Causal & Non-Causal Feature Selection for Ridge Regression

Gavin C. Cawley GCC@CMP.UEA.AC.UK
School of Computing Sciences
University of East Anglia
Norwich, Norfolk, NR4 7TJ, United Kingdom

Editors: I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov

Abstract

In this paper we investigate the use of causal and non-causal feature selection methods for linear classifiers in situations where the causal relationships between the input and response variables may differ between the training and operational data. The causal feature selection methods investigated include inference of the Markov Blanket and inference of direct causes and of direct effects. The non-causal feature selection method is based on logistic regression with Bayesian regularisation using a Laplace prior. A simple ridge regression model is used as the base classifier, where the ridge parameter is efficiently tuned so as to minimise the leave-one-out error, via eigen-decomposition of the data covariance matrix. For tasks with more features than patterns, linear kernel ridge regression is used for computational efficiency. Results are presented for all of the WCCI-2008 Causation and Prediction Challenge datasets, demonstrating that, somewhat surprisingly, causal feature selection procedures do not provide significant benefits in terms of predictive accuracy over non-causal feature selection and/or classification using the entire feature set.

Keywords: regularisation, feature selection, causal inference

1. Introduction

A common assumption underpinning the majority of classical statistical pattern recognition techniques holds that the training data represent an independent and identically distributed (i.i.d.) sample drawn from the same underlying distribution as the operational or test data. Unfortunately, in many practical applications this assumption may not be valid. For example one might train a classifier to diagnose lung cancer using historical data from a particular hospital. However, through changes in referral procedures, diet and lifestyle (for example through government initiatives to restrict smoking in enclosed public places), the distribution of symptoms presented by patients may become progressively more and more different from that of the training sample; a phenomenon known as covariate shift (Quiñonero Candela et al., 2009). Nevertheless, the classifier may still be of diagnostic value, especially if designed from the outset to be robust to covariate shift, for instance through careful feature selection. It seems a reasonable assumption that features in a close causal relationship with the target are likely to remain more reliable under covariate shift than those with a more tenuous link. A model that is robust in this sense would also be valuable in predicting the effects of interventions, for instance in planning more effective referral procedures. Therefore there are practical reasons for attempting to infer the causal relationships between the explanatory and target variables in order to improve predictive performance rather than uncovering the structure of the data. In this paper we evaluate the effectiveness of several causal and non-causal feature selection procedures in such situations, using ridge regression as the base classifier, using all of the datasets comprising in the WCCI-2008 Causation and Prediction Challenge.

2. Method

Ridge regression ensembles are used as the base classifier for the empirical study. Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\ell}$ represent the training sample, where $\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ is a vector of explanatory features for the i^{th} sample, and $y_i \in \{+1, \ -1\}$ is the corresponding response indicating whether the sample belongs to the positive or negative class respectively. Ridge regression provides a simple and effective classifier that is equivalent to a form of regularised linear discriminant analysis. The output of the ridge regression classifier, $\hat{y}_i$, and vector of model parameters, $\boldsymbol{\beta} \in \mathbb{R}^d$, are given by

$$\hat{y}_i = \boldsymbol{x}_i \cdot \boldsymbol{\beta} \qquad \text{and} \qquad \left[\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right]\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{y}, \tag{1}$$

where $\boldsymbol{X} = [\boldsymbol{x}_i]_{i=1}^{\ell}$ is the data matrix, $\boldsymbol{y} = (y_i)_{i=1}^{\ell}$ is the response vector and the ridge parameter, λ, controls the bias-variance trade-off (Geman et al., 1992). Note that classifiers used throughout this study included an unregularised bias parameter, which has been neglected here for notational convenience. Careful tuning of the ridge parameter allows the ridge regression classifier to be used even in situations with many more features than training patterns (i.e. $d \gg \ell$) without significant over-fitting (e.g. Cawley, 2006). Fortunately the ridge parameter can be optimised efficiently by minimising a closed-form leave-one-out cross-validation estimate of the sum of squared errors, i.e. Allen's PRESS statistic (Allen, 1974),

$$P(\lambda) = \frac{1}{\ell}\sum_{i=1}^{\ell}\left[\hat{y}_i^{(-i)} - y_i\right]^2 \qquad \text{where} \qquad \hat{y}_i^{(-i)} - y_i = \frac{\hat{y}_i - y_i}{1 - h_{ii}}, \tag{2}$$

$\hat{y}_i^{(-i)}$ represents the output of the classifier for the i^{th} training pattern in the i^{th} fold of the leave-one-out procedure and h_{ii} is an element of the principal diagonal of the hat matrix $\boldsymbol{H} = \boldsymbol{X}\left[\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}\right]^{-1}\boldsymbol{X}^T$. The ridge parameter can be optimised more efficiently in canonical form (Weisberg, 1985) via eigen-decomposition of the data covariance matrix $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{V}^T\boldsymbol{\Lambda}\boldsymbol{V}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues. The normal equations and hat matrix can then be written as

$$[\boldsymbol{\Lambda} + \lambda\boldsymbol{I}]\,\boldsymbol{\alpha} = \boldsymbol{V}^T\boldsymbol{X}^T\boldsymbol{y} \quad \text{where} \quad \boldsymbol{\alpha} = \boldsymbol{V}^T\boldsymbol{\beta} \qquad \text{and} \qquad \boldsymbol{H} = \boldsymbol{V}\,[\boldsymbol{\Lambda} + \lambda\boldsymbol{I}]^{-1}\,\boldsymbol{V}^T \tag{3}$$

As only a diagonal rather than a full matrix need now be inverted following a change in λ, the computational expense of optimising the ridge parameter is greatly reduced. For problems with more features than training patterns, $d > \ell$, the kernel ridge regression classifier (Saunders et al., 1998) with a linear kernel is more efficient and exactly equivalent. The ridge parameter for KRR can also be optimised efficiently via an eigen-decomposition of the kernel matrix (Saadi et al., 2007).

2.1 Non-Causal Feature Selection

The feature selection methods most frequently used in practical applications aim to determine a small subset of features that are predictive of the target variable, without any consideration of causal relationships. For a survey of conventional feature selection methods, see Guyon and Eliseeff (2003). In this study, we adopt an embedded feature selection method, known as BLogReg (Cawley and Talbot, 2006), based on logistic regression with Bayesian regularisation using a Laplace prior. As the usual regularisation parameter is integrated out analytically in this approach using an uninformative hyper-prior, the number of features is determined automatically, without the need for additional cross-validation. Rather than make predictions with BLogReg directly, it is used to select features for a ridge regression model so that the comparison of feature selection techniques is not obscured by the differences due to the classifier.

2.2 Finding the Markov Blanket

The most basic form of causal feature selection aims to determine the Markov blanket, $MB(T)$, the set of features such that the target, T, is conditionally independent of all other features, conditioned on the features comprising $MB(T)$. Under the faithfulness assumption (Pearl, 1988), the Markov blanket consists of the set of features representing the direct causes (parents) and direct consequences (children) of the target, and also any other causes directly affecting the consequences of the target (spouses). In this study, we use the HITON algorithm (Aliferis et al., 2003) to infer the Markov blanket of the *unmanipulated* distribution throughout. If information is available regarding which features have been manipulated, it is in principle possible to infer the Markov blanket of the *manipulated* distribution, however this was not investigated (due to the ignorance of the investigator at the time!). Once the Markov blanket of the unmanipulated distribution has been determined, the manipulated children of the target can be deleted (provided they are not also a spouse via an unmanipulated child) as the direct causal link has been broken and similarly spouses related only via manipulated children can also be deleted, forming the Markov blanket of the manipulated distribution.

2.3 Discerning Causes and Effects

More advanced causal inference methods attempt to construct a directed graph representing the causal relationships between variables. This can be used to identify direct causes and direct effects of the target variable, forming alternative feature sets for ridge regression classifiers. In this study, the PC and MMHC algorithms of the Causal Explorer package (for further details, see Aliferis et al., 2003) were used throughout. As these algorithms are computationally expensive, they are applied to the subset of features already identified as belonging to the Markov blanket. As the PC algorithm can accommodate problems with continuous attributes, no discretisation of continuous features was necessary.

2.4 Use of Ensembles

As feature selection algorithms are unstable, i.e. a perturbation of the data is likely to result in a different subset of features being selected, an ensemble of 100 ridge regression classifiers is used in all experiments to minimise the distracting effects of this source of variability. For each component classifier, a different random partition of the available data is used to form training and test sets (in proportions of 9:1) and feature selection performed separately for each partition. The prediction is then made using the arithmetic mean of the 100 component classifiers. As the regularisation parameter is also tuned separately for each of the component classifiers, this approach also addresses over-fitting the model selection criterion so some degree (Cawley and Talbot, 2007; Hall and Robinson, 2009). It should be noted that as feature selection is performed separately for each component classifier, the number of features used by the ensemble is generally much larger than the number used by any individual member of the ensemble.

3. Results

In this section, we compare the performance of different causal and non-causal feature selection procedures for ridge regression classifiers on two illustrative benchmarks and each of the four challenge datasets, before discussing the absolute performance of the base ridge regression classifier. The results are given for a number of different models, in most cases labelled according to the feature selection process used:

- **True MB:** Ensembles trained using the true Markov blanket of the target variable for each variant of the benchmark. For REGED and MARTI, manipulated features are only retained if they are parents or spouses of the target. For CINA and SIDO, where the feature set is comprised of real variables and probes, the true Markov blanket is unknown, and hence the union of the set of all real variables and probes belonging to the true Markov blanket is used instead, for further details, see Guyon et al. (2008).

- **Inferred MB:** Features identified by HITON_MB as forming the Markov blanket of the *unmanipulated* distribution are retained.

- **None:** No feature selection is performed, relying purely on regularisation to prevent over-fitting the training data.

- **Non-causal:** Feature selection performed using the BLogReg algorithm (Cawley and Talbot, 2006), providing an example of the performance of traditional embedded feature selection methods.

- **Causes & effects:** Members of the Markov Blanket of the unmanipulated distribution, identified by HITON_MB, that are determined using the MMHC or PC algorithms to be direct causes and effects of the target variable. It should be noted that the pre-filtering to remove features not belonging to the Markov blanket of target for the unmanipulated distribution probably made it very difficult for the causal discovery algorithm to correctly orient the edges of the causal graph, and this perhaps explains the poor performance observed.

- **Causes only:** Members of the Markov Blanket of the unmanipulated distribution, identified by HITON_MB, that are determined using the MMHC or PC algorithms to be direct causes of the target variable.

- **Winner:** This is the scores achieved by the winning entries for each benchmark in the WCCI-2008 Causation and Prediction Challenge. For REGED and CINA, these are described in Chang and Lin (2008); for SIDO and MARTI, these are described here and in the supplementary material[1]. Note that the entries for REGED and CINA are not based on ridge regression ensembles and so are not directly comparable to the other methods.

- **Best TSCORE:** Results for models achieving the best TSCORE, whether or not they were the winning entries, for each variant of each benchmark during the WCCI-2008 Causation and Prediction Challenge.

- **Yin *et al.*:** Feature selection based on the method by Yin et al. (2008), which was identified as the most successful algorithm for local structure determination.

3.1 LUCAS - LUng CAncer Simple Dataset

The LUCAS dataset represents a synthetic medical diagnosis problem, where the task is to identify patients with lung cancer from a set of explanatory variables of putative causal relevance. As the data are generated artificially using a set of simple Bayesian network models (see Guyon et al., 2008, Figure 1 a-c), the true nature of the underlying causal relationships is known, and so this benchmark is useful in illustrating the value of different approaches in ideal conditions. The results obtained[2] on this benchmark are shown in Table 1.

Regularisation proves satisfactory in suppressing the influence of uninformative features in the absence of external manipulation (LUCAS0), and so feature selection does not improve predictive performance (although selection of the Markov blanket is only marginally inferior). In the presence of mild manipulation (LUCAS1), the benefit of selecting only the variables comprising the Markov blanket of the target becomes more apparent, achieving the best TSCORE as the manipulation of causally irrelevant variables is ignored. It is interesting to note, however, that the result obtained is only marginally better than that for a ridge regression model without any form of feature selection, showing that regularisation is effective in suppressing the influence of irrelevant variables. However other explanations are plausible. For instance, it could be that including redundant variables is better than deleting important variables (Guyon et al., 2008, §6.2); alternatively it may be the case in many applications that the most relevant features are simply those best correlated with the target. For LUCAS2, only the direct causes are relevant, and for this simple dataset the causal discovery algorithms (HITON_MB and MMHC) are effective in identifying them so the *causes only* model performs significantly better than the others.

3.2 LUCAP - LUng CAncer with Probes Dataset

The LUCAP benchmark extends the medical diagnosis problem introduced in LUCAS to include *probes*, artificial variables that are noisy functions of the existing variables (see

[1] http://theoval.cmp.uea.ac.uk/~gcc/projects/causal

[2] Non-causal feature selection performed using BLogReg (tolerance = 1×10^{-9}), identification of the Markov blanket using HITON_MB (“g2” statistic, threshold = 0.05, maximum size of conditioning set = 4), identification of direct causes and effects using MMHC (with default parameter settings).

Table 1: Results obtained for the LUCAS benchmark: FNUM – number of features used, FSCORE – area under the receiver operating characteristic (AUROC) statistic for the detection of causally related features, DSCORE – AUROC on the training set, TSCORE – AUROC on the test set, AUC – AUROC using 100-fold repeated hold-out validation.

Dataset	Selection	FNUM	FSCORE	DSCORE	TSCORE	AUC
LUCAS0	None	11	1.0000	0.9139	0.9170	0.9079
	Inferred MB	6	1.0000	0.9102	0.9168	0.9082
	True MB	5	1.0000	0.9103	0.9167	0.9082
	Non-causal	11	0.8070	1.0000	0.9139	0.9079
	Causes & effects	4	0.9000	0.8911	0.8992	0.8910
	Causes only	2	0.7000	0.7782	0.7968	0.7832
LUCAS1	True MB	4	1.0000	0.9026	0.9041	—
	Inferred MB	6	1.0000	0.9102	0.9012	—
	None	11	1.0000	0.9139	0.9005	—
	Non-causal	11	1.0000	0.9139	0.9005	—
	Causes & effects	4	0.8571	0.8911	0.8808	—
	Causes only	2	0.7500	0.7782	0.7910	—
LUCAS2	True MB	2	1.0000	0.7782	0.7913	—
	Causes only	2	1.0000	0.7782	0.7913	—
	Causes & effects	4	0.9444	0.8911	0.7579	—
	Inferred MB	6	0.9444	0.9102	0.7410	—
	None	11	0.8333	0.9139	0.7348	—
	Non-causal	11	0.9444	0.9139	0.7342	—

Guyon et al., 2008, Figure 1 d-e). The results obtained[3] on this benchmark are shown in Table 2. The probes appear to obfuscate the task of discovering the true causal structure of the data, and the models with non-causal feature selection fare conspicuously better than those with causal feature selection on the manipulated datasets. The organizers suggest that selecting features that are direct causes of the target may be an attractive approach; however the causal discovery algorithm (MMHC) found 26 features that may be direct causes, over the 100 random partitions of the data, when in fact there are only two genuine direct causes. The *causes only* approach therefore performed very poorly. It is a rather discouraging result that the causal feature selection procedures perform so poorly, albeit perhaps in the hands of an inexpert user.

3.3 REGED - REsimulated Gene Expression Dataset

The REGED dataset represents a re-simulated gene expression microarray classification problem, where the task is to diagnose lung cancer on the basis of gene expression pro-

[3] BLogReg: tolerance = 1×10^{-9}, HITON_MB: "g2" statistic, threshold = 0.05, maximum size of conditioning set = 4, MMHC: default parameter settings.

files, classifying samples as malignant (adenocarcinoma) or benign (squamous). The results obtained[4] on this benchmark are shown in Table 3 and statistical significance diagrams (adapted from the critical difference diagrams introduced by Demšar (2006)), are shown in Figure 1:

- For all three datasets, True MB achieves a TSCORE that is statistically indistinguishable from the best obtained during the challenge (Best TSCORE), demonstrating that the linear ridge regression ensemble is a competitive base classifier for this benchmark.

- For all three datasets, the TSCORE performance obtained using the non-causal feature selection procedure (BLogReg) was statistically indistinguishable from that obtained using the best overall causal feature selection procedure (Yin *et al.*). This is perhaps because the BLogReg algorithm was originally developed with this particular application (Cawley, 2006) in mind, although the approach is generally applicable and without specific adaption to microarray classification.

- The total number of features used by the inferred Markov Blanket ensemble for REGED0 (78) is much larger than the average number of features used by the individual component classifiers (24.85), providing an indication of the instability of the causal feature selection methods. Note that the average size of the inferred Markov blanket for each component is however close to the true value (21). The average

[4] BLogReg: tolerance = 1×10^{-6}, HITON_MB: "z" statistic, threshold = 0.05, maximum size of conditioning set = 2, PC: 'z' statistic, threshold = 0.05, $k = 16$.

Table 2: Results obtained for the LUCAP benchmark, see caption of Table 1 for details.

Dataset	Selection	FNUM	FSCORE	DSCORE	TSCORE	AUC
LUCAP0	Non-causal	42	0.5930	0.9749	0.9711	0.9681
	True MB	105	1.0000	0.9757	0.9692	0.9698
	None	143	0.7381	0.9768	0.9686	0.9695
	Inferred MB	77	0.8466	0.9726	0.9684	0.9674
	Causes & effects	45	0.7143	0.9703	0.9675	0.9664
	Causes only	26	0.6238	0.9486	0.9382	0.8089
LUCAP1	True MB	11	1.0000	0.9139	0.9126	—
	Non-causal	42	0.6832	0.9749	0.8564	—
	Causes & effects	45	0.5561	0.9703	0.8317	—
	Inferred MB	77	0.5344	0.9726	0.8121	—
	None	143	0.6109	0.9768	0.7744	—
	Causes only	26	0.5090	0.9486	0.6504	—
LUCAP2	True MB	11	1.0000	0.9139	0.9165	—
	Non-causal	42	0.6832	0.9749	0.6578	—
	Inferred MB	77	0.5344	0.9726	0.5634	—
	Causes & effects	45	0.5561	0.9703	0.5575	—
	None	143	0.6109	0.9768	0.5100	—
	Causes only	26	0.5090	0.9468	0.4344	—

number of features used by individual component classifiers are shown in Table 8, demonstrating that a degree of instability is to be expected when using both causal and non-causal feature selection methods.

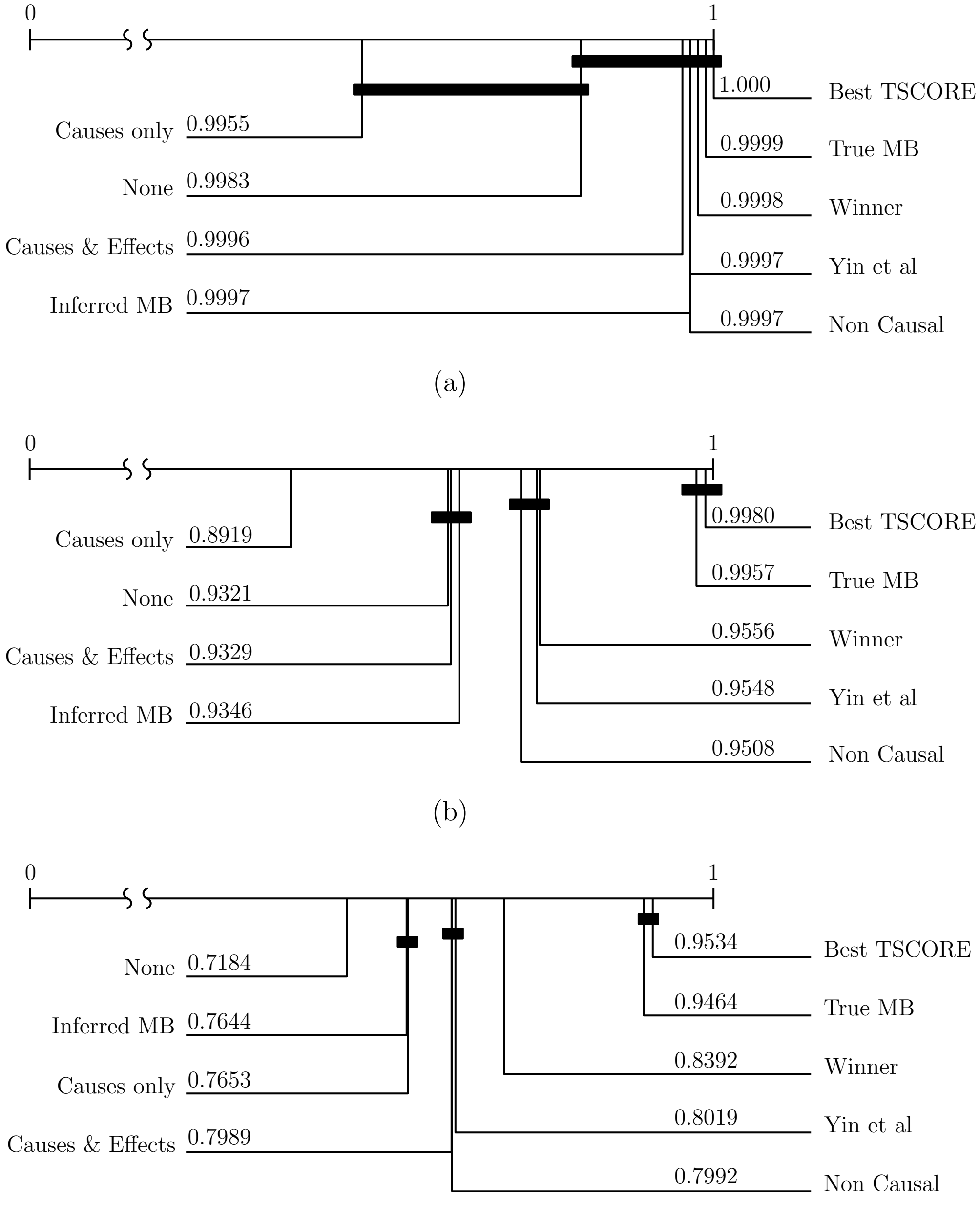
- None of the causal feature selection algorithms, with the exception of reference entries, proved statistically superior to non-causal selection procedures for any of the three REGED datasets, a rather disappointing and challenging result.

No use was made of the information regarding the manipulated variables for the methods introduced in this study; it is possible that better results might be obtained by taking advantage of this information, especially for causal feature selection approaches.

Table 3: Results obtained for the REGED benchmark, see caption of Table 1 for details.

Dataset	**Selection**	**FNUM**	**FSCORE**	**DSCORE**	**TSCORE**	**AUC**
REGED0	Best TSCORE*	122	0.8352	1.0000	1.0000 ± 0.0002	—
	True MB	21	1.0000	0.9999	0.9999 ± 0.0008	0.9997
	Winner‡	16	0.8526	1.0000	0.9998 ± 0.0009	—
	Yin *et al.*	15	0.8571	1.0000	0.9997 ± 0.0010	0.9998
	Non-causal	26	0.8070	1.0000	0.9997 ± 0.0009	0.9997
	Inferred MB	78	0.8988	0.9999	0.9997 ± 0.0012	0.9995
	Causes & effects	13	0.8095	0.9999	0.9996 ± 0.0011	0.9996
	None	999	0.9204	1.0000	0.9983 ± 0.0017	0.9962
	Causes only	9	0.7143	0.9984	0.9955 ± 0.0018	0.8961
REGED1	Best TSCORE*	122	0.7946	1.0000	0.9980 ± 0.0015	—
	True MB	14	1.0000	0.9926	0.9957 ± 0.0020	—
	Winner‡	16	0.8566	1.0000	0.9556 ± 0.0040	—
	Yin *et al.*	14	0.8185	0.9999	0.9548 ± 0.0036	—
	Non-causal	26	0.7798	1.0000	0.9508 ± 0.0036	—
	Inferred MB	78	0.8438	0.9999	0.9346 ± 0.0044	—
	Causes & effects	13	0.7822	0.9999	0.9329 ± 0.0037	—
	None	999	0.9078	1.0000	0.9321 ± 0.0036	—
	Causes only	9	0.7124	0.9984	0.8919 ± 0.0042	—
REGED2	Best TSCORE†	2	1.0000	0.9611	0.9534 ± 0.0042	—
	True MB	2	1.0000	0.9557	0.9464 ± 0.0041	—
	Winner‡	8	0.9970	0.9995	0.8392 ± 0.0052	—
	Yin *et al.*	11	0.9975	0.9997	0.8019 ± 0.0054	—
	Non-causal	26	0.9980	1.0000	0.7992 ± 0.0056	—
	Causes & effects	13	0.9970	0.9999	0.7989 ± 0.0057	—
	Causes only	9	0.9970	0.9984	0.7653 ± 0.0054	—
	Inferred MB	78	0.9975	0.9999	0.7644 ± 0.0057	—
	None	999	0.9950	1.0000	0.7184 ± 0.0059	—

*Reference “SNB(CMA), IID assumption”, †Reference “True model with parents”, ‡Yin-Wen Chang “final submission”.



(a)

(b)

(c)

Figure 1: Statistical significance diagrams for (a) REGED0, (b) REGED1 and (c) REGED2. The axis represents the TSCORE statistic, and the heavy bars denote groups of classifiers with statistically indistinguishable performance. The statistical significance of differences in TSCORE are determined using the two sample z-test at the 95% level of significance (a critical value of $z = 1.64$).

3.4 SIDO - SImple Drug Operation

The SIDO benchmark represents a problem in pharmacology, where the task is to identify small molecules that are active against the AIDS HIV virus on the basis of a large number of binary molecular descriptors. The results obtained[5] on this benchmark are shown in Table 4 and the statistical significance diagram in Figure 2:

- Again, for all three datasets, True MB achieves a TSCORE that is statistically indistinguishable from the best obtained during the challenge (Best TSCORE), demonstrating that the linear ridge regression ensemble is a competitive base classifier for this benchmark.
- For both manipulated datasets, the TSCORE achieved using no feature selection at all is statistically superior to the best overall causal feature selection method (Yin *et al.*). For the unmanipulated dataset, the differences in performance were statistically insignificant. This demonstrates that regularisation alone can be highly effective in suppressing the deleterious influence of uninformative features.
- The differences in TSCORE between the best causal feature selection procedure (Yin *et al.*) and non-causal feature selection using BLogReg were statistically insignificant for all three datasets.
- None of the causal feature selection algorithms investigated, with the exception of reference entries, proved statistically superior to an ensemble trained using all of the available features for any of the three SIDO datasets, again a rather disappointing result.

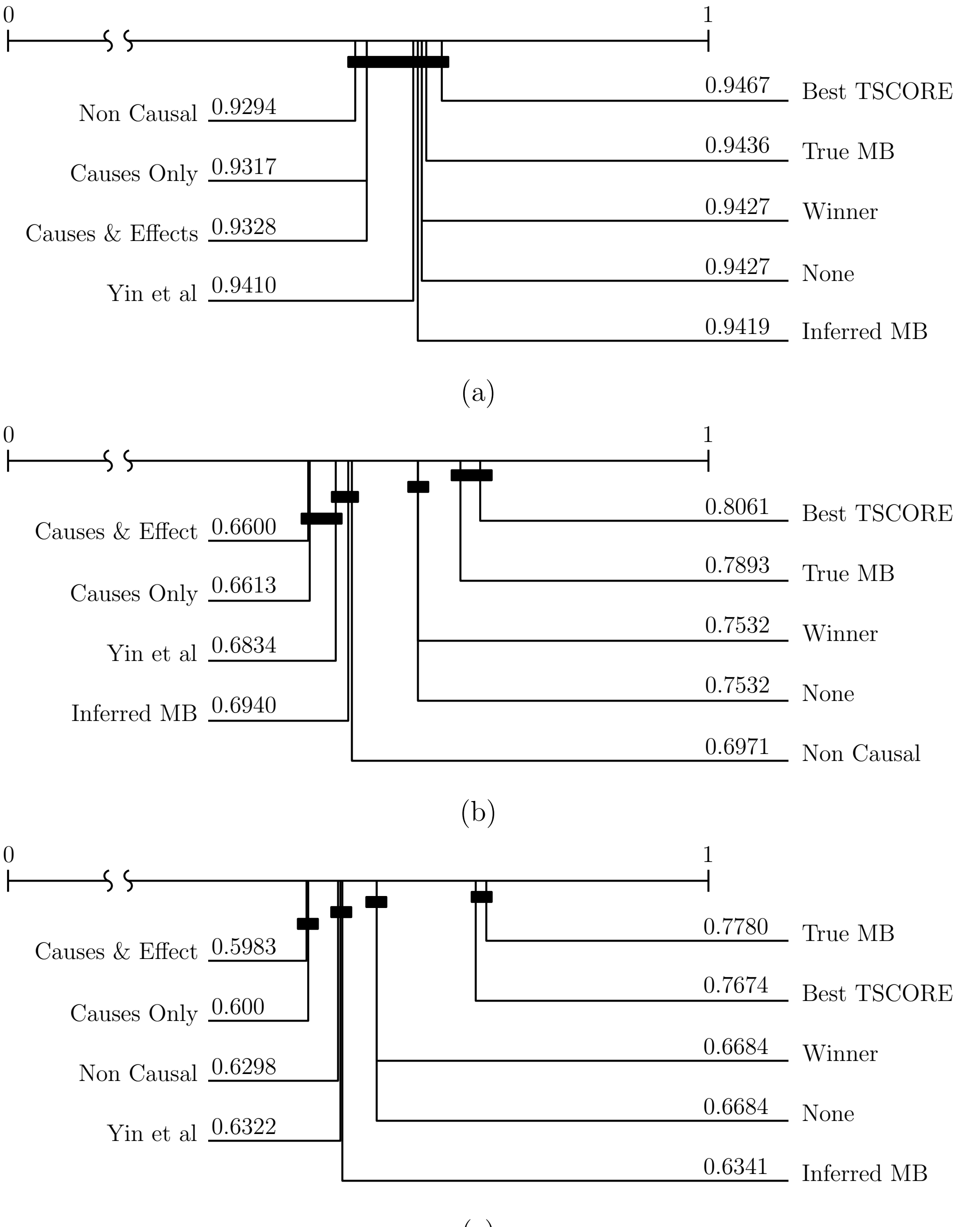
3.5 CINA - Census Is Not Adult

The CINA benchmark describes an econometrics problem, where the task is to discover the socio-economic factors affecting income (the positive class representing individuals with annual income in excess of $50K). The results obtained[6] on this benchmark are shown in Table 5 and the statistical significance of differences in TSCORE are depicted in Figure 3:

- True MB achieves a TSCORE that is statistically indistinguishable from the best obtained during the challenge (Best TSCORE), on both manipulated datasets, but not in CINA0. This suggests that linear ridge regression ensembles may not provide a genuinely competitive base classifier for this benchmark, especially as the difference is quite large for CINA0. Note that BLogReg was used as the base classifier for the final challenge submission as the mean AUC scores for the individual component classifiers was lower than that for linear ridge regression.
- For both manipulated datasets, the TSCORE achieved using no feature selection at all is statistically superior to the best overall causal feature selection method (Yin

[5] BLogReg: tolerance = 1×10^{-6}, HITON_MB: “g2” statistic, threshold = 0.05, maximum size of conditioning set = 3, PC: “g2” statistic, threshold = 0.05, $k = 8$.

[6] BLogReg: tolerance = 1×10^{-6}, HITON_MB: “z” statistic, threshold = 0.05, maximum size of conditioning set = 5, PC: “z” statistic, threshold = 0.05, $k = 4$.



(a)

(b)

(c)

Figure 2: Statistical significance diagrams for (a) SIDO0, (b) SIDO1 and (c) SIDO2. The axis represents the TSCORE statistic, and the heavy bars denote groups of classifiers with statistically indistinguishable performance.

Table 4: Results obtained for the SIDO benchmark, see caption of Table 1 for details.

Dataset	Selection	FNUM	FSCORE	DSCORE	TSCORE	AUC
SIDO0	Best TSCORE[†]	181	0.4940	0.9584	0.9467 ± 0.0073	—
	True MB	4301	0.9995[*]	0.9830	0.9436 ± 0.0072	0.9471
	Winner[§]	4928	0.5890	0.9840	0.9427 ± 0.0070	—
	None	4928	0.5890	0.9840	0.9427 ± 0.0070	0.9472
	Inferred MB	837	0.5834	0.9563	0.9419 ± 0.0075	0.9356
	Yin *et al.*	16	0.5019	0.9475	0.9410 ± 0.0074	0.9442
	Causes & effects	58	0.5067	0.9459	0.9328 ± 0.0085	0.8798
	Causes only	58	0.5067	0.9454	0.9317 ± 0.0089	0.8733
	Non-causal	138	0.5160	0.9482	0.9294 ± 0.0080	0.9226
SIDO1	True MB	1643	0.9997[*]	0.9098	0.8061 ± 0.0132	—
	Best TSCORE[‡]	1024	0.8114	0.9021	0.7893 ± 0.0135	—
	Winner[§]	4928	0.5314	0.9840	0.7532 ± 0.0137	—
	None	4928	0.5314	0.9840	0.7532 ± 0.0137	—
	Non-causal	138	0.4909	0.9482	0.6971 ± 0.0138	—
	Inferred MB	873	0.5351	0.9563	0.6940 ± 0.0138	—
	Yin *et al.*	16	0.5035	0.9475	0.6834 ± 0.0133	—
	Causes only	58	0.4989	0.9454	0.6613 ± 0.0138	—
	Causes & effects	58	0.4989	0.9459	0.6600 ± 0.0137	—
SIDO2	True MB	1643	0.9997[*]	0.9089	0.7780 ± 0.0130	—
	Best TSCORE[‡]	512	0.8114	0.8693	0.7674 ± 0.0129	—
	Winner[§]	4928	0.5314	0.9840	0.6684 ± 0.0130	—
	None	4928	0.5314	0.9840	0.6684 ± 0.0130	—
	Inferred MB	873	0.5351	0.9563	0.6341 ± 0.0124	—
	Yin *et al.*	16	0.5035	0.9475	0.6322 ± 0.0131	—
	Non-causal	138	0.4909	0.9482	0.6298 ± 0.0039	—
	Causes only	58	0.4989	0.9545	0.6000 ± 0.0129	—
	Causes & effects	58	0.4989	0.9459	0.5983 ± 0.0129	—

[*] Some features beneath the Markov blanket assigned a weight of zero and was not included.

[†] Gavin Cawley "Final #009", [‡] Reference "MB_LR_S", [§] Gavin Cawley "final models".

et al.). For the unmanipulated dataset, the difference in performance is statistically insignificant. In this case, it seems that, while regularisation is not that effective in supressing the influence of uninformative features, the instability of feature selection procedure means that better performance is only available given prior knowledge of the causal relationships.

- The differences in TSCORE between the best causal feature selection procedure (Yin *et al.*) and non-causal feature selection using BLogReg were statistically insignificant for all three datasets.

- None of the causal feature selection algorithms investigated, with the exception of reference entries, proved statistically superior to an ensemble trained using all of the available features for any of the three CINA datasets, again a rather disappointing result.

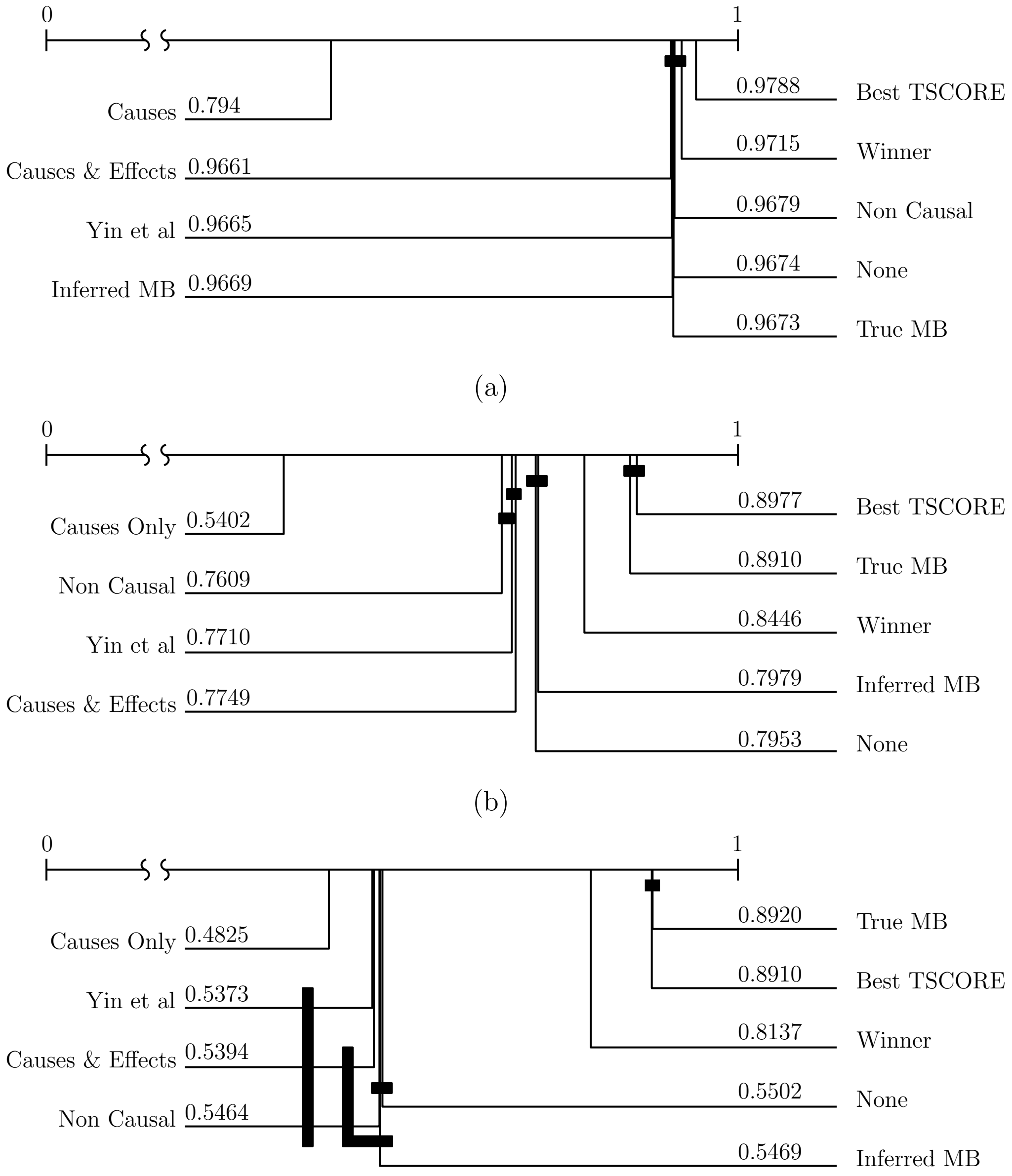
Table 5: Results obtained for the CINA benchmark, see caption of Table 1 for details.

Dataset	Selection	FNUM	FSCORE	DSCORE	TSCORE	AUC
CINA0	Best TSCORE[*]	90	0.8913	0.9794	0.9788 ± 0.0029	—
	Winner[‡]	64	0.6000	0.9721	0.9715 ± 0.0032	—
	Non-causal	67	0.5708	0.9682	0.9679 ± 0.0035	0.9660
	None	132	0.7908	0.9677	0.9674 ± 0.0035	0.9664
	True MB	115	1.0000	0.9674	0.9673 ± 0.0035	0.9663
	Inferred MB	70	0.7708	0.9669	0.9669 ± 0.0035	0.9660
	Yin *et al.*	22	0.5957	0.9657	0.9665 ± 0.0034	0.9657
	Causes & effects	42	0.6826	0.9654	0.9661 ± 0.0035	0.9653
	Causes only	4	0.5174	0.7923	0.7911 ± 0.0046	0.5351
CINA1	Best TSCORE[*]	90	0.4542	0.9794	0.8977 ± 0.0043	—
	True MB	44	1.0000	0.8915	0.8910 ± 0.0040	—
	Winner[‡]	64	0.7053	0.9721	0.8446 ± 0.0047	—
	Inferred MB	70	0.5261	0.9669	0.7979 ± 0.0052	—
	None	132	0.5865	0.9677	0.7953 ± 0.0050	—
	Causes & effects	42	0.5477	0.9654	0.7749 ± 0.0050	—
	Yin *et al.*	24	0.5823	0.9652	0.7710 ± 0.0048	—
	Non-causal	67	0.6436	0.9682	0.7609 ± 0.0053	—
	Causes only	4	0.5114	0.7923	0.5402 ± 0.0056	—
CINA2	True MB	44	1.0000	0.8915	0.8920 ± 0.0043	—
	Best TSCORE[†]	32	1.0000	0.8909	0.8910 ± 0.0042	—
	Winner[‡]	4	0.7053	0.8137	0.8157 ± 0.0052	—
	None	132	0.5865	0.9677	0.5502 ± 0.0043	—
	Inferred MB	70	0.5261	0.9669	0.5469 ± 0.0041	—
	Non-causal	67	0.6436	0.9682	0.5464 ± 0.0039	—
	Causes & effects	42	0.5477	0.9654	0.5394 ± 0.0038	—
	Yin *et al.*	18	0.5794	0.9636	0.5373 ± 0.0041	—
	Causes only	4	0.5114	0.7923	0.4825 ± 0.0035	—

[*] Reference “SNB(CMA), IID assumption”. [†] Reference “CINA Test”, [‡] Yin-Wen Chang “final submission”.

3.6 MARTI - Measurement ARTIfact

Like REGED, the MARTI benchmark represents a re-simulated microarray classification task, the aim of which is to identify genes that may be responsible for lung cancer. However, in this case additive zero-mean correlated noise has been added to the data to simulate measurement artifacts introduced by an instrument used to collect the training data that is substantially inferior to a more accurate instrument used to gather the test data. Figure 4 shows an example of the correlated noise corrupting a training sample from the MARTI benchmark. The correlated noise is likely to confuse both causal and non-causal feature



(a)

(b)

(c)

Figure 3: Statistical significance diagrams for (a) CINA0, (b) CINA1 and (c) CINA2. The axis represents the TSCORE statistic, and the heavy bars denote groups of classifiers with statistically indistinguishable performance.

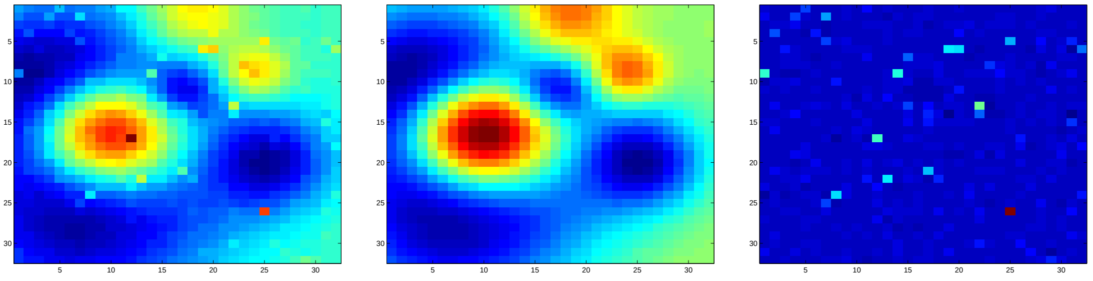
selection algorithms, and therefore MARTI differs from the other challenge datasets in that non-trivial pre-processing is required. We adopt a kernel ridge regression approach to try to estimate the noise for each training pattern as a function of the x- and y-co-ordinates of the spot on the microarray image. Let $\boldsymbol{X}$ represent the $d \times 2$ matrix, where each row, $\boldsymbol{x}_i$, gives the x- and y-co-ordinates of a spot on the microarray image, and $\boldsymbol{Y}$ represents the $d \times \ell$ matrix containing the expression levels for every gene, where each row, $\boldsymbol{y}_i$, represents a spot and each column represents a sample. We assume that the noise contaminating the expression levels can be approximated by a linear model in a feature space induced by a radial basis function kernel, with the expression levels themselves modelled by a Gaussian noise process,

$$\boldsymbol{y}_i = \boldsymbol{\phi}(\boldsymbol{x}_i) \cdot \boldsymbol{W} + \boldsymbol{\varepsilon_i}, \qquad \text{where} \qquad \boldsymbol{\epsilon}_i \sim \mathrm{N}\left(\boldsymbol{0}, \sigma_i^2 \boldsymbol{I}\right), \tag{4}$$

where $\boldsymbol{\phi}(\boldsymbol{x})$ represents the image of the data in the kernel induced feature space. Note that a heteroscedastic noise model is used (e.g. Cawley et al., 2004) as considerable variation is evident in the range of expression of different genes. The model (4) is equivalent to a multi-output weighted kernel ridge regression model (Saunders et al., 1998), with the weights given by the inverse noise variance for each spot, $\boldsymbol{\sigma}^{-1} = \left(\sigma_1^{-2}, \sigma_2^{-2}, \ldots, \sigma_d^{-2}\right)$. The iterative training algorithm alternates updates of the model parameters with re-estimation of the noise variance terms using the model residuals. The usual regularisation and kernel parameters were tuned via numerical minimisation of the cross-validation error. Estimates of the true expression profiles can then be obtained by simply subtracting from $\boldsymbol{Y}$ the estimate of the correlated noise given by the fitted model. The results obtained[7] on this benchmark are shown in Table 6, the corresponding statistical significance diagram is shown in Figure 5:

- The pre-processing steps described above proved quite satisfactory, as demonstrated by the similarity of results obtained on the REGED and MARTI benchmarks, shown in Tables 3 and 6 respectively, however no use was made of the calibrant features or knowledge of manipulated features so the results are likely to be somewhat sub-optimal.

[7] BLogReg: tolerance $= 1 \times 10^{-6}$, HITON_MB: "z" statistic, threshold $= 0.05$, maximum size of conditioning set $= 5$, PC: "z" statistic, threshold $= 0.05$, $k = 16$.



(a) (b) (c)

Figure 4: Example pattern from the training set of the MARTI benchmark (a) raw microarray image (b) estimate of correlated noise and (c) filtered expression levels.

- The TSCORE for a linear ridge regression ensemble using knowledge of the true Markov blanket exceeds that of the best TSCORE achieved by any challenges submission, by a statistically significant margin on the manipulated datasets. This suggest that linear ridge regression ensembles are competitive as a base classifier for this application.

- For both manipulated datasets, the TSCORE achieved using no feature selection at all is statistically superior to the best overall causal feature selection method (Yin *et al.*). For the unmanipulated dataset, the difference in performance is statistically insignificant. In this case, it seems that while regularisation is not that effective in suppressing the influence of uninformative features, the instability of feature selection procedure means that better performance is only available given prior knowledge of the causal relationships.

- The TSCORE achieved using non-causal feature selection was statistically indistinguishable from that achieved by the best all-round causal feature selection procedure (Yin *et al.*) on the unmanipulated data (MARTI0), was statistically superior on one manipulated dataset (MARTI1) and statistically inferior on the other (MARTI2), suggesting that causal feature selection does not improve overall on non-causal feature selection.

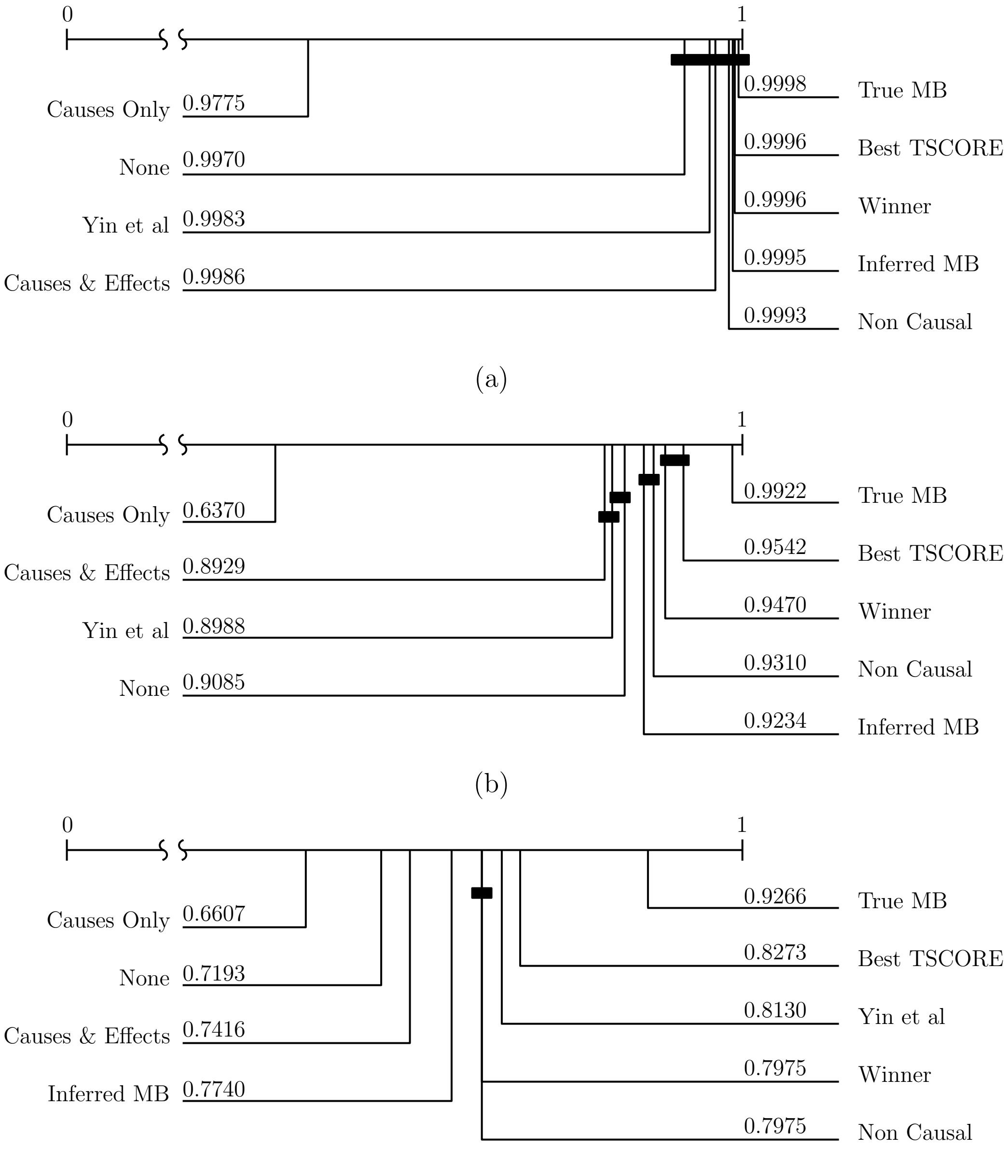
3.7 Final Challenge Submission

Table 7 shows the results for the final challenge submission. BLogReg was used as the base classifier for the CINA benchmark, as this gave slightly better performance under the 100-fold repeated hold-out procedure used for validation during the development phase of the challenge. The full set of models for the SIDO datasets was incomplete by the challenge deadline; the best models proved to be simple ridge regression models with no feature selection (note that there were four features in the training set with zero variance, hence only 4928 features were actually used by the classifier). The rankings indicate that the base classifiers were good choices for the benchmarks considered, and so the comparison of feature selection methods provides a good indication of their relative merits. Further details of the final challenge submission are available in the supplementary material.

4. Recommendations

The results of the investigation presented in the previous section suggest that further research is required in order for causal feature selection methods to approach more closely the superior performance that experimental "ground truth" evidence and qualitative arguments suggest are available. We are however in a position to make some recommendations for use in practical applications:

- Use regularisation: Regularisation is known to be a viable alternative to feature selection in applications with unmanipulated data, where predictive performance is the primary objective rather than discovering a compact set of informative features (Miller, 2002). It has also been argued that when faced with covariate shift it may be



(a)

(b)

(c)

Figure 5: Statistical significance diagrams for (a) MARTI0, (b) MARTI1 and (c) MARTI2. The axis represents the TSCORE statistic, and the heavy bars denote groups of classifiers with statistically indistinguishable performance.

better to include bad features rather than delete good features (Guyon et al., 2008, §6.2), in which case using a larger feature set with regularisation to avoid over-fitting seems a sensible strategy.

- Use Bagging: A comparison of the size of the true Markov blanket of the unmanipulated distribution with the number of determined to belong to the Markov blanket of individual component classifiers and the number of features used by the ensemble as a whole, suggests that identification of the Markov blanket using HITON_MB is unstable (i.e. the composition of the Markov blanket depends substantially on the sample of data from which it was inferred). Model selection, including the tuning of the regularisation parameter is also subject to over-fitting the selection criterion (Cawley and Talbot, 2007), and bagging will help to alleviate this also (Hall and Robinson, 2009).

Table 6: Results obtained for the MARTI benchmark, see caption of Table 1 for details.

Dataset	**Selection**	**FNUM**	**FSCORE**	**DSCORE**	**TSCORE**	**AUC**
MARTI0	True MB	21	1.0000	0.9997	0.9998 ± 0.0010	0.9991
	Best TSCORE*	148	0.9078	1.0000	0.9996 ± 0.0010	—
	Winner§	128	0.8697	1.0000	0.9996 ± 0.0012	—
	Inferred MB	131	0.8862	1.0000	0.9995 ± 0.0011	0.9994
	Non-causal	44	0.8029	0.9998	0.9993 ± 0.0014	0.9986
	Causes & effects	15	0.7849	0.9987	0.9986 ± 0.0016	0.9978
	Yin *et al.*	11	0.6896	0.9982	0.9983 ± 0.0018	0.9973
	None	1024	0.7980	1.0000	0.9970 ± 0.0019	0.9950
	Causes only	3	0.5714	0.9821	0.9775 ± 0.0031	0.9346
MARTI1	True MB	14	1.0000	0.9889	0.9922 ± 0.0024	—
	Best TSCORE†	8	1.0000	0.8992	0.9542 ± 0.0041	—
	Winner§	32	0.8064	1.0000	0.9470 ± 0.0039	—
	Non-causal	44	0.7752	0.9998	0.9310 ± 0.0039	—
	Inferred MB	131	0.8265	1.0000	0.9234 ± 0.0045	—
	None	1024	0.7923	1.0000	0.9085 ± 0.0047	—
	Yin *et al.*	11	0.6399	0.9982	0.8988 ± 0.0046	—
	Causes & effects	15	0.7820	0.9987	0.8929 ± 0.0049	—
	Causes only	3	0.5347	0.9821	0.6370 ± 0.0059	—
MARTI2	True MB	2	1.0000	0.9277	0.9266 ± 0.0049	—
	Best TSCORE‡	2	1.0000	0.8099	0.8273 ± 0.0060	—
	Yin *et al.*	11	0.9980	0.9982	0.8130 ± 0.0053	—
	Winner§	64	0.9956	0.9998	0.7975 ± 0.0059	—
	Non-causal	44	0.9976	0.9998	0.7975 ± 0.0059	—
	Inferred MB	131	0.9966	1.0000	0.7740 ± 0.0060	—
	Causes & effects	15	0.9956	0.9987	0.7416 ± 0.0063	—
	None	1024	0.9951	1.0000	0.7193 ± 0.0062	—
	Causes only	3	0.7485	0.9821	0.6607 ± 0.0062	—

*Gavin Cawley "marti001 part006", †Reference "MB_NB_F_S", ‡Reference "FMBLR", §Gavin Cawley "final models".

Table 7: Summary of results for the final challenge submission. Top Ts gives the best Tscore amongst all valid final submissions, Max Ts gives the optimal Tscore, given knowledge of the true causal relationships, estimated using reference submissions, see caption of Table 1 for further details.

Dataset	**Causal Discovery**		**Target Prediction**				**Rank**
	Fnum	**Fscore**	**Dscore**	**Tscore**	**Top Ts**	**Max Ts**	
CINA0	128	0.5166	0.9737	0.9743	0.9765	0.9788	3
CINA1	128	0.5860	0.9737	0.8691	0.8691	0.8977	
CINA2	64	0.5860	0.9734	0.7031	0.8157	0.8910	
MARTI0	128	0.8697	1.0000	0.9996	0.9996	0.9996	1
MARTI1	32	0.8064	1.0000	0.9470	0.9470	0.9542	
MARTI2	64	0.9956	0.9998	0.7975	0.7975	0.8273	
REGED0	128	0.9410	0.9999	0.9997	0.9998	1.0000	2
REGED1	32	0.8393	0.9970	0.9787	0.9888	0.9980	
REGED2	8	0.9985	0.9996	0.8045	0.8600	0.9534	
SIDO0	4928	0.5890	0.9840	0.9427	0.9443	0.9467	1
SIDO1	4928	0.5314	0.9840	0.7532	0.7532	0.7893	
SIDO2	4928	0.5314	0.9840	0.6684	0.6684	0.7674	

- Investigate alternative base classifiers: In this study, we investigated only two base classifiers, linear ridge regression and BLogReg (for CINA). It may be that the benefits of causal feature selection may be obscured by the use of a base classifier that is unable to take advantage of non-linear relationships between features.

- In orienting the edges in the causal graph, it would be better to pre-filter the features to include not only the Markov blanket of the target, but also the parents and children of all features within the Markov blanket (c.f. Yin et al., 2008).

- Like conventional feature selection procedures, causal feature discovery methods appear to exhibit significant instability. An empirical characterisation of this instability would be an interesting area for further research.

5. Summary

In this paper, we have evaluated causal and non-causal feature selection procedures for ridge regression under covariate-shift. The reference submissions generated with knowledge of the true causal relationships clearly demonstrate that causal feature selection is very effective in mitigating against covariate-shift. However the models with causal feature selection procedures investigated here generally failed to out-perform models with non-causal feature selection (or indeed without a feature selection step), except on the most basic toy benchmark (LUCAS). This is a surprising and disappointing result for datasets designed for causal inference. It should be noted that the causal feature selection procedures are also computationally expensive, for instance identification of the Markov blanket for the SIDO dataset using HITON_MB took on average 76 hours, 57 minutes 8 seconds, and orientation of causal links using the PC algorithm took on average 50 hours, 21 minutes and 26 seconds. This means that the SIDO experiments consumed approximately 18 processor-months, without providing any improvement in predictive accuracy! These results demonstrate that causal inference is a challenging task, where further theoretical and algorithmic advances are likely to bring substantial practical benefits and where a more detailed empirical study is clearly warranted.

Acknowledgments

I would like to thank the co-organizers for their efforts in staging a very interesting and, for myself at least, educational challenge. I would also like to thank Gareth Janacek, the editors and the anonymous reviewers for their helpful and constructive comments, and Nicola Talbot for her help in preparing the manuscript.

References

C. F. Aliferis, I. Tsamardinos, and A. Statnikov. HITON: A novel Markov blanket algorithm for optimal variable selection. In *Proc. AMIA Annual Symposium*, pages 21–25, 2003.

D. M. Allen. The relationship between variable selection and prediction. *Technometrics*, 16:125–127, 1974.

G. C. Cawley. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *Proc. IJCNN-06*, pages 1661–1668, July 16–21 2006.

G. C. Cawley and N. L. C. Talbot. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics*, 22(19):2348–2355, October 1 2006.

G. C. Cawley and N. L. C. Talbot. Preventing over-fitting during model selection via Bayesian regularisation of the hyper-parameters. *Journal of Machine Learning Research*, 8:841–861, April 2007.

G. C. Cawley, N. L. C. Talbot, R. J. Foxall, S. R. Dorling, and D. P. Mandic. Heteroscedastic kernel ridge regression. *Neurocomputing*, 57:105–124, March 2004.

Y.-W. Chang and C.-J. Lin. Feature ranking using linear SVM. *JMLR: Workshop and Conference Proceedings*, 3, WCCI-2008 Workshop on Causality:53–54, 2008.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, January 1992.

I. Guyon and A. Eliseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, March 2003.

I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Design and analysis of the causation and prediction challenge. *JMLR: Workshop and Conference Proceedings*, 3, WCCI-2008 Workshop on Causality:1–33, 2008.

P. Hall and A. P. Robinson. Reducing the variability of crossvalidation for smoothing parameter choice. *Biometrika*, 96(1):175–186, March 2009.

A. Miller. *Subset selection in regression*, volume 95 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, second edition, 2002.

J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988.

J. Quiñonero Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, editors. *Dataset shift in machine learning*. Neural Information Processing Series. MIT Press, 2009.

K. Saadi, G. C. Cawley, and N. L. C. Talbot. Optimally regularised kernel Fisher discriminant classification. *Neural Networks*, 20(7):832–841, September 2007.

C. Saunders, A. Gammermann, and V. Vovk. Ridge regression in dual variables. In J. Shavlik, editor, *Proc. ICML-98*. Morgan Kaufmann, 1998.

S. Weisberg. *Applied linear regression*. John Wiley and Sons, New York, 2^{nd} edition, 1985.

J. Yin, Y. Zhou, C. Wang, P. He, Zhengm C., and Z. Geng. Partial orientation and local structure learning of causal networks for prediction. *JMLR: Workshop and Conference Proceedings*, 3, WCCI-2008 Workshop on Causality:93–105, 2008.

Table 8: Mean number of features used, and hold-out set AUROC score, over the 100 models comprising each of the ensembles used to make predictions.

Benchmark	**Selection**	**Features**	**AUROC**
LUCAS	None	11.00	0.9079
	Non-causal	10.99	0.9079
	Markov blanket	5.01	0.9082
	Causes & effects	4.00	0.8910
	Causes only	2.00	0.7832
LUCAP	None	143	0.9695
	Non-causal	6.03	0.9426
	Markov blanket	47.83	0.9674
	Causes & effects	39.91	0.9664
	Causes only	2.06	0.8089
CINA	None	132.00	0.9664
	Non-causal	29.44	0.9660
	Markov blanket	55.30	0.9660
	Causes & effects	21.21	0.9653
	Causes only	1.02	0.5351
REGED	None	999.00	0.9962
	Non-causal	14.69	0.9997
	Markov blanket	24.85	0.9995
	Causes & effects	11.11	0.9996
	Causes only	2.39	0.8961
SIDO	None	4932.00	0.9472
	Non-causal	28.96	0.9226
	Markov blanket	136.27	0.9348
	Causes & effects	10.07	0.8798
	Causes only	9.95	0.8733
MARTI	None	1024.00	0.9950
	Non-causal	15.19	0.9986
	Markov blanket	26.86	0.9994
	Causes & effects	8.60	0.9978
	Causes only	1.56	0.9346