Deep Network Regression with Granger Causality and Quadratic Mean Loss Functions

Aneesh Sreevallabh Chivukula and Wei Liu
Advanced Analytics Institute, University of Technology Sydney, Australia
AneeshSrivallabh.Chivukula@student.uts.edu.au, Wei.Liu@uts.edu.au

Abstract

custom loss functions, feature learning and causal inference, deep learning and supervised learning, regression in multiple time series, finance data analysis

Keywords : Supervised learning, Data mining and knowledge discovery, Evolutionary learning, Adversarial learning, Deep learning, Genetic algorithms, Game theory

Introduction

TO DO : finish (i) experiments (ii) pseudocode
Causality inference is a central theme in computational sciences that construct mathematical models for causation. Statistics is a tool that is typically used to understand causation. In statistics, causality is defined over conditional dependencies modelled between features in the data (Pearl and others 2009). The conditional dependencies are used to construct probability distributions linking the causes with effects in a causal chain. The causal chain is used for estimating parameters in a machine learning model. The impact and risk of events in causal chain is then validated against domain knowledge (Lopez-Paz et al. 2016). Thus, changes in the causal chains create uncertainty in the estimated parameters. In this paper we propose a custom loss function to account for such uncertainty in the parameters. We use a deep learning model to estimate the parameters. The loss function in deep learning model is useful for simulating observations about the application domain with computations in the learning algorithms.

Deep learning models are a class of neural networks that learn hierarchical feature representations approximating non-linear functions defined on the analytics models outputs and inputs. In data-driven applications and predictive analytics, deep learning algorithms have been used to visualize, store, process and predict information. The information is modelled as statistical correlations between features and labels obtained from the application domain. Directly using such models without considering the meaning of the features and labels can lead to erroneous conclusions. Introducing causal methods into the deep learning model allows us to separate causal features from spurious features which inturn allows stakeholders to make well-informed decisions in the application domain.

Learning algorithms support decision making with behavioural statistics inferring hidden patterns in complex data-driven events. Loss functions are mathematical functions mapping a complex data-driven event to a real number that relates the algorithmic output to the theoretically expected analytics output (Liu and Chawla 2011). In decision theory, the theoretically expected analytics output is expressed in terms of the statistical significance tests used in error analysis. Normally, the difference between actual output and predicted output of the learning algorithm is measured by the estimation error in a loss function. Thus, the objective of a stable learning algorithm then is to reduce instability in the learning model by reducing the variance component of the estimation error. A stable learning model does not undergo large changes for small changes in the features due to adversarial or stochastic noise (Luo et al. 2015). In this context, a loss function can also be defined as a mathematical measure on linear dimension that quantifies degree of learning model's preference across alternatives. It explains the impact of uncertainty with reference to a distribution of outcomes. Then, an optimal solution for a loss function maximizes expected utility or average utility in the sense of decision-theoretic rationality.

In this paper we work with multivariate time series data obtained in the finance markets. Theoretically, we find sampled, constructed and extracted features in time series that are useful for finding probability distributions comparing time points in the local time window with the time points in the global time window. Such probability distributions are useful for pattern mining and feature learning in time series data. In this context, deep learning models optimize the custom loss function whose estimation error is defined over various types of parameters estimating class and cost distributions over the regularized input features. The output of optimization are both predictions and representations coupled with statistical tests combining observational data with randomized data in model training. Depending on the neural network architecture and loss function definition determining empirical risk in the deep learning model, we can use the information gained from factual scenarios in the observational data to assess the probabilities for counterfactual scenarios in the randomized data. In this paper, we assume the counterfactual scenario is represented by data skew patterns in multivariate time series data. Furthermore, the data skew pattern is analyzed for constructed features and causal features that can be used in training a deep neural network with custom loss layer.

Thus knowledge of errors in the deep learning model selection help us target the development efforts to decision making processes underlying predictive analytics restricted to associations in the data. In this paper, we use granger regression causality as the statistical significance test improving upon deep learning models with loss functions. The proposed loss function is implemented in Tensorflow. The proposed deep learning model is trained using a backpropagation algorithm. The gradients computation in the backpropagation algorithm is well supported by Tensorflow variables and operations. Feature learning is done by fully connected layers in the neural network.

Following are the major contributions of this paper.

- We formulate empirical risk based quadratic mean loss function to capture temporal dependence in deep learning for regression.
- We improve deep learning performance and simultaneously reduce regression error by including granger causal features on time points.
- We generate a causal graph in multivariate time series regression and apply deep learning regression over financial markets data.

Related Work

Causality Inference in Deep Learning

In machine learning and deep learning the objective is taken to be a predictive hypothesis on sensory data. Depending on definition of causality, different features may be identified as the causes of the same ground truth against which the predictive hypothesis is tested. The definition of causality is given based on logical formalizations of different classes of knowledge, reasoning and complexity. The statistically significant causes for predictive hypothesis are validated on features structure, context and content in a given application domain (Mirowski, Ranzato, and LeCun 2010). Here, noise would consist of anticausal features exploiting irregularities, dissimilarities and anomalies in the data. Causality methods have been applied to deep learning problems such as semi-supervised learning and transfer learning where informed priors retrieved from other networks are used to center the weights in hybrid deep learning architectures (Liu et al. 2016).

In thus comparing predicted output with actual output, deep learning models act as computational methods reasoning under uncertain environments. We assume the learning algorithm(or learner for short) is acting in an unknown environment. The environment is simulated by loss functions in complex systems. Such loss functions interpret causal features in the light of domain knowledge regarding the complex system. Furthermore, the environment maybe partially stochastic and observable. The learner's objective is not only to minimize validation error but also best cope with learning in the environment. This is our general regime for model based learning in deep learning. Here loss functions can also be defined over causal graphs whereas learners error functions are defined over constructed features. The causal graphs are estimated over the features and labels in the application domain. If the features are constructed from a carefully designed experiment, we can interpret statistical tests on changes in the features in the deep learning model as representing causal effects of the features. Moreover, if we intervene on these particular features, we can expect to observe the estimated causal effect. Formal causal analysis can help us better design the learner's environment having features interpretation or human intervention or both (Lipton 2016).

Model uncertainty can also be expressed as a comparison between associational inference on feature construction and causal inference on model construction in deep learning. The instance space for detecting such patterns consists of concept adapting data structures like strings, trees, networks and tensors (Zanin et al. 2016). Including causal features derived on such instances allow us to simulate not only prediction performance but also model selection in deep learning with prior probabilities. The backpropagation training algorithm used in deep learning has been improved by incorporating ideas for training probabilistic graphical models used for causal inference (Bishop 2006). The training approach is inherently Bayesian where priors inform and constrain the models to get uncertainty estimation as a posterior distribution (Spirtes, Glymour, and Scheines 2000). Alongwith a prediction output, such training algorithms output a causality graph that gives causal relationships between input and output to the deep learning model. Thus causal methods combined with deep learning can help build generative models over prior domain knowledge. Within a bayesian framework, causal methods also enhance the interpretability of features, representations, predictions and hypothesis estimated by deep learning models in an uncertain environment (Keohane 2009). Finally, causal features allow us to make structured predictions on the output of deep learning models. The predictions structure is explored by definitions on probability, causality and utility functions reasoning about the time varying uncertain environment (Kleinberg 2012).

Causality Inference in Time Series

In time series analysis, causal inference is the problem of identifying and classifying events in the time series such that some events are causes and some events are effects. Events are identified by mapping logic and structure of natural language to causal graphs and concept lattices. The relationship between causes and effects is determined either deterministically or probabilistically (Keogh et al. 2004), (Rakthanmanon et al. 2012).

Historically, causal reasoning in time series builds on the statistical analysis of covariance or correlation between two events in the time series. The calculated strength of the correlation is then used to predict the causal relation between the two events (Fu 2011). The disadvantage of this approach is that it cannot determine the direction of causation. It also cannot discover hidden causes and patterns for which the two observed events are effects.

Causal structures between two events in the time series can be derived based on one or more of statistical relations, human intervention, prior knowledge and temporal ordering. Statistical relations correspond to correlation information and conditional dependencies used to arrive at associative Causal Markov models on hidden or latent variables (Amblard and Michel 2012). Here the prediction cause and prediction mechanism are modelled as probabilities on events in the data. When the prediction cause is implicit rather than explicit, human intervention allows separation of the cause of prediction from the mechanism of prediction. In this case intervention mechanism affects the events in the causal structure. On the basis of the intervention or choice, an experiment can be conducted to find out the effect of the intervention on prediction outcome. The concept of prior knowledge and temporal ordering allows for intervention mechanisms that are better than trial-and-error learning. In temporal ordering mechanisms, events occurring prior to an event are said to be the cause of that event, but not vice versa. Causal structure learning then identifies significant causes for events that are desirable as distinguished from spurious causes for events that are undesirable. Constructing causal chains allows for time series analysis with temporal ordering. Prior knowledge on the formation of causal structure in given application domain is also useful as a hypothesis learning mechanism.

Granger Causality is a simple learning mechanism that allows us to explore all preceding ideas about causal methods in deep learning and time series (Bahadori and Liu 2013). 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. The insights obtained from granger causality regression can be used to study causal structures and non-linear regression for structural time series analysis (Bahadori and Liu 2012), multivariate time series analysis (Mohammad and Nishida 2010), dependency modelling (Cherkassky and Mulier 2007) and structured prediction (Laxman and Sastry 2006) in temporal data mining.

Causality Inference in Econometrics

Due to computational simplicity, Granger Causality is a popular method for causal inference in econometrics (Geweke 1984), (Gujarati and Porter 1999). Normally, Granger Causality is performed by fitting a vector autoregressive model (VAR) assuming linearity and stationarity to build causal structure on a multivariate time series (Eichler 2013). We follow the same approach in this paper. But, Granger Causality has been enhanced with Granger-causal Markov property to derive the causal structure over temporal causal graphs in econometrics. In this paper, we address the problem of deriving custom loss functions solving classification and optimization problems whereas temporal causal graphs focus on custom probability models solving sampling and optimization problems.

In cross-sectional econometrics, factor analysis and causal inference is used for estimating the impact of counterfactual policies. The focus is on estimating what would happen in the event of a change that may or may not actually happen. Systematic model selection in machine learning thus adds empirical value to econometrics studies (Claeskens, Hjort, and others 2008). However the purpose of such model selection is not simply optimization of the goodness of fit on a test sample. It must also be built for counterfactual predictions when the learning algorithm is operating in an uncertain environment. Such machine learning models may sacrifice predictive performance in the current environment to discover new causal features in a changing environment.

Granger causality has been extensively discussed in econometrics literature (Granger 1969). From the perspective of designing loss functions in deep learning, the related work is on sensitivity analysis over regression errors and model selection over regression residuals. In machine learning, Granger causality has been extended to time frequency methods, motif mining, change points detection, concept drifts mining and temporal data mining (Roddick and Spiliopoulou 2002). In probabilistic graphical models, Granger causality has been extended with markov chains and bayesian methods for non-linear regression (Arnold, Liu, and Abe 2007). In time series analysis, information theory based models in causal networks has also been used to assess Granger causality graphs of stochastic processes (Xu, Farajtabar, and Zha 2016).

We output causal structures from causal inference models alongwith prediction structures from associational inference models in deep learning. The counterfactual policies of econometrics can be used to define new anomaly detection and concept drift algorithms in deep learning model (Masnadi-Shirazi and Vasconcelos 2009).

Algorithm

TO DO : For illustration use histograms on absolute prices and do not give pearson skewness coefficient. Claim that QME captures Left/Right Skew with multiple peaks. For 4 stocks MSFT, T, ABT, CAT, ORCL. Give fstats, tstats and mae values to 3 decimal places precision. Give another table for stock acronymns and another table for t-stats on 4 models defined using MSE, QME over R and UR in granger sense. Use landscape potrait on the big tables. TO DO : To define empirical risk and loss function on target function, use expectations notation of Equation 8.1 in deep learning book. No need to use conditional probability notation. Also comment on the unsupervised learning, feature learning and parameters regularization by varying network structure for matching empirical risk from deep net(with dbn) with expected risk from regression model(with granger causality). To motivate further decomposition of abnormal points, include discussion like Equation 14.3 discussing autoencoders in deep learning book. TO DO : Also add equation and illustration on separating skewed data from normal data with mean and median in the finance time series. Also add features on variance based analysis that separate anomalous points from normal points in the loss function. TO DO : Give equation for empirical risk like Eq 3.3 R_empQ̂ in Wei's paper. Change eq 5 in ling's paper to have percentiles. Discussion on Quadratic mean vs Arithmetic mean as Expected value of squared error in skewed dependent variable with shifting mean. Quadratic mean is found on data partitioned by percentiles of the sorted dependent variable. y-predicted comes from complex nonlinear combination on y(t-i). y-actual is y(t) in the available data. mae are error measures for cumulative squared errors. the time window for y(t-i) is determined by lag set to a constant 100 and corresponding training epochs set to a constant 5. model order selection method gives better lag. TO DO : Change pseudocode to have equations implemented in code and pseudocode. Show plot of median vs mean for skew in the finance stock data plot. Show that median address smaller size partition to account for skewness compared to mean. TO DO Summary : illustration on separating skewed data from normal data, Quadratic mean vs Arithmetic mean in Empirical risk, empirical risk vs expected risk in regression model, input to risk in loss function definition, backpropogation algorithm of the pseudocode,

To predict a given stock price in financial markets, we design a Deep Neural Network (DNN) based regression model that uses temporal information present in multivariate time series. The DNN is assumed to satisfy its objective of univariate regression by operating in an uncertain environment of multivariate time series. The DNN model reduces both regression error and regression risk. The regression error is measured as mean squared error between predicted and actual regression values. It is reduced by granger causality which assesses statistically significant dependencies between regression values. Regression risk is formulated as a new loss function in DNN based on empirical risk. For normal data, we use a Mean Squared Error (MSE) as loss function to assess the model selection. For skewed data, we use a Quadratic Mean Error (QME) as loss function to assess the model selection.

If a particular feature value indicates the occurrence of an event in time series x preceding the occurrence of another event in time series y, then we define the following unrestricted model in Equation 1 and restricted model in Equation 2 to test for granger causality relation between lagged feature values of x and y with lags q and p respectively. $P(y(t)|y(t-i))$ is the conditional probability of predicting current regression value $y(t)$ for stock price y. The univariate prediction model for $y(t)$ is restricted to its past values $y(t-i)$. $P(y(t)|(y(t-i), x(t-j)))$ is the conditional probability of predicting $y(t)$ when bivariate prediction model for $y(t)$ is restricted to past values of not only $y(t-i)$ but also $x(t-j)$. Thus $y(t)$ is the dependent variable and $y(t-i)$, $x(t-j)$ are the independent variables for granger causality test. c_1, c_2 are noise terms(or residual terms) unaccounted for in the regression model error.

$$\begin{aligned} restricted\,model : y(t) =& P(y(t)|y(t-i)) + c_1, \\ & i = 1, 2, ..., p \end{aligned} \tag{1}$$

$$\begin{aligned} unrestricted\,model : y(t) =& P(y(t)|(y(t-i), x(t-j))) + c_2, \\ & i = 1, 2, ..., p\; and\; j = 1, 2, ..., q \end{aligned} \tag{2}$$

We use a DNN model dnn to estimate $P(y(t)|y(t-i))$ in Equation 3 and $P(y(t)|(y(t-i), x(t-j)))$ in Equation 4 parameterized by the regression coefficients tensors α and β representing feature learning layers in the DNN. The model error is evaluated over testing data whereas model fit is evaluated over train data. The prediction objective of the deep learning model is to maximize the model fit while minimizing the model error. Such optimization is achieved by training the DNN with a backpropagation algorithm implementing a stochastic gradient descent procedure. The objective of gradient descent procedure is to maximize α and β over a parameter space A and B that maximizes model fit for both univariate regression and multivariate regression.

$$\begin{aligned} P(y(t)|y(t-i)) = argmax_{\boldsymbol{\alpha} \in A} dnn(y(t-i), \boldsymbol{\alpha}), \\ i = 1, 2, ..., p \end{aligned} \tag{3}$$

$$\begin{aligned} & P(y(t)|y(t-i), x(t-j)) = \\ & \quad argmax_{\boldsymbol{\alpha} \in A, \boldsymbol{\beta} \in B} dnn(y(t-i), \boldsymbol{\alpha}; x(t-j), \boldsymbol{\beta}), \\ & \quad i = 1, 2, ..., p\; and\; j = 1, 2, ..., q \end{aligned} \tag{4}$$

The Squared Error (SE) Loss function l in Equation 5 is defined to be the DNN's loss layer relating the complex nonlinear features learnt on $y(t-i)$ and $x(t-j)$ to the predicted regression value $y(t)$ at time t. l uses a squared error function to matche the predicted regression value $y(t)$ with actual regression value $\hat{y}(t)$. The DNN is trained to minimize l for every training data point with actual regression value $\hat{y}(t)$. $P(y(t)|y(t-i))$ and $P(y(t)|y(t-i), x(t-j))$ predicting regression value $y(t)$ are inturn learnt by the DNN's chosen network structure.

$$SE\;Loss : l(y(t), \hat{y}(t)) = (y(t) - \hat{y}(t))^2 \tag{5}$$

On normal training data not exhibiting statistical skew in the actual regression value $\hat{y}(t)$, the DNN has following definition for Empirical Risk called Mean Squared Error (MSE) Loss in Equation 6 extending l in Equation 5 for SE loss.

$$MSE\;Loss : L(y(t), \hat{y}(t)) = \frac{\Sigma_{t=1}^{n} l(y(t), \hat{y}(t))}{n} \tag{6}$$

On training data exhibiting statistical skew in the actual regression value $\hat{y}(t)$, the DNN has following definition for Empirical Risk called Quadratic Mean Error (QME) Loss in Equation 7. It extends l in Equation 5 by partitioning the skewed regression variable $\hat{y}(t)$ into partitions p_k between $k-th\,percentile\;\hat{y}_{(k)}$ and $(k+10)-th\,percentile\;\hat{y}_{(k+10)}$.

$$QME\;Loss : L(y(t), \hat{y}(t)) = \sqrt{\frac{\Sigma_{k=1}^{10} (\frac{\Sigma_{m=1}^{|p_k|} l(y(t), \hat{y}(t))}{|p_k|})^2}{10}} \tag{7}$$

$$\left\{ \{\hat{y}(t)\} \in p_k \mid \hat{y}_{(k)} < \hat{y}(t) < \hat{y}_{(k+10)} \right\}, \quad k = 0, 10, 20, ..., 90 \tag{8}$$

Thus, the regression model computes empirical risk function in terms of the loss functions for deep learning and supervised learning. Deep learning models over time points allow us to design non-gaussian features (Rasmussen and Williams 2006) and non-linear models (Marinazzo, Pellicoro, and Stramaglia 2008) for causality inference. For various orders of the regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, a null hypothesis and alternative hypothesis is then formulated on the regression outputs $y(t)$. The regression coefficients are estimated with a deep learning model solving for the regression problem relating model output $y(t)$ with model input $y(t-i)$ and $x(t-j)$. The lags p, q determine only the dimensionality of the DNN's input time window but not its model selection. A F-test in Definition 1 below determines granger causality relation between x, y where Squared Error SE is computed for unrestricted regression as SE_{ur} and restricted regression model as SE_r over both MSE Loss and QME Loss in deep learning model.

Definition 1 $F\text{-}statistic = \frac{SE_r - SE_{ur}}{SE_{ur}}$

In the deep learning model, model error is obtained from squared errors relating output $y(t)$ with input $y(t-i)$ and $x(t-j)$ on testing data. The deep network structure for feature learning on x determines parameter tensors α and β on training data. The F-test is repeated for every pair of stock prices consisting of dependent variable $y(t)$ and dependent variables $y(t-i)$ and $x(t-j)$ on time points in the input multivariate time series $Z_t^i, i \in [1, N], t \in T$ with N time points being investigated for causal structure learning.

The deep neural model does supervised learning with both discrete feature values and continuous time points in multivariate time series. A sample of regression predictions are generated for every pair of stock prices in the multivariate time series. The null hypothesis states that the sample means of predictions are equal and the regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are zero in the unrestricted model. The alternative hypothesis states that there is significant variation between sample means of predictions with some non-zero $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The null hypothesis is rejected if p-value on F-test over model error has significance level of atleast 0.05. If the null hypothesis is rejected, we say that independent feature $x(t)$ granger causes dependent feature $y(t)$.

Algorithm 1 gives the training algorithm implementing Equation 1 to Equation 7. TO DO : Give line by line algorithm description for the backpropogation of errors in the algorithm iteration.

TO DO : For algorithm, use xj and xi inside the loop. Give current comments as descriptive comments on the equations. Remove discussion on regression residuals from algorithm section but keep in related work section. Instead of regression residuals, discuss training error, testing error and validation error for model selection. Also model validation is future work. - cover training iteration in the backpropogation algorithm of the pseudocode. - granger causality component of the qme regression code can be reused on the mse regression code

TO DO : Refer equations in description the algorithm pseudocode.

TO DO : L in the regression is SE or MSE. The MSE in current paper is AME. Define both AME and QME in algorithm description. Write down the theories of QME after reading the paper carefully.

TO DO : The QME illustration is on price values where p and q are features. Plot the y(t) for the stock in finance data and show skew points. Shift of skew points is different from mean in the stock. The anomaly in this paper is the skewed points. QME is measure of difference betwen y(t) and predicted y(t) when predicted y(t) is skewed.

Algorithm 1 Back Propagation Algorithm for Deep Network Regression

```
Training Input:
Multivariate time series : $\{Z_t^i\}, i \in [1, N], t \in T$
Training Output:
Predictive model output : Causal structure graphs $G_{MSE}, G_{QME}$; Prediction errors $SE_r, SE_{ur}$;
$SE_r = \Phi, SE_{ur} = \Phi, G_{MSE} = \Phi, G_{QME} = \Phi$
for $Y_t \in \{Z_t^i\}$ do
    for $X_t \in \{\{Z_t^i\} - Y_t\}$ do
        Begin
            $\hat{y}_{train}(t) = 0.7Y_t, \hat{y}_{test}(t) = 0.3Y_t$
            $\hat{x}_{train}(t) = 0.7X_t, \hat{x}_{test}(t) = 0.3X_t$
  ▷ Create cross-validatation data with a percentage split on every time series pair
            $y(t-i) = \hat{y}_{train}(t-i), \hat{y}(t) = \hat{y}_{train}(t)$
            $x(t-j) = \hat{x}_{train}(t-j), \hat{x}(t) = \hat{y}_{train}(t)$
    ▷ Past data from independent variables is used to predict the future values of dependent variable
            Construct restricted and unrestricted model $y(t)$ from Equation 1 and Equation 2 respectively.
            Construct MSE loss $L_{MSE}(y(t), \hat{y}(t))$ and QME loss $L_{QME}(y(t), \hat{y}(t))$ from Equation 6 and Equation 7 respectively.
        $y(t-i) = \hat{y}_{test}(t-i), \hat{y}(t) = \hat{y}_{test}(t)$
            $x(t-j) = \hat{x}_{test}(t-j)$
            From testing loss, update the testing error $SE_r$ and $SE_{ur}$ for restricted and unrestricted models
            Compute $F\text{-}statistic$ from Definition 8

            If $F\text{-}statistic > 0.05$ then
                Begin

                    If model is restricted then
                        $SE_r[Y_t][X_t] = SE_r$
                    else
                        $SE_{ur}[Y_t][X_t] = SE_{ur}$
                                                        ▷ Update Regression Errors

                    If risk is MSE then
                        $G_{MSE}[Y_t] = G_{MSE}[Y_t] \cup X_t \to Y_t$
                    else
                        $G_{QME}[Y_t] = G_{QME}[Y_t] \cup X_t \to Y_t$
                                                        ▷ Update Regression Losses
                End
        End
    return $SE_r, SE_{ur}, G_{MSE}, G_{QME}$
```

Experiments

Table 1: MSE Model's Mean Absolute Errors for Granger Causality

	AAPL		ABT		AEM		AFG		APA		CAT		LAKE		MCD		MSFT		ORCL		SUN		T		UTX		WWD	
	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR
AAPL	0.518	na	0.636	1.073	0.501	1.172	0.539	0.842	0.475	1.154	0.303	0.941	0.612	1.311	0.446	0.885	0.548	1.268	0.455	0.723	0.571	1.396	0.603	1.451	0.533	0.953	0.518	0.712
ABT	0.452	0.259	0.442	na	0.458	0.439	0.487	0.161	0.437	0.439	0.412	0.257	0.383	0.384	0.452	0.151	0.464	0.313	0.434	0.151	0.458	0.431	0.441	0.428	0.444	0.141	0.436	0.163
AEM	0.191	1.053	0.204	0.706	0.192	na	0.218	0.784	0.169	0.354	0.105	0.526	0.201	0.717	0.166	0.652	0.191	0.791	0.292	0.618	0.185	0.835	0.217	0.887	0.231	0.604	0.138	0.664
AFG	0.387	0.298	0.361	0.447	0.372	0.675	0.397	na	0.409	0.614	0.351	0.413	0.451	0.607	0.483	0.355	0.315	0.427	0.441	0.273	0.432	0.649	0.344	0.679	0.443	0.331	0.373	0.295
APA	0.091	0.918	0.136	0.672	0.145	0.311	0.093	0.756	0.096	na	0.111	0.421	0.101	0.631	0.079	0.588	0.076	0.772	0.083	0.572	0.081	0.663	0.095	0.781	0.071	0.521	0.091	0.569
CAT	0.185	0.685	0.137	0.387	0.073	0.463	0.141	0.436	0.056	0.348	0.102	na	0.138	0.511	0.112	0.319	0.135	0.647	0.111	0.361	0.058	0.604	0.066	0.679	0.057	0.282	0.058	0.331
LAKE	0.105	0.984	0.094	0.431	0.106	0.576	0.132	0.548	0.094	0.478	0.121	0.413	0.107	na	0.088	0.505	0.137	0.399	0.109	0.439	0.105	0.403	0.088	0.462	0.107	0.438	0.114	0.582
MCD	0.391	0.378	0.405	0.391	0.363	0.594	0.397	0.275	0.341	0.521	0.329	0.406	0.423	0.669	0.398	na	0.402	0.585	0.386	0.377	0.525	0.758	0.395	0.697	0.378	0.276	0.448	0.172
MSFT	0.204	0.266	0.251	0.211	0.226	0.451	0.231	0.146	0.251	0.431	0.218	0.275	0.235	0.328	0.233	0.196	0.231	na	0.282	0.201	0.200	0.438	0.241	0.258	0.225	0.205	0.211	0.159
ORCL	0.103	0.523	0.221	0.194	0.151	0.559	0.133	0.222	0.164	0.498	0.152	0.275	0.205	0.561	0.209	0.139	0.175	0.433	0.155	na	0.149	0.664	0.133	0.508	0.119	0.188	0.103	0.184
SUN	0.185	1.180	0.203	0.585	0.190	0.793	0.171	0.700	0.211	0.684	0.153	0.621	0.185	0.499	0.152	0.714	0.196	0.697	0.222	0.660	0.182	na	0.152	0.694	0.177	0.600	0.180	0.681
T	0.054	0.348	0.046	0.183	0.114	0.208	0.073	0.172	0.065	0.162	0.051	0.194	0.036	0.166	0.054	0.137	0.067	0.148	0.066	0.117	0.063	0.175	0.064	na	0.094	0.147	0.053	0.121
UTX	0.315	0.568	0.287	0.292	0.328	0.494	0.218	0.296	0.250	0.420	0.258	0.230	0.254	0.510	0.307	0.221	0.360	0.564	0.291	0.326	0.304	0.593	0.301	0.601	0.286	na	0.252	0.206
WWD	0.209	0.402	0.401	0.590	0.202	0.651	0.238	0.437	0.148	0.587	0.364	0.421	0.241	0.813	0.280	0.427	0.218	0.834	0.252	0.534	0.340	0.878	0.336	0.981	0.243	0.426	0.267	na
t-statistics	5.91e-04		1.61e-02		3.38e-04		1.83e-02		1.17e-03																			

Table 2: QME Model's Mean Absolute Errors for Granger Causality

	AAPL		ABT		AEM		AFG		APA		CAT		LAKE		MCD		MSFT		ORCL		SUN		T		UTX		WWD	
	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR	R	UR
AAPL	0.571	na	0.517	0.957	0.324	1.145	0.522	0.900	0.607	1.149	0.463	0.978	0.588	1.283	0.730	0.784	0.719	1.323	0.662	0.857	0.615	1.368	0.624	1.409	0.473	1.013	0.584	0.773
ABT	0.476	0.196	0.5	na	0.515	0.401	0.489	0.222	0.511	0.366	0.476	0.255	0.548	0.406	0.494	0.341	0.503	0.309	0.497	0.249	0.473	0.433	0.535	0.351	0.515	0.211	0.468	0.224
AEM	0.116	0.827	0.199	0.814	0.144	na	0.167	0.763	0.165	0.605	0.118	0.564	0.114	0.685	0.204	0.621	0.178	0.819	0.112	0.631	0.128	0.782	0.129	1.031	0.138	0.622	0.116	0.614
AFG	0.607	0.265	0.468	0.431	0.497	0.626	0.538	na	0.521	0.571	0.495	0.397	0.472	0.597	0.591	0.319	0.538	0.582	0.486	0.335	0.535	0.657	0.562	0.579	0.546	0.345	0.676	0.286
APA	0.101	0.768	0.092	0.681	0.128	0.383	0.133	0.726	0.119	na	0.135	0.481	0.111	0.639	0.143	0.574	0.119	0.867	0.111	0.591	0.128	0.632	0.141	0.821	0.112	0.559	0.105	0.584
CAT	0.319	0.475	0.261	0.489	0.474	0.441	0.251	0.434	0.194	0.443	0.288	na	0.226	0.528	0.287	0.374	0.306	0.716	0.201	0.429	0.363	0.555	0.291	0.756	0.231	0.391	0.345	0.321
LAKE	0.122	0.601	0.131	0.435	0.222	0.438	0.114	0.464	0.111	0.409	0.133	0.401	0.128	na	0.113	0.426	0.115	0.478	0.119	0.409	0.117	0.324	0.115	0.459	0.116	0.389	0.143	0.405
MCD	0.305	0.289	0.321	0.521	0.486	0.516	0.379	0.461	0.418	0.497	0.425	0.475	0.412	0.654	na	0.416	0.365	0.606	0.453	0.205	0.519	0.731	0.515	0.621	0.554	0.381	0.263	0.255
MSFT	0.177	0.183	0.206	0.203	0.141	0.399	0.146	0.181	0.134	0.346	0.123	0.276	0.151	0.317	0.195	0.219	0.145	na	0.117	0.222	0.129	0.351	0.126	0.269	0.111	0.237	0.132	0.157
ORCL	0.201	0.217	0.553	0.274	0.533	0.476	0.349	0.269	0.349	0.446	0.454	0.315	0.364	0.557	0.476	0.191	0.404	0.403	0.401	na	0.311	0.607	0.371	0.593	0.273	0.272	0.566	0.237
SUN	0.179	1.058	0.193	0.572	0.174	0.826	0.308	0.726	0.166	0.657	0.171	0.609	0.185	0.532	0.181	0.812	0.168	0.722	0.178	0.631	0.191	na	0.189	0.671	0.188	0.628	0.206	0.711
T	0.071	0.301	0.075	0.201	0.087	0.137	0.101	0.191	0.055	0.198	0.083	0.215	0.039	0.151	0.062	0.223	0.038	0.111	0.063	0.151	0.089	0.178	0.074	na	0.041	0.226	0.161	0.131
UTX	0.296	0.351	0.294	0.556	0.396	0.441	0.335	0.609	0.361	0.381	0.449	0.391	0.446	0.577	0.195	0.376	0.319	0.634	0.311	0.317	0.266	0.566	0.399	0.619	0.331	na	0.234	0.224
WWD	0.283	0.304	0.345	0.661	0.418	0.618	0.502	0.556	0.312	0.586	0.264	0.551	0.458	0.838	0.233	0.604	0.408	1.022	0.411	0.501	0.301	0.866	0.268	0.969	0.451	0.521	0.358	na
t-statistics	2.16e-04		2.16e-04		2.16e-04		2.16e-04		2.16e-04		2.16e-04		2.16e-04		2.16e-04		2.16e-04		2.16e-04		2.16e-04		2.16e-04		2.16e-04		2.16e-04	

We use granger causality test to assess for statistically significant dependencies between deep learning features. In the experiments we train the deep network with both MSE and QME loss functions on the network structure learning β_i and α_j. Moreover, the causal features that minimize model error are selected by testing the deep learning model on testing data consisting of causal features.

TO DO : Fine tune the unrestricted model for minimizing errors and maximizing causes. Divide by mean to normalize all the stocks. If required try more normalization. Select the stocks suitable for analysis. Increase training epochs in steps of 5 from 5 to 50. Increase DNN layers from 3 to 9 in steps of 2. Increase lags in steps of 100 from 100 to 500. Must keep p and q to the same value. Need to do missing value imputation before increasing lag size as we collect data from the past. Plot number of causes on y-axis. Make dataframes of changes in errors.

TO DO : After completing remaining draft of the paper, Make plots of tables and networks from the dataframe of QMEErrors, QMECauses and MSEErrors, MSECauses. Show three experiments output (i) MSE Causes and reduction in unrestricted model. QME Causes and reduction in unrestricted model (ii) QME Causes vs MSE Causes (iii) Least error of QME Causes vs MSE Causes (iv) Space permitting add comparision with quantile regression. No need to experiment with time lag and time windows in QME regression. (v) Also add graphs and tables varying the restricted model tuning parameters(for minimizing errors and maximizing causes) like Increase training epochs, Increase lags, Increase layers in DNN. Plot number of causes on y-axis. Make dataframes of changes in errors. After graphs and tables are available, put in PAKDD format and write abstract. (vi) If required, select the stocks suitable for analysis. TO DO : Show partial graph visualization showing extra causal causes for same stock price with and without mae from granger regression. Show best decrease in error in terms of training parameters : lag, epochs. TO DO : Show one table of mae errors for mse and another table for qme with and without effects of causal stocks. Unrestricted model should lead to decrease in error. Each row of table is stock and each coloumn is causal stocks with corresponding errors. Also discuss additional causes found from QME as compared to MSE as a partial graph. Also add a table of regression errors and corresponding f-test p-values. TO DO : Cross validation performance : Give custom cnn architecture diagram. Give links to keras site. The code must have three parts, namely, data loading, network structure, training and testing the computation graph. Dropout, normalization layers can be used. Feature learning with dense layers. Applying the custom loss function for every available batch in the qme loss. Percent split before training on past and testing on future prices. raw yahoo finance data description where Long term data is available only with daily data granularity. A batch of training and testing data are given for every iteration of the backpropagation training algorithm. The iteration must be executed in the tensorflow session environment with data input feeding train step and loss step. The testing data is fed to the predict step. Additional conditions on testing data that are related to granger causality. Regression model can be treated as feature based model like dnns, daes, cnns. A standard dnn must be used in the granger causality test. The same dnn must be used for both auto-regression and granger-regression. TO DO : Compute the error from regression model. test rejects the null hypothesis at the 5% significance level. Later, multiple data sources and baseline models(with lags, epochs, features, predictions in models) can be used to input the causal features addressing data skew in the network. Feature engineering on the causal model would involve further optimizations in deep learning features. Give max lag and more data size over more training epochs to get better distributed representations TO DO : Compare QME regression with Quantile Regression? Do comparision only for univariate prediction for benchmarking purposes if there is space. TO DO : Complete experiments asap after completing the coding tutorials on github and youtube. Add one more co-author once paper draft is ready. Acknowledge features contribution in the baseline model from the HDU students. Baseline Model - Multivariate Regression : Generate synthetic data and multivariate stock data. - Multivariate Classification : Data Sample, Classification Labels, Feature Construction, Feature Extraction - Concept drift, Anomaly detection and Imbalanced Classification : Variance analysis to get data labels and feature extraction to get classification models. - Also check to see if google finance data can be obtained at varying levels of granularity. Exploring LOBSTER data and SIRCA data. At present we can use both volume and price for getting features and labels of the data. - improve on mse regression (baseline) with qm regression on skewed data - show granger causality features for regression and classification(in journal extension). Experiment with different time lags to find best time lag in granger causality. - show improvment on baseline model with constructed, extracted and inferred features in deep learning loss function. - Discuss the data flow graph model visualization with points on data loading, network structure, training and testing, loss optimization visualization as given in mail. Also discuss experiments on feature construction on variance analysis, feature extraction on causal analysis and feature inference on deepnet analysis. - In model implementation, use tf.contrib api before using tf api(if required). Also check best feature learning layers in tf.contrib. try modifying contrib dbn on source rbm. if required, the tensorflow models zoo implementations can also be seen. The base line model for improving prediction is mse regression. The improved model improving prediction is qme regression. - In data schema get different types of stocks and make hypothesis in domain that are validated with features on time points. Need to include the variance based features in baseline model to motivate the anomalous time points separation from normal time points in the loss function. - network is trained and tested on both synthetic data and finance data. give diagram for network structure. - table of regression errors and corresponding f-test p-values must be given. check output on ready tensorflow code for MSE regression after implementing the QME regression. - if possible, use multiple finance data sources.

Conclusion and Future Work

We presented a deep learning model augmented with causal features for analyzing multivariate time series data sourced from the finance domain. The causal features are able to improve the model performance. The output causal graphs are useful to capture temporal dependence in skewed data. Our custom loss function can be used to uncover hidden granger causes and custom probability distributions in the data. Thus we facilitate features interpretation and human intervention to modify local probability distributions and causal graph neighbourhoods in causal structure learning.

We apply empirical risk in deep learning to a multivariate regression. As future work we shall combine multiple data sources to extract regularized features for big data mining (Chaudhry, Xu, and Gu 2017). The sensitivity analysis of loss function and model selection can be improved with features and patterns constructed on the application domain (Ye and Keogh 2009). Additionally, the custom loss function can be evaluated against a validation datasets and parametric models on higher order statistics useful for model validation.

Concept-adapting data structures, time frequency analysis methods and generative models can be used for feature engineering on errors and model selection on noise (Han, Pei, and Kamber 2011) in the temporal regression model. The proposed loss function can be extended for nonlinear algorithm-oriented approaches to robust regression. Change points detection and concept drift mining with kernel machines (Ristanoski, Liu, and Bailey 2013) can use our loss function for temporal data mining, empirical inference and learning generalisation.

References

Amblard, P.-O., and Michel, O. J. 2012. The relation between granger causality and directed information theory: A review. *Entropy* 15(1):113–143.

Arnold, A.; Liu, Y.; and Abe, N. 2007. Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 66–75. ACM.

Bahadori, M. T., and Liu, Y. 2012. Granger causality analysis in irregular time series. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, 660–671. SIAM.

Bahadori, M. T., and Liu, Y. 2013. An examination of practical granger causality inference. In *Proceedings of the 2013 SIAM International Conference on data Mining*, 467–475. SIAM.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. springer.

Chaudhry, A.; Xu, P.; and Gu, Q. 2017. Uncertainty assessment and false discovery rate control in high-dimensional granger causal inference. In *International Conference on Machine Learning*, 684–693.

Cherkassky, V., and Mulier, F. M. 2007. *Learning from data: concepts, theory, and methods*. John Wiley & Sons.

Claeskens, G.; Hjort, N. L.; et al. 2008. *Model selection and model averaging*, volume 330. Cambridge University Press Cambridge.

Eichler, M. 2013. Causal inference with multiple time series: principles and problems. *Phil. Trans. R. Soc. A* 371(1997):20110613.

Fu, T.-c. 2011. A review on time series data mining. *Engineering Applications of Artificial Intelligence* 24(1):164–181.

Geweke, J. 1984. Inference and causality in economic time series models. *Handbook of econometrics* 2:1101–1144.

Granger, C. W. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* 424–438.

Gujarati, D. N., and Porter, D. C. 1999. Essentials of econometrics.

Han, J.; Pei, J.; and Kamber, M. 2011. *Data mining: concepts and techniques*. Elsevier.

Keogh, E.; Chu, S.; Hart, D.; and Pazzani, M. 2004. Segmenting time series: A survey and novel approach. *Data mining in time series databases* 57:1–22.

Keohane, R. O. 2009. Counterfactuals and causal inference: Methods and principles for social research. *Social Forces* 88(1):466–467.

Kleinberg, S. 2012. *Causality, probability, and time*. Cambridge University Press.

Laxman, S., and Sastry, P. S. 2006. A survey of temporal data mining. *Sadhana* 31(2):173–198.

Lipton, Z. C. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Liu, W., and Chawla, S. 2011. A quadratic mean based supervised learning model for managing data skewness. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, 188–198. SIAM.

Liu, Q.; Jiang, H.; Evdokimov, A.; Ling, Z.-H.; Zhu, X.; Wei, S.; and Hu, Y. 2016. Probabilistic reasoning via deep learning: Neural association models. *arXiv preprint arXiv:1603.07704*.

Lopez-Paz, D.; Nishihara, R.; Chintala, S.; Schölkopf, B.; and Bottou, L. 2016. Discovering causal signals in images. *arXiv preprint arXiv:1605.08179*.

Luo, L.; Liu, W.; Koprinska, I.; and Chen, F. 2015. Discovering causal structures from time series data via enhanced granger causality. In *Australasian Joint Conference on Artificial Intelligence*, 365–378. Springer.

Marinazzo, D.; Pellicoro, M.; and Stramaglia, S. 2008. Kernel method for nonlinear granger causality. *Physical Review Letters* 100(14):144103.

Masnadi-Shirazi, H., and Vasconcelos, N. 2009. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in neural information processing systems*, 1049–1056.

Mirowski, P.; Ranzato, M.; and LeCun, Y. 2010. Dynamic auto-encoders for semantic indexing. In *Proceedings of the NIPS 2010 Workshop on Deep Learning*, 1–9.

Mohammad, Y. F., and Nishida, T. 2010. Mining causal relationships in multidimensional time series. *Smart information and knowledge management* 260:309–338.

Pearl, J., et al. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3:96–146.

Rakthanmanon, T.; Campana, B.; Mueen, A.; Batista, G.; Westover, B.; Zhu, Q.; Zakaria, J.; and Keogh, E. 2012. Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 262–270. ACM.

Rasmussen, C. E., and Williams, C. K. 2006. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge.

Ristanoski, G.; Liu, W.; and Bailey, J. 2013. A time-dependent enhanced support vector machine for time series regression. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 946–954. ACM.

Roddick, J. F., and Spiliopoulou, M. 2002. A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and data engineering* 14(4):750–767.

Spirtes, P.; Glymour, C. N.; and Scheines, R. 2000. *Causation, prediction, and search*. MIT press.

Xu, H.; Farajtabar, M.; and Zha, H. 2016. Learning granger causality for hawkes processes. In *International Conference on Machine Learning*, 1717–1726.

Ye, L., and Keogh, E. 2009. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 947–956. ACM.

Zanin, M.; Papo, D.; Sousa, P. A.; Menasalvas, E.; Nicchi, A.; Kubik, E.; and Boccaletti, S. 2016. Combining complex networks and data mining: why and how. *Physics Reports* 635:1–44.