Discovering Granger-causal Features from Deep Learning Networks

Aneesh Sreevallabh Chivukula[1], Jun Li[2], and Wei Liu[1]

[1] Advanced Analytics Institute,
[2] Centre for Artificial Intelligence,
University of Technology Sydney, Australia.
AneeshSrivallabh.Chivukula@student.uts.edu.au,
Jun.Li@uts.edu.au, Wei.Liu@uts.edu.au.

Abstract. In this research, we propose deep networks that discover Granger causes from multivariate temporal data generated in financial markets. We introduce a Deep Neural Network (DNN) and a Recurrent Neural Network (RNN) that discover Granger-causal features for bivariate regression on bivariate time series data distributions. These features are subsequently used to discover Granger-causal graphs for multivariate regression on multivariate time series data distributions. Our supervised feature learning process in proposed deep regression networks has favourable F-tests for feature selection and t-tests for model comparisons. The experiments, minimizing root mean squared errors in the regression analysis on real stock market data obtained from Yahoo Finance, demonstrate that our causal features significantly improve the existing deep learning regression models.

1 Introduction

Causal inference is a central theme in computational sciences that construct mathematical models for causation. In statistics, causality is defined over conditional dependencies modelled between features in the data. Such conditional dependencies are used to construct data distributions linking causes with effects in causal relations defined on data features. Such causal relations are useful for feature discovery in machine learning. The impact and risk of including causal relations or causal features is validated by domain knowledge.

The Granger-Sargent statistic and the Granger-Wald statistic are commonly used to discover Granger-causal features on time-domain and frequency-domain formulations of Granger causality [1]. In this paper we discover Granger-causal features by measuring model improvement in deep networks. Our models are useful for simulating time-dependent observations in application domains with neural computations in deep learning.

Deep learning is a class of neural networks that learn hierarchical feature representations approximating non-linear functions. In data-driven analytics applications, deep learning has been used to visualize, store, process and predict information. In supervised deep learning, the information is typically modelled as statistical correlations and variable associations. Introducing causality methods into supervised deep learning creates analytics models for data-driven decision making in an application domain where causal features are separated from spurious features.

In computational learning theory, loss functions are mathematical functions mapping a complex data-driven event in complex systems to real numbers. In decision and estimation theory, loss functions relate empirical risk defined on actual data to expected risk defined on predicted output of an analytics model.

In this paper, we analyze time series data distributions with the help of deep learning networks to discover causal relations and causal graphs from Granger causality tests [2]. To derive data representations, the deep networks are trained to optimize squared error loss functions between actual data and predicted output. The corresponding analytics predictions are tested and validated with statistical significance tests on regression errors. We also extend unrestricted models in Granger causality for supervised feature discovery with bivariate regression as well as supervised causal inference with multivariate regression. Theoretically, the deep network architecture and its squared error loss function determine empirical risk in our regression models.

Following are the major contributions of this paper:

- We identify Granger-causal features using deep networks that improve bivariate regression predictions amongst temporal dependencies in time series distributions.
- We discover Granger-causal graphs in time series distributions to improve multivariate regression in deep networks.
- We evaluate our theoretical model on Yahoo Finance data to solve causal inference problems defining stochastic processes found in financial markets.

The paper starts with related work in Section 2 comparing the new approach with existing approaches. Algorithms and experiments for the proposed method are presented in Section 3 and Section 4 respectively. The paper ends with Section 5 which summarises current and future work.

2 Related Work

Causality is generally defined on logical formalizations of different classes of knowledge, reasoning and complexity in data. Causality also depends on features and representations, patterns and noise from ground truth data generated in an application domain. Depending on a particular definition of causality, causal relations identify causal features for machine learning.

2.1 Causal Inference in Deep Learning

Causality methods have been applied to deep learning problems such as semi-supervised learning and transfer learning. In these problems informed priors retrieved from other networks are used to center the weights in hybrid deep learning networks. Such networks then construct statistically significant hypotheses and corresponding data representation on actual data from complex systems. An analytics model employing causality methods can then validate such hypotheses against causal features discovering patterns, structure, context and content in actual data [3]. In general, the instance space for learning causal features in actual data consists of concept adapting data structures like strings, trees, networks and tensors.

Backpropagation learning algorithms for deep networks have been improved by incorporating ideas for training probabilistic graphical models typically used in causal inference. Such training is inherently Bayesian where prior distributions inform and constrain analytics models predicting posterior distributions [4]. The improved deep learning algorithms result in a predicted output informed by a causality graph.

2.2 Causal Inference in Time Series Analysis

In time series analysis, causal inference is identifies and classifies events in time series such that the events have either deterministic or probabilistic causal relations. Events are identified by mapping logic and structure of natural language to concept lattices and causal graphs [5]. Historically, causal reasoning in time series builds on statistical analysis of covariance or correlation between two or more events in time series. The calculated correlation strength is then used to predict causal relation between two events [6]. The disadvantage of this approach is that it cannot determine the direction and significance of causation. It also cannot discover hidden causes and patterns for which observed events are effects.

Granger causality is a simple learning mechanism that allows us to explore all preceding ideas about causality methods in deep learning for time series analysis [7]. Here, Granger causality does not empirically prove actual causation between events but acts as a stepping stone to explore the phenomenon relating two events participating in a cause-effect relationship. Granger-causal features have been discovered with rule-based analytics models [8] and feature-based analytics models [9]. Our approach to causal inference also builds a feature-based analytics model.

3 Our Proposed Algorithms

We predict stock prices in financial markets with Deep Neural Networks (DNNs) for discriminative learning based regression models and Recurrent Neural Networks (RNNs) for sequence learning based regression models. Outputs from bivariate regression models are used to search Granger-causal features in multivariate time series data.

3.1 Empirical Risk training in Deep Learning Networks

Suppose a regression model for stock y having actual value $y(t)$ at time t predicts $\hat{y}(t;\boldsymbol{\alpha})$ parameterized by regression parameters $\boldsymbol{\alpha}$ belonging to parameter space A. In computational learning theory, the regression model is analyzed in terms of expected risk $E(L(\hat{y}(t;\boldsymbol{\alpha}), y(t)))$, which is defined as expected value of the loss function $L(\hat{y}(t;\boldsymbol{\alpha}), y(t))$, learning probability density function $P(\hat{y}(t;\boldsymbol{\alpha}), y(t))$ underlying the data [10]:

$$\text{Expected Risk} : E(L(\hat{y}(t;\boldsymbol{\alpha}), y(t))) = \int d(\hat{y}(t;\boldsymbol{\alpha}))d(y(t))L(\hat{y}(t;\boldsymbol{\alpha}), y(t))P(\hat{y}(t;\boldsymbol{\alpha}), y(t)) \tag{1}$$

The expected risk $E(L(\hat{y}(t;\boldsymbol{\alpha}), y(t)))$ is posed as a regression model when loss function $L(\hat{y}(t;\boldsymbol{\alpha}), y(t))$ is defined on squared errors computed between $\hat{y}(t;\boldsymbol{\alpha})$ and $y(t)$. If the regression model defining $L(\hat{y}(t;\boldsymbol{\alpha}), y(t))$ is learning a training dataset of finite size m, then expected risk $E(L(\hat{y}(t;\boldsymbol{\alpha}), y(t)))$ is called empirical risk [11] $\hat{E}(L(\hat{y}(t;\boldsymbol{\alpha}), y(t)))$.

$$\text{Empirical Risk : } \hat{E}_{y(t)\sim P(\hat{y}(t;\boldsymbol{\alpha}),y(t))}(L(\hat{y}(t;\boldsymbol{\alpha}), y(t))) = \frac{\Sigma_{i=1}^{m} L(\hat{y}(t;\boldsymbol{\alpha})^{(i)}, y(t)^{(i)})}{m} \tag{2}$$

The computational complexity of empirical risk $\hat{E}(L(\hat{y}(t;\boldsymbol{\alpha}), y(t)))$ is determined by the computational complexity of $L(\hat{y}(t;\boldsymbol{\alpha}), y(t))$ which in turn is determined by the regression model's feature selection and model validation. Thus, our intuition is that introducing causal features into deep networks not only minimizes empirical risk but also minimizes regression error.

In our deep network based regression models, regression error is minimized by the weights $\boldsymbol{\alpha}$ learnt on Squared Error (SE) Loss function $L(\hat{y}(t;\boldsymbol{\alpha}), y(t))$ as in Equation 3:

$$\text{SE Loss : } L(\hat{y}(t;\boldsymbol{\alpha}), y(t)) = (\hat{y}(t;\boldsymbol{\alpha}) - y(t))^2 \tag{3}$$

$L(\hat{y}(t;\boldsymbol{\alpha}), y(t))$ is determined by the deep network's data representation $P(\hat{y}(t;\boldsymbol{\alpha}), y(t)))$ of actual data $y(t)$. For training data of size m, the total loss function $L_{MSE}(\hat{y}(t;\boldsymbol{\alpha}), y(t))$ is given in Equation 4:

$$\text{MSE Loss : } L_{MSE}(\hat{y}(t;\boldsymbol{\alpha}), y(t)) = \frac{\Sigma_{i=1}^{m} L(\hat{y}(t;\boldsymbol{\alpha})^{(i)}, y(t)^{(i)})}{m} \tag{4}$$

By training a deep network model, we use either a DNN or RNN to minimize empirical risk in Equation 2. The backpropagation training algorithm solves for $\boldsymbol{\alpha}$ in Equation 4 with a stochastic gradient descent procedure finding best model fit on $P(\hat{y}(t;\boldsymbol{\alpha}), y(t))$.

3.2 Granger Causality testing in Deep Learning Networks

Causal features can be discovered by changing loss function $L(\hat{y}(t;\boldsymbol{\alpha}), y(t))$ in Equation 2 according to data representation $P(\hat{y}(t;\boldsymbol{\alpha}), y(t))$ in deep learning networks conditioned on actual past data $y(t-j), j = 1, 2, ..., p$ with p lags. In the deep network, $P(\hat{y}(t;\boldsymbol{\alpha}), y(t)) = P(\hat{y}(t;\boldsymbol{\alpha})|y(t-j))$ is the conditional probability of predicting regression value $\hat{y}(t)$ or it parameterized version $\hat{y}(t;\boldsymbol{\alpha})$ for stock y.

If another stock x at time point $x(t)$ with q lagged values $x(t-k), k = 1, 2, ..., q$, indicates the occurrence of $y(t)$ then we create a deep network conditioned on not only $y(t-j), j = 1, 2, ..., p$ but also $x(t-k), k = 1, 2, ..., q$. Then, $P(\hat{y}(t;\boldsymbol{\alpha};\boldsymbol{\beta}), y(t)) = P(\hat{y}(t;\boldsymbol{\alpha};\boldsymbol{\beta})|y(t-j), x(t-k))$ is conditional probability of predicting regression value $\hat{y}(t)$ or it parameterized version $\hat{y}(t;\boldsymbol{\alpha};\boldsymbol{\beta})$ for stock y parameterized by regression parameters tensors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ belonging to deep network parameter spaces A and B respectively.

From data representations $P(\hat{y}(t;\boldsymbol{\alpha}), y(t))$ and $P(\hat{y}(t;\boldsymbol{\alpha};\boldsymbol{\beta}), y(t))$ defined above, we devise following Granger causality test using Equation 3 to predict $\hat{y}(t;\boldsymbol{\alpha})$ and $\hat{y}(t;\boldsymbol{\alpha};\boldsymbol{\beta})$ as dependent test variables for $y(t-j), x(t-k)$ as independent test variables.

$$\text{restricted model: } \hat{y}(t;\boldsymbol{\alpha}) = L(P(\hat{y}(t;\boldsymbol{\alpha}), y(t))) = L(P(\hat{y}(t;\boldsymbol{\alpha})|y(t-j))) \tag{5}$$

$$\begin{aligned}\text{unrestricted model: } \hat{y}(t;\boldsymbol{\alpha};\boldsymbol{\beta}) &= L(P(\hat{y}(t;\boldsymbol{\alpha};\boldsymbol{\beta}), y(t))) \\ &= L(P(\hat{y}(t;\boldsymbol{\alpha};\boldsymbol{\beta})|y(t-j), x(t-k)))\end{aligned} \tag{6}$$

The null hypothesis of no Granger causality is rejected if and only if $x(t-k)$ has been retained along with $y(t-j)$ in the $\hat{y}(t)$ regression according to an F-test on Root Mean Squared Errors (RMSEs) between $\hat{y}(t)$ and $y(t)$. The F-test in Definition 1 [2] determines the Granger causality relation between stocks x and y where RMSE is computed for unrestricted regression as $RMSE_{ur}$ and restricted regression as $RMSE_r$.

Definition 1. *F-statistic* $= \frac{\frac{RMSE_r - RMSE_{ur}}{q-p}}{\frac{RMSE_{ur}}{n-q}}$

To compute causal features over N multivariate time series $X = \{X(t)^u\}, u \in [1, N], t \in [1, n]$ selected from N stock prices at n time points in financial markets, we repeat the F-test for every pair of stocks x and y. In each F-test, the null hypothesis is that the sample means of predictions are equal and the regression parameters $\boldsymbol{\beta}$ are zero. The alternative hypothesis is that there is significant variation between the sample means of predictions for some non-zero $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The null hypothesis is rejected if p-value on F-test has a significance level less than 0.05. If the null hypothesis is rejected, deep network features $y(t-j)$ and $x(t-k)$ Granger cause predicted output $\hat{y}(t)$ with actual stock price $y(t)$. In experiments with deep learning networks, a Granger-causal feature is represented by the causal relation $x \to y$ for stocks x and y.

3.3 Multivariate Regression validation with Deep Learning Networks

While we introduced theory to identify single Granger-causes in the previous subsection, in this subsection we explain the discovery of multiple Granger-causes for a given stock. Incorporating multiple Granger-causal features into the F-test allows us to improve deep learning with causal reasoning on multivariate time-dependent data.

Therefore, we discover multiple Granger-causal features with multivariate regression in an unrestricted model. The multiple Granger-causal features discovery process validates the single Granger-causal features and predictions. We extend the unrestricted model for bivariate regression in Equation 6 to the unrestricted model for multivariate regression as in Equation 7.

$$\begin{aligned}\text{multivariate unrestricted model: } \hat{y}(t;\{\boldsymbol{\alpha^w}\}) &= L(P(\hat{y}(t;\{\boldsymbol{\alpha^w}\}), y(t))) \\ &= L(P(\hat{y}(t;\{\boldsymbol{\alpha^w}\})|y(t-j), \{x(t-k)^w\}))\end{aligned} \tag{7}$$

Equation 7 predicts $\hat{y}(t)$ by discovering statistically significant Granger-causal features $\{x^w\} \to y,\ w \in [1, N]$ from multivariate regression. As detailed in Algorithm 2 in the next subsection, Granger causality test of Equation 5 is applied to all pairs of restricted and unrestricted models that differ in one independent variable x^w discovered by bivariate regression. A feature selection procedure for multivariate regression searches candidate feature sets in the power set of the set $\{x^w\}$. The optimal feature set is determined by $\{\boldsymbol{\alpha^w}\}$ with minimum RMSE $RMSE_{mv}$.

In bivariate regression, the single Granger-causal features are discovered by a DNN-based and RNN-based regression model. In multivariate regression, multiple Granger-causal features are discovered by a DNN-based regression model.

Algorithm 1 Discovery of Granger-causal features using deep learning networks

```
Input: Multivariate time series : $X = \{X(t)^u\}, u \in [1, N], t \in [1, n]$; Granger causality lags $p, q \in \mathbb{Z}$;
Output: Predictive model output : Bivariate Granger-causal features graph $G_{MSE}$; Multivariate Granger-causal features set $C_{mv}$; Bivariate regression errors $RMSE_r$, $RMSE_{ur}$ for restricted and unrestricted model; Multivariate regression errors $RMSE_{mv}$ for unrestricted model;
1: $G_{MSE} = C_{mv} = \Phi, RMSE_r = RMSE_{ur} = RMSE_{mv} = \Phi$
2: for $u \in [1, N]$ do
3:     $y(t) = X(t)^u$
4:     for $v \in [1, N]$ and $v \neq u$ do
5:         $x(t) = X(t)^v$
6:         Create preprocessed and lagged cross validation data $y(t-j), x(t-k)$ with lags $p, q$ from time series $y(t), x(t), t \in [1, n]$
7:         Construct restricted and unrestricted regression model on actual data $y(t), x(t)$ according to Equation 5 and Equation 6.
8:         Construct MSE loss predictions $\hat{y}(t)$ from Equation 4 for DNN as well as RNN networks.
9:         Calculate regression errors $RMSE_r$ and $RMSE_{ur}$ for each $\hat{y}(t)$ and $y(t)$.
10:        From Definition 1, compute $F\text{-}statistic$ over $RMSE_r$ and $RMSE_{ur}$.
11:        if $F\text{-}statistic > 0.05$ then
12:            if model is restricted then
13:                Update bivariate regression error, $RMSE_r[u][v] = RMSE_r$, for restricted model
14:            else
15:                Update bivariate regression error, $RMSE_{ur}[u][v] = RMSE_{ur}$, for unrestricted model
16:            Update Granger-causal features, $G_{MSE}[u] = G_{MSE}[u] \cup x(t) \rightarrow y(t)$, for bivariate regression
17: for $u \in [1, N]$ do
18:     $y(t) = X(t)^u$
19:     Retrieve bivariate Granger-causal features $\{x(t)^w\}$ for $u$ from $G_{MSE}$
20:     $RMSE_{mv}[u], C_{mv}[u]$ = multivar_granger($y(t)$, $\{x(t)^w\}$, $RMSE_{ur}$) to compute multivariate regression outputs.
21: return $RMSE_r, RMSE_{ur}, RMSE_{mv}, G_{MSE}, C_{mv}$
```

3.4 Deep Learning Networks based Regression Models

Algorithm 1 gives learning algorithm implementing Equation 5 and Equation 6 for loss function in Equation 4. The algorithm requires a multivariate time series $X = \{X(t)^u\}$ to predict regression model's causal graph G_{MSE} of Granger-causal features and corresponding regression errors $RMSE_r$, $RMSE_{ur}$, $RMSE_{mv}$ for the restricted model, the unrestricted model and the multivariate model participating in Granger causality.

Algorithm 1 executes from Line 1 to Line 16 for every pair of time series $y(t), x(t) \in X$ with lags p, q. Line 6 prepares crossvalidation data for training deep network on Line 8 which depends on the prediction $\hat{y}(t)$ from Granger causality models in Line 7. $\hat{y}(t)$ is predicted as a complex nonlinear combination of features $y(t-j)$ and $x(t-k)$ in Line 9. On Line 13 and Line 15, bivariate regression errors $RMSE_r$, $RMSE_{ur}$ are computed on actual time point $y(t)$ and predicted time point $\hat{y}(t)$. Line 11 applies F-test to discover Granger causality relations in Line 16. The null hypothesis of not finding Granger-causal features is rejected at 5% significance level. The corresponding Granger-causal graph G_{MSE} is searched on Line 19 to improve multivariate regression errors $RMSE_{mv}$ on Line 20. In Algorithm 1, while loop from Line 2 to Line 16 discovers single Granger-causal features G_{MSE} with bivariate regression, loop from Line 17 to Line 20 discovers multiple Granger-causal features C_{mv} with multivariate regression. Algorithm 1 ends on Line 21 by returning Granger-causal features G_{MSE}, C_{mv} as well as their regression errors $RMSE_r$, $RMSE_{ur}$ and $RMSE_{mv}$.

Algorithm 2 called on Line 20 of Algorithm 1 gives the search procedure implementing Equation 7. Algorithm 2 requires unrestricted model error $RMSE_{ur}$ found for bivariate regression predicting $y(t)$ from single Granger-causal features $\{x(t)^w\}$. The causal relations discovered between $\{x(t)^w\}$ are in Granger-causal graph G_{MSE}. Algorithm 2 then returns unrestricted model error $RMSE_{mv}$ from multivariate regression as well as corresponding multiple Granger-causal features set c_{mv} discovered by multivariate regression network. For all predicted $\{X(t)^u\}$, Granger-causal feature sets C_{mv} stored on Line 20 are the optimal Granger-causal feature sets discovered across many multivariate regression networks. The loop from Line 4 to Line 19 in Algorithm 2 uses two sets of selected_causes and candidate_causes to generate and evaluate candidate Granger-causal feature sets for unrestricted model in multivariate regression. On Line 2, selected_causes are initialized to Granger-causal features $\{x(t)^w\}$ discovered in G_{MSE} of Algorithm 1. On Line 6, Cartesian product of selected_causes and bivariate Granger-causal features $\{x(t)^w\}$ generates candidate_causes. $Candidate\_RMSE_r$, $Candidate\_RMSE_{ur}$ are used to track regression errors of restricted and unrestricted models built from candidate_causes. On Line 10, initially a restricted model in multivariate regression is assumed to be the same as the unrestricted model in bivariate regression. Later as the loop from Line 4 to Line 19 crosses more than one iteration as tracked by counter $iter$, the restricted model is evaluated against Granger-causal features $c \setminus \{x(t)^w\}$ on Line 12 while the unrestricted model is evaluated against Granger-causal features c on Line 13. In any giver iteration $iter$, the restricted and unrestricted models differ by only one of the Granger-causal features present in $\{x(t)^w\}$. The multivariate regression error $Candidate\_RMSE_{ur}$ is computed for each candidate c at Lines 13-15. If the corresponding F-statistic is greater than a predefined threshold on Line 17, then the candidate c is found to be a legitimate Granger-causal feature for subsequent processing with multivariate regression. Such a c is updated to selected_causes on Line 18. For every new iteration $iter$, selected_causes are reset to the empty set on Line 7 immediately after being used to generate candidate_causes on Line 6. This loop convergence condition ensures that larger Granger-causal feature sets are generated across iterations. On convergence, no further selected_causes are available for processing. Algorithm 2 terminates the search procedure by returning the optimal Granger-causal feature set c_{mv} that minimizes multivariate regression error $RMSE_{mv}$.

4 Experiments

Table 1: Companies Listing

Abbreviation	Company Name	Abbreviation	Company Name
AAPL	Apple Inc.	MCD	McDonald's Corporation
ABT	Abbott Laboratories	MSFT	Microsoft Corporation
AEM	Agnico Eagle Mines Limited	ORCL	Oracle Corporation
AFG	American Financial Group, Inc.	WWD	Woodward, Inc.
APA	Apache Corporation	T	AT&T Inc.
CAT	Caterpillar Inc.	UTX	United Technologies Corporation

In this section we discuss the empirical validation of Granger-causal features in deep learning networks regression models. Table 1 lists the stocks from different financial sectors in Standard & Poors 500 - a stock market index based on the market capitalizations of 500 large companies having common stock listed on the NYSE or NASDAQ.

Algorithm 2 Search procedure for constructing multivariate Granger-causal graphs

Input: Effect time series $y(t)$; Bivariate Granger-causal features $\{x(t)^w\}$; Bivariate Regression errors $RMSE_{ur}$ for unrestricted model

Output: Optimal Granger-causal feature set c_{mv} and multivariate regression error $RMSE_{mv}$ for unrestricted model

```
function MULTIVAR_GRANGER($y(t), \{x(t)^w\}, RMSE_{ur}$)
    Initialize selected_causes to Bivariate Granger-causal features $\{x(t)^w\}$
    $iter = 0, Candidate\_RMSE_r = Candidate\_RMSE_{ur} = \Phi$
    while selected_causes $\neq \Phi$ do
        $iter \mathrel{+}= 1$
        Generate candidate_causes, candidate_causes = $\{x(t)^w\}\times$ selected_causes, from previous iteration's selected_causes
        Reset selected_causes to $\Phi$ in current iteration
        for each candidate cause $c \in$candidate_causes do
            if $iter == 1$ then
                Set restricted model error $Candidate\_RMSE_r[c] = RMSE_{ur}[c]$
            else
                Set restricted model error $Candidate\_RMSE_r[c] = Candidate\_RMSE_{ur}[c \setminus \{x(t)^w\}]$
            Construct multivariate unrestricted regression model on actual data $y(t)$ and $\{x(t)^w\}$ according to Equation 7
            Construct MSE loss predictions $\hat{y}(t)$ from Equation 4 for DNN networks.
            Calculate regression error $Candidate\_RMSE_{ur}[c]$ for all $\hat{y}(t)$ and $y(t)$.
            From Definition 1, compute $F\text{-}statistic$ over $Candidate\_RMSE_r[c]$ and $Candidate\_RMSE_{ur}[c]$.

            if $F\text{-}statistic > 0.05$ then
                Update Granger-causal features: selected_causes = selected_causes $\cup$ $c$
    end while
    Among unrestricted models $Candidate\_RMSE_{ur}$, find optimal Granger-causal feature set $c_{mv}$ with minimum multivariate regression error $RMSE_{mv}$
    return $RMSE_{mv}, c_{mv}$
end function
```

The stocks daily closing prices were obtained from Yahoo Finance website [3]. The data is obtained for a period of 21 years from 26-07-1996 to 25-07-2017.

The regression model's feature learning is determined by deep network structure weights $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\{\boldsymbol{\alpha^w}\}$ with MSE loss function. Deep network structure is designed to minimize bivariate/multivariate regression errors and maximize significant Granger causes in the unrestricted model. We treat the regression model as a time-dependent data-based model with causal lags p, q set to a default value of 200 days. 5285 days of time points are used to create the crossvalidation data. Each data record has delayed prices time series predicting current price of a given stock. For fair comparison of baseline models, we split 30% of crossvalidatiton data into testing data while remaining 70% of crossvalidatiton data is taken to be training data.

On bivariate data, we treat regression modelling problem as a discriminative learning problem in DNNs as well as a sequence learning problem in RNNs to show that discovered Granger-causal features are not specific to a given network structure. On multivariate data, we treat regression modelling problem as a discriminative learning problem in DNNs to validate generalization capability of proposed feature discovery procedure. The regression errors for discovering Granger-causal features are also compared with those from a Autoregressive Integrated Moving Average (ARIMA) regression model. A grid search procedure is used to select ARIMA training parameters. Number of training epochs in DNN is set to a default value 50 over a total of 12 stocks. The DNN has three

[3] `https://finance.yahoo.com/`

Table 2: RMSEs with MSE loss for bivariate regression. DNN is selected as the best network structure for Granger causality.

Abbreviation	ARIMA	LSTM	GRU	DNN
AAPL	0.807	1.449	1.475	0.504
ABT	0.748	0.461	0.469	0.626
AEM	1.643	1.115	1.107	0.143
AFG	0.795	0.580	0.588	0.485
APA	2.795	1.558	1.520	0.145
CAT	1.254	1.474	1.452	0.106
MCD	0.319	0.981	0.994	0.425
MSFT	1.555	0.597	0.606	0.361
ORCL	0.190	0.521	0.520	0.497
T	0.786	0.335	0.339	0.078
UTX	0.209	1.110	1.113	0.297
WWD	0.237	0.817	0.819	0.311
t-test	1.24×10^{-2}	2.21×10^{-4}	1.89×10^{-4}	Base

hidden layers consisting of dense activation units and dropout regularization units. It is implemented in Keras [4] – a Tensorflow based API for deep learning. All time series are subject to min-max normalization before training.

We experiment with two variants of RNN with Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) activation units. The number of training epochs in RNN is set to a default value 15. The LSTM has one hidden layer consisting of LSTM activation unit with 50 neurons. The GRU has three hidden layers consisting of GRU activation units with 25 neurons. Dropout units are the regularization units. LSTM as well as GRU state is reset after each training epoch. The LSTM and GRU are trained for 200 time steps - one record at a time - over lagged data. The time series data is differenced and scaled to a range of [-1,1]. For multivariate regression, all the identified single Granger-causal features are used as input. On multivariate testing data, regression values are predicted one time step at a time.

4.1 Single Granger-causes validation

For each company's price time series, autoregression models RMSEs are reported in Table 2. From t-test statistics in Table 2, we find DNN generally has better performance than competitive models. So we choose DNN as the regression model for discovering Granger-causal features with bivariate regression in Table 3 as well as multivariate regression in Figure 1. For experimental validation of our algorithms, we also report Granger-causal features discovered by a GRU model in Table 4.

Table 3 and Table 4 report RMSEs for restricted model $RMSE_r$ and unrestricted model $RMSE_{ur}$. $RMSE_{ur}$ is consistently lower than $RMSE_r$ for Granger causality models given in Equation 5 and Equation 6 respectively. Each row in Table 3 and Table 4 shows pairwise causal relations and their RMSEs. From t-test p-value statistic comparing RMSEs with and without Granger-causal features in Table 3 and Table 4, we conclude that unrestricted model shows non-trivial reduction in RMSE compared to restricted model for any random pair of stocks involved in Granger causality. Figure 1(a) represents Granger-causal features discovered from bivariate regression as a causal graph between time series of stock prices represented by vertices where F-test statistics represented by edges show the strength of Granger causality.

[4] `https://www.tensorflow.org/api_docs/python/tf/contrib/keras`

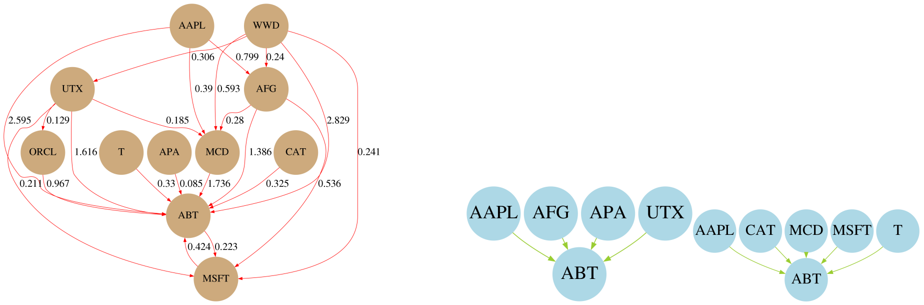
Table 3: RMSEs with MSE loss for Granger-causal feature discovery. The rows show causal relations with the restricted model and the unrestricted model RMSEs $RMSE_r$ and $RMSE_{ur}$ in bivariate regression with DNN.

Causal Relation	$RMSE_r$ Restricted model (DNN without causes)	$RMSE_{ur}$ Unrestricted model (our model with single cause)	Causal Relation	$RMSE_r$ Restricted model (DNN without causes)	$RMSE_{ur}$ Unrestricted model (our model with single cause)
AAPL → ABT	0.626	0.198	AAPL → AFG	0.485	0.293
AFG → ABT	0.626	0.191	WWD → AFG	0.485	0.396
APA → ABT	0.626	0.477	AAPL → MCD	0.425	0.315
CAT → ABT	0.626	0.372	AFG → MCD	0.425	0.418
MCD → ABT	0.626	0.261	UTX → MCD	0.425	0.365
MSFT → ABT	0.626	0.501	WWD → MCD	0.425	0.353
ORCL → ABT	0.626	0.362	ABT → MSFT	0.361	0.295
T → ABT	0.626	0.535	AFG → MSFT	0.361	0.249
UTX → ABT	0.626	0.271	UTX → MSFT	0.361	0.297
WWD → ABT	0.626	0.184	WWD → MSFT	0.361	0.183
WWD → UTX	0.297	0.219	UTX → ORCL	0.497	0.202
t-test	1.17×10^{-6}	Base	t-test	1.17×10^{-6}	Base

Table 4: RMSEs with MSE loss for Granger-causal feature discovery. The rows show causal relations with the restricted model and the unrestricted model RMSEs $RMSE_r$ and $RMSE_{ur}$ in bivariate regression with RNN.

Causal Relation	$RMSE_r$ Restricted model (RNN without causes)	$RMSE_{ur}$ Unrestricted model (our model with single cause)	Causal Relation	$RMSE_r$ Restricted model (RNN without causes)	$RMSE_{ur}$ Unrestricted model (our model with single cause)
ABT → AAPL	1.475	0.403	AFG → APA	1.529	0.788
AFG → AAPL	1.475	0.781	MCD → APA	1.571	0.944
MCD → AAPL	1.475	0.936	MSFT → APA	1.522	0.859
MSFT → AAPL	1.475	0.851	ORCL → APA	1.551	0.732
ORCL → AAPL	1.475	0.726	T → APA	1.527	0.741
T → AAPL	1.475	0.734	UTX → APA	1.522	1.021
UTX → AAPL	1.475	1.012	WWD → APA	1.526	0.846
WWD → AAPL	1.475	0.839	ABT → CAT	1.445	0.474
ABT → AEM	1.107	0.458	AFG → CAT	1.445	0.918
ABT → AFG	0.588	0.303	MSFT → CAT	1.445	1.001
ABT → APA	1.545	0.407	ORCL → CAT	1.444	0.853
ABT → MCD	0.994	0.382	T → CAT	1.445	0.863
ORCL → MCD	0.994	0.687	WWD → CAT	1.444	0.986
T → MCD	0.994	0.695	ABT → UTX	1.113	0.421
ABT → ORCL	0.521	0.257	ORCL → UTX	1.113	0.756
ABT → T	0.338	0.231	T → UTX	1.114	0.765
ABT → MSFT	0.606	0.256	ABT → WWD	0.821	0.397
t-test	3.21×10^{-11}	Base	t-test	3.21×10^{-11}	Base

Thus Table 3 and Table 4 validate our proposal to use Granger causality in feature selection for deep networks based regression models. We also observe that the proposed feature discovery process and supervised learning process are robust to any particular deep network structure.



(a) Granger-causal graph with F-test statistics computed on RMSEs $RMSE_r$,$RMSE_{ur}$ in the restricted and unrestricted models for bivariate regression

(b) Two sets of Granger-causal features discovered by Algorithm 2 for predicting stock price ABT with multivariate regression. They reduce restricted model $RMSE_r$ from 0.626 to unrestricted model $RMSE_{mv}$ 0.141 and 0.178 respectively.

Fig. 1: Granger-causal features, F-statistics on RMSEs $RMSE_r$,$RMSE_{ur}$ and multivariate regression RMSEs $RMSE_{mv}$ for the unrestricted model with DNN. The edge directions indicate the causal relations between pairs of stocks and the edge weights show the corresponding F-test statistic given in Definition 1.

4.2 Multiple Granger-causes validation

Figure 1(a) shows the Granger-causal graph with directed weighted edges that are outcomes of Definition 1. It indicates Granger-causal relations discovered for bivariate regression. The edge weights are F-test statistics for all the unrestricted models that reduce RMSEs in bivariate regression. For example, the causal relation $UTX \rightarrow ORCL$ indicates that UTX causes ORCL or ORCL is caused by UTX with F-test statistic 0.129. This relation has been selected in the Granger-causal graph because the unrestricted model including UTX prices in ORCL price prediction leads to a RMSE reduction from 0.497 to 0.202 according to Table 3. Figure 1(a) shows all the causalities identified on the training data. We do not assume causalities change at every time point. From Figure 1(a), we not only can identify causal features but also indicate the strength of causality.

Figure 1(b) shows the top ranked Granger-causal features discovered by Algorithm 2 from Figure 1(a). These causal features are suitable for multivariate regression. In Figure 1(a), vertices like APA without Granger-causes, ORCL and MSFT with one and two Granger-causes result in no output from Algorithm 2. For vertices like ABT with non-zero single Granger-causes, Algorithm 2 identifies multiple Granger-causes. For ABT, Algorithm 2 outputs a total of 69 Granger-causes which reduces RMSE $RMSE_{mv}$ in multivariate regression models from $RMSE_r = 0.626$ in the restricted model to $RMSE_{mv} \in [0.141, 0.541]$ in the unrestricted model. In Figure 1(b), the multivariate Granger-cause $\{$AAPL, AFG, APA, UTX$\}$ has regression error of $RMSE_{mv} = 0.141$ while $\{$AAPL, CAT, MCD, MSFT, T$\}$ has regression error of $RMSE_{mv} = 0.178$ in the unrestricted model. Of the 69 causes, the longest but not optimal Granger-causes are found to be $\{$AAPL, APA, CAT, MCD, ORCL, T, UTX, WWD$\}$ with $RMSE_{mv} = 0.162$ and $\{$AAPL, APA, CAT, MCD, MSFT, T, UTX, WWD} with $RMSE_{mv} = 0.174$ in the unrestricted model. We also find two Granger-causes of length 8, six Granger-causes of length 7 and ten Granger-causes of length 6. From Figure 1(b), we observe that multivariate regression on Granger-causal features results in a better unrestricted model than bivariate regression on Granger-causal features. In bivariate regression as well as multivariate regression, while F-test statistics on RMSEs validate our feature selection on regression errors, t-test statistics on RMSEs support our model validation on regression errors.

5 Conclusion and Future Work

We presented deep networks based regression models to augment and discover Granger-causal features analyzing multivariate time series data from finance domain. Our Granger-causal features are able to significantly improve multivariate regression performance. We also constructed Granger-causal graphs to capture temporal dependencies in multivariate data. On real stock market data we demonstrate that our theoretical model significantly outperforms existing deep learning regression models. As future work we shall combine multiple data sources to extract regularized features for cost sensitive concept learning and big data pattern detection.

6 Acknowledgements

This research is partially funded by The Capital Markets Cooperative Research Centre, Australia.

References

1. Guo, S., Ladroue, C., Feng, J. In: Granger Causality: Theory and Applications. Volume 15. Springer-Verlag London Limited (2010) 83
2. Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society (1969) 424–438
3. Mirowski, P., Ranzato, M., LeCun, Y.: Dynamic auto-encoders for semantic indexing. In: Proceedings of the NIPS 2010 Workshop on Deep Learning. (2010) 1–9
4. Spirtes, P., Glymour, C.N., Scheines, R.: Causation, prediction, and search. MIT press (2000)
5. Keogh, E., Chu, S., Hart, D., Pazzani, M.: Segmenting time series: A survey and novel approach. Data mining in time series databases **57** (2004) 1–22
6. Fu, T.c.: A review on time series data mining. Engineering Applications of Artificial Intelligence **24**(1) (2011) 164–181
7. Bahadori, M.T., Liu, Y.: An examination of practical granger causality inference. In: Proceedings of the 2013 SIAM International Conference on Data Mining, SIAM (2013) 467–475
8. Kleinberg, S.: Causal inference with rare events in large-scale time-series data. In: Proceedings of the 2013 International Joint Conference on Artificial Intelligence. (2013)
9. Li, Z., Zheng, G., Agarwal, A., Xue, L., Lauvaux, T.: Discovery of causal time intervals. In: Proceedings of the 2017 SIAM International Conference on Data Mining, SIAM (2013) 804–812
10. Ancona, N., Marinazzo, D., Stramaglia, S.: Radial basis function approach to nonlinear granger causality of time series. Phys. Rev. E **70** (Nov 2004) 056221
11. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016) `http://www.deeplearningbook.org`.