Entropy **2013**, *15*, 1690-1704; doi:10.3390/e15051690

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

An Estimate of Mutual Information that Permits Closed-Form Optimisation

Raymond Liu * and Duncan F. Gillies

Department of Computing, Imperial College, London SW7 2RH, UK; E-Mail: d.gillies@imperial.ac.uk

* Author to whom correspondence should be addressed; E-Mail: rl708@imperial.ac.uk;
Tel.: +44-7725-949148; Fax: +44-(0)-20-7594-8932.

Received: 1 February 2013; in revised form: 18 April 2013 / Accepted: 28 April 2013 / Published: 8 May 2013

Abstract: We introduce a new estimate of mutual information between a dataset and a target variable that can be maximised analytically and has broad applicability in the field of machine learning and statistical pattern recognition. This estimate has previously been employed implicitly as an approximation to quadratic mutual information. In this paper we will study the properties of these estimates of mutual information in more detail, and provide a derivation from a perspective of pairwise interactions. From this perspective, we will show a connection between our proposed estimate and Laplacian eigenmaps, which so far has not been shown to be related to mutual information. Compared with other popular measures of mutual information, which can only be maximised through an iterative process, ours can be maximised much more efficiently and reliably via closed-form eigendecomposition.

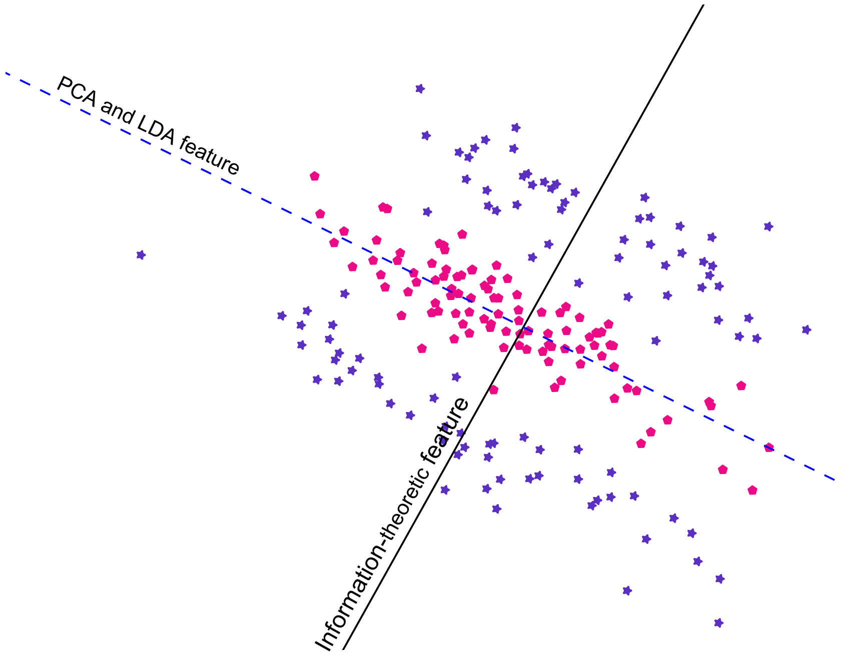
Keywords: mutual information; dimensionality reduction; feature extraction; pattern recognition; machine learning

1. Introduction

Mutual information has been receiving increased attention in the recent years, in the field of *machine learning* and *statistical pattern recognition*, more specifically in *dimensionality reduction*, also known as *feature extraction*. In this paper we focus on the problem of *classification*, in which a dataset of (usually high-dimensional) vectors are categorised into several distinct classes and presented as training data, and the task is to classify new data points into their correct categories. For this to be done feasibly and efficiently, a dimensionality reduction procedure needs to be carried out as a pre-processing step, where the dataset is projected onto a lower-dimensional space, whose basis vectors are often known as *features*. The search for these features is known as *feature extraction* in the machine learning literature [1], and sometimes also known as *dimensionality reduction* in the context of data visualisation and representation [2].

The features are found by declaring an objective value to maximise or minimise, and those features with the highest objective values are retained as "good" features. For example, given any feature (a unit vector in the input space, which can be viewed as a line), we can project the training data points onto it, and estimate the mutual information between these projected data points and the class label (our target variable), and use this as the objective value to determine the quality of the feature. Indeed, Torkkola's method [3] maximises Kapur's *quadratic mutual information* (QMI) [4] through gradient-ascent-type iterations. Another example of an objective value of a feature is the variance of the data along the feature, whose maximisation has a closed-form, eigenvalue-based solution, known as *principal component analysis* (PCA) [5]. In this paper however we want to focus on information-theoretic objective values, whose benefits over variance-based (or distance-based) objective values, at least for artificial datasets and low-dimensional datasets, have been demonstrated [3,6,7]. The main advantage of using information-theoretic measures is illustrated in Figure 1.

Figure 1. This artificial 2D data is categorised into two classes: pink pentagons and blue stars. We can see that the black line is a better line of projection for the data, in terms of class-discrimination, than the blue dashed line. The blue dashed line is the feature computed using Fisher's linear discriminant analysis (LDA) [5]; PCA produces a very similar feature in this case. The black line is computed using our eigenvalue-based mutual information method [8].



There are many difficulties associated with measuring mutual information from data. First of all, we do not know the true distribution of the data in the input space, which itself needs to be estimated from the training data. Secondly, Shannon's measure of mutual information [9,10], when applied to continuous data, requires either numerical integration, which is extremely computationally expensive, or discretisation, which may lead to inaccurate estimation and does not lend itself to efficient optimisation techniques. Several attempts have been made to circumvent some of these difficulties through the use of alternative measures of information [3,4,6,11] and the popular technique of kernel density estimation [12]. While these have proven to be effective techniques for measuring mutual information between the data along a feature and the class label, they can (so far) only be maximised through *iterative* optimisation. This has its own complications with regards to dimensionality reduction. More specifically, they have many free variables, and the algorithms are very computationally expensive, much more so than non-information-theoretic dimensionality reduction techniques like PCA. A technique was proposed [8] that avoids iterative optimisation, and instead has a closed-form, eigenvalue-based solution, similar in methodology to PCA. In this paper we will look at some more detailed properties of the underlying measure used therein, and show some connections with Laplacian eigenmaps [2,13].

2. Measures of Mutual Information

For the application of information theory to dimensionality reduction for classification problems, we study the mutual information between a continuous random variable Y and a discrete random variable C. Refer to Table 1 for notation. Shannon's mutual information in this case is

$$I(Y;C) = \sum_{c=1}^{K} \int_Y p_{Y,C}(y,c) \log_2 \frac{p_{Y,C}(y,c)}{p_Y(y)p_C(c)}\, dy \tag{1}$$

We want to avoid numerical computation of integrals because of their high computational cost. A common way of doing this is to discretise the continuous random variable Y.

Mutual information as defined in Equation (1) is the Kullback–Leibler (KL) divergence between a joint distribution $p_{Y,C}(y,c)$ and the product of the respective marginal distributions $p_Y(y)p_C(c)$, or equivalently the joint distribution under the assumption of independence. The KL divergence can be characterised by a set of postulates [11]. Kapur [4] argued that if the aim is to maximise the divergence, and not so much to calculate its value precisely, then the set of postulates can be relaxed [3]. A family of measures for the divergence of two distributions were proposed, and one of them results in the following *quadratic mutual information* (QMI).

$$I_Q(Y;C) := \sum_{c=1}^{K} \int_Y \Big(p_{Y,C}(y,c) - p_Y(y)p_C(c) \Big)^2 \, dy \tag{2}$$

This takes a particularly convenient form when the densities $p_{Y|C}(y|c)$ and $p_Y(y)$ are estimated using Gaussian kernel density estimation [12]. More specifically, the integral in equation (2) can be evaluated analytically using the convolution property of Gaussian densities [3,8]. We refer the reader to [3,8] for algebraic details, and only present the key results here. Equation (2) can be re-written as follows.

$$I_Q(Y;C) = \sum_{n=1}^{N} \sum_{m=1}^{N} \rho_{nm} G_{nm} \tag{3}$$

where we use the short-hand notation

$$\rho_{nm} := \frac{1}{N^2} \Big(\mathbb{I}[c_n = c_m] + \big(\sum_{c=1}^{K} \frac{N_c^2}{N^2} \big) - \frac{2N_{c_n}}{N} \Big) \tag{4}$$

$$G_{nm} := \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_n - y_m)^2}{2\sigma^2}\right) \tag{5}$$

With this, the maximisation of QMI can (theoretically) be solved by differentiating Equation (3) with respect to $\mathbf{w}$ and using a gradient-ascent-type algorithm. Note that in contrast, if we approximate $I(Y;C)$ as in Equation (1) by discretising the continuous random variable Y as mentioned, then we cannot differentiate it with respect to $\mathbf{w}$ analytically. However there are methods that approximate the gradient of Shannon's mutual information even when discretisation is used, and gradient ascent can be done with Shannon's mutual information [14].

Table 1. Notation.

$\mathbf{w}$	Unit vector representing a feature.
$\mathbf{X}$	Random vector variable representing a data point in the input space.
$\mathbf{x}_n$	The n^{th} training data point.
Y	Random variable representing a data point projected onto a feature $\mathbf{w}$. More precisely, $Y = \mathbf{w}^{\mathrm{T}}\mathbf{X}$.
y_n	$:= \mathbf{w}^{\mathrm{T}}\mathbf{x}_n$, the n^{th} training data point projected onto the feature $\mathbf{w}$.
C	A discrete random variable representing the class label of a data point.
c_n	The correct class label of the n^{th} training data point.
$\mathbf{x}, y, c$	Realisations of $\mathbf{X}$, Y, and C.
K	Total number of classes in the dataset.
N	Total number of training data points.
N_c	Size of training data class c.
$p_Y(y)$	Probability density function of Y at realisation y. Likewise for other variables.
σ	$\sqrt{2}$ × (bandwidth of the Gaussian kernel density estimator).

Torkkola uses QMI as a measure of quality of features in his iterative feature extraction algorithm [3]. It was shown that this technique can give superior classification results for some low-dimensional datasets, while the conventional feature extraction methods PCA and LDA perform better on others. No high-dimensional datasets were tested on however, possibly due to the high computational complexity of iterative algorithms. QMI is a theoretically elegant and practically applicable measure of mutual information. Regarding its optimisation however, there are practical drawbacks, which can be summarised as follows.

- The computational complexity (cost) of any iterative optimisation algorithm is very high, and QMI in its current form Equation (3) can only be maximised iteratively. Table 2 shows a comparison of the computing times of PCA, LDA, Torkkola's iterative QMI algorithm, and our proposed eigenvalue-based MI method (EMI).
- The current iterative algorithms have many free parameters, including the learning rate and the stopping criterion, for which there is not yet a principled method of estimating.

- Experiments show that a straightforward application of gradient ascent, a popular algorithm used to maximise QMI, can be unstable and unpredictable in maximising QMI in high-dimensional spaces. In particular, it does not always maximise what it is designed to maximise. Figure 2 illustrates this, for the AT&T Dataset of Faces. Note that this is not an intrinsic deficiency of QMI, but rather of gradient ascent as applied to maximising QMI.

Table 2. Times taken, in seconds, for algorithms PCA, LDA, eigenvalue-based MI, and iterative QMI to extract one feature and 39 features respectively, from the AT&T Dataset of Faces. The computing times for eigenvalue-based algorithms (PCA, LDA, EMI) are independent of the number of extracted features, due to the closed-form nature of the eigenvalue problem. This was a Python 2.7 implementation with packages Numpy, MatPltoLib, and PIL (Python Imaging Library), executed on an Intel Core i7-2600K processor at 3.40 GHz with 16 GB RAM.

	Time to extract	
	first feature	**39 features**
PCA	0.406	(0.406)
LDA	1.185	(1.185)
EMI	4.992	(4.992)
QMI	**265.034**	**10330.358**

In Figure 2, the reader may ask why LDA produces high values of QMI, similar to that produced by the EMI method, despite that LDA is not designed to maximise QMI. To answer this question, first note that Equation (2) can be re-written as

$$I_Q(Y;C) = \sum_{c=1}^{K} p_C(c)^2 \int_Y \left(p_{Y|C}(y|c) - p_Y(y)\right)^2 dy \tag{6}$$

Let us consider a simple 2-class example, where both classes have equal prior probabilities $P_C(c) = \frac{1}{2}$. In this case, Equation (6) tells us that $I_Q(Y;C)$ is proportional to a sum of two summands, each of which is the square of the area between the graphs of the class-conditional distribution $p_{Y|C}$ and the overall distribution p_Y. Let us suppose further that the classes are normally distributed with the same variance but different means. Then Figure 3 shows us that maximising QMI is equivalent to minimising the overlap between the 2 class-conditional distributions, or equivalently maximising the separation between the classes.

Figure 2. Plot of the QMI between the class label and each of the first 10 features, computed by various methods. The horizontal axis indexes the features, while the vertical axis measures the QMI along each feature. We see that Torkkola's iterative QMI algorithm [3] gives lower values of QMI than the EMI algorithm, which gives values similar to LDA. Note that LDA is not designed to maximise QMI. PCA is also included for the sake of comparison. In the experiment, the data was whitened and the initial features for the iterative QMI algorithm were set to random features.

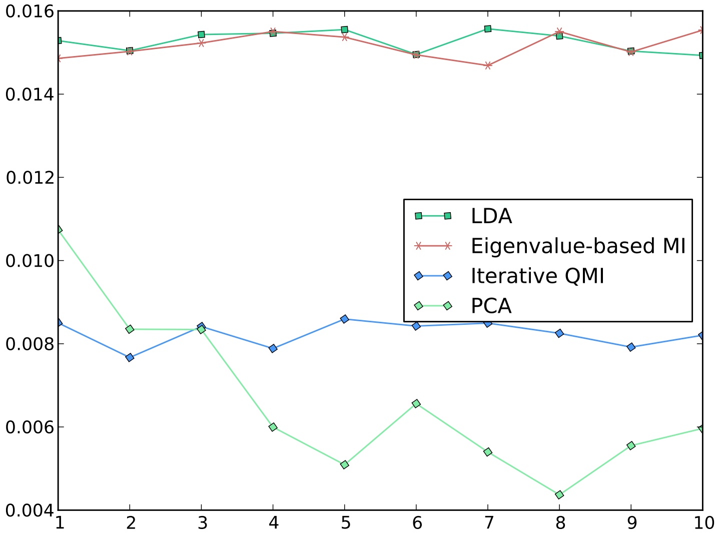
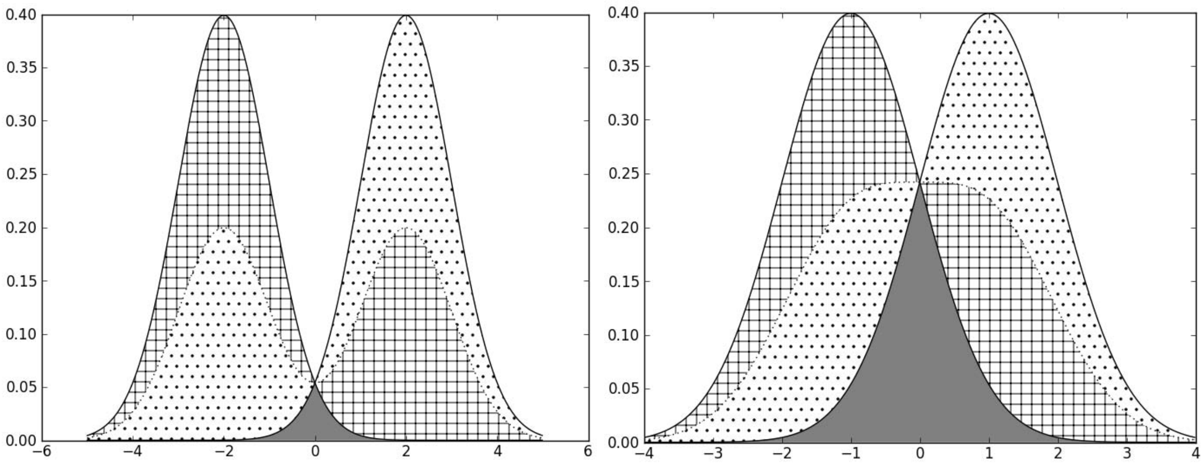
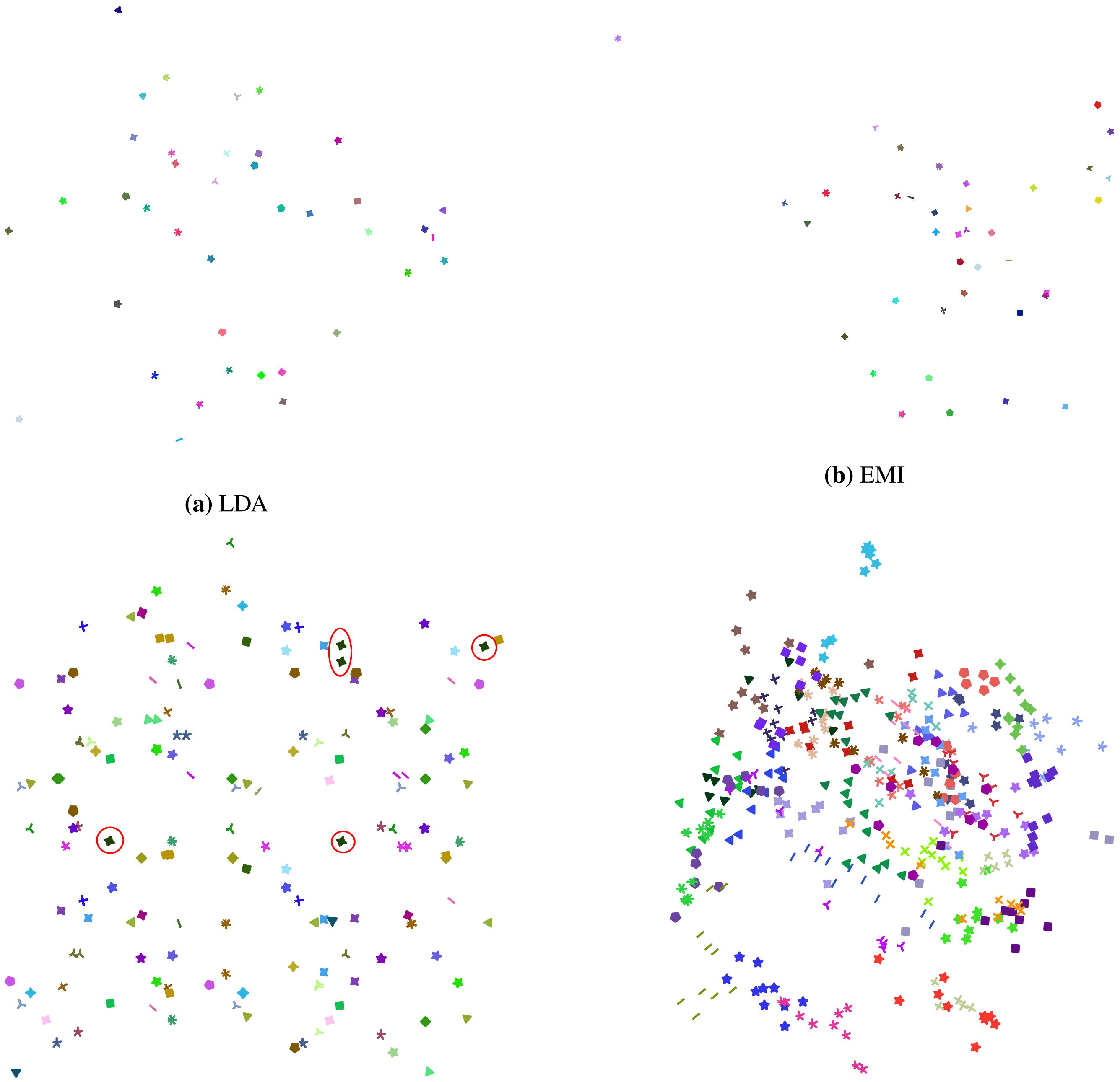


Figure 3. The solid lines are the class-conditional distributions for the 2 classes, and the dotted line is the overall distribution p_Y. The checkered and the dotted areas correspond to the value of $|p_{Y|C}(y|c) - p_Y(y)|$ for each of the 2 classes respectively, and the squares of only these areas contribute to the value of QMI, not the grey-shaded area. The grey-shaded area is the overlap between the two class-conditional distributions. We see that the smaller the overlap, the larger the value of the dotted and checkered areas, and therefore the larger the value of QMI. (**a**) Good class separation, high QMI; (**b**) Bad class separation, lower QMI.



(**a**) (**b**)

Figure 4. 2D projections of the AT&T Dataset of Faces, computed using (**a**) LDA; (**b**) our EMI method; (**c**) Torkkola's iterative QMI method; and (**d**) PCA for comparison. In (**a**) we see that each of the 40 classes in the dataset is in a compact cluster, well separated from others. In fact, each class is so tightly packed that at this resolution it looks like a singleton point. Similar observations can be made from (**b**). In (**c**) however, each class seems to be scattered into 2 or more clusters, as exemplified by the class marked by red circles. Note that each class of the dataset has 10 data points, and the class marked by red circles seems to be in 4 clusters. For all of (**a**), (**b**) and (**c**), the data was whitened as a pre-processing step.



(**a**) LDA

(**b**) EMI

(**c**) QMI using gradient ascent

(**d**) PCA

The intuition shown in Figure 3 can be validly extrapolated to higher dimensions, general class-conditional distributions, and multi-class datasets. LDA maximises the class-separation of the data in the sense that it maximises the between-class distances and minimises within-class distances between data points. Now depending on the class-conditional distributions and input dimensionality, LDA will not always succeed, as illustrated in Figure 1. However, where LDA does succeed, different classes will be well-separated and so the QMI along the features produced by LDA will be high. The AT&T Dataset of Faces is one in which LDA does succeed in maximising the separation between classes, as shown in Figure 4a. In fact, as a side-note, our experiments show that LDA succeeds in many high-dimensional face datasets. The reasons for this have little to do with mutual information but rather a lot to do with the high input dimensionality and the problem of over-fitting, and therefore are not within the scope of this paper. The similarity in the values of QMI obtained by LDA and our EMI method is explained by the similarity in their respective 2D projections shown in Figures 4a and 4b. Figure 4c, in contrast, does not exhibit good separation between classes, which explains the relatively low values of QMI obtained by the iterative method as shown in Figure 2. Furthermore, Figure 4c suggests that the iterative QMI method in this case might have sought a local optimum, due to the scattering nature of individual classes into more than one cluster.

2.1. Eigenvalue-Based Mutual Information (EMI)

The practical drawbacks of QMI with respect to its maximisation, as mentioned previously, can be circumvented through the use of another measure of mutual information. This measure was implicitly employed in [8] to address the practical problems with QMI. The maximisation of this measure of mutual information, which for now we will call EMI (eigenvalue-based mutual information), has a closed-form solution that is a set of eigenvectors of an objective matrix. Before we introduce EMI, we will briefly review the *pairwise interaction* interpretation of QMI.

Mutual information is often interpreted as the *difference* or *similarity* between two probability distributions: the true joint distribution and the joint distribution under the independence assumption. In the context of estimating mutual information *from data* however, Equation (3) uncovers an alternative view of mutual information, one in terms of *pairwise interactions* G_{nm} between each pair $\{\mathbf{x}_n, \mathbf{x}_m\}$ of training data points, weighted by the ρ_{nm}. This view is especially applicable and intuitive in classification problems. Each pairwise interaction G_{nm} is monotonically decreasing in the distance $|y_n - y_m|$ between two data points along the feature $\mathbf{w}$, as is clear from Equation (5). So for example if we simply wanted to maximise the sum $\sum_{n=1}^{N} \sum_{m=1}^{N} G_{nm}$ (rather than Equation (3)), then we will obtain a feature along which the data points are as close to each other as possible, which is obviously not desirable from a classification perspective. However the weights ρ_{nm} can be negative, in which case the corresponding pairwise distance is maximised. Let us conceive of a simple example in which there are 2 classes ($c = 2$), and each class has 5 training data points ($N_1 = N_2 = 5$ and $N = 10$). Then the reader may verify from Equation (4) that if two data points $\mathbf{x}_n$ and $\mathbf{x}_m$ are in the same class, then $\rho_{nm} = \frac{1}{200}$; and if they are in different classes, then $\rho_{nm} = -\frac{1}{200}$. This means that in the maximisation of $I_Q(Y; C)$ as in Equation (3), the within-class distances are minimised while the between-class distances are maximised. On a side-note, we see that in this pairwise interaction view of mutual information maximisation, there is significant resemblance to LDA. However, unlike LDA, QMI has advantageous information-theoretic properties as illustrated by Figure 1.

Another consequence of this view of mutual information is that we can now generalise the pairwise interactions between data points. In QMI with Gaussian kernel density estimation, each pairwise interaction G_{nm} is a Gaussian function of $(y_n - y_m)$. The salient characteristics of G_{nm} that give $I_Q(Y;C)$ its information-theoretic properties are as follows. Note that since $y_n = \mathbf{w}^{\mathrm{T}}\mathbf{x}_n$ by definition and that $\|\mathbf{w}\| = 1$, we always have $0 \leq |y_n - y_m| \leq \|\mathbf{x}_n - \mathbf{x}_m\|$.

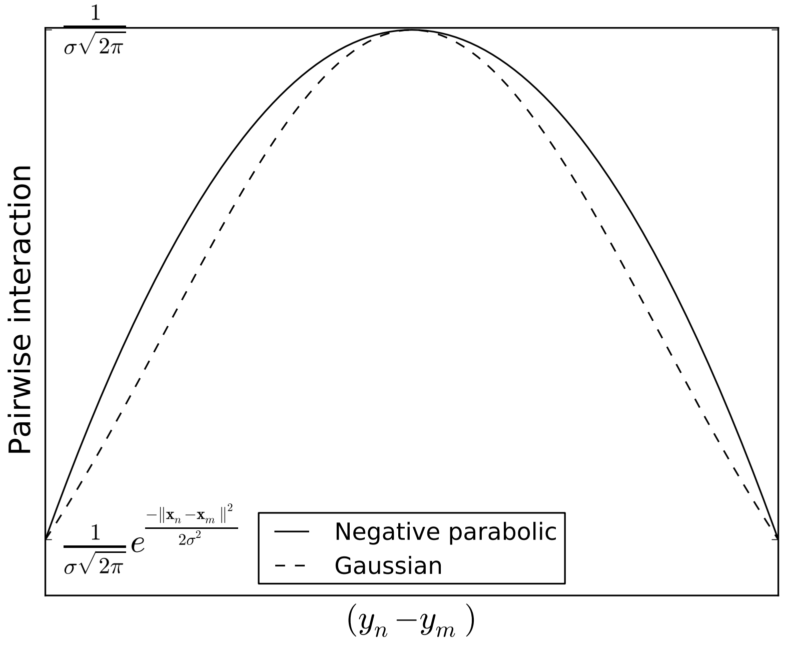
- It is symmetric in $(y_n - y_m)$ and monotonically decreasing in $|y_n - y_m|$.
- It reaches its maximum when $y_n - y_m = 0$, where the maximum is $\frac{1}{\sigma\sqrt{2\pi}}$.
- It reaches its minimum when $|y_n - y_m| = \|\mathbf{x}_n - \mathbf{x}_m\|$, where the minimum is $\frac{1}{\sigma\sqrt{2\pi}} \exp\Big(- \frac{\|\mathbf{x}_n - \mathbf{x}_m\|^2}{2\sigma^2}\Big)$.

All of these properties can be preserved by using an alternative, *negative-parabolic* (as opposed to Gaussian) pairwise interaction, in the form of $a - b(y_n - y_m)^2$. More precisely, define

$$e_{nm} := \frac{1}{\sigma\sqrt{2\pi}}\Big(1 - \frac{1 - e^{-\frac{\|\mathbf{x}_n - \mathbf{x}_m\|^2}{2\sigma^2}}}{\|\mathbf{x}_n - \mathbf{x}_m\|^2}(y_n - y_m)^2\Big) \tag{7}$$

We see that if we view $(y_n - y_m)$ as the abscissa, then the graph of e_{nm} in Equation (7) is a negative parabola, hence the name *negative parabolic* pairwise interaction. Figure 5 illustrates the differences and similarities between e_{nm} and g_{nm}.

Figure 5. Graphs of e_{nm} as in Equation (7), and g_{nm} as in Equation (5), where we view $(y_n - y_m)$ as the abscissa. The two pairwise interactions agree at their maximum (in the middle) and at their minima (two sides).



Now we can measure the mutual information between the data and the class label by using the following, instead of QMI as in Equation (2).

$$I_E(Y;C) := \sum_{n=1}^{N}\sum_{m=1}^{N} \rho_{nm} e_{nm} \tag{8}$$

where ρ_{nm} is the same as in Equation (4). For the example dataset in Figure 1, the black line is the feature along which the maximum value of $I_E(Y;C)$ is obtained. We call $I_E(Y;C)$ *eigenvalue-based mutual information* (EMI), for reasons that will become clear shortly. Figure 6 demonstrates the similarities between EMI, QMI and Shannon's MI (Equation (1)).

The real advantage of using EMI instead of QMI is that it can be optimised analytically, obviating the need for any iterative procedure. e_{nm} can be written as $\mathbf{w}^{\mathrm{T}} E_{nm} \mathbf{w}$ where

$$\begin{aligned} E_{nm} &:= \frac{1}{\sigma\sqrt{2\pi}} \exp\Big[-\frac{(\mathbf{x}_n - \mathbf{x}_m)(\mathbf{x}_n - \mathbf{x}_m)^{\mathrm{T}}}{2\sigma^2}\Big] \\ &= \frac{1}{\sigma\sqrt{2\pi}} \Big[I - \frac{(1 - e^{-\frac{\|\mathbf{x}_n - \mathbf{x}_m\|^2}{2\sigma^2}})}{\|\mathbf{x}_m - \mathbf{x}_m^2\|} (\mathbf{x}_n - \mathbf{x}_m)(\mathbf{x}_n - \mathbf{x}_m)^{\mathrm{T}}\Big] \end{aligned} \tag{9}$$

where the second line follows from evaluating the matrix exponential in the first line. Thus, if we define a matrix E by

$$E := \sum_{n=1}^{N}\sum_{m=1}^{N} \rho_{nm} E_{nm} \tag{10}$$

then we see that EMI can be written as $I_E(Y;C) = \mathbf{w}^{\mathrm{T}} E \mathbf{w}$. Now finding a feature that maximises EMI is equivalent to maximising $I_E(Y;C)$ in $\mathbf{w}$, and we see that the maximising $\mathbf{w}$ are just the largest eigenvectors of the matrix E.

The reader may notice some similarities between e_{nm} and the *Epanechnikov kernel*, which is defined as follows.

$$h_{nm} := \frac{3}{4}(1 - (y_n - y_m)^2)\mathbb{I}[|y_n - y_m| \leq 1] \tag{11}$$

This also has a negative parabolic shape, but there are several fundamental differences between this and our pairwise interaction e_{nm}. First, e_{nm} has a variable width that depends on $\|\mathbf{x}_n - \mathbf{x}_m\|$, while in contrast h_{nm} does not, and is fixed-width. In particular, h_{nm} does not take into account any pairs of points for which $|y_n - y_m| > 1$. Moreover, the indicator function $\mathbb{I}[|y_n - y_m| \leq 1]$ cannot be encoded in a matrix in the way that e_{nm} can be encoded in the matrix E_{nm} (Equation (9)) via $e_{nm} = \mathbf{w}^{\mathrm{T}} E_{nm} \mathbf{w}$. It is the ability of a pairwise interaction (or kernel) to be encoded in a matrix that allows the associated dimensionality reduction algorithm to be formulated as an eigenvalue-problem.

We end this section with a brief illustration of a practical application of EMI. More experiments using EMI can be found in [8]. Figure 7 shows the average classification error rates of the nearest-neighbour classifier through 10 repeats of 5-fold cross-validation, for the subspaces computed by PCA, the EMI method, and the iterative QMI method respectively, using the Pima Indians Diabetes Dataset available from the UCI machine learning repository. Figure 8 shows the 2D projections computed by the three methods. Note the similarity between the projections computed by the two information-theoretic methods, in contrast to that of PCA. Note also that a 2D projection cannot be computed using LDA since this is a 2-class problem and LDA would only be able to compute one feature.

Figure 6. The values of the various measures of mutual information along a feature, as the feature is rotated π radians. The dataset used is the one shown in Figure 1. Note that some of the measures are rescaled so that all measures are visually comparable. However, it is not the actual value of the measure that matters in the context of optimisation, but rather the shape of the graph. We see that EMI, QMI and traditional MI peak at almost the same place. For the sake of comparison, a non-information theoretic measure is included, that is, Fisher's discriminant (LDA). We see that Fisher's discriminant measure does not peak at the "right" place. The Fisher-optimal feature (computed by LDA) is shown as the blue dashed line in Figure 1.

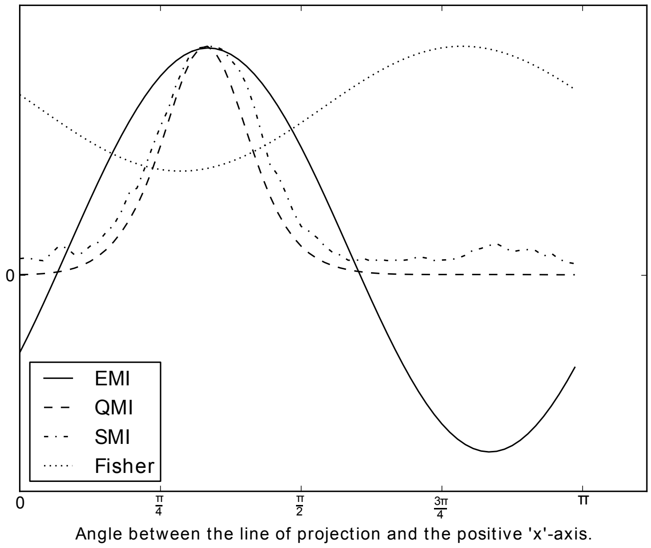


Figure 7. Average classification error rates of the nearest-neighbour classifier through 10 repeats of 5-fold cross-validation, for the subspaces computed by PCA, EMI and the iterative QMI method respectively. The dataset used is the Pima Indians Diabetes Dataset, available from the UCI Machine Learning Repository. LDA was not included in this evaluation because the dataset only has 2 classes and LDA would only be able to extract one feature. We see that EMI has lower error rates than PCA and Torkkola's iterative QMI method.

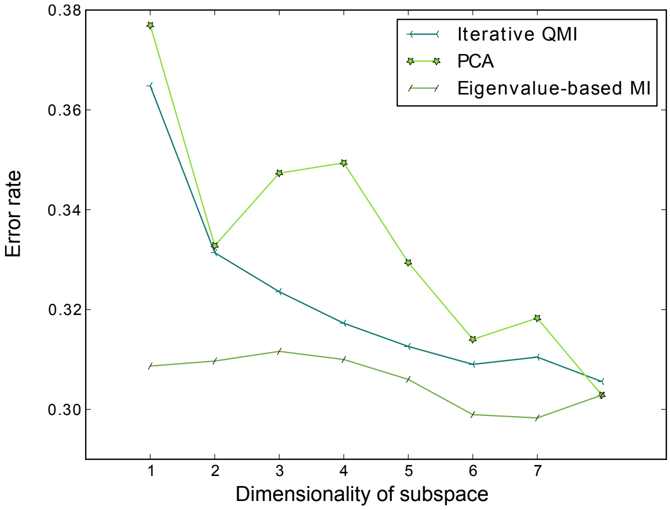
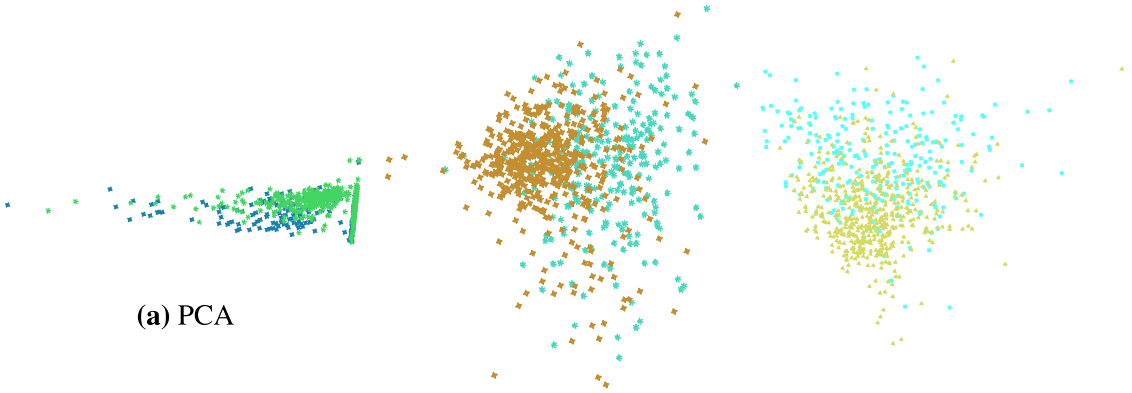


Figure 8. 2D projections of the Pima dataset computed by three feature extraction techniques.



(a) PCA

(b) EMI

(c) Iterative QMI

3. Relationship between EMI and Laplacian Eigenmaps

Laplacian eigenmaps [2,13] is a dimensionality reduction technique that is based on preserving local properties of the data. While information-theoretic techniques aim to find low-dimensional subspaces that maximise the mutual information between the data and the class label, Laplacian eigenmaps find low-dimensional subspaces that minimize the distances between nearby data points. This is achieved by using pairwise weights ω_{nm} on pairs of data points $\{\mathbf{x}_n, \mathbf{x}_m\}$. The general problem is to maximise (or minimise) (see Table 3 for notation)

$$\sum_{n=1}^{N}\sum_{m=1}^{N}\omega_{nm}\mathrm{tr}[(\mathbf{y}_n - \mathbf{y}_m)(\mathbf{y}_n - \mathbf{y}_m)^{\mathrm{T}}] \tag{12}$$

As an aside, this can be viewed as a generalisation to PCA, since the problem of PCA can be viewed as the maximisation of a special case of (12) where all the weights ω_{nm} are 1. The solution to the general problem (12) are the (largest or smallest) eigenvectors of the matrix $\mathcal{X}^{\mathrm{T}}L\mathcal{X}$. L is called a *Laplacian matrix*, it is symmetric and has the property that each row (column) sums to 0. We refer the reader to [13] for more algebraic details regarding Laplacian eigenmaps.

Table 3. Extra notation.

$\mathcal{X}$	*Design matrix*, whose n^{th} row is the n^{th} data point $\mathbf{x}_n^{\mathrm{T}}$.
$\mathbf{y}_n$	The n^{th} data point projected onto a low-dimensional subspace. If the dimensionality of the subspace is 1, then we use the notation y_n (Table 1).
$\mathcal{Y}$	The *design matrix* in a low-dimensional subspace, whose n^{th} row is the projected data point $\mathbf{y}_n^{\mathrm{T}}$.
$\mathbf{e}$	A vector, all of whose elements are 1.
ω_{nm}	Pairwise weights used in Laplacian eigenmaps.
Ω	The $N \times N$ weight matrix, whose $(n,m)^{\text{th}}$ element is ω_{nm}.
L	The *Laplacian matrix*, defined by $\operatorname{diag}(\Omega\mathbf{e}) - \Omega$.

From Equations (7) and (8), we can re-write the EMI along a feature $\mathbf{w}$ as the following.

$$I_E(Y;C) = \left(\sum_{n=1}^{N}\sum_{m=1}^{N} \frac{\rho_{nm}}{\sigma\sqrt{2\pi}}\right) - \left(\sum_{n=1}^{N}\sum_{m=1}^{N} \rho_{nm} \frac{(1 - e^{-\frac{\|\mathbf{x}_n - \mathbf{x}_m\|^2}{2\sigma^2}})}{\sigma\sqrt{2\pi}\|\mathbf{x}_n - \mathbf{x}_m\|^2} (y_n - y_m)^2\right) \tag{13}$$

Note that the first term on the right-hand-side of Equation (13) does not involve $\mathbf{w}$, and so is irrelevant in the maximisation. If we want a low-dimensional subspace spanned by M features $\mathbf{w}_1, \ldots, \mathbf{w}_M$ that maximises the EMI along each feature, then the quantity we aim to maximise can be written as follows.

$$I_E(\mathbf{Y};C) = \left(M\sum_{n=1}^{N}\sum_{m=1}^{N} \frac{\rho_{nm}}{\sigma\sqrt{2\pi}}\right) - \left(\sum_{n=1}^{N}\sum_{m=1}^{N} \rho_{nm} \frac{(1 - e^{-\frac{\|\mathbf{x}_n - \mathbf{x}_m\|^2}{2\sigma^2}})}{\sigma\sqrt{2\pi}\|\mathbf{x}_n - \mathbf{x}_m\|^2} \operatorname{tr}[(\mathbf{y}_n - \mathbf{y}_m)(\mathbf{y}_n - \mathbf{y}_m)^{\mathrm{T}}]\right) \tag{14}$$

Now if we define the pairwise weights

$$\omega_{nm} := \rho_{nm} \frac{(1 - e^{-\frac{\|\mathbf{x}_n - \mathbf{x}_m\|^2}{2\sigma^2}})}{\sigma\sqrt{2\pi}\|\mathbf{x}_n - \mathbf{x}_m\|^2} \tag{15}$$

noting that the first term on the right-hand-side of Equation (14) does not involve $\mathbf{w}_1, \ldots, \mathbf{w}_M$, we see that maximising EMI as in Equation (14) is equivalent to minimising expression (12), whose solution is given by the smallest eigenvectors of the matrix $\mathcal{X}^{\mathrm{T}} L \mathcal{X}$.

In Section 2.1 we saw that we can view EMI as an alternative to QMI where we use negative parabolic pairwise interactions instead of Gaussian pairwise interactions between training data points. In this section we see that another view of EMI maximisation is a special case of Laplacian eigenmaps, where the weights are chosen as in Equation (15). While the Laplacian eigenmaps method has been shown to bear some relation to maximum entropy methods in unsupervised dimensionality reduction [15], so far it has not been shown to be related to mutual information. Hence, it is interesting to see that a set of weights ω_{nm} can be discovered (Equation (15)) that produces a special case of Laplacian eigenmaps maximising an estimate of mutual information.

The difference between EMI maximisation and the original formulation of Laplacian eigenmaps is that in the original formulation of Laplacian eigenmaps, the weight matrix Ω, and consequently the Laplacian matrix L, is sparse. In contrast, the weight matrix and the Laplacian matrix for EMI maximisation are dense.

4. Conclusions and Further Research

We have introduced a measure of mutual information between a dataset and a discrete target variable (class label) that can be maximised analytically and has practical advantages over the current state-of-the-art QMI. The motivation for using information-theoretic measures in dimensionality reduction stems from the fact that classic non-information-theoretic techniques, such as PCA and LDA, deteriorate for some data distributions, as shown in Figure 1. We have studied the pairwise interaction view of QMI, which led to the formulation of EMI. We have shown some similarities and differences between EMI and other measures of mutual information, and have briefly demonstrated the practical applicability of EMI in dimensionality reduction for classification. Finally, we have shown some relationships between EMI and Laplacian eigenmaps, which is a widely used dimensionality reduction algorithm.

The behaviour of information-theoretic algorithms for dimensionality reduction can be counterintuitive in high-dimensional spaces. The dataset used in our experiment in Figure 7 is relatively low-dimensional. High dimensional datasets pose a computational challenge for information-theoretic algorithms, due to the high computational complexity of iterative algorithms for QMI maximisation. With the introduction of EMI, it is now possible study the behaviour of information-theoretic dimensionality reduction for high-dimensional datasets, such as face and image recognition, at a significantly lower computational cost. However, for high-dimensional data, our experiments have shown that while EMI maximisation is good for data visualisation and representation, it gives poor classification results. Our current experiments indicate that this is due to *over-fitting* [5,16]. A recent review of dimensionality reduction algorithms [2] has found that despite the sophistication of more modern algorithms, the best classification performance for real-world data is typically observed with PCA. Our current experimental results agree with this. It seems that the reason for this lies in the fact that the benefits offered by MI-based methods over traditional methods (Figure 1) become less relevant in high-dimensional spaces, where classic non-information-theoretic methods such as LDA often succeed in maximising class-discrimination. We briefly discussed this in Section 2 (Figure 4). Future research into the exact mechanisms that generate this phenomenon and whether we can *reliably* improve on PCA for real-world data (as opposed to for only a small subset of applications) will be of great practical importance.

Acknowledgements

This research is funded by UK EPSRC, through the Departmental Teaching Award scholarship of Department of Computing at Imperial College.

References

1. Guyon, I.; Gunn, S.; Nikravesh, M.; Zadeh, L. *Feature Extraction: Foundations and Applications*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 207.
2. Van der Maaten, L.; Postma, E.; van den Herik, H. Dimensionality reduction: A comparative review. *J. Mach. Learn. Res.* **2009**, *10*, 1–41.

3. Torkkola, K. Feature extraction by non parametric mutual information maximization. *J. Mach. Learn. Res.* **2003**, *3*, 1415–1438.
4. Kapur, J. *Measures of Information and Their Applications*; Wiley: New Delhi, India, 1994.
5. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer-Verlag New York, Inc.: Secaucus, NJ, USA, 2006.
6. Qiu, X.; Wu, L. Info-margin maximization for feature extraction. *Pattern Recognit. Lett.* **2009**, *30*, 1516–1522.
7. Hild, K.; Erdogmus, D.; Torkkola, K.; Principe, J. Feature extraction using information-theoretic learning. *Pattern Anal. Mach. Intell. IEEE Trans.* **2006**, *28*, 1385–1392.
8. Liu, R.; Gillies, D.F. An Eigenvalue-problem Formulation for Non-parametric Mutual Information Maximisation for Linear Dimensionality Reduction. In Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition, Las Vegas, USA, 16–19 July 2012; Volume 2, pp. 905–910.
9. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, Volume 27, pp. 379–423, 623–656.
10. Cover, T.; Thomas, J. *Elements of Information Theory*; Wiley-interscience: Hoboken, NJ, USA, 2006.
11. Renyi, A. *On Measures of Entropy and Information*; University of California Press: Berkeley, CA, USA, 1961; Volume 1, pp. 547–561.
12. Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076.
13. Belkin, M.; Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **2003**, *15*, 1373–1396.
14. Sharpee, T.; Rust, N.C.; Bialek, W. Analyzing neural responses to natural signals: Maximally informative dimensions. *Neural Comput.* **2004**, *16*, 223–250.
15. Lawrence, N. Spectral Dimensionality Reduction via Maximum Entropy. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011; pp. 51–59.
16. Jain, A.; Duin, R.; Mao, J. Statistical pattern recognition: A review. *Pattern Anal. Mach. Intell. IEEE Trans.* **2000**, *22*, 4 –37.