Panel Article **Open Access**

Eric Zheng*, Yong Tan, Paulo Goes, Ramnath Chellappa, D.J. Wu, Michael Shaw, Olivia Sheng, Alok Gupta

When Econometrics Meets Machine Learning

https://doi.org/10.1515/dim-2017-0012
received June 24, 2017; accepted August 20, 2017.

Keywords: econometrics, machine learning, information system, prediction, explanation, integration.

1 Introduction

Econometrics and machine learning used to perform well in their own orbits separately. They differ in purpose, focus, and methodologies. However, due to the abundant supply of "big data" and the demand for solving complex problems, there emerges a trend of applying econometrics and machine learning in an integral manner. At the 11th China Summer Workshop on Information Management (CSWIM 2017) in Nanjing, China, the meeting's chairs assembled a panel of senior information system (IS) scholars to discuss the theme "When Econometrics Meets Machine Learning". The panel consisted of Paulo Goes, Yong Tan, Ramnath Chellappa, D.J. Wu, and Michael Shaw. Eric Zheng was the panel coordinator. Olivia Sheng and Alok Gupta also participated in the discussion of the panel. The panel discussed the guidelines and key issues of combining econometrics and machine learning in IS research. This paper is a narrative of what the panelists discussed in the panel discussion.

***Corresponding author: Eric Zheng:** Jindal School of Management, University of Texas at Dallas, Richardson, TX, USA. E-mail: ericz@utdallas.edu
Yong Tan: Foster School of Business, University of Washington, Seattle, WA, USA
Paulo Goes: Eller College of Management, University of Arizona, Tucson, AZ, USA
Ramnath Chellappa: Goizueta Business School, Emory University, Atlanta, GA, USA
D.J. Wu: Ernest Scheller Jr. College of Business, Georgia Institute of Technology, Atlanta, GA, USA
Michael Shaw: Gies College of Business, University of Illinois at Urbana-Champaign, Champaign, IL, USA
Olivia Sheng: David Eccles School of Business, University of Utah, Salt Lake City, UT, USA
Alok Gupta: Carlson School of Management, University of Minnesota, Minneapolis, MN, USA

2 Topic 1: What is the meaning of combining econometrics with machine learning?

2.1 Eric Zheng

In the past few years, we have seen growing interests in combining econometrics with machine learning in research. In this panel, we would like to discuss what we mean by saying "when economics meets machine learning". We will answer why this is important and what we could do in a research. We are happy to have the great panel consisting of Paulo Goes, Yong Tan, Ramnath Chellappa, D.J. Wu, and Michael Shaw.

My first question to our panelists is what we mean by saying "combining econometrics into machine learning". What would qualify as a good paper that combines econometrics with machine learning? I would like to invite Yong Tan to help answer the question first.

2.2 Yong Tan

2.2.1 Idea 1: Integration of machine learning and econometrics will be of great value

Actually, I have been talking about this topic many times from the methodological point of view. However, this time I would like to start with a disclaimer that methodology itself is just a tool. We want to advance the tool in order to solve more interesting business problems. The objective of using econometrics is to find out the causal relationship of data, which is our main purpose. On the other hand, from the machine learning's point of view, we want to achieve predictive accuracy. Therefore, if we probe into the two approaches deeper, it can be realized that there are probably the same philosophical ideas that accurately describe the data generation process. Obviously, once we get to know the way in which data are generated, we can get more ideas about the causal relationship in this process. On the other hand, if we can get to know how the data are generated and where it comes from, we can

achieve higher predictive accuracy. Therefore, that is my point of view about the integration of machine learning and econometrics.

2.2.2 Idea 2: Technical integration will be a substantial integration

Now I would like to answer Eric's first question about "qualification" of the research combining these two areas. Apparently, we have seen papers using these two approaches separately. For example, we have papers that use deep learning to extract features from pictures and then convert unstructured data to structured data, which can be used in the econometric models. As a result, we can find out the economic meaning of those extracted features. Therefore, this will be a valuable research. Even though we can use the two methods separately, the combination of both will give us an opportunity to make use of otherwise-unusable data. For example, we can put the actual data, picture data, or video data directly into econometrical models. Consequently, I perceive this as an extremely necessary step to start.

However, we also need to think about how to integrate the two approaches in an essential way, which means technically combining them to a significant degree. It is not only a methodological branch but also a challenge for us to see the econometric sense. Although this is at an early stage, there could be opportunities for fruitful collaboration between econometrics and machine learning in the future. So the one thing for us to consider is that once you build an econometrical model, can you use machine learning or deep learning as a component, not just a separate component, but sort of an endogenous component, which means the real technological integration for the process of data generation? This will be a true melding of the two areas.

2.3 Paulo Goes

2.3.1 Idea 1: Machine learning is predictive, while econometrics is explanatory

First, I would like to know how many of you have used machine learning, how many of you have used econometrics, and how many of you have used both or neither? This is a very interesting topic. We are going to discuss very different approaches, very different objects. However, they are still compatible in some way, and they could be consistent at some level.

In the past few years, we have had a typical paradigm in machine learning. It is just like making a choice in a black box, in which causal relationships and explanatory features have been de-emphasized. Black boxes are actually fine since we do not care about the underlying relationships inside. What we are interested in is developing a model that can provide us a solution. All we need is solution and prediction, and all the data are used and trained to do the validation, cross-validation to be exact. And the model can be adjusted until the prediction accuracy of the model satisfies us. Machine learning is a way of problem-solving, which is designed in the paradigm to find a solution, a recommendation system, a computation system, or a classification. It is usually in a predictive mode, very computational, not necessarily in a statistical way. However, it is totally different in econometrics, the main idea of which is to develop a model that can help explain the world and find out the true status of the world. By hypothesis testing and other traditional social science research methods, econometrics is trying to explain what is going on. In other words, econometrics explores the causal effects in a statistical way. During my research, I found out that machine learning usually goes along with design science paradigm, while econometrics always goes along with the economics of IS paradigm. They are very different. In machine learning, you use various computational techniques (some of which are periodically computational with several layers), classification trees, random forest, and deep learning models, and some of which have statistic features, such as logistic regression. Machine learning is used as a pattern-searching tool. The interesting thing is that when you develop your models, you are prone to combining different models if they can help. You may try five different models and make an aggregation. Since all you want is achievement of great predictive power, there is rarely a linear relationship in machine learning. However, econometrics basically explores linear relationships. Sometimes, you can also find a way to incorporate nonlinear ones. It is primarily a take-away from regression, with different sophistication levels to pursue these linear relationships.

To sum up, prediction of machine learning has nothing to do with causal influence, which is the key point that you should keep in mind. Finding a good prediction model with high accuracy of machine learning is different from inferring the true underlying structure of econometrics. Additionally, predictive power is different from explanatory power. They are two very different angles.

2.3.2 Idea 2: Combination of machine learning and econometrics is achievable and challenging

There seem to be four ways of combining machine learning with econometrics.

1. The first one has been mentioned in the opening remarks already. You use a machine-learning model to feed in the variables that are going to be used in an econometric model. More examples will be put forward later on.
2. The second one is you have the machine-learning algorithm. Then, you can borrow some of the fundamental techniques from econometrics to enhance those of machine learning.
3. Thirdly, you can do it the other way around, which means to borrow machine-learning techniques to enhance econometrics models.
4. The fourth one, which is also the most interesting one, is that you use both of them to explain the same phenomenon. You can take different approaches then. For example, by doing text mining and sentiment analysis, you can define different variables, and then they become part of your econometric models.

Even so, there is still one challenge. When you use machine learning techniques, you can always get to know the precision of the predictive power in the end. For instance, you gain predictive power with accuracy of 85%. When you use that in econometrics models, how can you consider that prediction accuracy when you explain the error of the econometric model? Maybe you do not know how to deal with that or just ignore that. As far as I am concerned, this is a question that should be lucubrated. Therefore, this is a common way of incorporating machine learning with econometrics. There is a great paper by Hal Varian (2014), which is about machine learning and econometrics, and it suggests that econometric models can be improved upon by incorporating the features, including training sets, validating the paradigm, and overfitting the address that have been formed in machine learning. According to this paper, predictive accuracy can be increased by more than 50% after the integration. Econometrics helps you build a model without checking the adequacy of the model. As a result, there is still something that needs to be incorporated, such as *k*-fold cross-validation. Once you build a model, there will be many robust tests for evaluation, which will provide you a big table with several different columns. However, you can also put the different models together and integrate them to see how they perform and how the big data tools can be used. Machine learning works the other way around, which reduces careful, experimental, and natural treatment, or even eliminates it. There should be some other way for enhancement of machine learning. Finally, the paper published in *Management Information Systems (MIS) Quarterly* is strongly recommended, which describes the differences between predictive models and explanatory models (Shmueli & Koppius, 2011). It also describes how to integrate them together. As far as I am concerned, I would like to have the approach of using machine learning first and then try to get the explanation of what is going on, as well as gain an understanding of the underlying correlation for the co-occurrence. Then, it is time to find their causal relationships. It is believed that probably, the inner loop of the hierarchical models is machine learning driven, while the outer loop is econometrics driven. In my view, they are just techniques and tools for us to attack all these interesting problems.

2.4 Ramnath Chellappa

2.4.1 Idea 1: Machine learning and econometrics are different in perception and understanding of problems

I have used some components of machine learning when I was an engineer. According to my experience and understanding of machine learning and econometrics, the biggest difference between the two camps of people is the way they perceive problems. Econometricians are basically from one side, and statisticians and computer scientists are from another side. Take econometricians as an example. Since econometrics needs theories from economics, econometricians develop econometrical models according to these theories, which can be called theoretical models. However, with the assistance of various techniques, statisticians and computer scientists use data to develop models. Take a friend of mine as an example. As a statistician, he uses data to solve problems without any analysis of structures or backgrounds of issues. By using bootstraps and other auxiliary techniques, he can develop statistical models and then find solutions for problems. To be specific, statisticians and computer scientists do not make assumptions when they develop models. They will not suppose variables to be dependent or independent before it can be proved. The features and relationships of data have to be figured out by computer techniques, which is called bootstrapping. Thus, computer scientists and econometricians are totally different in their treatment of data. With the help of computers, computer scientists can do data processing, such as storage, analysis, organization, and classification, which can finally lead to a model. However, based on economic, econometric, and inferential

statistical foundation, econometricians always start with an assumed structured-stochastic model, the unknown and unobservable parameters of which are estimated from a relevant sample by using observed data, and they can be used for inference and prediction (Judge, 2016).

2.4.2 Idea 2: Econometrics is theoretical foundation, and machine learning is technical assistance

Therefore, the way I look at it fundamentally is that one need a theory to put a structure on it, and then you can estimate the components of machine learning, which otherwise could not be done.

As a result of the great power of computers, nowadays, we can make good use of machine learning techniques. In econometrics, estimating the average response of the independent variable to the dependent variable is the main purpose of the prediction and regression process. One goal of the model is to eliminate all the things that cannot really explain the dependent variables in the dependent variables. For example, if you want to hide in the crowd, you had better look like an average person. In other words, the primary reason that all these methods have been used is that they can avoid high degree of complication in econometrics, and they can also be helpful to reveal the commonality and relationship. Nevertheless, they ignore and lose all the individual components in the data, which machine learning can exactly address.

Take myself as an example. Once I was asked to forecast the sales of music albums for a large label. As usual, we ran a traditional statistical model and got nice statistics that can fit the model well. We also did some face-to-face interviews and research with some audience and published a book thereafter. We were very confident about the model and strongly believed that it is a one-approach-fits-all model. However, the customer told us that our prediction was wrong. As an MBA graduate, I was convinced that if I could explain 80% of the issues, the model would work. And I would not take into consideration the remaining 20%. Then, I realized that I just provided a prediction that looked good but could not explain everything. The remaining 20% should not be ignored or omitted. Otherwise, the prediction will fail in some cases. Therefore, my understanding is that even though the terms of econometrics may actually be used, they depend very much on what kind of problems they are used to solve. Generally speaking, if there is no structured problem, econometrics cannot explain everything.

In my opinion, a successful combination of machine learning and econometrics is that you can use econometric models with the bootstrap of machine learning. After all, machine learning cannot start from a vacuum, and it also needs a theoretical foundation. For example, we can use theory for bootstrap so as to obtain good results much faster than when working without it.

2.5 D. J. Wu

Why should we integrate those two? In my opinion, we should thank Susan Athey, who has been trying to start the theory about econometrics and machine learning. And I also participated in a similar panel held by the National Bureau of Economic Research (NBER), which was mainly about the integration of econometrics and machine learning. The series of work by Susan Athey and her colleagues will be a good start of reading for you to do theoretical research. All the references mentioned will be of great value to you, and they are worth deep reading. Moreover, I encourage you to think about the underlying theories and read the cutting-edge papers, whereby you can make great progress.

2.5.1 Idea: Both econometrics and machine learning have advantages and disadvantages

Susan Athey has already proved that there are close connections between the two disciplines indeed. Currently, the integration of econometrics and machine learning seems appealing to us. However, there is something that has not been fully discovered or totally developed, and it seems to be a new opportunity for IS researchers to make achievements.

1. Machine learning is good at cross-validation, which can be used for model evaluation. As a model validation technique, cross-validation is used to assess the performance of models in practice, which is really hard for econometrics to accomplish.
2. Econometrics can be used to reveal the fundamental underlying process, which is helpful for grasping the economic structure. However, machine learning cannot help contribute to the fundamentals, due to which you are not satisfied. Computer scientists would like to resort to computers when dealing with changes and new problems. It is necessary for theoretical researchers to know about the fundamental process and underlying relation, which are important for simulation, what-if analysis, and decision-making effect analysis. All these can be helpful for better decision-making. Professor Yong Tan, who has

also joined this panel, is also trying to find out the structure, as well as to understand the fundamental underlying process, which will broaden the horizon of IS researchers.

There are so many important problems, such as health care, government issues, and so on. Can we use both techniques in a new way and in a high level? There are some theories that we do not know even if you do randomized experiments. We want to know about the theories in terms of randomized experiments, and this is what we want to explore.

2.6 Michael Shaw

2.6.1 Idea: Econometrics and machine learning depend on each other

My job is to give everybody a quiz. Take the bank lending process as an example. Assume there are eight individuals with different features, such as height, income, and education level. One person is a tall guy, with high income and high school education. Another person is a tall guy, with low income and undergraduate education. The third one is a tall guy, with medium income and undergraduate education. Therefore, we just want to find out the logical relationship behind the data. Is height relevant with income? Is education relevant with income? Alternatively, is there any other connection among these features? What are the decision causes in the eight individuals? What are the dependent variables? And how can we get a decision tree with dependent variables? It can be inferred that income is the most relevant in the decision tree from the samples above. The tree I showed is not a tree example, but an acrobatic example. So what exactly is machine learning? It is a process that derives insights from a large set of data. We try to understand the distribution, which is a classification. By conducting a survey, we find a pattern, which is fully coincidental to different factors or variables. Then we try to find out the causal effect of a decision. After that, we can evaluate the decision and its effects. So what essentially are we doing with machine learning and econometrics? In general, we tend to develop explanations inside knowledge through certain data. The examples mentioned earlier are very simplistic, and you need to acquire more in-depth knowledge by yourself.

When dealing with millions of data, people have to use machine learning, data mining, as well as a variety of econometric models to discover valuable information, which we have been doing during the past few decades. Take my research with a commercial bank as an example. What we found from the decision tree generated to explain the commercial loaning process for medium-sized companies is that the loan is based on the financial attribute and the risk level of the applicant. The numbers one to five represent the risk level perceived for the small- and medium-sized businesses, with the score one meaning the most secure and five meaning the riskiest levels. In classification, we use one to five as the dependent variables, and the financial characters of the company applying for loan as the independent variables. Consequently, machine learning not only gives you the causal relationship and explanatory power, but also depicts the structure, which is called tree structure. For example, the decision-making process of the most secure company is very short. And we have a longer decision-making process for the less-risky company. Finally, we definitely have a very long decision-making and evaluation process for the riskiest company.

As a result, this process is about generating not only knowledge that we can gain from machine learning but also the knowledge about knowledge, which is called meta-knowledge. When you look at the tree, you can see characteristics about the decision tree. Therefore, you are able to reevaluate the decision and the policy effects. In summary, that is why we need econometrics. We never know whether the tree generated or any other machine learning model generated is stable or reliable. Therefore, we have to do validation. Over the years, we have used many different techniques for validation. I think that is where econometrics can really play an important role. Cross-validation is also necessary. However, the reality is that the use of machine learning usually brings a massive result, which is a trigger for the training. The tree training can make the tree more understandable, which means you need some sort of techniques to provide a confidence level of the validity about the machine learning model generated. There are also some other useful techniques, such as statistics and econometrics.

3 Topic 2: When do econometrics and machine learning need each other?

3.1 Eric Zheng

Michael Shaw has just proposed a very important question. That is, when do machine learning and econometrics

need each other? Maybe we can have more discussions on this topic. We usually see econometrics combined with experiments since identification is really important for econometrics, and randomized experiments can be helpful to address the identification issues. However, it is still not clear why machine learning and econometrics need each other, and when can they be helpful mutually? Let us have the panelists give us their opinions.

3.2 Paulo Goes

3.2.1 Idea: Machine learning and econometrics can be beneficial to each other

I think there may be several ways in which machine learning and econometrics could be used together.

1. Use machine learning models to feed in the components of econometric models. Just as Michael mentioned, the classification results of machine learning can be used as variables in econometric models, with which we can make explanation.
2. Enhance econometric models with machine learning techniques, or enhance machine learning models with econometric techniques. Another approach is to borrow techniques from one of the two areas to enhance those of the other one. Consequently, cross-validation, as a component of machine learning, is usually used when developing an econometric model. Additionally, it is possible to use econometric models when dealing with overfitting in machine learning.
3. Use both machine learning and econometrics for comprehensive understanding of a phenomenon. Sometimes, there may be a tough problem, which is difficult to begin. Then you could use machine learning for coherences, which emerge from exercises such as building a tree, or something similar to a decision rule. Then, you may have some ideas to determine the approach to be used for the causal influence or some kind of analysis.

Therefore, I think that after integration, they can enhance techniques mutually, and they also can be used in a parallel or hierarchical way to solve the problems.

3.3 Yong Tan

3.3.1 Idea 1: Deep learning is a good approach for revealing the correlation between endogenous variables and instrument variables

Discovery and explanation of endogeneity are crucial in writing an econometric paper of high quality. Although instrument variables can be very helpful for that, it is also quite difficult to get the instrument variables, especially when short of variables. Moreover, instrument variables could be somewhat weak, which means the correlation between the endogenous variables and instrument variables is not close. In this situation, instrument variables are incapable of describing the variation in the endogenous variables. Then, it is time to use deep learning as a new method. The actual relationship between endogenous variables and instrument variables cannot be conjectured or assumed easily but can be discovered and caught through deep learning. Consequently, deep learning is able to improve the predictive power of endogenous variables and instrument variables.

3.3.2 Idea 2: Exploitation and exploration are essential to each other

Another example of recent research is how exploitation can help exploration. In the process of exploration, you get out of exploitation. Knowing how exploration actually works and how it improves the chance of exploitation is a very complicated process that involves the methodology of how we actually read and understand the exploration. Then finally, we can achieve exploration. Here, you need an intermediate variable, which is unobservable. This is the gap between where you are and where you can make the exploitation. Traditionally, without assuming the latent variables, it is very hard to capture the entire process. That is the common understanding of deep learning, and we can use the short-term or long-term memory to basically capture the immediate process. Then, instead of just purely using machine learning to train the model parameters, because we do not have the intermediate variable, we actually use the latent variable to develop a tree, which comes from econometric models at the later part. There are actually two different connected processes from deep learning to econometrics. Then the entire model can be achieved, as well as the estimation of the entire parameters on both components. So in my opinion, the methodology of deep learning can be used

to enrich the understanding of the generation of data, as well as that of the whole process.

3.4 Ramnath Chellappa

3.4.1 Idea: We need to know both what and why

With the abundance of theoretical knowledge and the mastery over causal relationships and effects, you will find econometric models to be a good start. Moreover, if the past data generation is consistent with the future data generation, the prediction has been proved to be extremely good, which is perfect. However, it is not always the case. That is the reason why your prediction is not always so good compared to what the model explains. There are basically two fundamental reasons for that. The first one is that the data you are using to generate the predictive model is not consistent with the future data generation. The second reason is that the theory does not apply.

On the one hand, when the underlying process is unperceived, machine learning can help with identification and recognition, and it also can be useful in the process of theory building potentially. If you want to explain why it is, you have to tell what it is first. Something like when Amazon says people who like this would also like something else, there is no reason for why, there is no explanation for it is what it is.

Basically, this is just what we happened to observe in data. When doing explanation, people always use the theory, which contains many constraints. For instance, when doing a movie classification, people assume that thriller movies are similar to horrible movies. This may come from some kind of category, which machine learning techniques are good at and can be of great help. On the other hand, as I have mentioned, econometrics can be helpful to machine learning with the quick converging of process. When the processes are based on economic theories, econometrics helps the process converge quickly, which is followed by good identification and bootstrap.

The last thing I want to explain here is what D.J Wu has mentioned just now. Since more and more machine learning methods and certain practices have been inducted into traditional econometrics, such as the out-sample indexing in the holdout, some sort of prediction seems to be coming. Although you have a prediction model, it does not mean that you can explain everything, even if your model has a high level of prediction accuracy. Prediction is not only to do with machine learning. Increasingly, because of the use of all-sample indexing and holdout samples, as well as doing for particular purposes, it seems to gain some increasing attraction.

3.5 D.J. Wu

3.5.1 Idea: Why do machine learning and econometrics need each other?

From my understanding of the panel, I would like to use the reducible version. The reducible version is to use the computer to write a paper and be accepted by *Information Systems Research* (ISR). That means using machine learning to generate models in a systematic way. Machine learning is very good at generating models systematically. However, it is uncertain whether the cause of inference can be maintained, as Paulo just emphasized. Since each of the two disciplines has strength in models and inference separately, they need each other most of the time.

Can we use machine learning to generate models systematically while maintaining the causal inference? Nobody knows the answer, and it is a new challenge, which you are encouraged to take. Another interesting example is to use machine learning to mimic the decision process of the judges. Sometimes, we think it is a little bit too much to solve the underlying societal problems. For example, in the US, many people are potential crime suspects. The judges have to decide who should go to trial. Imagine, the majority of the people ending up in trial have been found to be innocent. It would have been a waste of resources if the person had been kept in prison earlier. Therefore, the judge is trying to use intuition, experience, and all that kind of stuff in the trial process. This research involves the use of machine learning to mimic the judge, instead of the outcomes of the judge, such as their emotion, their eyes, and many other aspects. Then, a decision is made about who is going to be devoiced. All you have to do is to use machine learning to learn, to train the observables. And then, you use a robot or machine for a better decision. When you challenge yourself with these theory problems, you can finally solve some societal problems. In China, there are also many issues that can be explored, such as health applications, government issues, and so on. Consequently, there are many opportunities for you to make an impact.

3.6 Michael Shaw

3.6.1 Idea: A cocktail approach can be used in the combination of machine learning and econometrics

In modern medicine, there is an approach called the cocktail method, which has been used to make quite a few breakthroughs. This kind of approach combines two or more techniques for a treatment, which contributes

significantly to applicable results. The way we combine machine learning and econometrics is just like a cocktail approach from different disciplinary perspectives. It is not necessary to emphasize that they are not equal because they originate from different research foundations. Econometrics seems familiar to us since it has been around for a while. Machine learning comes from a tradition that started from the so-called symbolic pattern recognition, and it has already combined with statistical pattern recognition. Both engineers and scientists want to solve problems as much as possible. It should not be overemphasized that these are two different approaches. Let us look at the problem at hand and try to solve the problems with reliable results that can validate theoretical hierarchical hypotheses, thus laying out the path for important scientific discovery in this way. I always think of them as research approaches for whatever research end I am trying to reach. We should also notice the pros and cons of different approaches. In this case, econometrics obviously has a strong foundation that we are all familiar with. What will be the foundation and benefits of using machine learning? Specifically, with symbolic processing capabilities, machine learning gives you more structured information, as well as texture information. You know a little bit about the content, and perhaps it provides explanation from a different angle. Consequently, the combination of the two can be quite useful. For example, you want to provide a decision maker, no matter what the field is, with the result of your analysis. With pure econometrics, most of the time in statistics, the base is a numerical base. With machine learning, you can provide explanation from many different angles, and you are able to provide the insightful information for the decision maker involved. I do think that the current high-level machine learning can prevent the shortcoming. Once it is actually used in any industry, you can generate some decision by scientists, and then the decision will lead to other decisions. It does not converge to a visible solution, but explores because of the unpredictable results that nobody can explain, which would cause problems. As a result, I think using different approaches to help each other out and then understanding the underlying decision-making process is important, not just relying on the predictive ability or the explanatory power. Moreover, I do appreciate such a combination of various approaches where you can be satisfied with the feeling that it is a very nice approach.

4 Topic 3: How to get prepared for the combination of econometrics and machine learning as researchers?

4.1 Eric Zheng

In this panel, we have not only the panelists, but also a great group of editors. On behalf of the audience, I want to ask them what they would like to recommend to authors here regarding how they could prepare themselves for research that combines econometrics with machine learning. What they should and should not do?

4.2 Olivia Sheng

4.2.1 Idea: It is necessary to master domain knowledge and use tools skillfully

In my opinion, econometrics and machine learning are just methods, and their routes are always joined together with statistics upstream. Within computer science, there is inductive learning. Besides, there are predictive methods, or methods that can make predictions or discover some common patterns that might be taught in data mining classes and machine learning classes, as well as statistical methods. Statistics takes one direction and machine learning takes one direction, while deep learning handles large scales of data in the process. Here, I have some advice and then I will come back with the question. We can use certain tools that we have trained and that can be applied better than the tools that we do not know.

If you truly care about the benefits of combining these two, you must know the tools very well. Since I am trained in machine learning, I have to learn econometrics and many other tools that remain unknown to me. When I apply econometrics to deal with problems, I need to be very cautious. As a serious researcher doing practical research, one has to master the tools, as well as the usage and application methods. Additionally, it is necessary to compare them with other tools and find out the similarities. All of these are challenging, and I can only probably comment on that. When you solve a predictive analysis problem, you either use the machine learning method or use the econometrics method combined with cross-validation; you either evaluate model fit or evaluate predictive errors: all that depends on your goal. Different communities have different methods that can predict better. My expectation is that predictive analysis

researchers from a machine learning background have the attitude of improving performance. However, the field has evolved very long, and getting good performance is not an easy task. If you want to have good performance, you need to know the problem very well, as well as the data, the method, and the process. It is necessary for you to do numerous evaluations and try different things.

I will take the decision tree as an example. Data-driven method already gave me the highest evaluation. If I come from a theoretical background, for credit or for voting, then I will develop a model and hypothesis that are only related to income and so on. These two ways may find the same result. You can use these to do cross-validation very easily. So my suggestion is that we can no longer get away with good performance, we have to explain. It does not mean that all of us need to be very knowledgeable with behavior theory, but at least you have to possess some good domain knowledge, which can help us explain well. I will not stop immediately after I have good performance, and I will go out and find some reasons to help explain. Today, I am paid to do so because I am a consultant, I definitely have to find some explanation because your stakeholders or your clients ask you that your deep learning predicts so well and why, and whether it can predict better and have hierarchical correlation relationship. And benchmark well. As far as I know, the skill of the problems and the skill of the applications are so big that we cannot explain everything about them. We can just explain some features. However, there are so many analyses and predictions that we need to explain. Therefore, there is not only predictive analytics, but we also need to use all kinds of training and methodological backgrounds to help us develop good explanations.

4.3 Alok Gupta

4.3.1 Idea: It is important to choose appropriate methodologies for problems

From my perspective, it is very important to choose the appropriate methodology for the problem. This panel has made some good observations about where they can be enhanced, the gaps that can be filled, and the challenges that econometrics techniques present and that can be overcome. You will see papers that actually do interesting things: they do not actually include these things in the modeling, but when they are looking at or doing the estimations, they show all the explanations. So if you cannot do it theoretically, you can use machine learning techniques, because in my view, it is not true that machine learning does not provide explanations; machine learning models do provide explanations, they provide explanations through logic, you know, through logic or business domain knowledge. Therefore, you can actually provide good explanations. The question is whether you can interpret them with theoretical explanations. When you are using them, it is important whether or not you can explain why and what it is. One of the things related to the pruning of trees I want to mention is that it gives every other exception. It provides opportunities for explanations. The branches that you do can be potentially good opportunities for explanations. They may not have some statistical relation, such as those of the sample, but they do provide you some insight that might be some potential explanations of certain things and it might be some interesting research opportunities. From my perspective, I hope you do all kinds of methodologies.

Acknowledgments: We would like to acknowledge Professor Qinghua Zhu from Nanjing University of China, Professor Ming Fan from University of Washington of the USA, Professor Zhenhui (Jack) Jiang from National University of Singapore and Professor Han Zhang from Georgia Institute of Technology of the USA for their great support to this paper.

References

Judge, G. (2016). Some Comments on the Current State of Econometrics. Annual Review of Resource Economics, 8(1), 1.

Shmueli, Galit, & Koppius, O. (2010). Predictive analytics in information systems research. Social Science Electronic Publishing, 35(3), 553-572.

Susan Athey. (n.d.). Retrieved December 27, 2017, from: https://www.gsb.stanford.edu/faculty-research/faculty/susan-athey

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. Journal of Economic Perspectives, 28(2), 3–28. https://doi.org/10.1257/jep.28.2.3