Variable Selection for Regression Problems Using Gaussian Mixture Models to Estimate Mutual Information

Emil Eirola, Amaury Lendasse, and Juha Karhunen

***Abstract*— Variable selection is a crucial part of building regression models, and is preferably done as a filtering method independently from the model training. Mutual information is a popular relevance criterion for this, but it is not trivial to estimate accurately from a limited amount of data. In this paper, a method is presented where a Gaussian mixture model is used to estimate the joint density of the input and output variables, and subsequently used to select the most relevant variables by maximising the mutual information which can be estimated using the model.**

I. Introduction

In machine learning, variable selection (or feature selection) is an important phase, since discarding irrelevant variables not only results in a simpler model, but often also leads to improved generalisation accuracy [1].

Mutual information (MI) is a measure of dependence between variables. Applied to regression problems, estimating the mutual information between inputs and outputs is an effective procedure for identifying useful variable sets [2]. In general, mutual information accounts for all forms of dependence between variables, and is as such an ideal criterion for variable selection in machine learning. Its use is only restricted by the practical difficulties in estimating it accurately from data.

A multivariate Gaussian mixture model (GMM) [3] is a parametric model of a probability density, and with a sufficient number of components can be used to approximate any arbitrary continuous distribution. The EM algorithm is an efficient and accurate method to fit the model to a given set of data.

There are several reasons which make Gaussian mixtures an appealing method for estimating mutual information for feature selection:

1) After fitting the mixture model to the full set of variables, the model can directly be used to calculate the mutual information for any subset of variables. This is useful in variable selection.
2) Estimates for different variable sets seem to behave more consistently than with using other estimators. In particular, the estimate of the mutual information always increases when adding variables, as it should.
3) As the Gaussian mixture can be fit to data with missing values, the estimator works for such incomplete data sets as well.

Mutual information estimators based on density estimates have generally been discouraged in the literature due to the difficulty of obtaining accurate estimates, but should not be ignored entirely. This paper shows that Gaussian mixture models can be used very effectively for this purpose.

This paper is structured as follows: after reviewing related work on the topic in Section II, the proposed approach for mutual information estimation by Gaussian mixtures is detailed in Section III. Experimental results on synthetic data with missing values and on several benchmark regression tasks are presented in Section IV with comparisons to other methods.

II. Related work

A. Estimating mutual information

Early attempts to estimate mutual information from data with an unknown structure have been based on binning and histograms. That approach is not feasible for more than a couple of variables, as the amount of data required for accurate estimation grows exponentially.

More recently, Kraskov et al [4] proposed a method to estimate mutual information by considering nearest neighbours of each point in the input and output spaces separately and together. This approach has proven effective, and gained popularity.

Maximum Likelihood Mutual Information (MLMI) [5], [6] is another recent development promising accurate estimates.

B. Variable selection by mutual information

The seminal work on feature selection with mutual information [2] formulates the problem as follows: find the subset with k features that maximises the mutual information, for some a priori fixed value k. The MI is only estimated from histograms and binning.

For classification problems, a method involving kernel density estimation for the conditional distribution of each class has been used to estimate mutual information for variable selection [7], and later extended to dimensionality reduction [8].

A suggestion for regression problems is to find variables which maximise Kraskov's mutual information estimator [9]. To gauge the uncertainty of Kraskov's estimator, a resampling strategy has been proposed which can also help in determining how many variables to select in a forward search [10]. Extending variable selection to datasets with missing

Emil Eirola and Juha Karhunen are with the Department of Information and Computer Science, Aalto University, Finland

Amaury Lendasse is with the Department of Information and Computer Science, Aalto University, Finland, Arcada University of Applied Sciences, Helsinki, Finland, IKERBASQUE, Basque Foundation for Science, Spain, and the Department of Mechanical and Industrial Engineering, The University of Iowa, USA.

values, the partial distance strategy (PDS) has been used to find nearest neighbours for Kraskov's estimator [11].

While optimising mutual information generally leads to accurate models in more concrete performance measures (classification rate, mean squared error), it has been shown that pathological examples exist where this is not true [12]. The adequacy of mutual information for estimating prediction accuracy is more precisely detailed in [13].

C. Gaussian mixture models and mutual information

Gaussian mixture models have previously been used to estimate mutual information in a speech processing application [14]. The joint density of the variables is estimated by GMM, and separate GMMs for the marginal distributions. For estimating the mutual information, the authors interpret the integral as an expectation, and estimate it as the sample mean over the available data. In another example from speech recognition [15], a low-order Gaussian mixture model of the distribution is used to derive MI estimates by numerical integration.

A method for dimensionality reduction [16] uses a Gaussian mixture with only two components to find an expression for the mutual information which can then be used as the target function for an optimisation scheme.

For classification, the GMM-MI method [17] conducts feature selection with MI, by using GMMs to model the conditional distributions of each class. A new mixture model is estimated whenever considering new features. The same idea has also been applied to image classification in computer vision [18].

In contrast with all previous approaches, we suggest to train only one Gaussian mixture model on the joint space of the output and all input variables, and the mutual information for any combination of variables can be extracted from that.

III. Mutual Information by Mixture of Gaussians

A. Definition

The mutual information is a measure of dependence between two random variables. It can be defined through the Shannon entropy:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{1}$$

where the entropy H is the expected information content which can be written in terms of a random variable's probability density function $p_x(X)$:

$$H(X) = \mathrm{E}[-\log(p_x(X))] \tag{2}$$

For continuous random variables X, Y with a joint distribution described by the density $p(x,y)$, the definition is equivalent to the integral below.

$$I(X;Y) = \int_Y \int_X p(x,y) \log\left(\frac{p(x,y)}{p_x(x)p_y(y)}\right) dx\, dy \tag{3}$$

Here p_x and p_y are the marginal probability densities of the random variables.

Another interpretation of mutual information is that it is the Kullback–Leibler (KL) divergence of the product of the marginal distributions $(p_x(x)p_y(y))$ from the joint distribution $p(x,y)$. The KL divergence is a measure of the difference between the distributions. If X and Y are independent random variables, the joint density is separable as the product $p(x,y) = p_x(x)p_y(y)$, and the divergence is 0. The more dependent the variables are, the larger the divergence, and the higher the value of the mutual information is.

B. Estimating Mutual Information

The main idea is to use a Gaussian mixture model to estimate the densities of the variables. However, instead of directly calculating Eq. 3, we consider Eq. 2, and interpret the integral as an expectation.

$$\begin{aligned} I(X;Y) &= \int_Y \int_X p(x,y) \log\left(\frac{p(x,y)}{p_x(x)p_y(y)}\right) dx\, dy && (4) \\ &= \mathrm{E}\left[\log p(x,y) - \log p_x(x) - \log p_y(y)\right] && (5) \end{aligned}$$

Given a sample of data $\{x_i, y_i\}_{i=1}^N$, the expectation can be approximated by the arithmetic mean over the data:

$$\hat{I}(X;Y) = \frac{1}{N}\sum_{i=1}^{N}\left(\log p(x_i,y_i) - \log p_x(x_i) - \log p_y(y_i)\right) \tag{6}$$

The proposed approach is based on this expression, requiring only estimates of the density and marginal density at each point of data. By fitting a Gaussian mixture model to the joint space $X \times Y$, the resulting model directly provides an estimate of $p(x_i, y_i)$. To calculate the marginal probability densities, the *same* Gaussian model is used, restricted to the appropriate variables. The marginal model is easily acquired by only including the appropriate elements from the means and covariances of each Gaussian component.

As the goal is to evaluate differences between the joint density $p(x,y)$ and the product $p_x(x)p_y(y)$, the same model should be used to estimate all the quantities. It might seem reasonable to separately optimise another mixture model in the space for X to estimate p_x instead, and this could result in a more accurate estimate for p_x itself, but could also lead to spurious differences causing an inflated KL divergence. Having consistent estimates is particularly important for variable selection, where mutual information estimates for different variable sets are compared to each other.

C. Feature selection

In machine learning, the goal is to find a model that can predict an output variable Y from several input variables X. As X can be high-dimensional, discarding redundant and irrelevant parts both simplifies the model and increases its interpretability

When selecting variables for a machine learning task, the mutual information with the output is an intuitive choice for a relevance criterion. However, the mutual information never decreases when adding irrelevant variables. Thus an exhaustive search over all feature sets is meaningless, as the criterion is maximised when all variables are included. The forward search is a more practical approach; here variables are added one by one, at each step selecting the variable which leads to the largest increase in MI when considered together with the previously selected variables. The order of successive selection then leads to a ranking of variables: the first selected variable can be seen as the most important, and so on.

In the present approach, the first step is to fit a Gaussian mixture model to the joint space $X \times Y$. Then, for a given subset of variables of X, the marginal mixture model for those variables is easily realised. Having a GMM with K components in the $X \times Y$ space with mixing coefficients π_k, means $\boldsymbol{\mu}_k$, and covariances $\boldsymbol{\Sigma}_k$ for each component k ($0 < \pi_k < 1$, $\sum_{k=1}^{K} \pi_k = 1$), the parameters can be partitioned as below:

$$\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_k^X \\ \boldsymbol{\mu}_k^Y \end{bmatrix}, \quad \boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_k^{XX} & \boldsymbol{\Sigma}_k^{XY} \\ \boldsymbol{\Sigma}_k^{YX} & \boldsymbol{\Sigma}_k^{YY} \end{bmatrix}. \qquad (7)$$

The marginal model for X is directly determined as a GMM of K components with the same mixing coefficients π_k, but means $\boldsymbol{\mu}_k^X$ and covariance matrices $\boldsymbol{\Sigma}_k^{XX}$. The marginal GMM is similarly found for Y, and for any subspaces of X corresponding to different sets of selected variables.

D. Fitting the Gaussian mixture model

The EM algorithm can be used to fit a Gaussian mixture model to a given dataset. The only parameter to determine beforehand is the number of components to use. A low number of components is unable to retain all the relevant properties of the distribution, whereas using too many components often leads to exaggerating spurious features of the data.

The number of components can be selected according to the Akaike information criterion (AIC) [19], expressed as a function of the log-likelihood $\mathcal{L}(\theta)$ of the converged mixture model:

$$\text{AIC} = -2 \log \mathcal{L}(\theta) + 2P, \qquad (8)$$

where P is the number of free parameters. Several alternative criteria are discussed in [20, Ch. 6].

E. High-dimensional data

A limitation of the conventional Gaussian mixture model is that the number of parameters grows quadratically with the number of dimensions. However, cases with high-dimensional data is precisely when variable selection is most crucial.

High-dimensional data clustering (HDDC) [21] is a variant of Gaussian mixture models for high-dimensional data that works by restricting the structure of the covariance matrices. In essence, HDDC retains the shape of a cluster in directions corresponding to the largest eigenvalues of its covariance matrix, and assumes it is spherical in other directions. Thereby the number of free parameters to learn is reduced significantly.

The extension makes it possible to still use the EM algorithm to fit a mixture model on data with a large number of variables, and the model is directly usable for feature selection through MI estimation.

F. Missing values

Datasets with missing values are a common occurrence in machine learning tasks. The EM algorithm for fitting a Gaussian mixture model has been extended to handle such data in a natural way [22], [23], [24].

An assumption here is that data are Missing-at-Random (MAR) [25]:

$$P(M \mid x_{\text{obs}}, x_{\text{mis}}) = P(M \mid x_{\text{obs}}), \qquad (9)$$

i.e., the event M of a measurement being missing is independent from the value it would take (x_{mis}), conditional on the observed data (x_{obs}).

This implies that samples with partial information can still be included in the estimation, and need not be discarded. Making maximal use of all available data is important for accurately finding the most relevant set of variables.

G. MI and mean squared error

While concerns have been raised over the use of MI as representative of prediction error [12], [13], there is a clear connection between the two measures. In the general case, MI implies a lower bound for the mean squared error (MSE) of an arbitrary estimator $\hat{Y}(X)$:

$$\text{E}[(Y - \hat{Y}(X))^2] \geq \frac{1}{2\pi e} e^{2H(Y|X)} \qquad (10)$$

[26, Thm. 8.6.6] if the entropy H is in nats (base e logarithm). Here equality is achievable only for the optimal estimator $\hat{Y}(X) = \text{E}(Y|X)$ *and* if the residual $Y - \hat{Y}(X)$ is Gaussian.

Since $H(Y|X) = H(Y) - I(X;Y)$ we have that

$$\begin{aligned} \text{E}[(Y - \hat{Y}(X))^2] &\geq \frac{1}{2\pi e} e^{2(H(Y) - I(X;Y))} \\ &= C e^{-2I(X;Y)} \end{aligned} \qquad (11)$$

where $C = \frac{1}{2\pi} e^{2H(Y)-1}$ is a constant that does not depend on the chosen variables X. Increasing the MI reduces the lowest achievable error.

IV. Experiments

A. Synthetic experiment with missing values

In this section, we study the use of the mutual information estimator as a variable selection method in the presence of missing values in the data set. In order to accurately assess the quality of the selection, a synthetic data set is generated. Following [10], [11], a dataset of 100 samples with 10 input variables is generated. The input variables X_i for $i \in \{1, 10\}$ are independently uniformly distributed between 0 and 1, and the output Y uses only the first five variables:

$$Y = \sin(X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon, \quad (12)$$

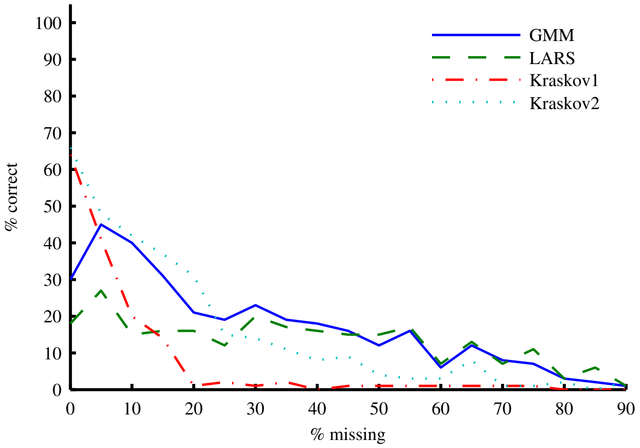

Fig. 1. Comparison of the ratio of correct variable selection between the Gaussian Mixture Model (solid blue), LARS (dashed green) and two variants of Kraskov's MI estimator (dash-dot red, dotted cyan) on Eq. 12 with 100 samples.

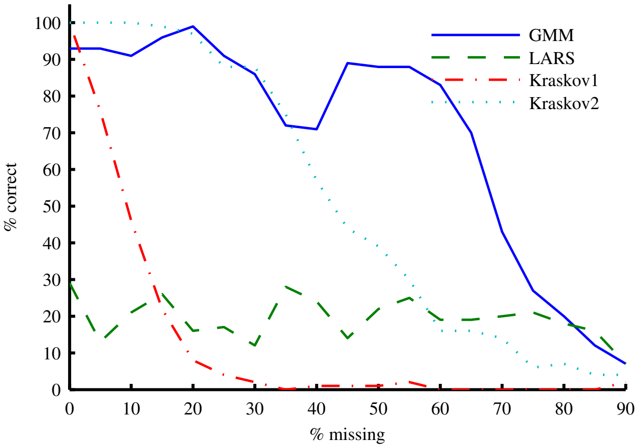

Fig. 2. Comparison of the ratio of correct variable selection between the Gaussian Mixture Model (solid blue), LARS (dashed green) and two variants of Kraskov's MI estimator (dash-dot red, dotted cyan) on Eq. 12 with 500 samples.

where ε is Gaussian noise with zero mean and unit variance. Obviously, only the first five variables are useful for estimating the output, and variables 6–10 are irrelevant. Missing values are introduced by randomly and independently removing values from the input variables until a certain fraction of the data is missing, for values from 0% up to 90%.

The quality of a ranking of variables is then determined by the fraction of cases when the "correct" variables are found. In this case, correct means that features 1–5 are ranked as the first five variables in any order.

The Gaussian mixture model for mutual information estimation is compared to three other methods for variable selection with missing data:

1) LARS: Least angle regression [27]. Missing data are imputed by the mean of each variable.
2) Kraskov1: Following [11], the algorithm of [4] using Euclidean distance and PDS to deal with missing data.
3) Kraskov2: Following the software of [4], this is using the maximum norm to determine nearest neighbours. Missing data are imputed by the mean of each variable.

For LARS, imputation by the mean is justified since values equal to the mean of a variable do not contribute to the estimation of linear correlations. In the method Kraskov2, imputation by the mean is reasonable, considering this has a conservative effect on the maximum norm.

Results with 100 samples are shown in Fig. 1 for a missing value ratio ranging from 0–90%. The lines represents the ratio of cases where all five relevant variables are correctly identified. Both variants of Kraskov's method display good accuracy with no or few missing values, but the quality quickly deteriorates when the ratio is increased. The Gaussian mixture model is more consistent, but the low number of samples is an issue as it occasionally leads to too few components being used.

As all methods display disappointing performance with such a low number of samples, the amount is increased to 500. The results are displayed in Fig. 2, and the accuracy has improved as expected. In particular, the Gaussian mixture model can here identify all variables correctly with a high certainty even with 50–60% of missing data.

The relatively low performance of LARS is explained by it being unable to identify the importance of variable X_3, as it only considers linear dependencies. The other four variables are reliably identified.

In both experiments, the maximum norm variant of Kraskov's estimator consistently outperforms the Euclidean version. Increasing the ratio of missing values rapidly reduces the accuracy of both variants. While MI using estimators based on nearest neighbours can work well, there is a risk that they focus too strongly on local effects, ignoring strong global trends.

B. *Real-world data regression problems*

To assess the performance on more realistic problems as well, several benchmark regression tasks are studied. The datasets are presented in Table I, and the proposed approach is compared with four other methods for variable selection:

1) LARS: Least angle regression [27].
2) RReliefF: a method to rank variables, based on nearest neighbours and how local differences in each variable correspond to changes in the output [28].
3) Mutual Information by Kraskov's estimator [4] with the maximum norm.
4) Variable selection by Maximum Likelihood Mutual Information (MLMI) estimation [5], [6].

The comparison criterion is the mean squared error of a least squares support vector machine (LS-SVM) [31] regression model, as this model is known to be sensitive to redundant variables. The model is trained using the selected variable set, and the median (over repeated runs of optimising hyperparameters) leave-one-out error is calculated. This can be considered a fair criterion for comparing the *selections*

TABLE I

DATA SETS USED FOR THE EXPERIMENTS, WITH THE NUMBER OF SAMPLES (N), NUMBER OF VARIABLES (d), AND SOURCE.

Name	N	d	Source
Auto-price	159	15	[29]
Stocks	950	9	[29]
Housing	506	13	[30]
Breast Cancer Wisconsin (Prognostic)	194	32	[30]
Tecator	240	100	[1]
Triazines	186	60	[29]
Anthrokids	1019	53	[2]

[1] http://www.dm.unibo.it/~simoncin/tecator

[2] http://ovrt.nist.gov/projects/anthrokids/

TABLE II

THE SELECTED INPUTS AND LOO ERRORS FOR THE AUTO-PRICE DATA. BOLD VALUES REPRESENT OPTIMAL CHOICES IN THE SENSE OF THE LOWEST ERROR WITH THE SMALLEST SET OF VARIABLES.

LARS		RReliefF		Kraskov		MLMI		GMM	
7	**0.1638**	8	0.2734	7	**0.1638**	8	0.2734	7	**0.1638**
5	0.1581	5	0.1690	5	0.1581	10	0.2774	12	**0.1522**
8	0.1205	7	0.1205	12	**0.1136**	2	0.2563	5	**0.1136**
12	**0.0994**	12	**0.0994**	14	0.1065	4	0.2099	2	0.1184
2	0.1226	2	0.1226	4	0.1278	7	0.1694	6	**0.0993**
11	0.1281	3	0.1179	15	0.1315	14	0.1616	4	0.1060
3	0.1261	10	0.1405	8	0.1329	13	0.1437	9	0.1089
13	0.1255	9	0.1432	10	0.1342	6	0.1292	3	0.1174
10	0.1484	13	0.1278	9	0.1397	15	0.1306	11	0.1074
9	0.1427	15	0.1245	11	0.1503	1	0.1471	14	0.1255
4	0.1420	6	0.1172	3	0.1601	5	0.1218	13	0.1046
1	0.1370	4	0.1090	13	0.1377	9	0.1088	10	0.1273
6	0.1262	11	0.1317	6	0.1141	11	0.1185	8	0.1218
14	0.1171	14	0.1219	2	0.1219	12	0.1235	1	0.1171
15	0.1183	1	0.1183	1	0.1183	3	0.1183	15	0.1183

of variables. As a preprocessing step, all variables including the target variable are standardised to zero mean and unit variance before the variable selection process.

The HDDC method [21] is used for fitting the Gaussian mixture model on data with twenty or more variables.

All the criteria are used with a forward search approach for selecting variables. This results in a ranking of variables, and the variable sets formed by successively selecting the selected variables are evaluated by the resulting LS-SVM accuracy.

Results on each data set for up to 30 selected variables are presented in Tables II–VIII, where values typeset in bold represent optimal choices in the sense of achieving the lowest error with the smallest set of variables among the presented methods.

For the Auto-price data (Table II) it is clear that only a few of the 15 variables are relevant for the prediction. The variables selected by GMM generally lead to a lower prediction error than those selected by the other methods.

Table III presents results on the Stocks data, which is an example of a problem where all variables are important for predicting the target. GMM and Kraskov's method perform the best in finding smaller sets leading to reasonable prediction accuracy.

TABLE III

THE SELECTED INPUTS AND LOO ERRORS FOR THE STOCKS DATA. BOLD VALUES REPRESENT OPTIMAL CHOICES IN THE SENSE OF THE LOWEST ERROR WITH THE SMALLEST SET OF VARIABLES.

LARS		RReliefF		Kraskov		MLMI		GMM	
1	**0.1714**	1	**0.1714**	1	**0.1714**	1	**0.1714**	1	**0.1714**
6	0.0824	4	0.0717	2	0.0719	2	0.0719	5	**0.0699**
9	0.0228	9	0.0296	7	0.0259	6	0.0194	9	**0.0172**
5	0.0133	3	0.0153	5	**0.0119**	5	0.0130	6	0.0133
7	0.0109	7	0.0124	4	**0.0096**	4	0.0109	2	0.0112
2	0.0096	8	0.0121	6	0.0092	9	**0.0091**	3	0.0110
8	0.0094	5	0.0094	3	**0.0085**	8	0.0096	7	0.0095
4	0.0090	2	0.0086	9	**0.0084**	7	0.0090	4	**0.0084**
3	**0.0083**	6	**0.0083**	8	**0.0083**	3	**0.0083**	8	**0.0083**

TABLE IV

THE SELECTED INPUTS AND LOO ERRORS FOR THE HOUSING DATA. BOLD VALUES REPRESENT OPTIMAL CHOICES IN THE SENSE OF THE LOWEST ERROR WITH THE SMALLEST SET OF VARIABLES.

LARS		RReliefF		Kraskov		MLMI		GMM	
13	**0.3236**	6	0.4257	13	**0.3236**	13	**0.3236**	13	**0.3236**
6	**0.2323**	13	**0.2323**	6	**0.2323**	11	0.2416	6	**0.2323**
11	0.2037	5	0.1901	10	**0.1516**	5	0.2018	11	0.2037
12	0.1909	8	0.1739	5	**0.1476**	10	0.1918	8	0.1531
4	0.1772	4	0.1829	9	**0.1305**	9	0.1946	5	0.1359
1	0.1544	10	0.1571	1	**0.1158**	3	0.1893	4	0.1434
8	0.1435	12	0.1379	12	0.1205	6	0.1167	12	0.1366
5	0.1331	2	0.1316	7	0.1176	8	**0.1129**	2	0.1395
2	0.1360	9	0.1150	8	**0.0964**	7	0.1094	1	0.1360
3	0.1290	11	0.1067	3	**0.0892**	4	0.1148	9	0.1237
9	0.1161	3	0.1054	4	0.0991	12	0.0956	10	0.1062
10	0.1048	7	0.0953	2	0.0982	2	0.0953	3	0.1048
7	0.0926	1	0.0926	11	0.0926	1	0.0926	7	0.0926

For the Housing data in Table IV, performance is improved by adding new variables in nearly all cases, and including all the variables leads to an accurate model. The only exception is found by Kraskov's estimator, which manages to find a better performing set of variables by excluding 2, 4, and 11.

The results for the Breast cancer data in Table V reveal that only a few of the 32 variables are relevant, and the variables selected by GMM lead to the most accurate model.

Tecator is the most high-dimensional data set studied here, and the results in Table VI show that GMM again finds the best variables.

The Triazines data set (Table VII) represents a challenging regression problem with varying success for each variable selection method. The lowest prediction error is eventually achieved by RReliefF. Due to some collinearity between the variables, our implementation of LARS identified only 16 variables.

The Anthrokids data contains several variables which are irrelevant for the prediction task, but many of them are still useful, as evidenced by the best result in Table VIII being achieved with a set of 21 variables. Both GMM and Kraskov's estimator successfully identify useful variables at first, but in the end GMM leads to the best performance.

C. Computational times

The computational times for each method are shown in Table IX. The experiments are conducted in MATLAB on a

TABLE V

THE SELECTED INPUTS AND LOO ERRORS FOR THE BREAST CANCER DATA. BOLD VALUES REPRESENT OPTIMAL CHOICES IN THE SENSE OF THE LOWEST ERROR WITH THE SMALLEST SET OF VARIABLES.

LARS		RReliefF		Kraskov		MLMI		GMM	
4	**0.8968**	3	0.9425	5	0.8972	2	0.8975	4	**0.8968**
31	0.8732	26	0.9168	3	0.8466	25	0.8895	3	**0.8429**
3	**0.8342**	23	0.9223	30	0.8410	24	0.8846	31	**0.8342**
30	0.8330	30	0.9041	21	0.8450	16	0.8978	28	**0.8223**
13	0.8253	31	0.8887	2	0.8446	13	0.8537	13	**0.8145**
11	0.8270	27	0.8538	27	0.8509	3	0.8432	10	**0.8081**
1	0.8262	11	0.8597	20	0.8299	18	0.8493	16	0.8106
10	0.8304	7	0.8402	4	0.8284	15	0.8541	22	**0.8051**
18	0.8303	10	0.8461	31	0.8291	5	0.8490	15	0.8060
28	0.8255	28	0.8362	32	0.8314	22	0.8473	1	0.8061
16	0.8172	16	0.8430	24	0.8344	14	0.8504	27	0.8124
32	0.8180	22	0.8529	25	0.8331	19	0.8548	26	0.8307
6	0.8217	24	0.8574	22	0.8319	7	0.8495	6	0.8115
19	0.8263	2	0.8464	15	0.8339	23	0.8537	18	0.8338
8	0.8202	6	0.8516	12	0.8367	11	0.8389	8	0.8292
7	0.8236	4	0.8467	9	0.8417	32	0.8406	32	0.8252
26	0.8271	21	0.8507	19	0.8491	29	0.8444	19	0.8356
20	0.8258	29	0.8539	17	0.8495	10	0.8396	14	0.8327
25	0.8303	20	0.8443	14	0.8504	6	0.8430	9	0.8388
22	0.8306	18	0.8475	16	0.8483	12	0.8448	20	0.8501
27	0.8339	8	0.8429	7	0.8512	9	0.8444	12	0.8435
5	0.8402	19	0.8510	18	0.8474	20	0.8492	23	0.8544
17	0.8354	25	0.8468	26	0.8614	27	0.8536	17	0.8536
15	0.8412	5	0.8475	10	0.8609	4	0.8508	7	0.8554
14	0.8615	9	0.8476	1	0.8638	17	0.8543	11	0.8583
21	0.8542	17	0.8486	23	0.8618	8	0.8519	21	0.8625
9	0.8584	32	0.8584	11	0.8650	21	0.8464	24	0.8598
23	0.8455	12	0.8690	8	0.8479	30	0.9402	25	0.8506
29	0.8584	14	1.0052	13	0.8654	28	1.0052	2	1.0052
24	0.8602	13	1.0052	28	1.0052	31	0.8998	5	1.0052

TABLE VI

THE SELECTED INPUTS AND LOO ERRORS FOR THE TECATOR DATA. BOLD VALUES REPRESENT OPTIMAL CHOICES IN THE SENSE OF THE LOWEST ERROR WITH THE SMALLEST SET OF VARIABLES.

LARS		RReliefF		Kraskov		MLMI		GMM	
41	**0.7137**	41	**0.7137**	40	0.7149	1	0.8643	99	0.7146
7	0.0914	40	0.6625	8	**0.0855**	98	0.4214	7	0.4224
8	0.0882	42	**0.0488**	53	0.0513	61	0.3906	36	0.0688
63	0.0403	39	0.2505	41	0.0459	72	0.2547	28	**0.0359**
62	0.0524	38	0.0433	42	0.0479	59	0.2546	50	**0.0143**
56	0.0512	43	0.1812	52	0.0158	65	0.2548	27	**0.0131**
100	0.0247	37	0.0454	51	0.0216	31	0.1619	29	0.0141
59	0.0348	44	0.0958	43	0.0170	35	0.0176	53	0.0147
55	0.0357	36	0.1486	44	0.0158	69	0.0178	55	0.0181
64	0.0378	35	0.1451	47	0.0164	73	0.0183	51	0.0163
9	0.0161	45	0.0384	45	0.0188	53	0.0186	67	0.0158
54	0.0295	34	0.1214	46	0.0166	23	0.0172	56	0.0135
99	0.0157	33	0.1206	48	0.0157	74	0.0180	49	**0.0130**
53	0.0150	32	0.1196	7	0.0154	57	0.0179	57	0.0158
15	0.0190	46	0.0610	49	0.0190	32	0.0184	26	0.0163
42	0.0139	9	0.0141	50	0.0163	3	0.0178	52	0.0146
17	0.0141	10	0.0142	6	0.0179	5	0.0189	54	0.0173
5	0.0144	8	0.0148	54	0.0164	93	0.0274	24	0.0242
16	0.0178	11	0.0146	36	0.0166	58	0.0241	25	0.0250
52	0.0178	7	0.0153	5	0.0170	30	0.0289	58	0.0298
18	0.0194	100	0.0172	39	0.0186	81	0.0314	65	0.0329
21	0.0213	6	0.0167	37	0.0195	60	0.0431	23	0.0348
22	0.0221	12	0.0232	38	0.0226	11	0.0502	59	0.0758
51	0.0278	99	0.0311	4	0.0333	17	0.0606	22	0.0750
98	0.0484	98	0.0416	56	0.0352	46	0.0689	60	0.0892
91	0.0647	5	0.0733	55	0.0477	79	0.0842	64	0.1096
40	0.0691	13	0.0863	2	0.0747	85	0.1073	63	0.1569
19	0.1611	97	0.0994	3	0.0804	6	0.2136	61	0.3081
20	0.2143	4	0.1161	1	0.1309	41	0.1421	62	0.4833
43	0.2786	14	0.4118	66	0.2435	49	0.7466	66	0.7237

workstation with a quad-core Intel Xeon E3–1230 processor at a clock rate of 3.20 GHz. No particular effort was made to optimise the code for fast computation, so the reported times should be considered tentative and only roughly indicative of relative performance. Some parts of some methods may run as parallel threads, due to several built-in MATLAB functions being multi-threaded.

LARS is the fastest method here, which is not surprisingly since it is also the simplest. RReliefF is the second fastest by a clear margin. Of the MI estimators, MLMI is clearly the slowest, whereas Kraskov's method and GMM are roughly equivalent; GMM is faster for the datasets with many variables and few samples.

The computational load for GMM can be separated in two parts: fitting the mixture model (including finding the number of components), and using it to determine the ranking of variables. The second part tends to be fast, whereas fitting the model can be time-consuming for data sets with a large number of samples, since this allows more components to be used and requires more iterations to converge.

D. Selected GMM components

The number of components selected for each data set is: Auto price (2), Stocks (11), Housing (1), Breast cancer (1), Tecator (4), Triazines (2), Anthrokids (3). It can be seen that for the Housing and Triazines data sets where the GMM procedure performance was poor, only 1 and 2 components were used, respectively. Perhaps better results could have been obtained by having more components, suggesting that using AIC for model selection might be unnecessarily restrictive in these cases.

V. CONCLUSIONS

The Gaussian mixture model is shown to be an effective and versatile method for estimating mutual information. Using a single mixture of Gaussians to derive estimates for different sets of variables leads to a useful method for variable selection for regression problems.

The experiments show that the proposed method works well overall, at least on par with all the other methods. Particularly for some data sets with a larger number of variables (Breast cancer, Tecator, Anthrokids), the GMM approach leads to the best performance by a clear margin.

In the context of variable selection with missing data, it outperforms competing methods clearly in cases with a high number of missing values.

REFERENCES

[1] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, Eds., *Feature Extraction: Foundations and Applications*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[2] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *Neural Networks, IEEE Transactions on*, vol. 5, no. 4, pp. 537–550, 1994.

[3] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

TABLE VII

The selected inputs and LOO errors for the Triazines data. Bold values represent optimal choices in the sense of the lowest error with the smallest set of variables.

LARS		RReliefF		Kraskov		MLMI		GMM	
8	0.9372	10	0.9413	1	0.9386	2	**0.8658**	9	0.9478
10	**0.8269**	5	0.9472	32	0.8802	31	0.8428	32	0.8984
11	0.7987	4	0.8783	11	**0.7807**	15	0.8593	8	0.7956
40	0.7909	6	0.8235	9	0.7537	32	**0.7380**	10	0.7396
33	0.6320	2	0.7981	33	**0.5622**	49	0.8244	33	0.6149
42	0.6502	3	0.7928	10	0.5695	8	0.8009	11	0.5989
54	0.6509	8	0.7951	54	0.5719	52	0.8053	6	0.5906
26	0.6497	31	0.7834	7	0.5755	1	0.7095	42	0.6110
36	0.6669	32	0.6405	43	0.5797	43	0.7162	49	0.6339
37	0.6517	33	**0.5219**	6	0.6014	4	0.6965	31	0.6040
5	0.6386	36	0.5532	12	0.6134	36	0.7113	36	0.6168
1	0.6494	11	0.5532	35	0.5905	50	0.7469	54	0.6190
15	0.6575	37	0.6132	18	0.5911	16	0.7554	7	0.6444
23	0.6565	9	0.5868	51	0.5983	38	0.7419	4	0.7614
24	0.6549	21	0.5550	52	0.5862	37	0.7459	34	0.6205
25	0.6596	38	0.5311	19	0.5816	6	0.7507	26	0.6165
		12	0.5718	17	0.5867	40	0.7729	48	0.6249
		1	0.5693	13	0.5879	33	0.6988	17	0.7596
		15	0.5846	14	0.6108	29	0.7217	44	0.6187
		40	0.5690	16	0.6108	3	0.7215	47	0.6319
		16	0.5680	20	0.6109	47	0.7512	16	0.6590
		13	0.5753	15	0.6108	17	0.7563	13	0.6604
		14	0.5720	5	0.6074	28	0.7659	12	0.6548
		18	0.6011	42	0.6400	41	0.7630	57	0.6948
		20	0.5715	8	0.6574	21	0.7545	52	0.6807
		39	0.6042	37	0.7792	11	0.7606	3	0.6902
		19	0.6206	31	0.6996	45	0.7667	5	0.8207
		7	0.6533	3	0.6871	46	0.8047	35	0.7939
		17	0.6584	2	0.8676	34	0.7973	21	0.8220
		49	0.9771	4	1.0054	10	0.7927	27	0.9771

TABLE VIII

The selected inputs and LOO errors for the Anthrokids data. Bold values represent optimal choices in the sense of the lowest error with the smallest set of variables.

LARS		RReliefF		Kraskov		MLMI		GMM	
35	**0.0473**	38	0.1075	37	0.0660	37	0.0660	35	**0.0473**
39	0.0283	36	0.0858	20	**0.0249**	43	0.0305	39	0.0283
21	0.0273	37	0.0505	21	**0.0155**	52	0.0312	37	0.0211
37	0.0189	19	0.0327	1	**0.0140**	35	0.0217	21	0.0189
17	0.0171	3	0.0329	39	**0.0126**	16	0.0206	2	0.0148
2	0.0140	39	0.0263	17	**0.0114**	39	0.0174	19	0.0147
3	0.0140	35	0.0196	35	**0.0109**	23	0.0159	17	0.0139
36	0.0131	21	0.0179	38	**0.0102**	3	0.0156	3	0.0139
19	0.0131	17	0.0167	4	**0.0101**	45	0.0147	20	0.0109
33	0.0124	52	0.0169	2	**0.0100**	20	0.0120	48	0.0105
20	0.0100	2	0.0132	36	**0.0094**	21	0.0117	1	0.0100
48	0.0098	4	0.0097	3	0.0094	10	0.0116	36	**0.0093**
44	0.0098	33	0.0094	18	0.0093	24	0.0116	33	**0.0093**
49	0.0097	41	0.0092	19	0.0092	26	0.0116	4	**0.0091**
51	0.0096	24	0.0092	40	0.0092	29	0.0117	38	**0.0087**
53	0.0097	1	0.0091	45	0.0093	18	0.0116	18	**0.0087**
52	0.0099	18	0.0092	42	0.0093	28	0.0116	50	0.0087
46	0.0099	53	0.0092	22	0.0093	41	0.0116	49	**0.0085**
16	0.0100	26	0.0091	26	0.0091	19	0.0112	26	**0.0084**
40	0.0100	30	0.0091	52	0.0093	27	0.0113	22	**0.0084**
15	0.0100	32	0.0091	41	0.0092	42	0.0113	7	**0.0084**
14	0.0100	20	0.0090	24	0.0091	9	0.0114	40	0.0084
10	0.0100	40	0.0090	23	0.0093	14	0.0114	23	0.0085
38	0.0096	22	0.0091	32	0.0092	30	0.0115	43	0.0102
41	0.0122	28	0.0097	28	0.0096	1	0.0109	11	0.0089
9	0.0107	45	0.0098	30	0.0099	4	0.0108	46	0.0100
42	0.0137	42	0.0126	29	0.0116	40	0.0126	32	0.0104
27	0.0146	23	0.0133	27	0.0143	38	0.0121	29	0.0159
43	0.0178	6	0.0179	25	0.0155	22	0.0199	34	0.0241
26	0.0314	29	0.0309	34	0.0270	8	0.0329	24	0.0313

TABLE IX

Computational times for the full variable selection procedure for each method and dataset (seconds).

Name	LARS	RReliefF	Kraskov	MLMI	GMM
Auto-price	0.02	0.10	1.08	163.46	2.32
Stocks	0.01	0.46	0.86	106.94	19.57
Housing	0.02	0.31	1.89	183.50	6.27
Breast Cancer	0.03	0.17	6.32	120.56	9.85
Tecator	0.09	0.47	150.70	8870.04	120.21
Triazines	0.07	0.25	27.32	2117.25	20.82
Anthrokids	0.05	1.33	92.70	4825.93	208.25

[4] A. Kraskov, H. Stögbauer, and P. Grassberger, “Estimating mutual information,” *Phys. Rev. E*, vol. 69, p. 066138, Jun. 2004.

[5] T. Suzuki, M. Sugiyama, J. Sese, and T. Kanamori, “Approximating mutual information by maximum likelihood density ratio estimation,” *JMLR Workshop and Conference Proceedings*, vol. 4, pp. 5–20, 2008.

[6] T. Suzuki, M. Sugiyama, and T. Tanaka, “Mutual information approximation via maximum likelihood estimation of density ratio,” in *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, 2009, pp. 463–467.

[7] N. Kwak and C.-H. Choi, “Input feature selection by mutual information based on parzen window,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 12, pp. 1667–1671, 2002.

[8] N. Kwak, “Feature extraction based on direct calculation of mutual information,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 21, no. 07, pp. 1213–1231, 2007.

[9] F. Rossi, A. Lendasse, D. François, V. Wertz, and M. Verleysen, “Mutual information for the selection of relevant variables in spectrometric nonlinear modelling,” *Chemometrics and Intelligent Laboratory Systems*, vol. 80, no. 2, pp. 215–226, 2006.

[10] D. François, F. Rossi, V. Wertz, and M. Verleysen, “Resampling methods for parameter-free and robust feature selection with mutual information,” *Neurocomputing*, vol. 70, no. 7–9, pp. 1276–1288, 2007.

[11] G. Doquire and M. Verleysen, “Feature selection with missing data using mutual information estimators,” *Neurocomputing*, vol. 90, pp. 3–11, 2012.

[12] B. Frénay, G. Doquire, and M. Verleysen, “Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification,” *Neurocomputing*, vol. 112, pp. 64–78, 2013.

[13] ——, “Is mutual information adequate for feature selection in regression?” *Neural Networks*, vol. 48, pp. 1–7, 2013.

[14] M. Nilsson, H. Gustaftson, S. Andersen, and W. Kleijn, “Gaussian mixture model based mutual information estimation between frequency bands in speech,” in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, 2002, pp. I–525–I–528.

[15] D. P. W. Ellis and J. A. Bilmes, “Using mutual information to design feature combinations,” in *6th International Conference on Spoken Language Processing: ICSLP 2000, the proceedings of the conference, Oct. 16-Oct. 20, 2000, Beijing International Convention Center, Beijing, China*, 2000.

[16] J. M. Leiva-Murillo and A. Artés-Rodríguez, “A gaussian mixture based maximization of mutual information for supervised feature extraction,” in *Independent Component Analysis and Blind Signal Separation*, ser. Lecture Notes in Computer Science, C. G. Puntonet and A. Prieto, Eds. Springer Berlin Heidelberg, 2004, vol. 3195, pp. 271–278.

[17] T. Lan, D. Erdogmus, U. Ozertem, and Y. Huang, “Estimating mutual information using gaussian mixture model for feature ranking and selection,” in *Neural Networks, 2006. IJCNN ’06. International Joint Conference on*, 2006, pp. 5034–5039.

[18] M. A. Kerroum, A. Hammouch, and D. Aboutajdine, “Textural feature selection by joint mutual information based on gaussian mixture model

for multispectral image classification," *Pattern Recognition Letters*, vol. 31, no. 10, pp. 1168–1174, 2010.

[19] H. Akaike, "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716–723, Dec. 1974.

[20] G. McLachlan and D. Peel, *Finite Mixture Models*, ser. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 2000.

[21] C. Bouveyron, S. Girard, and C. Schmid, "High-dimensional data clustering," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 502–519, 2007.

[22] Z. Ghahramani and M. Jordan, "Learning from incomplete data," Lab Memo No. 1509, CBCL Paper No. 108, MIT AI Lab, Tech. Rep., 1995.

[23] L. Hunt and M. Jorgensen, "Mixture model clustering for mixed data with missing information," *Computational Statistics & Data Analysis*, vol. 41, no. 3–4, pp. 429–440, 2003.

[24] E. Eirola, A. Lendasse, V. Vandewalle, and C. Biernacki, "Mixture of gaussians for distance estimation with missing data," *Neurocomputing*, vol. 131, pp. 32–42, 2014.

[25] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed. Wiley-Interscience, 2002.

[26] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, ser. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, 2006.

[27] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–499, 2004.

[28] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine Learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[29] L. Torgo, "Regression datasets," 2012, University of Porto. [Online]. Available: http://www.liaad.up.pt/~ltorgo/Regression/DataSets.html

[30] A. Asuncion and D. J. Newman, "UCI machine learning repository," 2012, University of California, Irvine, School of Information and Computer Sciences. [Online]. Available: http://archive.ics.uci.edu/ml/

[31] J. A. K. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.