Communication Theoretic Inference on Heterogeneous Data

Kwang-Cheng Chen*†, Baturalp Mankir*, Shao-Lun Huang*†, Lizhong Zheng†‡ and H. Vincent Poor§
*Graduate Institute of Communication Engineering, National Taiwan University
Email: ckc@ntu.edu.tw, bmankir@gmail.com
†Research Laboratory of Electronics, Massachusetts Institute of Technology
Email: {chenkc, shaolun}@mit.edu
‡Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology
Email: lizhong@mit.edu
§Department of Electrical Engineering, Princeton University
Email: poor@princeton.edu

***Abstract*—Statistical learning has attracted considerable recent research interest due to the wide-ranging demands of big data analytics. The recent introduction of communication theory and information coupling theory into this area suggests a new perspective on statistical learning and inference for data analytics. This paper investigates inference of one data variable from heterogeneous data variables, a problem that plays an increasingly important role in the emerging applications of big data analytics. To generalize the existing conceptual approach, information coupling filtering under hidden data structure or unknown knowledge of interactions among data variables is developed. A least-mean-squares (LMS) filtering approach for non-stationary data similar to an equalizer is suggested, while the training data gives the depth of the filter analogously to model selection in learning theory. The information combining in diversity communication is extended to fuse more data variables for even greater precision of inference. Extending from multiuser detection, an algorithm based on Multiple Signal Classification (MUSIC) is demonstrated to identify useful data variables for inference, as a novel solution to knowledge discovery. A series of examples illustrate the effectiveness of this framework, suggesting that statistical communication theory and statistical signal processing can substantially contribute to statistical learning theory.**

I. Introduction

To extract meaningful information from very large data sets, data analytic tools such as machine learning have emerged as one of the most prominent technologies. The spirit of machine learning is to delineate the relationships among data. Existing methodologies and subsequent algorithms can be found in the literature [1]. Communication theory and information theory have been adopted into machine learning and data analytics, usually to serve as a means to enrich mathematical meaning and data analysis performance, particularly in roles like measuring distance between data [2] or information criteria for mode selection.

Exploring further the frontier of statistical communication theory and network theory, entities bearing relationships can form a generalized social network [3], and such entities include data. Facilitation of this concept has been demonstrated in [4] for data having probabilistic relationships, which generally hold for most data analysis problems by considering statistical properties. We may consider two series of data (or two data variables/vectors), $x[n]$ and $y[n], n = 0, 1, \ldots$. The existence of conditional probabilities $P\{y[n]|x[n]\}$ suggests an equivalent communication channel to transfer information from $x[n]$ to $y[n]$. In a large data set, this involves multi-layer multiple data variables, which may be generally depicted as in Figure 1. Current active research is trying to discover the knowledge structure and to fuse heterogeneous data to enable useful inference (a.k.a. learning). Based on information coupling [5] and its communication theoretic realization, examples are given to illustrate its potential effectiveness in [4]. However, detailed facilitation from the concept to working mechanisms and thus algorithms is still needed. In this paper, we will demonstrate a way to generalize results from communication theory into data processing schemes and algorithms, with illustrations. In particular, we will focus on processing heterogeneous data as an analog to statistical signal processing in point-to-point physical transmission and multi-user communications, to show the knowledge discovery and information fusion capability on heterogeneous data under such a communication theoretic approach. Here, heterogeneous data means that we are using data variable(s) quite different from the target variable. For example, we may wish to infer the stock price of a company based on an exchange rate rather than a common market composite index in one example, or infer an index of public health from environmental data as another example.

Each data variable in Figure 1 represents a sequence or a homogeneous collection of data, such as a time series in financial data or an indexed sample vector from a stochastic source. The grey area represents unknown and unobserved variables and their linear/nonlinear interactions as dynamic systems. We aim at inference based on heterogeneous data variables, which is of great interest in big data analytics. Understanding of nonlinear or dynamic systems of data variables is also a topic of interest in big data analytics [13][14]. In Section II, we re-visit the subject of inference between heterogenous data variables by introducing the new concept of an information coupling filter (ICF). In Section III, we leverage the concept of information combining in diversity communi-

978-1-4799-6664-6/16/$31.00 ©2016 IEEE

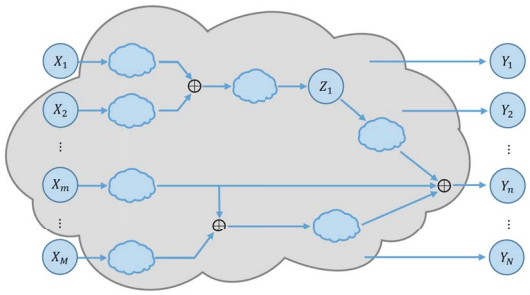

Fig. 1. Graphical model of a large data set with networked probabilistic relationships, where the grey area indicates unknown or unobserved variables and their interactions.

cation to realize ICF for more heterogeneous data variables. Subsequently, knowledge discovery to identify effective data variables in machine learning is shown to be equivalent to identification of active users in multiuser communications. The MUSIC algorithm used in multiuser detection on top of singular value decomposition is therefore demonstrated as an effective methodology to determine data variables for information combining.

II. Inference Between Heterogeneous Data Variables

As a starting point of our communication theoretic view of inference on data, let us look at two data variables in Figure 1, $X_m \rightarrow Y_n$, where the conditionally probabilistic relationship between the two data variables can be thought of being analogous to a "communication" channel, and each data variable is actually an indexed data vector.

A. Equalizer to Optimize Information Transfer

For pedagogical reasons, we briefly review the intuition of [4] in this sub-section. To infer Y_n based on X_m, an intuitive idea is to construct a "good" receiver, while "good" may typically imply minimum mean square error (MMSE). Such a communication channel is obviously noisy and very nonlinear, and likely very dynamic in many cases. Assuming the channel is static for a period of time, say our observation period, any digital communication system textbook suggests adopting an equalizer for pre-detection in an optimal receiver. Consequently, an inference of Y_n can be represented in shorthand notation by

$$\hat{Y_n} = X_m * \mathbf{h} \tag{1}$$

where $\mathbf{h}$ is the response of an *equalization filter* and both X_m and Y_n are data vectors. The convolution operation of (1) is essentially a sliding window filtering operating on the data vector. To further elaborate upon the insight and mechanism behind [4], please note that if the *equalizer* is implemented by a tapped-delay-line finite-response filter, the coefficients of the equalizer, as a vector $\mathbf{h}$, can be obtained by the well-known *Wiener-Hopf* equation for stationary signals or data, as long as we know the appropriate length of the filter. A longer filter suggests better performance due to longer observation and training for stationary data. Unfortunately, in (big) data analytics, two fundamental conditions are not satisfied here: known filter length and stationary data. Unless some critical domain knowledge is available, a simple extension of the equalizer concept for the purpose of machine learning is not feasible for our general purpose. For non-stationary linear filtering, the Kalman filter is well known to be optimal but requires a one-step system equation to describe Markovian dynamics, which is not possible nor feasible for general data analytics, although it can be successful in certain classes of data such as financial data and some sensor data for control [6]. To resolve the dilemma regarding unknown filter depth and non-stationary data, a new supervised learning methodology has been introduced as [4], based on the equalizer concept. During a training period, an appropriate filter length is selected and subsequently the filter coefficients are determined, based on the MMSE criterion. To infer non-stationary data, we maintain the filter length but update the filter coefficients online. As indicated in [7] to practically design *support vector regression*, ε-deviation is allowed in the training period. Following the same spirit, we determine the filter depth and fix it for future inference, while adapting the coefficients of the filter online. We will show later, in a practical case that the mean-square error (MSE) is indeed around the same level near the optimal depth of the filter, such that our communication theoretic implementation succeeds.

B. Adaptive Filtering via Information Coupling

The above equalizer approach for statistical (machine) learning can be more precisely viewed from the perspective of information coupling. Figure 2 illustrates the entire concept, while the grey cloud indicates an unknown and unobserved environment.

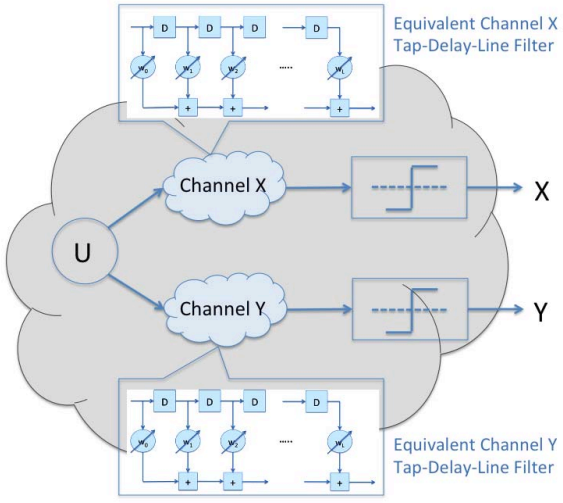

Fig. 2. Information coupling view of inference between two data variables, where the grey area indicates hidden or unknown structure and relationships. The channels representing dynamics of unknown interactions of data variables are generally non-stationary and non-linear, but are approximated by linear filters and possible quantization to represent decision or observation data.

Here, U denotes a hidden data variable representing a simplified hidden structure analogous to an information source

in communications. We treat U as being one-dimensional here, but it could be of a finite but usually not large dimension. The information from U goes through an equivalent *data channel* that represents disturbing dynamics and introduces noise. After the threshold device or equivalent to yield a corresponding decision, the data variable X becomes our observation. The dynamics of the data channel can be modeled (more precisely, approximated) as a linear time-invariant filter with an unknown response $\mathbf{h_X}$ of weighting coefficients, while the weighting coefficients can be time varying by changing the weighting coefficients and filter length. Similarly, we can obtain another data variable Y through a different data channel with response $\mathbf{h_Y}$ and different decision mechanism. Here, the decision mechanism translates noisy data to a discrete format (like a quantizer) or a continuous format of observations. Now, we are interested in investigating *inference on heterogeneous data*. That is,

$$\hat{y}[n] = f(x[n-1], x[n-2], \dots). \tag{2}$$

This problem is a typical statistical learning, regression, or estimation problem, and the goal is to identify an appropriate f to form a *statistic* for optimal performance. However, we develop a new scenario to re-visit this problem by considering the equivalent communication system. Though we will focus on inference on heterogeneous data in this paper, we mention another related problem, the *prediction of heterogeneous data*, as

$$\hat{y}[n+1] = f(x[n], x[n-1], \dots). \tag{3}$$

For computational efficiency, we wish to use a linear function for f and therefore

$$\hat{y}[n] = X * \mathbf{h} \tag{4}$$

where $\mathbf{h}$ denotes the tap-delay-line filter as a realization of the earlier mentioned *equalizer*, with filter depth l and weighting coefficient vector $\mathbf{w}_l = (w_0, w_1, \dots, w_l)$. From two data channels in Figure 2, ignoring noise, we have

$$X = U * \mathbf{h_X} \tag{5a}$$

$$Y = U * \mathbf{h_Y}. \tag{5b}$$

Assuming nice mathematical structure and ignoring causality of filtering,

$$U = X * \mathbf{h_X}^{-1}. \tag{6}$$

Substituting into the equation linking U and Y, we have

$$Y = X * \mathbf{h_X}^{-1} * \mathbf{h_Y} = X * \mathbf{h}. \tag{7}$$

More precisely, this implies that the filtering of the following response successfully establishes the inference of Y from X:

$$\mathbf{h} = \mathbf{h_X}^{-1} * \mathbf{h_Y} \tag{8}$$

where we ignore the causality in strict mathematical treatments. As $\mathbf{h}$ intuitively finds the relationship of information generation, this suggests that we name this technique *information-coupling filtering*, similar to the role of a preprocessing equalizer of an optimal receiver. Information-coupling filtering serves as a potential mechanism to infer Y based on X, though we do not actually know the hidden structure at all. Where $\mathbf{h}$ therefore serves as the optimal filtering for the pseudo communication $X \to Y$, $U \to X \to Y$ gives precisely the scenario from information theory described in [5]. Now we need to identify an appropriate performance measure to design such filtering. If we adopt MMSE as the performance criterion of inference or statistical learning, the optimal filter $\mathbf{h}$ satisfies

$$\min_{\mathbf{h}} E|X * \mathbf{h} - Y|^2, \tag{9}$$

where the expectation is facilitated by the average of empirical squared errors.

Under the MSE measurement criterion, the consequent inference based on this view of statistical learning can be generally realized using the terminology of statistical learning theory and statistical signal processing as follows [7][8]:

1) Approximation: We approximate the impact from U to X through the filter $\mathbf{h}_X$, and similarly from U to Y through the filter $\mathbf{h}_Y$. We then wish to find the approximation of the *information coupling filter* with coefficient vector $\mathbf{w}_l$.
2) Model Selection: Given the observation depth N, and the depth L_0 of the training period, we select the filter depth l^* from $\{1, \dots, N - L_0 + 1\}$ by the MMSE criterion:

$$l^* = \underset{l}{\operatorname{argmin}} \min_{\mathbf{w}_l} \frac{1}{L_0+1} \sum_{i=0}^{L_0} |(x[N-i-l], \dots, x[N-i]) * \mathbf{w}_l - y[N-i]|^2, \tag{10}$$

   where for a given l, the coefficients $\mathbf{w}_l$ of the least-mean-squares (LMS) filter are updated by the steepest descent method under a convergence criterion for step size [6].
3) Statistical Inference: By keeping the filter depth l^*, we update $\mathbf{w}_l$ online to infer/estimate and to track $y[\cdot]$ based on $x[\cdot]$.

One immediate feature of this methodology is no requirement of a data model for the data variables X and Y, that is, niether *a priori* statistics nor a stationarity assumption is needed. As long as these two data variables are generated from some partially common hidden and thus unknown mechanisms, we can apply this methodology to achieve the purpose of inference. Now, we use some widely available financial data to illustrate the nature of our proposed *information-coupling filtering* (ICF) approach. Taiwan Semiconductor Manufacturing Company (TSMC) is the world's largest contract integrated circuit manufacturer and is publicly listed on the Taiwan Stock Exchange and NASDAQ in the US. We wish to infer its stock price, $\hat{y}[n]$, based on another data variable (actually a sequence of observations), $x[n-1], x[n-2], \dots$ as in (2). We consider several examples for X, such as the NASDAQ index, a financial stock Morgan-Stanley, a global telecommunication operator Vodafone, and the exchange rate between the US Dollar and Taiwan currency (NTD). We use the data during

2009-2013 as the training period and test the data for inference from 2014 to the first half of 2015 (i.e., June 2015). The construction of an information-coupling filter is just like an LMS equalizer, e.g. [6], for such non-stationary data. The primary purpose of ICF during the training period is to determine an effective filtering depth and to train the corresponding coefficients of this adaptive filter. From the intuition of statistical learning, this step is similar to identifying the support vector or kernel of proper observation length in regression [1][7]. Figure 3 shows the root MSE (RMSE) for the above-mentioned heterogeneous data variables vs. filtering depth. Please note that the calculation of RMSE weights the most recent MSE more, due to the non-stationarity assumption. With *a priori* knowledge or common sense from domain knowledge, NASDAQ is expected to be highly related and requires short depth of filtering to be effective. All three other data variables demonstrate similar levels of depth for optimal filtering, and slight deviations from the optimal depth affect the performance little. Since these data are non-stationary, the average MSE (ASE) rises gradually after reaching an optimal value, which is fundamentally different from the case of stationary signals. To better reflect the nature of non-stationary data, the ASE used to determine the optimal depth of filter is evaluated by the average MSE in the last year of the training period (i.e., L_0=251).

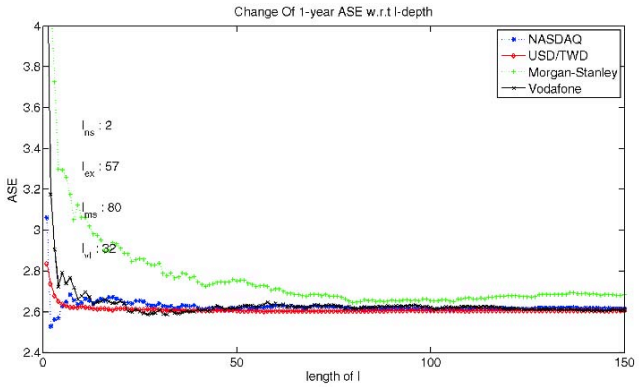

Fig. 3. Average MSE Corresponding to the Depth of Filter for Training on Non-Stationary Data

The MSE performance of inference is shown in Figure 4. Unsurprisingly, the NASDAQ with short depth filtering performs well and its MSE is clearly better than the "yesterday estimator" that is famous as a straightforwardly simple but effective predictor. The other three data variables result in similar performance to the yesterday estimator, while the exchange rate surprisingly performs best among these data variables and better than the yesterday estimator. As expected, the NASDAQ indeed performs the best among the four data variables. However, the apparently irrelevant exchange rate between USD and NTD outputs the second best inference among the four. Later in Section III, we will find a more useful role of the exchange rate. Such a process can be used to identify useful data variables in inference, which can be viewed as another realization of knowledge discovery in machine learning.

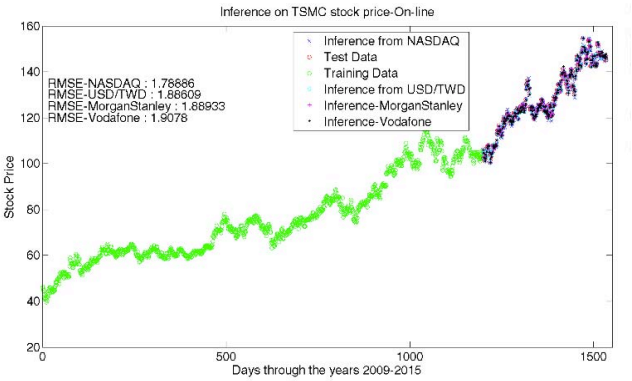

Fig. 4. MSE in the Inference Stage with Filter Length from Figure 3

Remark. *Extending from information coupling [5], linear facilitation is able to address many cases of interest. As a matter of fact, as seen in Figure 4, all four financial data variables are pretty good for inferring the TSMC stock price using the proposed information-coupling filtering, much more precise than other approaches [1][6] such as least squares, Bayesian linear regression, ridge regression, or principal component regression, at a similar level of filtering/regression complexity.*

Remark. *Another interesting similarity to support vector regression [7][8] is that we do not use all the training data to generate the regression model and we only use the effective length of filtering with weighting coefficients to count on a certain portion of training data, where the correspondingly small magnitude of filter coefficients implies light influence on the inference. However, all training data are used to minimize the empirical risk function.*

Remark. *Selecting the step size in the steepest descent algorithm to obtain filter coefficients can result in slightly different outcomes from Figures 3 and 4, under the convergence criterion. However, our extensive numerical tests suggest similar levels of filter depth and subsequent MSEs.*

III. Information Combining Based on Diversity Communications and Multiuser Communications

Due to the unknown and unobserved (or hidden) structure of data, we always try to incorporate more heterogeneous data variables or more heterogeneous factors into the inference process to result in greater precision. Regression methods increase data dimensions. However, information coupling filtering suggests a different approach as shown in Figure 5, which is extended from Figure 2 to illustrate the simplest case of using two heterogeneous data variables for inference.

A. Information Combining Inspired by Diversity Communications

By looking at the right-hand side of Figure 5 while ignoring the grey hidden part of the system, the information combining

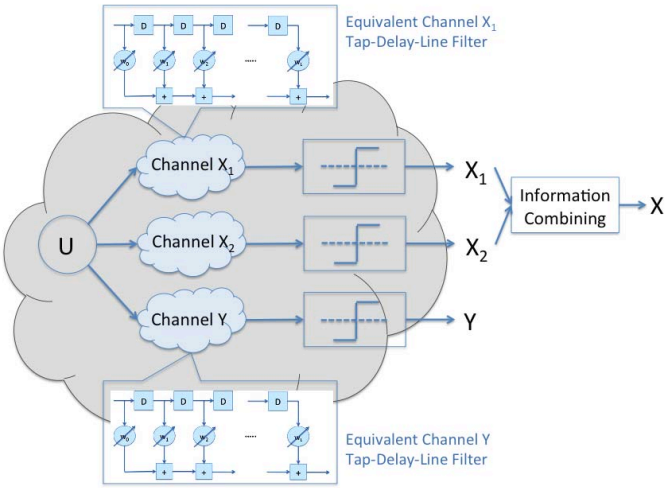

Fig. 5. Illustration of Using Two Heterogeneous Data Variables, X_1,X_2 to infer Y.

of X_1, X_2 into X is like signal combining in receive-diversity communications, in which we hope to take advantage of independent observations X_1, X_2 to maximize the signal-to-noise ratio (SNR) in inference [9]. In [4], it is suggested that information combining just as diversity combining improves the precision of inference. Following the concept of a communication system as illustrated in Figure 5, the information combining can be implemented by the linear operation $\circ$ between two data variables or more, i.e.,

$$Y = (X_1 \circ X_2) * [(\mathbf{h}_{X_1} \circ \mathbf{h}_{X_2})^{-1} * \mathbf{h}_Y] \qquad (11)$$

where $(\mathbf{h}_{X_1} \circ \mathbf{h}_{X_2})^{-1} * \mathbf{h}_Y$ serves as an information coupling filter. Typical combining techniques for receiving diversity include equal-gain combining, selective combining, and maximal-ratio-combining (MRC), while MRC is known to be optimal in additive white Gaussian noise (AWGN) channels. Although we assume non-stationary data, we can still use least squares to obtain the weighting coefficients in MRC. The linear operator $\circ$ might serve as an approximation of a nonlinear information combining operation, say selective combining. Generally speaking, the intuitive meaning of ICF still holds. Here, Table I documents our experiment for information combining, which is a unique feature for ICF in learning. As illustrated in Table I, MRC, as an effective information combining scheme, indeed demonstrates benefits by introducing extra data variables other than NASDAQ, even though those extra variables individually are not providing better performance than the primary NASDAQ data. However, equal-gain information combining and selective information combining may not bring effective information for inference, as indicated by the higher RMSE than the benchmark of using only the NASDAQ. Moreover, as shown in the fourth row of Table I, introducing even more data variables for the MRC information combining might not necessarily result in better performance of inference. This suggests the next critical exploration, namely methodology to identify effective data variables for inference, in Section III-B.

Combining of NASDAQ & Exchange Rate	Online RMSE	Combining of All 4 Data Variables	Online RMSE
Equal-Gain	1.8206	Equal-Gain	1.8306
Selective	1.8451	Selective	1.8451
MRC	1.7815	MRC	1.7824
Benchmark: NASDAQ of Optimal Depth	1.7888		

TABLE I
RMSE FOR INFORMATION COMBINING

Remark. *Without knowing the hidden structure, the information coupling filter can be realized as a tap-delay-line LMS filter as in Section II. The training data and resulting MSE determine the ICF depth and the weighting coefficients of MRC. Then, we execute inference by an online update of coefficients.*

B. Source Identification Inspired by Multiuser Communications

We have discussed the success of communication theoretic information combining in statistical learning, yet we still observe that introducing inappropriate data variables might not help. For example in Table I, MRC based on all four data variables does not yield better MSE than the case of MRC of only the NASDAQ and exchange rate. In particular, the NASDAQ and exchange rate are the most useful two data variables as shown in Figure 4, which means the exchange rate supplies some useful "information" that is not supplied by NASDAQ. This suggests the need of combining heterogeneous data in inference. Consequently, a new technology challenge arises: to identify the useful data variables for inference, which is actually equivalent to *knowledge discovery* in machine learning.

We propose another surprising approach from communication theory. Instead of trying to modify the well-known Akaike information criterion (AIC) or Bayesian information criterion (BIC), which can be found in any standard textbook in statistical learning [1], let us generalize Figure 5 to have data variables $X_1, \ldots, X_K$ to infer Y, under the hidden system structure. Now, we want to know which of these K data variables are useful. This is equivalent to the blind identification of active users in multiuser detection (MUD) [10][11] with methods fundamentally related to singular value decomposition. In this paper, we utilize the MUSIC algorithm similar to [11] to design an algorithm for data analytics.

We start by sorting the features or sources by using the maximum dependence minimum redundancy (MDMR) method to define subsets and use dimension reduction methods to obtain a desirable subset of data variables. The idea follows from the MDMR criterion [15], which originally used mutual information. Now, we consider a new data variable formed by linear combinations of observed data variables $X_1, \ldots, X_K$,

which is analogous to noiseless MUD:

$$X = \sum_{i=1}^{K} \alpha_i X_i, \tag{12}$$

where the α_i's are real numbers. Omitting the derivations to extend results in [11], we modify the MUSIC algorithm to identify useful data variables to the following steps:

1) Use the data to form the empirical covariance matrix.
2) Apply eigenvalue decomposition to obtain the eigen-structure of the covariance matrix, $\hat{\mathbf{E}}$.
3) Determine the number of useful variables, K_u, by the eigenvalue ratio $\sum_{i=1}^{K_u} \lambda_i / \sum_{i=1}^{K} \lambda_i$ exceeding a pre-selected threshold that is close to 1 and this threshold may be treated as the level of completeness of information for inference.
4) Select such K_u data variables corresponding to the maxima of $\left\| X_i \hat{\mathbf{E}} \right\|^2$ (i.e., data variables that deliver more accurate MSE in a training period).

Back to our example, via this framework while we order $X_1, \ldots, X_4$ according to the accuracy in Figure 4 (i.e., X_1 is the data of the NASDAQ index), we obtain the results and compare with some well-known methods in Table II, where the benchmark methods include principal component analysis (PCA), sparse-PCA, and least absolute shrinkage and selection operator (LASSO). PCA, sparse-PCA, and MUSIC select the NASDAQ and exchange rate as useful data variables, while LASSO includes one more data variable. The proposed MUSIC and MDMR algorithm well matches information coupling filtering to yield the best performance of inference, over other benchmark methods in statistical learning.

Method	Selected Data Variables	RMSE
PCA	X_1, X_2	1.7843
Sparse-PCA	X_1, X_2	1.7842
LASSO	X_1, X_2, X_3	1.7894
MUSIC & MDMR	X_1, X_2	1.7815

TABLE II
COMPARISON OF METHODS TO DETERMINE USEFUL DATA VARIABLES AND RESULTING RMSE

The overall implementation of this communication theoretic inference on heterogeneous data is, during the training period, to identify the "useful" data variables and the ICF structure, then to fuse these data based on MRC after finding the weighting coefficients, using to the information coupling filter for the final inference. Note also that this information coupling filtering approach is realized by tapped-delay-line filters and typical signal processing for wireless communication systems, and thus can be implemented using digital signal processors that are usually much faster than general-purpose processors.

IV. CONCLUDING REMARKS

Using ideas from digital communications, this paper has described an approach to statistical learning on heterogeneous data, with experiments to verify the applicability of statistical communication theory and multiuser detection, connected with the newly developed theory of information coupling. We summarize some interesting open issues:

- Here, we consider one-dimensional data. Obviously, each data variable could be treated as a hyper-plane (i.e., higher dimensional data) without obstacles.
- Noise and causality in filtering require investigation in the future.
- We consider linear systems and filtering to approximate general non-linear system responses in developing the approach. The theory of information coupling suggests effective linear realization if the involved data variables are effectively influenced by some common and hidden mechanism (e.g., correlations in probabilistic systems).
- Via proper pre-filtering and mapping into the feature domain, such methodology can be extended to non-data applications, say handwriting or OCR recognition.
- Certainly, more potential applications to thoroughly investigate information coupling filtering are needed.

ACKNOWLEDGMENT

This research was supported in part by the U.S. Air Force under Grant AOARD-14-4053, the Taiwanese Ministry of Science and Technology under Grant 104-2119-M-002-040, and the U. S. National Science Foundation under Grant CCF-1420575.

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
[2] J. V. Davis, B. Kulis, P. Jain, S. Sra, I. S. Dhillon, "Information-Theoretic Metric Learning," *in Proc. 24th International Conference on Machine Learning*, 2007.
[3] K.-C. Chen, M. Chiang, H. V. Poor, "From Technological Networks to Social Networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 9, pp. 548-572, September 2013
[4] K.-C. Chen, S.-L. Huang, L. Zheng, H. V. Poor, "Communication Theoretic Data Analytics," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 4, pp. 663-675, April 2015.
[5] S.-L. Huang and L. Zheng, "Linear Information Coupling Problems," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2012, pp. 1029-1033.
[6] S. O. Haykin, *Adaptive Filter Theory*, 5th ed., Prentice-Hall, 2013.
[7] A. J. Smola, B. Schlkopf, "A Tutorial on Support Vector Regression," *Statistics and Computing*, vol. 14, pp. 199-222, 2004.
[8] V. N. Vapnik, "An Overview of Statistical Learning Theory," *IEEE Trans. Neural Networks*, vol. 10, no. 5, pp. 988-999, September 1999.
[9] D.G. Brennan, "Linear Diversity Combining Techniques," *Proc. IEEE*, vol. 91, no. 2, pp. 331-356, Feb. 2003.
[10] X. Wang, H. V. Poor, "Blind Multiuser Detection: A Subspace Approach," *IEEE Trans. Information Theory*, vol. 44, no. 2, pp. 677-690, March 1998.
[11] W.-C. Wu, K.-C. Chen, "Identification of Active Users in Synchronous CDMA Multiuser Detection," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 9, pp. 1723-1735, Dec. 1998.
[12] S.L. Huang, L. Zheng, "Spectrum Decomposition to the Feature Spaces and the Application to Big Data Analytics," *in Proc. IEEE Int. Symp. Info. Theory*, 2015.
[13] Z. Li, U. Kruger, L. Xie, A. Almansoori, H. Su, "Adaptive KPCA Modeling of Nonlinear Systems," *IEEE Trans. Signal Processing*, vol. 63, no. 9, pp. 2364-2376, May 2015.
[14] R. Talmon, S. Mallat, H. Zaveri, R.R. Coifman, "Manifold Learning for Latent Variable Inference in Dynamic Systems," *IEEE Trans. Signal Processing*, vol. 63, no. 15, pp. 3843-3856, Aug. 2015.
[15] H. Peng, F. Long, C. Ding, "Feature Selection Based on Mutual Information Criteria of Max-dependency, Max-relevance, and Min-redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, Aug 2005.