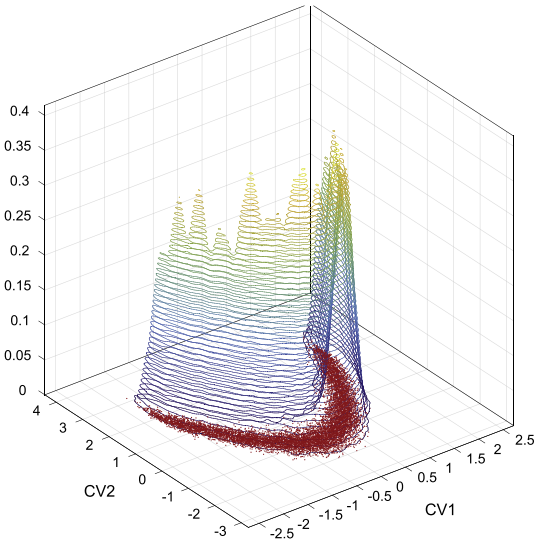
Consider now this numeric example with a variable x_1 sampled equidistantly on the interval [0, 1]. Let another variable x_2 be random Gaussian noise with mean zero and standard deviation one. Let y_1 be x_1^2 with random Gaussian noise with mean zero and standard deviation one tenth added. Let y_2 be 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 projection vectors 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 projection vectors 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 contour and scatter plots of the leading canonical variates are shown in Fig. 2a (correlation based) and b



(a) Canonical correlation analysis



(b) Canonical information analysis

Fig. 2. Toy example. (a) Correlation based canonical variates and (b) mutual information based canonical variates for toy example with variables sampled equidistantly on the interval $[-1, 1]$.

(mutual information based). Fig. 2a reveals no structure but in Fig. 2b we clearly recognize the noisy parabola originally in variables x_1 and y_1. Unlike maximization of correlation of linear combinations of the two sets of variables, maximization of mutual information gives meaningful results in both cases.

We compare CIA to the 'explicit' (e.g., Yin, 2004) estimation of maximal mutual information projections performance in terms of accuracy and computation time. The accuracy is evaluated in terms of the geometric mean

$$\mu = \sqrt{|\rho_1||\rho_2|} \tag{12}$$

of the absolute correlations $\rho_1 = \mathrm{corr}(x_1, U)$ and $\rho_2 = \mathrm{corr}(y_1, V)$. E.g., the correct $\mathbf{a}^\star = \mathbf{b}^\star = [1, 0]^T$ would yield $\rho_1 = \rho_2 = \mu = 1$. Fig. 3b shows the difference in geometric mean $\mu_{\mathrm{CIA}} - \mu_{explicit}$ for three different sample sizes $N = \{500, 1000, 5000\}$ and for ten values of the standard deviation σ for the noise added to x_1^2 to form y_1. We see that in low-noise cases ($\sigma < 0.6$) the difference in geometric mean is negligible, while both estimation procedures have difficulties for larger noise levels and sample sizes < 5000. Fig. 3b shows the computation times as a ratio ('explicit'/CIA) of the time it has taken the genetic algorithm to converge. Note that the y-axis is in logarithmic units. For a sample size of $N = 500$ the speed is comparable, slightly in favor of the explicit estimation, for $N = 1000$ CIA is 1.4 times faster and for $N = 5000$ it is approximately 20 times faster. To put the computation time ratio into perspective, we note that for, e.g., $\sigma = 0.89$ and $N = 5000$ the explicit estimation takes 194.5 min to converge, while CIA takes 3.8 min to converge to an equally good solution with an average of 18.37 s and 0.36 s per function evaluation respectively. In the supplementary material we supply similar comparison plots for three other simulation scenarios suggested by Yin (2004).

7.2. Weather radar and Meteosat data

This data set consists of satellite and radar imagery from 20 August 2007, where extreme downpour intensities (53 mm in 10 min) were recorded in some regions of Denmark.

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 3.9 μm to 13.4 μm, and these bands monitor cloud top reflectance properties. The radar data are recorded three minutes before the satellite image using the Danish Meteorological Institute (DMI) weather radars and consists 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 $N = 7577$ observations in the radar imagery exhibiting reflectance from precipitation.

This case has also been treated by Vestergaard and Nielsen (2012), where an elaborate geometric and temporal alignment was needed to ameliorate the CCA solution. As will be shown below, this is entirely unnecessary when using the method suggested here.

The motivation for fusion of these two data sources is twofold: First, weather radars have a limited coverage of approximately 240 km from their position while satellites cover almost the entire planet. A fusion of these two could be a way of using satellite data as a proxy for radar data. Second, the two types of data come from very different types of sensors, wherefore the distributions of the data are very different. Therefore this is an illustrative example of using an information theoretic approach rather than a method based on assumptions of distributions.

The first mutual information canonical variate (MICV) is shown in Fig. 4b where the eight infrared bands from the satellite data are projected onto the projection direction $\mathbf{a}$ determined by canonical information analysis. As the second set of variables consists of only a single variable, $\mathbf{b} = b = 1$. Therefore only the MICV related to the satellite data is shown. For comparison, the solution to the same problem determined by canonical correlation analysis is shown in Fig. 4a. An area has been marked with a dashed red rectangle in both figures; extreme precipitation is known to occur in the dark region inside the rectangle in Fig. 4b. A viable solution would therefore accentuate the cloud tops in this particular area. It is seen that this is the case for canonical information analysis, where a contrast with the surroundings is evident, while the correlation based result shows less contrast.

A correlation of 0.344 and 0.303 between the leading pair of canonical variates was obtained using CCA and CIA respectively. Mutual information between the two mutual information based canonical variates is 0.101 while it is 0.088 between the two correlation based variates.

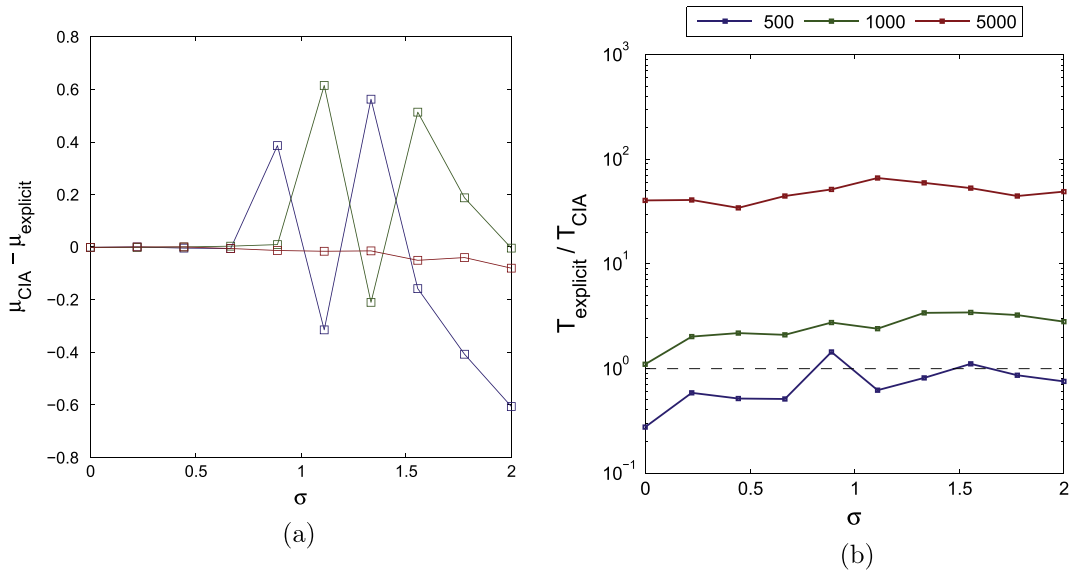


Fig. 3. Simulation studies. Comparison of accuracy and speed for CIA and 'explicit' estimation. σ is the noise level, μ_{CIA} and $\mu_{explicit}$ are defined as in Eq. (12) Values above 0 indicate a higher correlation between the found components and the true components (a better solution) for CIA, while values below 0 indicate a better solution yielded by the explicit estimation. Speed is shown as $\frac{T_{explicit}}{T_{CIA}}$ on a logarithmic scale, thus CIA is slower for values below 1 and faster for values above 1. The three colored lines represent results obtained with $N = \{500, 1000, 5000\}$ simulated observations.

Quantitative comparison of correlation based and mutual information based analysis can, for example, be done by calculating spatial autocorrelation over the marked region in Fig. 4a and b. We have chosen to calculate the autocorrelation over spatial lags of [0 1], [1 1], [1 0] and [−1 1] to capture spatial correspondences in all directions. For both analysis methods these values are shown in Table 1.

The average value for the mutual information based analysis is 0.950 compared to 0.897 for the correlation based analysis. These values confirm the subjective evaluation that the spatial coherence is larger in the mutual information based solution compared to the correlation based analysis.

7.3. DLR 3K data

The images used in this example were recorded with the airborne DLR 3K camera system (Kurz et al., 2007a,b) 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 s apart cover a busy motorway. These data have previously been treated by Nielsen and Canty (2009), Nielsen (2011, 2007) where the original RGB images can be seen. 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.

Fig. 5b shows the difference image between the first set of MICVs whereby canonical information analysis acts as a tool for change detection. Previously, a method for change detection based on canonical correlation analysis has been proposed (Nielsen et al., 1998). Comparing with the solution obtained by canonical correlation analysis in Fig. 5a it is evident that a much larger amount of change information is gained by using CIA: the background is much smoother and clearly distinguishable from the areas of change (the cars) and the extreme values are only present where change has actually occurred. The difference image between the second set of MICVs is included in the supplementary material. Since relevant changes are due to the moving cars on the motor way only, higher order CVs in this case do not contain further information.

To quantify the different quality of the solutions, a region in the difference image has been selected. This region is known not to change between the two acquisition times and is assumed to be constant over the region in an ideal difference image. The variance in this region will therefore represent the unwanted noise in the difference image and is denoted var(N) below. The ratio R between the signal-to-noise ratios for the two solutions is defined as

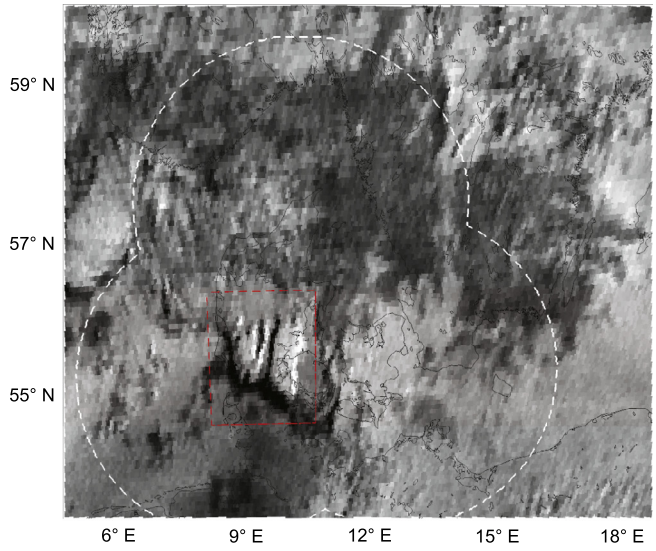
$$R = \frac{\mathrm{SNR}_{\mathrm{CIA}}}{\mathrm{SNR}_{\mathrm{CCA}}} = \frac{\frac{\mathrm{var}(S)}{\mathrm{var}(N_{\mathrm{CIA}})}}{\frac{\mathrm{var}(S)}{\mathrm{var}(N_{\mathrm{CCA}})}} = \frac{\mathrm{var}(N_{\mathrm{CCA}})}{\mathrm{var}(N_{\mathrm{CIA}})} \tag{13}$$

and is independent of the signal variance, when assuming that the true signal S is equal in the two solutions. 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., $R = 3.319$. This verifies the subjective evaluation that a more homogeneous no-change background is obtained using the proposed mutual information based method.

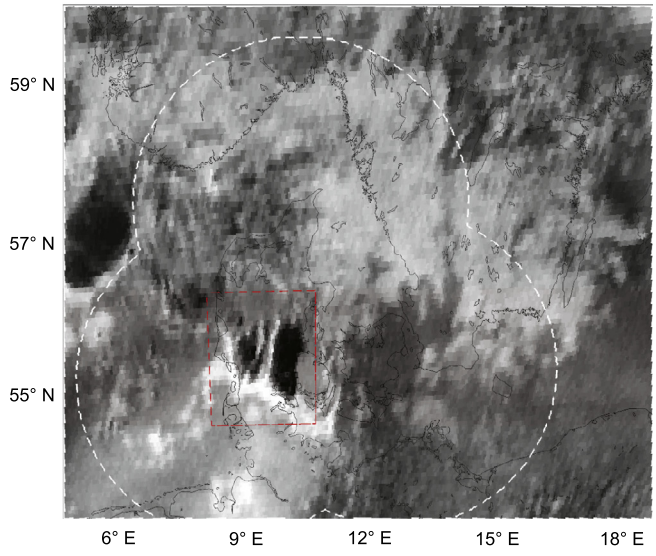
A correlation of 0.982 and 0.945 between the leading pair of canonical variates was obtained using CCA and CIA respectively, which demonstrates that a high correlation is not always the best measure for similarity. A mutual information of 1.034 and 1.335 between the leading pair of canonical variates was obtained using CCA and CIA respectively.

7.4. Summary

Table 2 summarizes the results for all three cases using canonical information analysis. Co-inspection of table and Figs. 2, 4, 5 clearly shows that the solution with the largest mutual information is superior to that with the largest correlation. Second order MICVs, MI between input bands and MICVs and a matrix of MI between pairs of MICVs are included as supplementary material



(a) Canonical correlation analysis



(b) Canonical information analysis

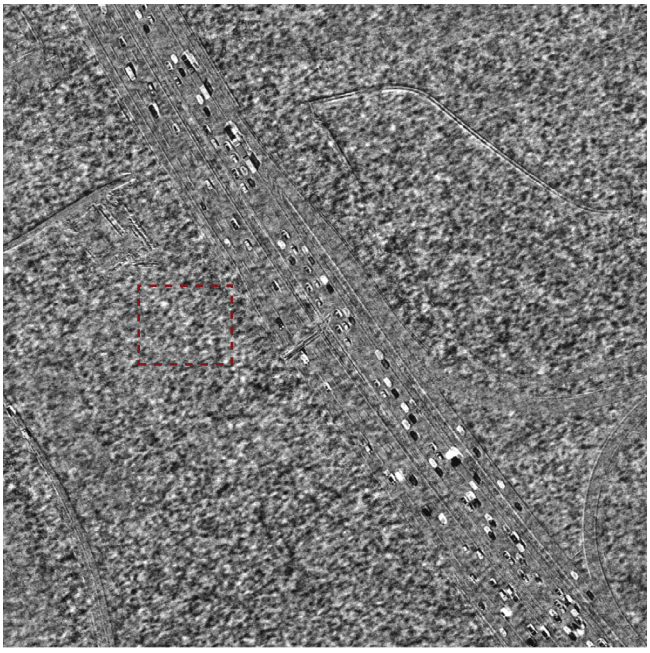
Fig. 4. Weather. The first CV determined by canonical correlation analysis and canonical information analysis for the *weather* data set. The marked rectangular area is known – from radar imagery – to exhibit extreme rain at this particular point in time. The display range of the intensity values is within ± three standard deviations of the mean. The dashed white line marks the extent of the radar coverage.

Table 1
Results for the *weather* data set evaluated in terms of spatial autocorrelation in the region of interest. Marked in bold are the average spatial autocorrelation over the four directions.

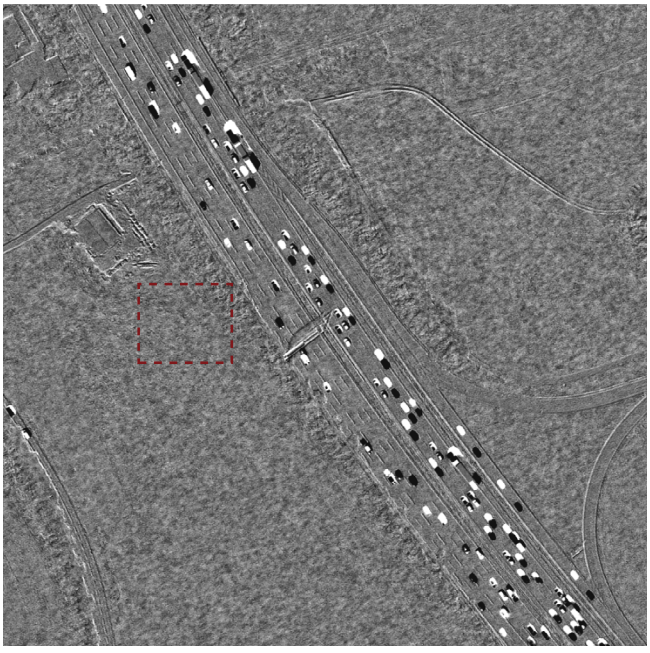
Method	→	↘	↓	↙	Average
CIA	0.973	0.932	0.950	0.943	**0.950**
CCA	0.952	0.859	0.892	0.886	**0.897**

for the *weather* and cars cases. Additional simulation studies suggested by Yin (2004) are detailed in the supplementary material, where the geometric mean using CIA, explicit estimation or CIA are shown.

In the *weather* case all 7,577 observations having a value in the radar data were used, while a random sample of 10,000 observations were used in the *cars* case were used for the optimization



(a) Canonical correlation analysis



(b) Canonical information analysis

Fig. 5. Cars. Difference image of the first set of MICVs for the *cars* data set using (a) canonical correlation analysis and (b) canonical information analysis respectively. The display range of the intensity values is within ± three standard deviations of the mean. The marked region is used to quantify the no-change noise variance.

of mutual information. The determined linear transformations were applied to all observations in the two sets of variables. Each computation was done on a 64-bit Linux system with 2 X5650 6-Core processors, 2.66 GHz, 48 GB RAM.

In all three cases visual inspection of the resulting scatter plots and imagery clearly show the superior behavior of the mutual information based canonical analysis: the solution to the toy example illustrates that the CIA solution recovers the latent signal (the noisy parabola), while the CCA solution fails to do the same. The solution

Table 2
Summary of results for each of the three cases: *toy* is the toy example from Section 7.1, *weather* is the satellite/radar case from Section 7.2 and *cars* is the DLR 3K change-detection case from Section 7.3. I is mutual information as in Eq. (8), ρ is correlation, # is the number of function evaluations needed and sec. is the time in seconds.

		I	ρ	#	sec.
toyexample	CIA	0.127	0.010	4160	669
$(k,\ell)=(2,2)$	CCA	0.018	0.016	<1	<1
Cars	CIA	1.335	0.945	9360	1165
$(k,\ell)=(3,3)$	CCA	1.034	0.982	<1	<1
Weather	CIA	0.101	−0.303	21060	1672
$(k,\ell)=(8,1)$	CCA	0.088	0.344	<1	<1

for the weather satellite data provides a representation of these data, which carry the most similar information to the weather radar data. This can be useful for, e.g., visualization purposes for meteorologists, or providing pseudo-radar coverage outside of the radar's range. In the change detection case, the background noise in the CCA solution looks almost similar to the signal, i.e., the cars. This is not the case for the CIA solution, where the noise in the difference image is suppressed and the cars stand out. This is clearly beneficial for any kind of application of these data.

8. Conclusions and future work

In this paper mutual information successfully replaces correlation to find canonical variates for two sets of multivariate observations. Unlike correlation which allows for second order statistics only, mutual information allows for the actual density of the variables at hand. An illustrative toy example with zero correlation between strongly associated variables proves the usefulness of the idea. Optical satellite data and weather radar data are successfully fused using the proposed method to accentuate precipitating clouds in the satellite data. This illustrates the benefit of mutual information when working with data sets of different modalities. Optical airborne (DLR 3K) data from two acquisition times 0.7 s apart are included to illustrate the use of the proposed method in the context of change detection.

Canonical information analysis employs approximate marginal and joint entropy estimation. A simulation study shows that this approximation is as accurate as and much faster than previously presented algorithms, making the method feasible for image analysis problems and other large sample problems. Small sample applications ($N \leqslant 500$) do not benefit from this approach.

MATLAB software will be made available on the first author's homepage.

Acknowledgments

The authors would like to thank Researcher Dr. Thomas Bøvith and Aviation Meteorologist Birgitte Knudsen at Danish Meteorological Institute, DMI, for selecting and providing the weather radar and optical geostationary satellite data for the heavy precipitating case.

Thanks to Dr. Peter Reinartz and coworkers, German Aerospace Center, DLR, Oberpfaffenhofen, Germany, for letting us use the geometrically coregistered DLR 3K camera data

AAN initially started work on this subject during a sabbatical leave to the University of Oxford, Department of Statistics. Thanks to Professor Brian D. Ripley for hosting.

Appendix A. Comparison of bandwidth estimators

Here we motivate the choice of the maximal smoothing principle (Terrell, 1990) for kernel bandwidth estimation by comparing its properties with the Sheather–Jones estimator (Sheather and Jones, 1991) and a diffusion based estimator (Botev et al., 2010).

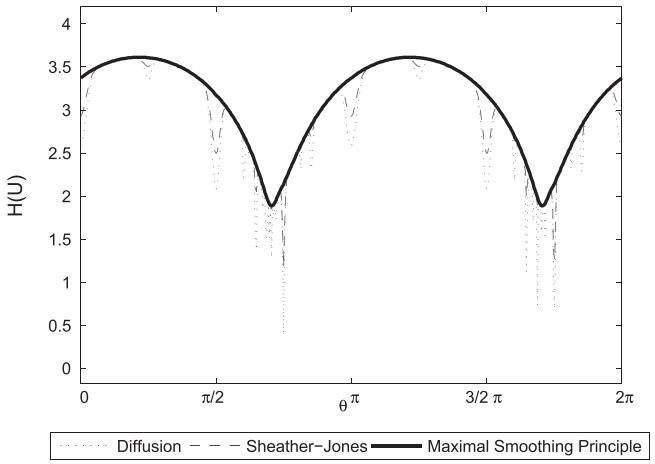


Fig. A.6. Comparison of entropy estimates based on kernel density estimates using three different bandwidth estimators: A diffusion based estimator, the "solve-the-equation plug-in" estimator by Sheather–Jones and the maximal smoothing principle.

The desirable properties of the maximal smoothing principle for kernel bandwidth estimation can best be illustrated by an example. For illustration purposes we consider a single set of a two-dimensional stochastic variable $\mathbf{X}$. We wish to estimate the entropy of the linear combination $U = \mathbf{a}^T\mathbf{X}$ using a kernel density estimator. The entropy becomes a function of the bandwidth estimate $H(\hat{\sigma}_X(\mathbf{a}|\mathbf{X}))$. The bandwidth is estimated based only on the linear combination U and is thereby a function of the projection direction $\mathbf{a}$ given the data $\mathbf{X}$.

We let $\mathbf{a}$ be a vector on the unit circle and it can thus be fully described in spherical coordinates as $\mathbf{a}(\theta) = (1, \theta)$ by the angle θ. In the following experiment we vary the angle over the range $\theta \in [0, 2\pi]$ and estimate the bandwidth $\hat{\sigma}_X$ for each value of θ. This bandwidth is used for calculating the entropy.

Fig. A.6 shows the entropy $H(U)$ as a function of the projection direction angle θ for three different bandwidth estimators. It is immediately seen that the entropy estimate is smoother and avoids local minima when using the maximal smoothing principle, while the Sheather–Jones estimator and the diffusion based estimator fluctuate much more. The average computation times over 500 estimations of the bandwidth is 0.09, 72.03 and 0.04 s for the diffusion based estimator, the Sheather–Jones and the maximal smoothing principle, respectively.

Based on these observations, we find the maximal smoothing principle best suited for estimation of bandwidth in the context of optimizing mutual information of linear combinations. Though this behavior is illustrated in two-dimensional data only, we employ this principle for higher dimensional data as well.

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.isprsjprs.2014.11.002.

References

Bach, F.R., Jordan, M.I., 2002. Kernel independent component analysis. J. Mach. Learn. Res. 3, 1–48.

Beirlant, J., Dudewicz, E.J., Györfi, L., Van der Meulen, E.C., 1997. Nonparametric entropy estimation: an overview. Int. J. Math. Stat. Sci. 6, 17–40.

de Bie, T., de Moor, B., 2002. 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.
Bishop, C.M., 2007. Pattern Recognition and Machine Learning (Information Science and Statistics), first ed. Springer.
Botev, Z.I., Grotowski, J.F., Kroese, D.P., 2010. Kernel density estimation via diffusion. Ann. Stat. 38, 2916–2957.
Canty, M.J., 2010. Image Analysis,, .. Classification, and Change Detection in Remote Sensing, second ed. CRC Press/Taylor and Francis.
Conese, C., Maselli, F., 1993. Selection of optimum bands from TM scenes through mutual information analysis. ISPRS J. Photogramm. Rem. Sens. 48, 2–11.
Ehlers, M., 1991. Multisensor image fusion techniques in remote sensing. ISPRS J. Photogramm. Rem. Sens. 46, 19–30.
Erten, E., Reigber, A., Ferro-Famil, L., Hellwich, O., 2012. A new coherent similarity measure for temporal multichannel scene characterization. IEEE Trans. Geosci. Rem. Sens. 50, 2839–2851.
Fletcher, R., 1970. A new approach to variable metric algorithms. Comput. J. 13, 317–322.
Friedman, J., 1987. Exploratory projection pursuit. J. Am. Stat. Assoc. 82, 249–266.
Haber, E., Modersitzki, J., 2007. Intensity gradient based registration and fusion of multi-modal images. Methods Inform. Med. 46, 292–299.
Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28, 321–377.
Hyvärinen, A., Karhunen, J., Oja, E., 2004. Independent Component Analysis, vol. 46. John Wiley & Sons.
Jones, M., Marron, J., 1996. A brief survey of bandwidth selection for density estimation. J. Am. Stat. 91, 401–407.
Karasuyama, M., Sugiyama, M., 2012. Canonical dependency analysis based on squared-loss mutual information. Neural Netw. 34, 46–55.
Kraskov, A., Stögbauer, H., Grassberger, P., 2004. Estimating mutual information. Phys. Rev. E 69, 066138.
Kurz, F., Charmette, B., Suri, S., Rosenbaum, D., Spangler, M., Leonhardt, A., Bachleitner, M., Stätter, R., Reinartz, P., 2007a. 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, Munich, Germany. pp. 09–19.
Kurz, F., Müller, R., Stephani, M., Reinartz, P., Schroeder, M., 2007b. Calibration of a wide-angle digital camera system for near real time scenarios. In: Proc. of ISPRS Hannover Workshop 2007 – High Resolution Earth Imaging for Geospatial Information, pp. 1682–1777.
Lai, P.L., Fyfe, C., 2000. Kernel and nonlinear canonical correlation analysis. Int. J. Neural Syst. 10, 365–377.
Mackay, D.J.C., 2003. Information Theory,, .. Inference and Learning Algorithms, first ed. Cambridge University Press.
Modersitzki, J., 2004. Numerical Methods for Image Registration. Oxford University Press, USA.
Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Comput. J. 7, 308–313.
Nielsen, A.A., 2007. The regularized iteratively reweighted MAD method for change detection in multi- and hyperspectral data. IEEE Trans. Image Process. 16, 463–478, <http://www.imm.dtu.dk/pubdb/p.php?4695>.
Nielsen, A.A., 2011. Kernel maximum autocorrelation factor and minimum noise fraction transformations. IEEE Trans. Image Process. 20, 612–624, <http://www.imm.dtu.dk/pubdb/p.php?5925>.
Nielsen, A.A., Canty, M.J., 2009. Kernel principal component and maximum autocorrelation factor analyses for change detection. Proc. SPIE 7477, 74770T-1-74770T-2. <http://www.imm.dtu.dk/pubdb/p.php?5757>.
Nielsen, A.A., Conradsen, K., Simpson, J., 1998. Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: new approaches to change detection studies. Rem. Sens. Environ. 64, 1–19, <http://www.imm.dtu.dk/pubdb/p.php?1220>.
Parzen, E., 1962. On estimation of a probability density function and mode. Ann. Math. Stat. 33, 1065–1076.
Pohl, C., Van Genderen, J.L., 1998. Review article multisensor image fusion in remote sensing: concepts, methods and applications. Int. J. Rem. Sens. 19, 823–854.
Rosenblatt, M., 1956. Remarks on some nonparametric estimates of a density function. Ann. Math. Stat., 832–837.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423.
Sheather, S., Jones, M., 1991. A reliable data-based bandwidth selection method for kernel density estimation. J. Roy. Stat. Soc. Ser. B (Meth.) 53, 683–690.
Shwartz, S., Zibulevsky, M., Schechner, Y., 2005. Fast kernel entropy estimation and optimization. Signal Process. 85, 1045–1058.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis, vol. 26. Chapman & Hall/CRC.
Studholme, C., Hill, D., Hawkes, D., 1999. An overlap invariant entropy measure of 3D medical image alignment. Pattern Recogn. 32, 71–86.
Suri, S., Reinartz, P., 2010. Mutual-information-based registration of TerraSAR-X and Ikonos imagery in urban areas. IEEE Trans. Geosci. Rem. Sens. 48, 939–949.
Terrell, G., 1990. The maximal smoothing principle in density estimation. J. Am. Stat. Assoc. 85, 470–477.
Vestergaard, J.S., Nielsen, A.A., 2012. Automated invariant alignment to improve canonical variates in image fusion of satellite and weather radar data. J. Appl. Meteorol. Climatol. <http://journals.ametsoc.org/doi/abs/10.1175/JAMC-D-12-05.1>.
Viola, P.A., 1995. Alignment by Maximization of Mutual Information. Ph.D. Thesis. Massachusetts Institute of Technology.
Viola, P., 1997. Alignment by maximization of mutual information. Int. J. Comput. Vis. 24, 137–154.
Yin, X., 2004. Canonical correlation analysis based on information theory. J. Multivariate Anal. 91, 161–176.