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