Hey everybody, I’ve got noon, so let’s get started.

So today I’m pleased to introduce Professor Yiwen Liu. Professor Liu earned her BS and MS in Statistics from the Central University of Finance and Economics in China and her PhD in Statistics from the University of Georgia.

Today, she’s an Assistant Professor of Practice in the Department of Epidemiology and Biostatistics at the Mel and Enid Zuckerberg, Zuckerman College of Public Health at the University of Arizona.

Her research primarily focuses on developing statistical methods and theory to harness a variety of issues in analyzing the high dimensional data or the complex data set. More specifically, her research interests include developing model-free dimension reduction methods, which are high dimensional data regression and integration methods for multiple source data.

Today, she’s gonna talk to us about a model-free variable screening method based on leverage score. Let’s welcome Professor Liu.

Thank you, Robert,
for your nice introduction and it’s my great honor to be invited and present my work here.

So in today’s talk, I will introduce a model-free variable screening method based on leverage score, and we named the method as the weighted leverage score.

So as we know, this is a joint work with Dr. Wenzuan Zhong from the University of Georgia and Dr. Peng Zeng from Auburn University.

So as we know, as we’ve heard there’s big data error, there are numerous data produced almost in every field of science including biology.

So we are facing data extremely with high dimensionality and also data with really complex structures.

Thank you.

And we are facing data of extremely high dimensionality and really complex structures and how do we effectively extract information from such large and complex data.

So to motivate my research, let us see an example first.

So currently cancer has graduated from the primary cause of death across the world.

Nowadays cancer is diagnosed by an expert
who has to look at the tissue samples under
the microscope. You can imagine that there are millions
of new cancer cases each year and this often means that those doctors
will find themselves looking at hundreds of images each day.
And this is really tedious work. And because of, you may find that,
because of the shortage of qualified doctors, there could be a huge lag time
before those doctors can even figure out what is going on with the patient.
So detect cancer using only manpower, looking at images is not enough.
And we intend to build a statistical and mathematical model to identify, detect
cancer in a more accurate, less expensive way. Okay, so the the second generation sequencing
makes this becomes possible and promising. And so a typical research,
critical inference is to find the markers that related to cancer.
Right now there’s new sequencing technology called spatial transcriptomics.
You know that for bulk I sequencing data, it just sequence the whole tissue
and it generate a average the gene expression data.
But with this new technology called spatial transcriptomic,
This kind of cancer tissue will be sliced into several thin sections. And within each section, the grid point in the section will be sequenced simultaneously. So you can see that here we have two areas of invasive cancers, okay, all the dot points within these two sections will be invasive cancer areas for invasive cancer patients. The other six areas, they are noninvasive cancer areas. The grid points in these locations will be noninvasive cancer areas, but for other parts, they’re normal part, okay? And the data that we will have is because this new technology will sequence the whole tissue, all those grid points simultaneously. This data matrix that we will have, each row corresponds to a location within the section and the other columns, each column corresponding to the expressions for certain genes. And the data matrix like this, the Y’s are labels for those patients, the normal noninvasive or invasive. And we will get a gene expressions for all those genes for each location, okay? So this is the data that we have.
and our goal then comes to identify marker genes for those noninvasive and invasive cancer areas.

As showed in this figure, this is the tissue sections.

There are points, color dots here with the color of the dots showing the expression levels.

Okay?

So the dots with a yellow color shows a higher expression.

We intended to build the models to identify such genes.

These genes are show remarkable differential express the levels across issue sections, okay?

These two genes have higher expression in invasive cancer areas.

Okay, we intended to build a status quo model to identify such genes but there exist several challenges here.

Usually the data that we have, the samples or take the locations here is only our label the data is only around hundreds,

but the number of genes could be tens of thousands.

This is so-called a large piece modern problem.

Usually for any traditional methods, there’s no way to utilize those traditional methods to solve this problem.
And the talk mentioned that there is a further layer of complication between the gene expression levels and the cancer or normal types, okay? Usually how the gene expression levels would influence, could affect different types of cancer, this mechanism is largely unknown and the association between them is beyond linear.

So these are the two challenges. That means that we’re going to, what we need is a statistical methods that can do variable screening in a more general model set up. So we, to achieve this goal, we choose to build our efforts under this so called general index model. In a general index model it describes a scenario that $Y_i$ will have relation to pay linear combinations of $X_i$. So that’s beta one transpose $X_i$ to beta K transpose $X_i$ through some anomaly function F. So this is the general index model and we know that here, $X_i$ is a P directional vector and if K is a value that is much smaller than P, then we actually achieved the goal of vanishing reduction.
because the original P directional vector is projected onto a space of a pay dimensional, pay beta one X, beta one transpose X-i to beta eight transpose X-i. And we choose this general index model because it actually is a very general model framework.

If we map this general index model to our problem here, the Y-i could be the label for location i. And, for example, the non-invasive location. Then X-i is a key dimensional vector could be the gene expression levels for location i of those P genes and then beta one transposed X-i to beta k transpose X-i could be those K coregulated gene K groups of coregulated genes. Okay so this is the linear model.
and the error term is additive. So linear model is one special case for it. The number per match model is another special case for the general index model. That is where $K$ equal to $P$ and $\beta_1$ to $\beta_P$, it forms an identity matrix. Thank you. And then the third one, the single index model is another special case for the general index model. That is when $K$ equal to one and that error term is additive. So the reason that I show these three special cases, just to let everyone know that general index model is a very general model framework. In this case, using this model framework to do a variable screening or variable selection, we can say for those, this is determined by the, whether we should screen or whether we should remove certain variables is determined by the coefficients here. Say, for a specific variable, if the coefficient across those $K$, if it’s coefficient across those $K$, that $K$ different factors are all zero, then we say this value, this variable, is redundant.
Okay so this is how we utilize the model in an estimated coefficient to do a variable screening or, say, variable selection. So the question becomes how can we estimate data under this model framework, right? Just like making estimating beta in a simple linear refreshing model. So let’s see a simple case.

That is when F function can be invertible. So that is to say, we have the model becomes F inverse of Y-i equal to beta transposed X-i plus epsilon-i. Using this optimization problem we can recover beta, okay, given that F is invertible and F function is known. And if we want to estimate beta, we can just simply maximize the correlation between inverse of Y-i and the equal transpose X-I.

Using this optimization problem we can recover beta, okay, given that F is invertible and F function is known. But we know in real case, F function is unknown and sometimes it is unconvertible. And then what can we do to estimate beta when F function is unknown? So when F function is unknown, we can consider all the transformations of Y-i and we can solve beta through the following optimization problem.
We consider all transformations of \( Y_i \) we know as \( E(Y_i) \) and we define our square of \( \eta \) which is a function of \( \eta \) as the maximized correlation between \( NA \) tran \( E(Y_i) \) and \( \eta \) transposed \( X_i \).

This maximization is taken over any transformations \( E \), okay?

Using this function, \( \beta \) basically is the solution for this maximization problem and with certain conditions satisfied, we can simplify this objective function with respect to \( \eta \), we can say \( R \) transform this, \( R \) square of \( \eta \) into a this really nice quadratic form.

Okay, in the numerator, it’s \( \eta \) transposed times, okay, this conditional variance times \( \eta \) and in the denominator, it’s either transposed the variance of \( X_i \), \( \eta \).

This is a very nice projected form.

Basically, the solution of this, the solution \( \beta \), is just taking factors, okay,

is just taking factors, okay,
in the sigma X is a P value matrix.

Sigma X given Y is also P value matrix.

And we know that we are dealing with a case of P is larger than N, in this scenario, it would be really difficult to generate a consistent estimate based on a scenario when P is larger than N.

And we also have this inverse here for a very large matrix, it would be really time consuming to produce an inverse of the matrix. That alone, this matrix is not a consistent estimate, okay?

So this matrix in the middle, if we want to estimate that in the P brought up in this scenario, it would be really problematic, right?

And in the following, I will show how we gonna use the weighted leverage score, the method that we proposed to bypass the estimation of these two matrix and then perform the variable selection once again if the reduction under the general index model.

So we call our method the weighted leverage score. Let us first take a look at what is leverage score and what is weighted leverage score, okay?

So let’s consider a simple case that is the linear regression model.
and then the rhD single value competition of X
as X equal to U log to T-transpose.
This is the singular value competition X.
And then, so in statistics,
the leverage score basically is defined as
the diagonal element of the hat matrix.
And then the hat matrix,
we use the further least the singular value competition.
It can be simplified to UU transpose.
And then which means that the diagonal element
or the hat matrix basically is the real norm
of the U matrix, okay?
And then actually this leverage score
has a very good interpretation.
It is the partial directive of Y-i hat with respect to Y-i.
Okay, which means that if the leverage score
is larger and closer to one,
it would be more influential in predicting Y-i hat.
So there is a recent work of Dr. Pima
who’s using this leverage score to do a sub-sampling
in big data.
As you can see that again,
that message here is that if the U-i norm is larger,
if the leverage score is larger
than we say this point is more influential.
Think the motivating example is like this,
in the first figure, this black dots, they are original data and the solid black is the actual model. And if we want to do a linear regression, usually sometimes if the really big data, the data is really large, so it is hardly possible to utilize all the data points to generate the line here. So a typical strategy is just to do sub-sampling from the such big data and then performing a linear regression model. Right now, you can see that the regression line produced by a random sub sample from the population, those data is represented by the screen crosses. So we'll generate a linear regression line that largely deviates from the true model. So that is when the random sampling does not work in this case. However, if we do a sub-sampling according to its leverage score, okay, you will see that, in the second graph, these red crosses on the data sub sample we're using utilize the so-called leverage score and the red dashed line is the model. attempt value in those sub samples, okay? So we can see that using the leverage score to the sub sample can help us to generate a line.
that is very good, can approximate the true model.

So what I want to say using these graphs is that the leverage score, the UI norm, can be used, say, as an indicator of how you fully ensure the data point is to the prediction. Okay so UI is the role norm of the left single matrix but we are talking about variable selection. So UI norm can be used to select the roles. Intuitively, to select the columns of X, we can just do a transpose of X. So X transpose equal to the VAU transpose. To select the columns of X basically is to select the roles of X transpose. Intuitively we can just use the rule map of V matrix, which is the right single matrix to do a selection, to select the influential columns effects, okay, right? And then, so we call the rho nu of U, we call the U as the left singular matrix V as the right singular matrix. And we call the rho nu of U as the left leverage score, rho of B as the right leverage score. So I want to say, use the previous two slides is that basically, the raw information, intuitively, the raw information is contained in the U, the column information of X
is contained in the V matrix and we know that there is a fertile complication between X and Y, which is unknown link function F.

So how do we utilize the information from the column from the rho and also the anomaly function to generate a same method that can help us to the variable selection that is to select influential columns for X.

Okay, let us get back to the matrix we derived in the previous slides. We have the conditional variance in the denominator and we have variance X in the numerator and the variance of X in the denominator.

And with simple statistics, this can be simplified to variants of expectation of Z given Y where Z is a standardized X.

Okay, further we used a singular variety competition, Z can be simplified to UV transpose and then it’s IJ element basically is in the product of Ui and Vj, so Vi basically contains both raw information and column information.

And then we proposed the weighted leverage score, which is defined in this equation.

And the interpretation of the Wj,
which is the weight leverage score for J’s predictor

So first of all you can see it contains both the column information and the raw information of X.

And we know and thus, in the second fold, you can see in the middle, basically contains the information from the unknown function F

because we have the conditional expectation here,

expected of Ui given Y,

basically it’s a kind of a reflection of the anomaly function F.

And third, this method is viewed under the general index model and it is model three,

in the case that the general index model encompasses many different model types.

Okay so this is kind of a population version of the weighted leverage score.

In terms of estimation,

you will see we only need to estimate in the matrix in the middle,

which is the variance of the expectation of Ui given Y.

To estimate this matrix,

there’s a, we can see that Ui is actually three dimensional

because this is a directly single value composition.

Ui is a three dimensional vector.
Y is only one dimensional.
This is a function of one dimensional variable.
So it can be easily approximated by dividing, okay,
the range of Y into h slices as much as h
and within each slice, okay?
Within each slice we calculate the slice mean
for all roles of U.
Lastly, illustrated inn this graph,
we can first the slice into Y into edge slices
and then within each slice, if those Yi
fall into the same slice,
we find out their corresponding use
and do, calculate its mean for each U.
And in that way, we can simplify,
we can simply estimate expectation of Ui given Y.
And then further we can estimate the variance
of those averages.
Basically just taking that the variance of U
one bar
to U edge bar.
So this is the way how we estimate the variance
of the expectation of Ui given Y.
Okay so in that way we actually generate our estimate
weighted leverage score
we define as the right leverage score
weighted by the matrix in the middle.
Okay so first of all, this weighted leverage score
is built, say, upon the general index model,
It is considered as model free because $Y_i$, the response is connected with hitting combination of $X$ through anomaly function $F$.

And this model is general to encompass many different model types. So we can consider it as model free.

And second, this to generate this weighted leverage score where there is no need to estimate the covariance matrix and there is no need to estimate anomaly function $F$. So we can bypass all those procedures to calculate this weighted leverage score.

This weighted leverage score actually encompass a very good feature that is it is an indicator of how influential of the columns are and we can basically run our predictors according to the weighted leverage score. The higher the score is, the more influential the predictor will be. And later I will show why this ranking properties would help or even with the leverage score.

So this is a basic procedures for giving weighted leverage score to a variable selection or variable screening. So given that we have this matrix, which is an impart matrix and we have the responses, the labels, for each of the location,
we only need one time singular value composition, okay?

This is a rank D singular value composition of X.

And then we can just calculate the weighted leverage score

According to the equations, rank those weighted leverage score from the highest to the lowest. Select the predictor that we use, the highest weighted leverage scores.

This is the basic screening procedure using the weighted average score and there is still implementation issue that we will later address. First one, how can we determine the number of D?

So given the data which is an IP, how can we determine, how many, say, spiked or how many singular values to be included in the model?

And then the second implementation issue is determine the number of variables to be selected in the model.

So you can see that the weight leverage score procedure is screening procedure only include one type of singularity, competition is quite efficient.

Okay, in the next, let us using two slides to discuss a little but, basically just one slides to discuss the ranking properties
So as I mentioned, the weighted leverage score has a very nice property. It is guaranteed by the theorem here. We show that, given certain conditions are satisfied, we have the minimum value of the weighted leverage score from the true predictors will always rank higher than the maximum value, maximum weighted leverage score of those redundant predictors. And this holds for the population with leverage score. In terms of the estimation weighted leverage score, we utilize the two step procedure. So first of all, we show that the estimate weighted leverage score is very close to the population version of the weighted leverage. Okay and then the estimated weighted leverage score will also have the ranking property that is the estimated weighted leverage score of the active predictors or important predictors ranks higher than the estimated weighted leverage score of the down predictors with probability turning to one. Okay so this, the ranking properties
of the weighted leverage score. Basically is guaranteed by these two properties.

And then further, we also know that there are two implementation issue. The first one is determine the number of spiked singular values \( d \). So how many single values we need to include in our model?

This is a question and it’s quite crucial because we need to know how many of the signals contain the data and we need to remove all those redundant or noise information. Okay so here I develop a criterion based on the properties of those aken values.

\[ DR = f(R) \] is a function of \( R \) and the \( \theta_1 \) is the ratio between the highest aken value and the largest aken value, number one hat. Then, you can see that as we include more, say, single values in the model, the first term will decrease and then tier to some point. The first, the decreasing of the first term is smaller than the increasing of the second term. Then DR starts to increase. And then we can use the criterion to find \( \hat{D} \)
we show that we had is very close to a true D.
Okay, you meet this criterion,
we can select the true number of signals in the model.
And the second implementation issue
is about how many predictors, how many true predictors
we need to include in our model.
Okay again, we’re ranking our weighted leverage scores
and here we utilize the criterion here based on the properties of the weighted leverage score.
Okay, as we include more predictors into the active set,
the first term will decrease, okay?
The summation of the weighted leverage score will increase,
but the increment will decrease
and then the second term will increase, okay?
So, as we include more predictors in the model,
there’s some changing point
when the increment is smaller than the increment
of the second penalty term.
We show that, using this criteria,
the set we selected using this criteria,
which is A, will always we’ll say can include all the true predictors with probability pending to one.
Okay so that’s how we using this criterion
to determine the number of predictors
to be selected in the model.
Okay in the next step, let me show some empirical study results of using the weighted leverage score to do the variable selection in the model. In the example one, as I mentioned, we are utilizing, we are proposing our method under the general index model framework. So the first model is the general index model. Why there’s two directions? The first one is in the numerator, so this is so called a beta one transpose X. The second direction beta two transpose X is in the variable system, the term within the variant, this is beta two transpose X. Okay so this is called a general index model and we assume that X is generated from a very normal distribution and we submit our zero and covariant structure back in this way, okay? We let rho equal to 0.5 which will generate a matrix with moderate correlations. So in this way, we generate both X and Y, let’s see how the performance of variable selection give you the weighted leverage score. In our scenarios, we let N equal to 1000 and the rho equal 0.5. In example one, there are four different scenarios.
For scenario one, we let $P$ as 200 and then we increase $P$ to 200, sorry 22,000 and 2,500.

We also increase the variance of the error term as 1.5 this now to 1.3.

Okay, there are three criteria we used to evaluate the performance of the method. The first one is the false positive. False negative which shows how many variables are falsely selected?

False negative which shows how many variables falsely excluded, how many true predictors are falsely excluded.

The last one, the last criterion is because, is basically is our model size because we have this ranking properties of all the methods here, okay?

You will see that weighted leverage score is basically have a better performance in terms of the false positive and the false negative.

I want to say a bit more about model size.

We can see that when $N$ is 1000, $P$ is 200, the minimum model size is 6.14, meaning that we, in total,
we only have six true predictors, okay?
We only need to include the six variables
to encompass all true predictors.
Basically all those six variables are rank higher than all our other novel variables, right?
In a second model, when \( P \) increases to 2,200, we only need seven predictors
in order to, on average, in order to include all the true predictors.
Meaning that overall, all the true predictors ranks higher than those redundant predictors.
And then when \( \sigma \) increases and when \( P \) increases, we still need only the minimum number of variables
just to include all true predictors.
Okay so this is for the first example
of the model index model.
The second is more challenging, it’s called a heteroscedastic model.
You will find that, in the first model, the active predictors only influence \( Y \) in its variance.
Okay so this is for the first example
of the model index model.
The second is more challenging, it’s called a heteroscedastic model.
You will find that, in the first model, the active predictors only influence \( Y \) in its variance.
But here, you can see four different types
for those variables, the average of \( Y \),
the mean of \( Y \) is zero
because error term is in the numerator.
and those \( X \), those active predictors will influence \( Y \)
in its variance.
So it’s a much challenging case.

In this case, we also assume that $X$ follows a very normal distribution, given mean is zero and a covariant structure like this.

In our scenarios, we let $N$ equal to 1,000.

So let’s see, in a heteroscedastic model, what are the behaviors of our methods?

Okay, sorry, I forget to introduce the method that we are compare these with true independence ranking and screening and the distance correlation. So both of these methods can be utilized to measure, say, the association between the response and the predictors.

And both of the methods have the ranking properties.

So we can compare the minimum model size for all the methods.

Regards the false positive of these two methods, there’s no criteria proposed to select the number of predictors.

And in all those methods, I just use a harder stretch holding in order to be included in the model.

Okay let’s see, in a heteroscedastic model, what are the behaviors of those three methods?

Okay, when we have $N$ greater the number of predictors,
Here, you will find that there are slightly larger, with false negative for the weighted leverage score. This is because both the methods within the threshold, they select around 140, more than 140, true predictors out of 200 from the model. So that’s why they have very small false negative, but if you look at the minimum model size, you will find that our weighted leverage score still maintains a very good performance. Okay, it has a smaller value of the minimum model size. Okay in general, we only need 46 variables in order to include our two predictors in the model. And then as P diverges, as CSP increased to 2,500, basically the weighted leverage score will measure 1.3 variable true predictors from the model and every method have a really hard time to identify all the true predictors. They have really large minimum model size. So this is basic a performance of the using with the leverage score to perform a variable screening under general index model. So I only present two examples here for interest in odd scenarios. We can talk about that at the top.
Okay so let’s get back to our real data example.

So in a motivating example, as I mentioned, we utilize this spatial transcriptomics data. We are sequencing the grid point within each section, okay?

Basically these locations are invasive cancer areas, the other areas are the noninvasive cancer areas, and then these are the normal areas.

How to determine the invasive, non-invasive and the normal area? These are determined by qualified doctors and they’re utilizing some logical information of these locations.

Okay in general, for these two sections, we have identified 518 locations, 64 invasive areas and 73 are noninvasive areas. And there are rest of the areas, 381, they are normal.

And we have our gene, about 3,572 expressions, gene expressions across the section.

Okay so in general, basically we have our data matrix it is about 518 times 3,572.

So we trying to identify biomarkers, okay, within those three genes that can help us discriminate between invasive cancer, non-invasive cancer and normal areas.

So we utilize the weighted leverage score,
we apply the weight leverage score screening procedure for this data set. And we identified around 225 genes among all those P genes. In the plot, a heat map here show the results because just for the ease of presentation, I only printed around 20 genes here and with the top, say, weighted leverage scores, you can see that there are certain patterns here. This group of genes are more highly expressed for the non-invasive cancer area. There are certain group of genes right here. They are more highly expressed in the invasive cancer areas. Okay so this is the gene expression patterns of those top 20 genes. And then we also plot the expressions of those genes in these sections. Again, these are invasive areas, noninvasive and the normal areas. So we plot a group of genes, I can’t remember exactly what our genes are, but these genes have, you can see, have a higher expression on those noninvasive cancer areas. And we plug another group of genes, basically are these three genes, in the section, the expression shows a higher,
this means that these three genes have higher expression,

okay, in the invasive cancer areas, okay?

Basically this means that the genes that we selected

show a remarkable spatially differential expressed patterns

across the tissue sections.

And later we do a, say, pathway analysis and see that there are 47 functional classes for those all those gene that we have identified.

And there are several cancer hallmarks, for example,

38 of the genes that we identified enriched in the regulation of apoptotic process.

This is a kind of cancer hallmark.

And then another 41 gene that we have identified are involved in the regulation of cell death.

More specifically, because we are really interested in the invasive cancer,

so we identified these three, there are like three genes for example,

in the regulation of brain process,

they have many relations with the breast cancer, okay?

And later we can investigate or, say,

even adaptation of those, those genes

that are enriched in the revelation of apoptotic process.

So, in summary, that weighted leverage score
772 00:44:36.150 --> 00:44:39.767 that we have developed is a variable screening method

773 00:44:39.767 --> 00:44:43.613 and it is developed under the general index model.

774 00:44:44.613 --> 00:44:48.628 And this a very general model framework.

775 00:44:48.628 --> 00:44:52.120 It can be used to address the curse of dimensionality

776 00:44:52.120 --> 00:44:54.625 in regression and also,

777 00:44:54.625 --> 00:44:58.607 because we utilize both the leverage score,

778 00:44:58.607 --> 00:45:01.157 the left leverage score and the right leverage score

779 00:45:01.157 --> 00:45:03.927 to evaluate a predictor’s importance

780 00:45:03.927 --> 00:45:06.960 in the general index model,

781 00:45:06.960 --> 00:45:10.496 we provide a theoretical underpinning

782 00:45:10.496 --> 00:45:14.336 to that objectify that you need both the leverage scores

783 00:45:14.336 --> 00:45:16.260 of both the left and right leverage scores,

784 00:45:16.260 --> 00:45:18.930 we can evaluate the predict’s importance.

785 00:45:18.930 --> 00:45:21.090 Okay so this is kind of a new framework

786 00:45:21.090 --> 00:45:24.675 for analyzing those numerical properties,

787 00:45:24.675 --> 00:45:28.530 especially for the single matrixes

788 00:45:28.530 --> 00:45:30.093 under the general index model.

789 00:45:31.590 --> 00:45:34.097 Okay so, this is basically a summary

790 00:45:34.097 --> 00:45:36.150 of the weighted leverage score

791 00:45:36.150 --> 00:45:41.150 and I wanna stop here and to see if anyone has any questions

792 00:45:41.610 --> 00:45:43.983 or comments about weighted leverage score.

793 00:45:56.280 --> 00:45:57.330 <v Robert>Questions? </v>

794 00:45:59.850 --> 00:46:01.973 Anybody on Zoom have questions?

795 00:46:03.855 --> 00:46:05.040 <v Student>Can I ask a quick question</v>

796 00:46:05.040 --> 00:46:07.440 regarding this weighted leverage score?

797 00:46:07.440 --> 00:46:08.730 So when we look at results,

798 00:46:08.730 --> 00:46:09.900 this weighted leverage score
799 00:46:09.900 --> 00:46:11.130 has much better performance
800 00:46:11.130 --> 00:46:14.430 with less inverse regression regional one, right?
801 00:46:14.430 --> 00:46:17.340 So I wonder, is this correct
802 00:46:17.340 --> 00:46:19.710 that the reason why improves so much
803 00:46:19.710 --> 00:46:22.830 is because it utilize the information on the line,
804 00:46:22.830 --> 00:46:25.710 maybe, total make sense in a lot of applications
805 00:46:25.710 --> 00:46:28.020 that those important features,
806 00:46:28.020 --> 00:46:31.230 they may be like more contributing
807 00:46:31.230 --> 00:46:34.770 to like also leading to like variation of other features
808 00:46:34.770 --> 00:46:39.151 and as a result, maybe could show up in the top
809 00:46:39.151 --> 00:46:43.860 as vectors in the design matrix.
810 00:46:43.860 --> 00:46:44.853 Is this correct?
811 00:46:45.690 --> 00:46:47.237 <v Professor Liu>Yeah, thank you very much</v>
812 00:46:47.237 --> 00:46:49.020 for your question, it’s a very good question.
813 00:46:49.020 --> 00:46:52.263 So first of all, I want to clarify,
814 00:46:53.160 --> 00:46:55.023 maybe I’m not very clear about SIRS.
815 00:46:55.950 --> 00:46:58.140 Basically this is representing
816 00:46:58.140 --> 00:47:00.930 the true independence ranking and screening.
817 00:47:00.930 --> 00:47:02.250 So it’s also a method
818 00:47:02.250 --> 00:47:05.966 that is based on the slicing versus regression.
819 00:47:05.966 --> 00:47:10.440 So yeah, and the other one that why with the leverage score
820 00:47:10.440 --> 00:47:12.007 has a much better performance
821 00:47:12.007 --> 00:47:15.480 comparing these two methods is basically
822 00:47:15.480 --> 00:47:18.810 because, one of the reason is because
823 00:47:18.810 --> 00:47:22.513 the true independent screening and the distance correlation,
they all just utilize the partial and partial correlation. So between X and Y. Okay so it does not utilize any of the information within X. It’s kind of a marginal correlation between each variable X and then one, okay? However, the weighted leverage score will utilize both the raw information and also the variance information, the correlation structure within the model, which is the V matrix, as I mentioned, is derived from the covariance structure of X. The V matrix basically is the vector of the covariance structure, covariance of X. So it utilize the, say, kind of a correlation between all the X variables and a variable screening. So I’m not sure if this answers your question.

Yeah, thank you.

I think it is, you answered my question. Essentially, I’m thinking like if those important features, they are actually not the top contributors to the top other lectures, then we wouldn’t expect the weighted leverage score to aim true way.

Thank you.

Any other questions?

(students mumbling)
851 00:48:56.280 --> 00:48:58.050 <v Vince>Can I ask naive question?</v>
852 00:48:58.050 --> 00:48:59.910 <v Professor Liu>Yes, Vince.</v>
853 00:48:59.910 --> 00:49:03.660 <v Vince>So I’m wondering kind of, you know,</v>
854 00:49:03.660 --> 00:49:05.640 when I think about doing SVP on data,
855 00:49:05.640 --> 00:49:08.921 the first thing that I think of is easier
856 00:49:08.921 --> 00:49:12.870 and I keep coming back to that,
857 00:49:12.870 --> 00:49:15.563 and I can’t tell if there’s a relationship?
858 00:49:16.560 --> 00:49:18.720 <v Professor Liu>Yeah, basically, right,</v>
859 00:49:18.720 --> 00:49:22.105 we can generate U and V in many ways, right?
860 00:49:22.105 --> 00:49:23.986 We can using regression model,
861 00:49:23.986 --> 00:49:26.790 we can generate the U and V as well, right?
862 00:49:26.790 --> 00:49:28.803 We can generate, we can do a,
863 00:49:30.090 --> 00:49:31.980 so it’s basically, I think a lot of inhibition
864 00:49:31.980 --> 00:49:34.230 is that the right score can generate the left and right
865 00:49:34.230 --> 00:49:36.600 singular vectors, so we can use many different ways
866 00:49:36.600 --> 00:49:37.473 to generate that.
867 00:49:38.940 --> 00:49:40.453 Yeah, it’s not really.
868 00:49:40.453 --> 00:49:41.367 <v Robert>Thank you for that.</v>
869 00:49:41.367 --> 00:49:43.320 All right, well then, if there’s nothing further,
870 00:49:43.320 --> 00:49:45.470 let’s thank the teacher again.
871 00:49:45.470 --> 00:49:46.770 <v Professor Liu>Thank you everyone for having me on.</v>
872 00:49:50.803 --> 00:49:54.720 (students overlapping chatter)