Communication Theoretic Prediction on Networked Data

Tzu-Yu Chuang*, Jia-Pei Lu† and Kwang-Cheng Chen†, *Fellow, IEEE*,
Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan
Email : {d99942022* , r01942050† , ckc‡}@ntu.edu.tw

***Abstract*—Prediction based on observed data is one of the major purposes in (big) data analytics, and has shown great impacts in many applications, including engineering, social science, and medical treatments. Statistical machine learning has been widely adopted to deal with such problem. In this paper, we analogize the relationship among data variables as a sort of generalized social network [1], that is, networked data. Consequently, a direct causal relationship from one data variable to another is thus equivalent to information transfer over a communication channel. Prediction based on data variables is consequently to maximize utilizations of information conveyed over communication channels. Therefore, we introduce the concept of adaptive equalization to data analytics in this paper, which allows us to select appropriate data variables and optimum depth of observations for prediction. We illustrate by finance market data to show surprisingly good performance using this simple methodology. This result not only indicates a new direction to knowledge discovery and inference in big networked data analytics based on communication theory, but also shows the consistency with the newly developed information coupling.**

***Index Terms*—Statistical communication theory, networked data, data analytics, communication channel, equalizer, receiver diversity, knowledge discovery.**

I. Introduction

DATA nalysis is an emerging technology involving statistics, optimization, and computer science; one of the most promising approaches of data modeling is the *networked data*, which has a close relationship with social networks and thus communication networks. The concept of networked data, *i.e.*, linking data together to form a network by their mutual relationship, is equivalent to the model of *social networks*. It has been addressed that the general structure of social networks is equivalent to the communication networks [1]. Therefore, one can expect that the signal processing and communication theory can bring us more insights and opportunities for networked data analysis.

The key to relate data model and social/communication networks is the *channel*. In modeling of data, each attribute (feature) is represented by a random variable, then embedded into a network by a node. An arbitrary pair of variables (nodes) are linked together via their statistical dependency. Furthermore, we can extend each random variable into a stochastic process by considering their evolution and dynamics in time. Such relationship is consistent with the definition of *communication channel* [2], *i.e.*, a sequence of conditional probabilities. Therefore, a correspondence from data analysis to statistical communication theory can be established, and many fundamental limits can be understood and explained from the knowledge of communication networks.

In this paper, we address networked data prediction in the viewpoint of communication theory. A sequence of observed data point (observations) is given, and the objective is to predict a future outcome of a related variable (target). This task can be regarded as an *information transfer* over the communication channel with the target and observations taken as the information transmitter and the information receiver, respectively. The information transfer is surely under noisy distortion, and it is well known to use equalization to optimize the reception. Therefore, we are motivated to explore the data prediction optimized by the *adaptive equalization*, which is similar to the linear estimation, but allows the adaptation to the time-varying circumstance. Moreover, the multi-source prediction is equivalent to the *single-input multiple-output* (SIMO) communications, in which the receiver diversity techniques can be directly applied. The idea proposed in this paper is verified by an experiment of prediction of stock prices in the financial market. The result not only motivates us a novel viewpoint on the *knowledge discovery* [3], but also shows a consistency with the information coupling from the frontier of information theory [4].

There are a lot of literature touched the networked data analysis from different aspects. The idea of information transfer can be traced back in [5] and related works of the analysis in time series. Authors in [6] proposed the using of compressed sensing to process relational data. Besides, the constrained learning and estimation in networks were addressed in [7], and the decision and inference of networked data via local network topology was considered in [8]. Also, there were recent works discussing the signal processing over graphs on big data [9], and the knowledge discovery in databases [10]. On the shoulder of these excellent researches of networked data, we first relate the data analysis with social and communication networks, and provide a novel insight on data processing.

In this paper, we model a general data analysis by the communication channel first, and then the fundamental knowledge of liner predictors is provided through information transfer. The techniques of receiver diversity is introduced later to solve the multi-source prediction. We also present a selection criterion based on the information transfer. Finally, an experiment of stock price prediction is demonstrated and discussed.

This research was supported in part by the US Air Force under the grant AOARD-14-4053 and by the Ministry of Science and Technology under the grant MOST-103-2221-E-002-022-MY3.

II. Problem Formulation

A novel explanation of data prediction in the viewpoint of communication theory is provided in this section. Our work is based on the observation that large datasets have a distributional basis; *i.e.*, there exists an underlying (sometime implicit) statistical model for the data. We formulate the data prediction as the following.

A. Model for Data

Our proposed model for data focuses on data generated from a large system composed of many interacting units. Each unit is capable of generating data continuously or sporadically with time. Therefore, the dataset $\mathcal{D}$ generated by such system is modeled by a family of stochastic processes

$$\mathcal{D} \triangleq \left\{ \{X_{(i),t}\}_{t\in I} : i = 1, 2, \ldots \right\}, \tag{1}$$

where I is a proper index set corresponding to the scenario we are interested in. In this paper, we only consider the case of discrete time; *i.e.*, $I = \mathbb{N}$. We write $\{X_{(i),t}\}$ as $\{X_{(i),t}\}_{t\in\mathbb{N}}$ for simplicity. Furthermore, we use $[X_i]_a^b = [X_{(i),a}, X_{(i),a+1}, \ldots, X_{(i),b}]$ to denote an sub-sequence of stochastic process $\{X_{(i),t}\}$. In the following, we start from a simplified case to give a glimpse of communication theoretic prediction of networked data.

B. Data Analysis via Communication Channels

Consider $\mathcal{D} = \{\{X_t\}, \{Y_t\}\}$ only. Given a time instance n, the problem of prediction is stated as follow:

Problem 1. *Suppose we have two random sequences, $\{X_t\}$, $\{Y_t\}$. We observe $\{X_t\}$ in some set of time $n-L \le t \le n-1$ and we wish to estimate Y_n from these observations.*

The relationship between Y_n and $[X]_{n-L}^{n-1}$ is linked by the conditional probability $\mathbb{P}\{[X]_{n-L}^{n-1} | Y_n\}$. This dependency can be considered as a *channel*, which links the *information transmitter*, Y_n, and the *information receiver*, $[X]_{n-L}^{n-1}$. Therefore, the data prediction is a generalized signal detection with receiving a realization of $[X]_{n-L}^{n-1}$. The optimal predictor of Y_n, $\hat{Y}_n([X]_{n-L}^{n-1})$, can be derived by minimizing the measure of error. The most common one is in the mean-square error sense

$$\mathbb{E}\left\{ \left[Y_n - \hat{Y}_n\left([X]_{n-L}^{n-1}\right) \right]^2 \right\} \tag{2}$$

With this model, one can immediately identified that, the **Problem 1** is equivalent to the signal detection over a highly distorted channel, such as a *frequency selective channel*. The receiver can access to the information transmitted by Y_n up to the depth (delay in the communication, equivalently) of observation up to L, *i.e.*, $\{X\}_{n-L}^{n-1}$. The goal of signal detection is to find an optimum estimate of Y_n based on the reception of $[X]_{n-L}^{n-1}$.

It is well-known that the minimum value of (2), referred to as the minimum mean-square error or MMSE, is achieved by the conditional mean estimator:

$$\hat{Y}_n\left([X]_{n-L}^{n-1}\right) = \mathbb{E}\left\{ Y_n \mid [X]_{n-L}^{n-1} \right\}. \tag{3}$$

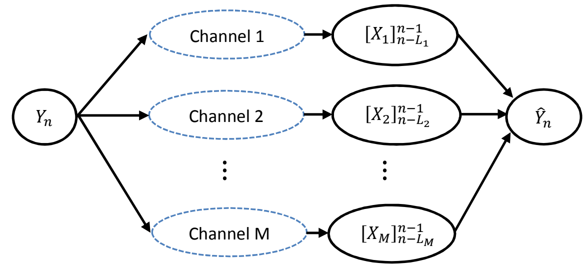

Fig. 1. A general model of prediction. The target is a single Information Transmitter Y_n and the observations is the multiple Information Receiver $[X]_{n-L}^{n-1}$. The statistical dependency of each pair is represented by the channel, the conditional probability $\mathbb{P}\{[X_i]_{n-L_1}^{n-1} | Y_n\}$ This is also the general model of SIMO communications.

Upon the observation of the receiver $[X]_{n-L}^{n-1}$, one would like to infer the information bearing transmitter Y_n. The *mutual information* between $[X]_{n-L}^{n-1}$ and Y_n is:

$$\mathcal{I}_Y(X) \triangleq I([X]_{n-L}^{n-1}; Y_n) = \mathbb{E}\left\{ \log \frac{p_{Y_n|[X]_{n-L}^{n-1}}(Y_n \mid [X]_{n-L}^{n-1})}{p_{Y_n}(Y_n)} \right\}$$

Or, equivalently, we call this mutual information the *information transfer* to emphasize the idea of modeling data prediction as a communication of information over the channel.

More generally, we can consider the data prediction with more than one source; that is, the prediction of Y_n with observation set $\left\{[X_i]_{n-L_i}^{n-1} : i = 1, 2 \ldots, M\right\}$. Since each pair of random sequences $(\{X_i\}, \{Y\})$ forms a channel, and thus a signal detection problem with single-input and multiple-output, or SIMO, can be formulated to this problem. (see Fig. 1) Under this network, the total information transfer is measured by

$$\begin{aligned} &\mathcal{I}_Y(X_1, X_2, \ldots, X_M) \\ &\triangleq I\left([X_1]_{n-L_1}^{n-1}, [X_2]_{n-L_2}^{n-2}, \ldots, [X_M]_{n-L_M}^{n-1}; Y_n\right) \end{aligned} \tag{4}$$

The following theorem is useful when we need to measure the information transfer over some specific channels.

Theorem 1. *Given the target and the data variables to be of the linear form; that is, $[X]_{n-L}^{n-1} = \mathbf{h}Y_n + N_n$, where $N_n \sim \mathcal{N}(0, \Sigma_n^2)$. For any Y_n independent of N_n with $\mathsf{Var}\{Y_n\} \le \sigma^2$, we have*

$$I([X]_{n-L}^{n-1}; Y_n) \le \frac{1}{2}\log\left(1 + \sigma^2 \mathbf{h}^\mathsf{T}\Sigma_N^{-1}\mathbf{h}\right) \tag{5}$$

Proof: Let $Y_\mathsf{G} \sim \mathcal{N}(0, \sigma^2)$, $X_\mathsf{G} = \mathbf{h}Y_\mathsf{G} + N_n$. It is very straightforward to show that

$$\begin{aligned} &I\left([X]_{n-L}^{n-1}; Y_n\right) \\ &= D\left(P_{[X]_{n-L}^{n-1}|Y_n} \| P_{Y_\mathsf{G}} \mid P_{[X]_{n-L}^{n-1}}\right) - D\left(P_Y \| P_{Y_\mathsf{G}}\right) \\ &= \mathbb{E}\left\{\imath_{Y_G;X_G}(Y, \mathbf{h}Y + N_n)\right\} - D(P_Y \| P_{Y_G}) \\ &\le \frac{1}{2}\log\left(1 + \sigma^2\mathbf{h}^\mathsf{T}\Sigma_N^{-1}\mathbf{h}\right) \end{aligned} \tag{6}$$

where $D(\cdot\|\cdot)$ is the relative entropy. The equality holds when $Y \sim \mathcal{N}(0, \sigma^2)$ □

In the following section, we will discuss a practical realization of data prediction based on communication theory.

We consider the single channel case first, and then give an extension to the prediction with multiple channels.

III. Data Analysis over One Channel

In order to give a computable estimate, we have to consider a constrained family of estimates. In the following, the linear estimate will be introduced, and then the selection of depth of observation will be discussed.

A. Equalization to Data Prediction

In **Problem 1**, of course, the optimum estimator (in the MMSE sense) is the conditional mean, $\hat{Y}_n = \mathbb{E}\{Y_{n+1}|[X]_{n-L}^{n-1}\}$. However, the computation of such estimator can be quite cumbersome unless the problem exhibits special structure. Furthermore, the determination of the conditional mean generally requires knowledge of the joint distribution of $Y_n, X_{n-L}, \ldots, X_{n-1}$, which may be impractical (or impossible) to obtain in practice.

One way of circumventing these problems is to constrain the signal of some computationally convenient form, and then to minimize the MSE over this constrained class. The *linear constraint* serves the purpose, in which we consider estimates $\hat{Y}_n$ of the form

$$\hat{Y}_n = \sum_{t=n-L}^{n-1} w_{n,t} X_t + c_n, \tag{7}$$

where $\{w_{n,t}\}_{t=n-L}^{n-1}$ and c_n are scalars. This setting is equivalent to the *MMSE Equalizer* in the communication theory [11]; that is, the coefficients $\{w_{n,t}\}$ are chosen to minimize (2). As such, finding the optimal filter coefficients $\{w_{n,t}\}$ becomes a standard problem in linear estimation. In fact, this problem is a standard *Weiner filtering problem*, whose solution is provided by the following proposition [11]:

Proposition 1. *$\hat{Y}_n$ minimizes (2) if and only if*

$$\mathbb{E}\{Y_n\} = \mathbb{E}\{\hat{Y}_n\} \tag{8}$$

and

$$\mathbb{E}\{(Y_n - \hat{Y}_n)X_l\} = 0 \ \forall \ n-L \leq l \leq n-1 \tag{9}$$

Proof: Please refer to [11]. □

Proposition 1 gives conditions that are necessary and sufficient for the set of coefficients $\{w_{n,t}\}$ and c_t to yield an optimum linear estimator of Y_n from $[X]_{n-L}^{n-1}$. An optimal linear estimate will always be of the form

$$\hat{Y}_n = \mathbb{E}\{Y_n\} + \sum_{t=n-L}^{n-1} w_{n,t}(X_t - \mathbb{E}\{X_t\}) \tag{10}$$

and

$$\boldsymbol{\sigma}_{YX}(n) = \Sigma_X \mathbf{w}_n, \tag{11}$$

where $\boldsymbol{\sigma}_{YX}(n) \triangleq [\mathsf{cov}(Y_n, X_{n-L}), \ldots, \mathsf{cov}(Y_n, X_{n-1})]$, and Σ_Y is the covariance matrix of the vector $(X_{n-L}, \ldots, X_{n-1})^{\mathsf{T}}$. Equation (11) is known as the *Wiener-Hopf equation*. Assuming that Σ_X is positive definite,

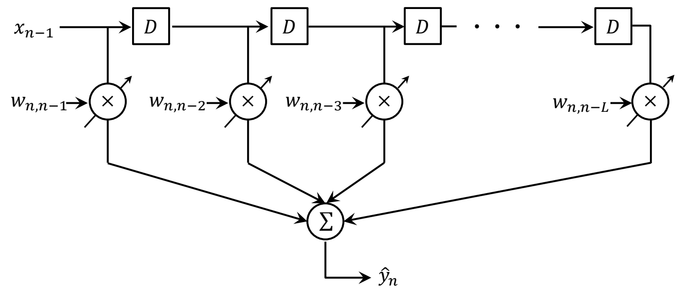

Fig. 2. Communication-inspired Equalizer for data prediction. The equalizer coefficient wet $\{w_{n,t}\}$ can be adaptive adjusted to fit the dynamic signals.

we see that the optimum estimator coefficients are given by

$$\mathbf{w}_n = \Sigma_X^{-1} \boldsymbol{\sigma}_{YX}(n) \tag{12}$$

The realization of this equalizer is depicted in Fig. 2. We can further implement the adaptive equalizer, *i.e.*, adjustable equalizer coefficients, to fit the dynamic environment better. The detail is omitted in this paper, and can be referred in [12].

B. Optimum Selection of Tap Numbers

There is an important difference between actual equalizers in communications and our communication-inspired equalizer for data prediction: the depth (delays) of observations. In the communication system, the tap delay is determined by the nature of fading channel itself, while the depth of observations in data prediction should be determined under some optimality constraints. Furthermore, data usually evolves with circumstance and they do not persist a consistent statistical properties for too long. Therefore, it is of importance to consider a prediction based on a proper depth of observations to avoid either a heavy load on computations or *overfitting* in statistics.

Generally, the larger L is, it suggests more information transfer over the channel. However, since

$$\mathcal{I}_Y(X) = I(X_{n-1}; Y) + \sum_{i=2}^{L} I(X_{n-i}; Y | X_{n-i+1}^{n-1}), \tag{13}$$

it is believed that the latter term is relatively small due to the weak correlation between $X_{n-i}|X_{n-i+1}^{n-1}$ and $Y_n|X_{n-i+1}^{n-1}$. An intuitive selection is constrained on the increasing of information transfer. That is

$$L^* = \min_{L \in \mathbb{N}} \left\{ L : I(X_{n-L}; Y | X_{n-L+1}^{n-1}) < \epsilon \right\}, \tag{14}$$

where ϵ is a predefined threshold. However, this increasing of information transfer is not always computable. Instead, we constrain on the maximum increasing of information transfer:

$$\begin{aligned} L^* &= \min_{L \in \mathbb{N}} \left\{ L : \max I(X_{n-L}; Y | X_{n-L+1}^{n-1}) < \epsilon' \right\} \\ &= \min_{L \in \mathbb{N}} \left\{ L : \frac{1}{2} \log \left(1 + \frac{\sigma^2_{Y|X_{n-L+1}^{n-1}}}{\sigma_n^2} \right) < \epsilon', \right\} \end{aligned} \tag{15}$$

where the analytical form is derived under the linear model with the power spectrum density of noise σ_n^2 and the conditional variance $\sigma^2_{Y|X_{n-L+1}^{n-1}}$. This is a direct result of **Theorem 1**, we omit the derivation here.

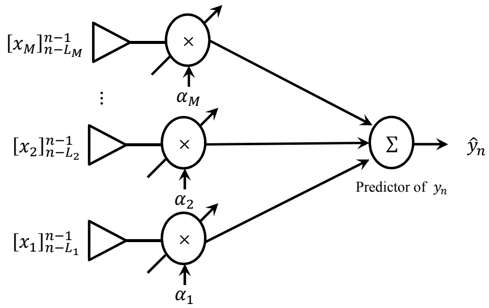

Fig. 3. The illustration of information diversity and combining. One can either combine different estimates of y_n from different sources to form a new estimate, or simply combine all observations via proper selections.

Another way to deal with the selection of depth is the *regularization*, which penalizes models with extreme number of parameters. The original problem (2) can be equivalently represented as the follow:

$$\mathbf{w}_n = \underset{\mathbf{w}\in\mathbb{R}^L}{\arg\min}\, \mathbb{E}\left\{\left(Y_n - \mathbf{w}^\mathsf{T} X_{n-L}^{n-1}\right)^2\right\}. \tag{16}$$

By regularization, the object in (16) to be minimized becomes

$$\mathbb{E}\left\{\left(Y_n - \mathbf{w}^\mathsf{T} X_{n-L}^{n-1}\right)^2\right\} + \lambda f(\mathbf{w}), \tag{17}$$

where $f(\cdot)$ is a measure of model complexity of $\mathbf{w}$. Many type of complexity measure has been shown to have good and consistent properties, such as *Akaike information criterion* (AIC) and *minimum description length* (MDL) in information theory [13], or LASSO and Ridge regression in statistics [14].

Remark. *In the viewpoint of communication, we do not wish L to be too large, because such large L implies heavy inter-symbol interference (ISI) in the channel. We refer the problem of selection of depth of observation as the mitigation of the cross-interference of information. It is potentially advantageous over many conventional methods.*

IV. Data Prediction over Multiple Channels

Consider a process $\{Y_t\}$ to be predict. Suppose we have M set of data variables that can help us to do prediction. That is, $\mathcal{D} = \{\{X_{1,t}\}, \{X_{2,t}\}, \ldots, \{X_{M,t}\}\}$. There exhibits a channel between each pair of data variable and the target. *i.e.*, $(\{Y_t\}, \{X_{m,t}\})$. This prediction of Y_n with multiple sources is equivalent to the signal detection in SIMO environment. Thus, the techniques of *receiver diversity* can be directly applied to data prediction over Multiple Channels to obtain the *information diversity*.

A. Information Diversity and Combining

So far we have established the basic framework for single-pair data prediction by equalizer. This idea can be easily extended to the case of multiple information sources. Consider M sets of data, $\{[X_m]_{t=n-L_m}^{n-1}\}_{m=1}^{M}$. We aim to find an estimate of Y_n based the the realization of these data. This problem is equivalent to the signal detection in SIMO. The techniques of receiver diversity can be directly applied in this scenario. One can construct an estimate of Y_n based on each set $[X_m]_{n-L_m}^{n-1}$ respectively, and then combine these estimates together to form a new estimate of Y_n. Such combination can be done in several ways, and all of them depends on the measure of prediction performance.

The adoption of linear prediction is based on the believe that the signal model is

$$[X]_{n-L}^{n-1} = \mathbf{h}_n Y_n + N_n, \tag{18}$$

where $N_n \sim \mathcal{N}(0, \sigma_n^2)$ at a particular time n. While the channel coefficients $\mathbf{h}_n$ is unknown, we can use equalizer and training sequences to find optimal coefficients of equalizer, $\{w_{n,t}\}$, to detect the signal. We also believe that the power spectrum of the embedded error in each channel is time-varying. Therefore, at every time instance, each equalizer output has different and time-varying performances, and this combination can leverage this property to obtain better predictor of Y_n.

The performance of each branch can be measured in the term of prediction error of every time instance; that is,

$$\gamma_{m,n} = (y_n - \hat{y}_{m,n})^2, \tag{19}$$

which is the unbiased estimator of the power spectral density of error; *i.e.*, σ_n^2. It is also of interest to develop other measure of performance of each branch, but it is not included in the scope of this paper. In the following subsections, we introduce two most common combining schemes: selection combing (SC) and maximal-ratio combining (MRC).

B. Selection Combining

In selection combining (SC), the combiner outputs the estimate on the branch with the highest performance.

$$\hat{Y}_n = \hat{Y}_{m^*,n}, \tag{20}$$

where $m^* = \arg\max_m \gamma_{m,n-1},\ m \in \{1,\ldots,M\}$. The performance of such combination relies on the belief that the condition of channel exhibits a strong time dependency.

C. Maximal-Ratio Combining

Since we believe that the error process is a time-varying process, it is somewhat risky and wasteful to select only one branch and abandon other estimates at each time. An alternative is that, we choose a convex combination of each branch to form a new estimate; that is, $\hat{Y}_n = \sum_{m=1}^{M} \alpha_{m,n}\hat{Y}_{m,n}$. The coefficients $\{\alpha_m\}$ can be determined by the following:

$$\alpha_{m,n} = \frac{\gamma_{m,n-1}^{-1}}{\sum_{i=1}^{M}\gamma_{i,n-1}^{-1}} \tag{21}$$

Also, a special case of MRC is the Equal-Gain Combining (EDG). It assume that both channel are suffered from noises with the same power spectrum density. The the coefficients to be multiplied is exactly the same. Generally, this situation is rare; however, the EGC is still of some values because it works as *robust* combining scheme. In the situation that the channel changes faster than we expect, and the estimate of noise density is believed to be inaccurate, the EGC is still valid because it is the average of all the estimate we have.

D. Equalizer Combining

The equalizer approach of prediction and the diversity combining can be all regarded as a type of linear prediction. Besides, unlike the communication system, which is limited by the antennas, there is no restriction that we have to separate these observations of different sources. The most straightforward way to find a linear predictor of Y_n is to construct an estimate of the following form:

$$\hat{Y}_n = \sum_{j=1}^{M} \sum_{t=n-1}^{n-L_j} w_{j,t} X_{j,t}. \tag{22}$$

In such case, we allow different numbers of taps of different sources to guarantee that each channel has the least ISI. This is also a tradeoff between prediction performance and computation complexity.

The selection of numbers of taps can also be established by either the minimization of conditional information transfer or the regularization method.

Generally speaking, this combining should dominate all methods discussed above, because it is consistent with the most general MMSE estimate of Y_n. However, the price to pay is the computational complexity. The number of candidates of model grows exponentially to the number of sources, while the number of coefficients we have to compute grows with the product of tap numbers, $\prod_{i=1}^{M} L_i$. This fact leads to the computation of cumbersome auto-correlations and cross-correlations among all sources and the target variables.

V. Experiment and Discussions

A. Experiment

In the financial market, it is of particular interest to understand the mechanism among international stock prices and the exchange rate of foreign currency. In this experiment, we select the NASDAQ-composite index (NASDAQ) and the stock price of Morgan Stanley (MS) from the U.S. stock market to predict the stock prices of Taiwan Semiconductor Manufacturing Company (TSMC) and the Hon Hai/Foxconn Technology Group (FOXCONN) from Taiwan, respectively. Also the exchange rate of New Taiwan Dollar to U.S. Dollar, (ExR), is included in the prediction. The period we sampled is form Jan. 1, 2009 to Aug. 14, 2014, including 1,382 samples of stock prices and exchange rates in total.

We predict the target stock prices (TSMC and FOXCONN) by the other data, respectively. Both the prediction over one channel and multi-channel are conducted. The selection of depth of observations is done both by the Akaike information criterion (AIC) and the information transfer (Eq. 15), respectively. Both selections show the consistent results. One of our result is visualized in Fig. 4. One can see that, even with one-channel prediction, the accuracy of prediction of equalizer-based approach is satisfactory. The overall performance of prediction is shown in Table I and Table II.

In the prediction of FOXCONN, we only use NASDAQ and ExR, while we adopt NASDAQ, ExR, and MS to predict the TSMC. Serval observations and insights can be found in this experiment. We will discuss them in the next subsection.

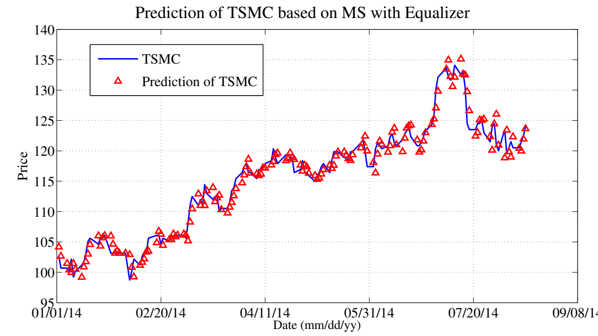

Fig. 4. The prediction of TSMC price based on Morgan Stanley (MS) with equalizer approach. The depth of observations is 1.

TABLE I
Data Prediction of Foxconn Prices

FOXCONN Data Prediction over One Channel

Source	Normalized NSE [a]	Depth of Obs.
ExR	0.99	2
NASDAQ	1.00	2

FOXCONN Data Prediction over Multi-Channel

Combining	Normalized MSE	Depth of Obs. [b]
EGC	0.99	$[2, 2]$
SC	0.98	$[2, 2]$
MRC	0.97	$[2, 2]$
Eq-MR	0.93	$[2, 1]$

[a]The normalized mean square error for each method. The reference MSE is 2.8306, which is the MSE of predicting TSMC by NASDAQ.
[b]The number of taps corresponding to ExR, NASDAQ, and MS, respectively.

TABLE II
Data Prediction of TSMC Prices

TSMC Prediction over One Channel

Source	Normalized MSE [a]	Depth of Obs.
ExR	0.99	2
NASDAQ	1.00	2
MS	0.83	1

TSMC Data Prediction over Multi-Channel

Combining	Normalized MSE	Depth of Obs. [b]
EGC	0.85	$[2, 2, 1]$
SC	0.86	$[2, 2, 1]$
MRC	0.84	$[2, 2, 1]$
Eq-MR	0.83	$[0, 0, 1]$

[a]The normalized mean square error for each method. The reference MSE is 2.9409, which is the MSE of predicting TSMC by NASDAQ.
[b]The number of taps corresponding to ExR, NASDAQ, and MS, respectively.

B. Discussion

The stock prices we select are believed to have dependency by our domain knowledge, but the exact influence may be too complicated such that we cannot explain the mechanism among them by a proper financial model. For example, we believe that ExR has strong impact on the TSMC, and the observation of ExR can help us to predict TSMC, but we do not know the exact relation. This problem is referred to the *causal inference* in time series [15]. Furthermore, the determination of the causal relationship, and the model of prediction are in the research field of *knowledge discovery* in machine learning [3].

In our experiment, all three data variables (NASDAQ, ExR, and MS) works great in the accuracy of predictions of targets (TSMC and FOXCONN). However, there are serval interesting observations worth of discussions.

- As what we expected, the Equal-Gain Combining (EGC) is a robust strategy of information combining. Its performance is simply the average of respecitve prediction. Generally, Maximal-Ratio Combining (MRC) outperforms Selection Combining (SC) and Equal-Gain Combining (EGC).
- The Selection Combining (SC) is not as well as we expected in the prediction of TSMC. The reason is that, in the single source test, we can discover that the prediction of stock price does not relies on a long length of historical data. This fact indicates that the channel condition changes very rapidly such that less dependency can be found between time series. In addition, the prediction based on MS outperforms all other sources, which make the selection combining meaningless. Likewise the MRC does not performs well when there exists strong information transfer (TSMC-MS).
- The Equalizer Combining is competitive to all the other combining schemes. The selection of depth of observations is equivalent to model selection in knowledge discovery. However, in our experience, the implementation complexity and computational cost are heavier than all the other combinings.

In this experiment, we believe that, given the information transferred to the MS, the other information on the NASDAQ and ExR would be not necessary for prediction. Too much information included in the prediction will only creates ISI and degrades the performance. It is a straightforward result since the information transfer (15) conditioning on the MS to all other sources are close to zero.

Above fact implies that, the information transfer can be an explanation, even criterion, for knowledge discovery. Not only can the condition of sufficient information in the data predicition be derived via information theory but also can the criterion be justified by the experiments. Furthermore, this viewpoint is rationalized by the *information coupling* [4], which studies the geometric structure on the space of probability distributions and finds a similarity measure for probability measures. Also, it should be noted that, the equalization of time series is equivalent to the *Kalman filtering* with deterministic model in a more general setting [16], which is one of the standard approaches in analysis of time series. This work not only provides a new methodology for data analytics, but also gives explanations to many useful and well-known techniques in data analysis.

VI. Conclusion

In this paper, we present a totally new understanding of data prediction in the viewpoint of communication theory and social networks. A prediction process is identified as a process of information transfer, which can be quantified by the mutual information and modeled by the communication channel. Based on this idea, many techniques, such as equalizer, receiver diversity and combining, are directly applied to predict data. An experiment based on stock prices and exchange rate is conducted. Many interesting observations can be found in this experiments, including the relation of our idea with the most advanced researches in machine learning and information theory. Of course, more in-depth study on the communication-inspired techniques of data analysis, such as combining and information transfer, are still open with more opportunities in statistical data processing and learning, where we wish this paper to be a cornerstone toward a new technology.

References

[1] K.-C. Chen, M. Chiang, and H. V. Poor, "From technological networks to social networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 548–572, Sep. 2013.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, Jul. and Oct. 1948.

[3] M. Last, Y. Klein, and A. Kandel, "Knowledge discovery in time series databases," *IEEE Trans. Syst., Man, Cybern. B*, vol. 31, no. 1, pp. 160–169, Feb 2001.

[4] S. Huang and L. Zheng, "Linear information coupling problems," in *Proceedings of the 2012 IEEE International Symposium on Information Theory, ISIT 2012, Cambridge, MA, USA, July 1-6, 2012*, 2012, pp. 1029–1033.

[5] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, pp. 461–464, Jul 2000.

[6] J. Haupt, W. Bajwa, M. Rabbat, and R. Nowak, "Compressed sensing for networked data," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 92–101, March 2008.

[7] S. Yang and E. D. Kolaczyk, "Target detection via network filtering," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2502–2515, May 2010.

[8] T.-Y. Chuang and K.-C. Chen, "Cognition on the networked data of stochastic topology," in *2014 IEEE International Conference on Communications (ICC)*, June 2014, pp. 3753–3757.

[9] A. Sandryhaila and J. M. F. Moura, "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 80–90, Sept 2014.

[10] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan 2014.

[11] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. New York: Springer, 1994.

[12] A. Goldsmith, *Wireless Communications*. New York, NY, USA: Cambridge University Press, 2005.

[13] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inf. Theory*, vol. 30, no. 4, pp. 629–636, Jul 1984.

[14] A. Neumaier, "Solving ill-conditioned and singular linear systems: A tutorial on regularization," *SIAM Review*, vol. 40, pp. 636–666, 1998.

[15] Z. Wang and L. Chan, "Learning causal relations in multivariate time series data," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, pp. 76:1–76:28, Sep. 2012.

[16] A. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1990.