WEBVTT

1 00:00:01.110 --> 00:00:04.980 <v Robert>Hey everybody, I've got noon,</v>

2 00:00:04.980 --> 00:00:06.120 so let's get started.

3 00:00:06.120 --> 00:00:09.627 So today I'm pleased to introduce Professor Yiwen Liu.

4 00:00:10.500 --> 00:00:13.260 Professor Liu earned her BS and MS in Statistics

5 00:00:13.260 --> 00:00:16.320 from the Central University of Finance and Economics

6 00:00:16.320 --> 00:00:19.110 in China and her PhD in Statistics

7 00:00:19.110 --> 00:00:20.940 from the University of Georgia.

8 00:00:20.940 --> 00:00:22.920 Today, she's an Assistant Professor of Practice

9 00:00:22.920 --> 00:00:25.890 in the Department of Epidemiology and Biostatistics

10 00:00:25.890 --> 00:00:29.040 at the Mel and Enid Zuckerberg, Zuckerman, sorry,

11 00:00:29.040 --> 00:00:32.243 College of Public Health at the University of Arizona.

12 00:00:32.243 --> 00:00:34.950 Her research primarily focuses on developing

13 00:00:34.950 --> 00:00:36.540 statistical methods and theory

14 00:00:36.540 --> 00:00:39.150 to harness a variety of issues in analyzing

15 00:00:39.150 --> 00:00:43.260 the high dimensional data or the complex data set.

16 00:00:43.260 --> 00:00:45.540 More specifically, her research interests

17 00:00:45.540 --> 00:00:48.300 include developing model-free dimension reduction methods,

18 00:00:48.300 --> 00:00:50.280 which are high dimensional data regression

19 00:00:50.280 --> 00:00:53.100 and integration methods for multiple source data.

20 00:00:53.100 --> 00:00:54.480 Today, she's gonna talk to us

21 00:00:54.480 --> 00:00:56.940 about a model-free variable screening method

22 00:00:56.940 --> 00:00:58.590 based on leverage score.

23 00:00:58.590 --> 00:00:59.990 Let's welcome Professor Liu.

24 00:01:03.960 --> 00:01:04.793 <v Professor Liu>Thank you, Robert</v>

25 00:01:04.793 --> 00:01:07.800 for your nice introduction and it's my great honor

26 00:01:07.800 --> 00:01:12.060 to be invited and present my work here.

27 00:01:12.060 --> 00:01:16.020 So in today's talk, I will introduce a model-free

28 00:01:16.020 --> 00:01:18.720 variable screening method based on leverage score,

29 00:01:18.720 --> 00:01:22.290 and we named the method as the weighted leverage score.

30 00:01:22.290 --> 00:01:24.480 So as we know, this is a joint work

31 00:01:24.480 --> 00:01:27.480 with Dr. Wenxuan Zhong from the University of Georgia

32 00:01:27.480 --> 00:01:30.633 and Dr. Peng Zeng from Auburn University.

33 00:01:31.530 --> 00:01:35.700 So as we know, as we've heard there's big data error,

34 00:01:35.700 --> 00:01:39.750 there are numerous data produced almost in every field

35 00:01:39.750 --> 00:01:42.030 of science including biology.

36 00:01:42.030 --> 00:01:47.030 So we are facing data extremely with high dimensionality

37 00:01:47.610 --> 00:01:51.457 and also data with really complex structures.

38 00:01:55.669 --> 00:01:56.502 Thank you.

39 00:01:57.776 --> 00:02:00.450 And we are facing data of extremely high dimensionality

40 00:02:00.450 --> 00:02:02.320 and really complex structures

41 00:02:03.780 --> 00:02:06.750 and how do we effectively extract information

42 00:02:06.750 --> 00:02:10.020 from such large and complex data

43 00:02:10.020 --> 00:02:12.120 pose new statistical challenge.

44 00:02:12.120 --> 00:02:17.120 So to motivate my research, let us see an example first.

45 00:02:17.700 --> 00:02:19.710 So currently cancer has graduated

46 00:02:19.710 --> 00:02:23.103 from the primary cause of death across the world.

47 00:02:24.270 --> 00:02:27.510 Nowadays cancer is diagnosed by an expert

48 00:02:27.510 --> 00:02:31.384 who has to look at the tissue samples under the microscope.

49 00:02:31.384 --> 00:02:33.300 You can imagine that there are millions

50 00:02:33.300 --> 00:02:35.610 of new cancer cases each year

51 00:02:35.610 --> 00:02:39.120 and this often means that those doctors

52 00:02:39.120 --> 00:02:43.800 will find themselves looking at hundreds of images each day.

53 00:02:43.800 --> 00:02:48.150 And this is really tedious work.

54 00:02:48.150 --> 00:02:51.900 And because of, you may find that,

55 00:02:51.900 --> 00:02:54.870 because of the shortage of qualified doctors,

56 00:02:54.870 --> 00:02:56.580 there could be a huge lag time

57 00:02:56.580 --> 00:02:59.100 before those doctors can even figure out

58 00:02:59.100 --> 00:03:00.800 what is going on with the patient.

59 00:03:02.130 --> 00:03:05.130 So detect cancer using only manpower,

60 00:03:05.130 --> 00:03:07.821 looking at images is not enough.

61 00:03:07.821 --> 00:03:11.520 And we intend to build a statistical

62 00:03:11.520 --> 00:03:15.963 and mathematical model to identify, detect cancer

63 00:03:15.963 --> 00:03:19.357 in a more accurate, less expensive way.

64 00:03:22.104 --> 00:03:25.170 Okay, so the the second generation sequencing

65 00:03:25.170 --> 00:03:28.680 makes this becomes possible and promising.

66 00:03:28.680 --> 00:03:32.730 And so a typical research,

67 00:03:32.730 --> 00:03:35.790 critical inference is to find the markers

68 00:03:35.790 --> 00:03:37.590 that related to cancer.

69 00:03:37.590 --> 00:03:39.420 Right now there's new sequencing technology

70 00:03:39.420 --> 00:03:41.797 called spatial transcriptomics.

71 00:03:41.797 --> 00:03:45.600 You know that for bulk I sequencing data,

72 00:03:45.600 --> 00:03:47.490 it just sequence the whole tissue

73 00:03:47.490 --> 00:03:51.000 and it generate a average the gene expression data.

74 00:03:51.000 --> 00:03:54.780 But with this new technology called spatial transcriptomic,

75 00:03:54.780 --> 00:03:57.420 this kind of cancer tissue will be sliced

76 00:03:57.420 --> 00:04:00.213 into several thin sections.

77 00:04:01.320 --> 00:04:05.940 And within each section, the grid point in the section

78 00:04:05.940 --> 00:04:08.250 will be sequenced simultaneously.

79 00:04:08.250 --> 00:04:11.430 So you can see that here we have two areas

80 00:04:11.430 --> 00:04:13.500 of invasive cancers, okay,

81 00:04:13.500 --> 00:04:16.530 all the dot points within these two sections

82 00:04:16.530 --> 00:04:21.420 will be invasive cancer areas for invasive cancer patients.

83 00:04:21.420 --> 00:04:25.320 The other six areas, they are noninvasive cancer areas.

84 00:04:25.320 --> 00:04:29.100 The grid points in these locations

85 00:04:29.100 --> 00:04:32.071 will be noninvasive cancer areas,

86 00:04:32.071 --> 00:04:35.220 but for other parts, they're normal part, okay?

87 00:04:35.220 --> 00:04:36.701 And the data that we will have

88 00:04:36.701 --> 00:04:39.510 is because this new technology

89 00:04:39.510 --> 00:04:41.763 will sequence the whole tissue,

90 00:04:42.840 --> 00:04:45.480 all those grid points simultaneously.

91 00:04:45.480 --> 00:04:46.950 This data matrix that we will have,

92 00:04:46.950 --> 00:04:50.220 each row corresponds to a location

93 00:04:50.220 --> 00:04:53.220 within the section and the other columns,

94 00:04:53.220 --> 00:04:55.740 each column corresponding to the expressions

95 00:04:55.740 --> 00:04:57.123 for certain genes.

96 00:04:58.980 --> 00:05:01.380 And the data matrix like this,

97 00:05:01.380 --> 00:05:04.560 the Y's are labels for those patients,

98 00:05:04.560 --> 00:05:07.290 the normal noninvasive or invasive.

99 00:05:07.290 --> 00:05:11.520 And we will get a gene expressions for all those P genes

100 00:05:11.520 --> 00:05:14.610 for each location, okay?

101 00:05:14.610 --> 00:05:17.277 So this is the data that we have

102 00:05:17.277 --> 00:05:22.277 and our goal then comes to identify marker genes

103 00:05:22.500 --> 00:05:26.544 for those noninvasive and invasive cancer areas.

104 00:05:26.544 --> 00:05:31.544 As showed in this figure, this is the tissue sections.

105 00:05:33.420 --> 00:05:37.170 There are points, color dots here

106 00:05:37.170 --> 00:05:42.045 with the color of the dots showing the expression levels.

107 00:05:42.045 --> 00:05:42.878 Okay?

108 00:05:43.920 --> 00:05:47.490 So the the dots with a yellow color

109 00:05:47.490 --> 00:05:48.903 shows a higher expression.

110 00:05:49.920 --> 00:05:51.840 We intended to build the models

111 00:05:51.840 --> 00:05:55.200 to identify such genes.

112 00:05:55.200 --> 00:05:57.623 These genes are show remarkable

113 00:05:57.623 --> 00:06:02.623 differential express the levels across issue sections, okay?

114 00:06:03.090 --> 00:06:05.820 These two genes have higher expression

115 00:06:05.820 --> 00:06:08.103 in invasive cancer areas.

116 00:06:10.470 --> 00:06:12.510 Okay, we intended to build a status quo model

117 00:06:12.510 --> 00:06:14.490 to identify such genes

118 00:06:14.490 --> 00:06:17.220 but there exist several challenges here.

119 00:06:17.220 --> 00:06:19.860 Usually the data that we have,

120 00:06:19.860 --> 00:06:21.720 the samples or take the locations here

121 00:06:21.720 --> 00:06:25.271 is only our label the data is only around hundreds,

122 00:06:25.271 --> 00:06:29.704 but the number of genes could be tens of thousands.

123 00:06:29.704 --> 00:06:33.543 This is so-called a large piece modern problem.

124 00:06:35.010 --> 00:06:38.100 Usually for any traditional methods,

125 00:06:38.100 --> 00:06:40.680 there's no way to utilize those traditional methods

126 00:06:40.680 --> 00:06:42.273 to solve this problem.

127 00:06:43.950 --> 00:06:46.140 And the talk mentioned that there is a further layer

128 00:06:46.140 --> 00:06:49.410 of complication between the gene expression levels

129 00:06:49.410 --> 00:06:52.800 and the cancer or normal types, okay?

130 00:06:52.800 --> 00:06:56.310 Usually how the gene expression levels

131 00:06:56.310 --> 00:06:59.730 would influence, could affect different types of cancer,

132 00:06:59.730 --> 00:07:01.940 this mechanism is largely unknown

133 00:07:01.940 --> 00:07:05.730 and the association between them is beyond linear.

134 00:07:05.730 --> 00:07:09.000 So these are the two challenges.

135 00:07:09.000 --> 00:07:11.193 That means that we're going to,

136 00:07:12.180 --> 00:07:15.303 what we need is a statistical methods

137 00:07:15.303 --> 00:07:17.680 that can do variable screening

138 00:07:18.900 --> 00:07:21.303 in a more general model set up.

139 00:07:25.980 --> 00:07:29.310 So we, to achieve this goal,

140 00:07:29.310 --> 00:07:32.040 we choose to build our efforts

141 00:07:32.040 --> 00:07:35.363 under this so called general index model.

142 00:07:35.363 --> 00:07:40.050 In a general index model it describes a scenario that Y-i

143 00:07:40.050 --> 00:07:41.254 which is the response

144 00:07:41.254 --> 00:07:46.254 will have relation to pay linear combinations of X-i.

145 00:07:46.890 --> 00:07:49.230 So that's beta one transpose X-I

146 00:07:49.230 --> 00:07:51.330 to beta K transpose X-I

147 00:07:51.330 --> 00:07:53.643 through some anomaly function F.

148 00:07:54.960 --> 00:07:57.480 So this is the general index model

149 00:07:57.480 --> 00:08:02.480 and we know that here, X-i is a P directional vector

150 00:08:03.480 --> 00:08:07.893 and if K is a value that is much smaller than P,

151 00:08:09.090 --> 00:08:13.320 then we actually achieved the goal of vanishing reduction

152 00:08:13.320 --> 00:08:16.213 because the original P directional vector

153 00:08:16.213 --> 00:08:21.213 is projected onto a space of a pay dimensional,

154 00:08:21.570 --> 00:08:24.300 pay beta one X, beta one transpose X-i

155 00:08:24.300 --> 00:08:25.713 to beta eight transpose X-i.

156 00:08:27.360 --> 00:08:29.730 And we choose this general index model

157 00:08:29.730 --> 00:08:33.877 because it actually is a very general model framework.

158 00:08:33.877 --> 00:08:38.460 If we map this general index model to our problem here,

159 00:08:38.460 --> 00:08:42.003 the Y-i could be the label for location i.

160 00:08:43.080 --> 00:08:47.223 And, for example, the non-invasive location.

161 00:08:48.070 --> 00:08:49.830 And then X-i is a key dimensional vector

162 00:08:49.830 --> 00:08:52.177 and could be the gene expression levels

163 00:08:52.177 --> 00:08:54.760 for location i of those P genes

164 00:08:56.610 --> 00:08:58.470 and then beta one transposed X-i

165 00:08:58.470 --> 00:09:00.660 to beta k transpose X-i

166 00:09:00.660 --> 00:09:04.800 could be those K coregulated gene K groups

167 00:09:04.800 --> 00:09:06.453 of coregulated genes.

168 00:09:07.320 --> 00:09:10.140 And those K groups of coregulated genes

169 00:09:10.140 --> 00:09:14.313 will affect the response through some anomaly function F.

170 00:09:16.350 --> 00:09:19.803 Okay so this is our general model setup.

171 00:09:24.480 --> 00:09:26.370 We utilize the general index model

172 00:09:26.370 --> 00:09:29.430 because it's a general model framework

173 00:09:29.430 --> 00:09:32.746 that encompasses many different model types.

174 00:09:32.746 --> 00:09:35.340 There is three special cases,

175 00:09:35.340 --> 00:09:39.840 for example, the linear model is one special case.

176 00:09:39.840 --> 00:09:44.840 Here, KY-i, that is when K equals one

177 00:09:44.940 --> 00:09:49.940 and F anomaly function acquires an identity form.

178 00:09:50.190 --> 00:09:51.660 Okay so this is the linear model

179 00:09:51.660 --> 00:09:53.853 and the error term is additive.

180 00:09:54.954 --> 00:09:58.380 So linear model is one special case for it.

181 00:09:58.380 --> 00:10:01.320 The number per match model is another special case

182 00:10:01.320 --> 00:10:02.850 for the general index model.

183 00:10:02.850 --> 00:10:06.960 That is where K equal to P and beta one to beta P,

184 00:10:06.960 --> 00:10:09.483 it forms an identity matrix.

185 00:10:12.270 --> 00:10:13.860 Thank you.

186 00:10:13.860 --> 00:10:15.000 And then the third one,

187 00:10:15.000 --> 00:10:16.230 the single index model

188 00:10:16.230 --> 00:10:20.520 is another special case for the general index model

189 00:10:20.520 --> 00:10:22.364 that is when K equal to one

190 00:10:22.364 --> 00:10:24.963 and that error term is additive.

191 00:10:26.190 --> 00:10:28.440 So the reason that I show these three special cases,

192 00:10:28.440 --> 00:10:31.680 just to let everyone know that general index model

193 00:10:31.680 --> 00:10:34.773 is a very general model framework.

194 00:10:35.910 --> 00:10:39.150 In this case, using this model framework

195 00:10:39.150 --> 00:10:43.323 to do a variable screening or variable selection,

196 00:10:45.000 --> 00:10:49.650 we can say for those, this is determined by the,

197 00:10:49.650 --> 00:10:51.360 whether we should screen

198 00:10:51.360 --> 00:10:53.730 or whether we should remove certain variables

199 00:10:53.730 --> 00:10:55.863 is determined by the coefficients here.

200 00:10:56.820 --> 00:11:01.820 Say, for a specific variable, if the coefficient beta one,

201 00:11:01.863 --> 00:11:04.323 if it's coefficient across those K,

202 00:11:04.323 --> 00:11:07.770 that K different factors are all zero,

203 00:11:07.770 --> 00:11:10.983 then we say this value, this variable, is redundant.

204 00:11:12.990 --> 00:11:15.633 Okay so this is how we utilize the model

205 00:11:15.633 --> 00:11:18.360 in a estimated coefficient

206 00:11:18.360 --> 00:11:21.183 to do a variable screening or, say, variable selection.

207 00:11:26.760 --> 00:11:30.000 So the question becomes how can we estimate data

208 00:11:30.000 --> 00:11:31.890 under this model framework, right?

209 00:11:31.890 --> 00:11:34.170 Just like made estimating beta

210 00:11:34.170 --> 00:11:36.660 in a simple linear refreshing model.

211 00:11:36.660 --> 00:11:38.640 So let's see a simple case.

212 00:11:38.640 --> 00:11:43.640 That is when F function can is invertible.

213 00:11:43.980 --> 00:11:45.090 So that is to say,

214 00:11:45.090 --> 00:11:48.990 we have the model becomes F inverse of Y-i

215 00:11:48.990 --> 00:11:51.993 equal to beta transposed X-i plus epsilon-i.

216 00:11:53.337 --> 00:11:57.360 Okay and this is very similar, looks similar model, right?

217 00:11:57.360 --> 00:11:59.130 And if we want to estimate beta,

218 00:11:59.130 --> 00:12:02.850 we can just simply maximize the correlation

219 00:12:02.850 --> 00:12:07.440 between inverse of Y-i and the equal transpose X-I.

220 00:12:07.440 --> 00:12:12.440 Using this optimization problem we can recover beta, okay,

221 00:12:12.570 --> 00:12:16.533 given that F is invertible and F function is known.

222 00:12:18.690 --> 00:12:21.190 But we know in real case, F function is unknown

223 00:12:23.700 --> 00:12:26.130 and sometimes it is unconvertible.

224 00:12:26.130 --> 00:12:30.540 And then what can we do to estimate beta

225 00:12:30.540 --> 00:12:31.983 when F function is unknown?

226 00:12:34.470 --> 00:12:36.600 So when F function is unknown,

227 00:12:36.600 --> 00:12:39.600 we can consider all the transformations of Y-i

228 00:12:39.600 --> 00:12:41.278 and we can solve beta

229 00:12:41.278 --> 00:12:45.210 through the following optimization problem.

230 00:12:45.210 --> 00:12:47.940 We consider all transformations of Y-i

231 00:12:47.940 --> 00:12:50.850 we know as E of Y-i

232 00:12:50.850 --> 00:12:54.150 and we define our square of eta

233 00:12:54.150 --> 00:12:58.950 which is a function of eta as the maximized correlation

234 00:12:58.950 --> 00:13:02.913 between NA tran E of Y-i and eta transposed X-i.

235 00:13:04.590 --> 00:13:07.154 And this maximization is taken over

236 00:13:07.154 --> 00:13:12.154 or any transformations E, okay?

237 00:13:13.380 --> 00:13:17.430 So using this function, beta basically is the solution

238 00:13:17.430 --> 00:13:19.680 for this maximization problem

239 00:13:19.680 --> 00:13:23.130 and with certain conditions satisfied,

240 00:13:23.130 --> 00:13:26.730 we can simplify this objective function

241 00:13:26.730 --> 00:13:31.730 with respect to eta, we can say R transform this,

242 00:13:33.510 --> 00:13:37.980 R square of eta into a this really nice quadratic form.

243 00:13:37.980 --> 00:13:42.980 Okay, in the numerator, it's eta transposed times, okay,

244 00:13:43.080 --> 00:13:46.860 this conditional variance times eta

245 00:13:46.860 --> 00:13:49.590 and in the denominator, it's either transposed

246 00:13:49.590 --> 00:13:51.167 the variance of X-i, eta.

247 00:13:52.357 --> 00:13:55.800 This is a very nice projected form.

248 00:13:55.800 --> 00:14:00.800 Basically, the solution of this, the solution beta,

249 00:14:02.083 --> 00:14:06.420 is just taking factors, okay,

250 00:14:06.420 --> 00:14:09.330 corresponding to the pay largest taken values

251 00:14:09.330 --> 00:14:10.863 of this matrix in the middle.

252 00:14:12.960 --> 00:14:15.183 That's how we solve beta in this case.

253 00:14:16.260 --> 00:14:18.330 But as we know, as I mentioned,

254 00:14:18.330 --> 00:14:21.330 that there is a really like big challenges here.

255 00:14:21.330 --> 00:14:24.200 One is that we have really large P here

256 00:14:24.200 --> 00:14:26.702 in the sigma X is a P value matrix.

257 00:14:26.702 --> 00:14:29.763 Sigma X given Y is also P value matrix.

258 00:14:31.200 --> 00:14:33.540 And we know that we are dealing with a case

259 00:14:33.540 --> 00:14:37.650 of P is larger than N, in this scenario,

260 00:14:37.650 --> 00:14:39.720 it would be really difficult

261 00:14:39.720 --> 00:14:43.750 to generate a consistent estimate based on a scenario

262 00:14:45.300 --> 00:14:47.520 when P is larger than N.

263 00:14:47.520 --> 00:14:52.520 And we also have this inverse here for a very large matrix,

264 00:14:52.920 --> 00:14:54.720 it would be really time consuming

265 00:14:54.720 --> 00:14:56.763 to produce an inverse of the matrix.

266 00:14:57.780 --> 00:15:01.800 That alone, this matrix is not a consistent estimate, okay?

267 00:15:01.800 --> 00:15:03.900 So this matrix in the middle,

268 00:15:03.900 --> 00:15:05.580 if we want to estimate that

269 00:15:05.580 --> 00:15:08.312 in the P brought up in this scenario,

270 00:15:08.312 --> 00:15:13.312 it would be really problematic, right?

271 00:15:13.320 --> 00:15:17.514 And in the following, I will show how we gonna use

272 00:15:17.514 --> 00:15:21.330 the weighted leverage score, the method that we proposed

273 00:15:21.330 --> 00:15:24.587 to bypass the estimation of these two matrix

274 00:15:24.587 --> 00:15:27.810 and then perform the variable selection

275 00:15:27.810 --> 00:15:32.163 once again if the reduction under the general index model.

276 00:15:35.250 --> 00:15:37.750 So we call our method the weighted leverage score.

277 00:15:38.910 --> 00:15:43.320 Let us first take a look at what is leverage score

278 00:15:43.320 --> 00:15:46.260 and what is weighted leverage score, okay?

279 00:15:46.260 --> 00:15:49.110 So let's consider a simple case

280 00:15:49.110 --> 00:15:51.120 that is the linear regression model

281 00:15:51.120 --> 00:15:54.540 and then the rhet D single value competition of X

282 00:15:54.540 --> 00:15:58.023 as X equal to U log to T-transpose.

283 00:15:59.280 --> 00:16:02.640 This is the singular value competition X.

284 00:16:02.640 --> 00:16:04.740 And then, so in statistics,

285 00:16:04.740 --> 00:16:08.370 the leverage score basically is defined as

286 00:16:08.370 --> 00:16:11.603 the diagonal element of the hat matrix.

287 00:16:14.429 --> 00:16:15.407 And then the hat matrix,

288 00:16:15.407 --> 00:16:18.360 we use the further least the singular value competition.

289 00:16:18.360 --> 00:16:20.973 It can be simplified to UU transpose.

290 00:16:22.020 --> 00:16:24.750 And then which means that the diagonal element

291 00:16:24.750 --> 00:16:29.430 or the hat matrix basically is the real norm

292 00:16:29.430 --> 00:16:31.253 of the U matrix, okay?

293 00:16:33.840 --> 00:16:37.200 And then actually this leverage score

294 00:16:37.200 --> 00:16:40.380 has a very good interpretation.

295 00:16:40.380 --> 00:16:45.150 It is the partial directive of Y-i hat with respect to Y-i.

296 00:16:45.150 --> 00:16:47.910 Okay, which means that if the leverage score

297 00:16:47.910 --> 00:16:50.730 is larger and closer to one,

298 00:16:50.730 --> 00:16:55.413 it would be more influential in predicting Y-i hat.

299 00:16:56.910 --> 00:17:00.600 So there is a recent work of Dr. Pima

300 00:17:00.600 --> 00:17:05.460 who's using this leverage score to do a subsampling

301 00:17:05.460 --> 00:17:06.962 in big data.

302 00:17:06.962 --> 00:17:08.940 As you can see that again,

303 00:17:08.940 --> 00:17:12.510 that message here is that if the U-i norm is larger,

304 00:17:12.510 --> 00:17:13.980 if the leverage score is larger

305 00:17:13.980 --> 00:17:16.923 than we say this point is more influential.

306 00:17:18.990 --> 00:17:20.780 Think the motivating example is like this,

307 00:17:20.780 --> 00:17:24.505 in the first figure, this black dots,

308 00:17:24.505 --> 00:17:28.330 they are original data

309 00:17:29.310 --> 00:17:33.603 and the solid black is the actual model.

310 00:17:34.680 --> 00:17:36.780 And if we want to do a linear regression,

311 00:17:36.780 --> 00:17:39.210 usually sometimes if the really big data,

312 00:17:39.210 --> 00:17:40.800 the data is really large,

313 00:17:40.800 --> 00:17:44.190 so it is hardly possible to utilize all the data points

314 00:17:44.190 --> 00:17:48.930 to generate the line here.

315 00:17:48.930 --> 00:17:53.730 So a typical strategy is just to do sub-sampling

316 00:17:53.730 --> 00:17:56.160 from the such big data

317 00:17:56.160 --> 00:17:58.460 and then performing a linear regression model.

318 00:17:59.430 --> 00:18:02.435 Right now, you can see that the regression line

319 00:18:02.435 --> 00:18:07.435 produced by a random sub sample from the population,

320 00:18:07.650 --> 00:18:11.640 those data is represented by the screen crosses.

321 00:18:11.640 --> 00:18:16.623 So we'll generate a linear regression line

322 00:18:16.623 --> 00:18:20.493 that largely deviates from the true model.

323 00:18:21.390 --> 00:18:26.390 So that is when the random sampling does not work

324 00:18:26.670 --> 00:18:27.780 in this case.

325 00:18:27.780 --> 00:18:32.580 However, if we do a sub-sampling

326 00:18:32.580 --> 00:18:35.700 according to its leverage score, okay,

327 00:18:35.700 --> 00:18:38.471 you will see that, in the second graph,

328 00:18:38.471 --> 00:18:43.471 these red crosses on the data sub sample we're using

329 00:18:43.710 --> 00:18:47.700 utilize the so-called leverage score

330 00:18:47.700 --> 00:18:51.660 and the red dashed line is the model

331 00:18:51.660 --> 00:18:55.440 attempt value in those sub samples, okay?

332 00:18:55.440 --> 00:18:59.430 So we can see that using the leverage score

333 00:18:59.430 --> 00:19:03.475 to the sub sample can help us to generate a line

334 00:19:03.475 --> 00:19:07.473 that is very good, can approximate the true model.

335 00:19:08.790 --> 00:19:10.380 So what I want to say using these graph

336 00:19:10.380 --> 00:19:14.430 is that the leverage score, the UI norm,

337 00:19:14.430 --> 00:19:18.630 can be used, say, as an indicator

338 00:19:18.630 --> 00:19:22.593 of how you fully ensure the data point is to the prediction.

339 00:19:24.799 --> 00:19:29.799 Okay so UI is the role norm of the left single matrix

340 00:19:31.560 --> 00:19:33.750 but we are talking about variable selection.

341 00:19:33.750 --> 00:19:37.887 So UI norm can be used to select the roles.

342 00:19:37.887 --> 00:19:42.887 Intuitively, to select the columns of X,

343 00:19:43.110 --> 00:19:45.150 we can just do a transpose of X.

344 00:19:45.150 --> 00:19:48.840 So X transpose equal to the VAU transpose.

345 00:19:48.840 --> 00:19:50.530 To select the columns of X

346 00:19:51.870 --> 00:19:55.080 basically is to select the roles of X transpose.

347 00:19:55.080 --> 00:19:59.190 Intuitively we can just use the rule map of V matrix,

348 00:19:59.190 --> 00:20:03.240 which is the right single matrix to do a selection,

349 00:20:03.240 --> 00:20:07.770 to select the influential columns effects, okay, right?

350 00:20:07.770 --> 00:20:11.460 And then, so we call the rho nu of U,

351 00:20:13.200 --> 00:20:16.110 we call the U as the left singular matrix

352 00:20:16.110 --> 00:20:18.180 V as the right singular matrix.

353 00:20:18.180 --> 00:20:23.180 And we call the rho nu of U as the left leverage score,

354 00:20:24.090 --> 00:20:25.983 rho of B as the right leverage score.

355 00:20:29.820 --> 00:20:33.100 So I want to say, use the previous two slides

356 00:20:34.454 --> 00:20:36.090 is that basically, the raw information,

357 00:20:36.090 --> 00:20:39.600 intuitively, the raw information is contained in the U,

358 00:20:39.600 --> 00:20:41.100 the column information of X

359 00:20:41.100 --> 00:20:44.100 is contained in the V matrix

360 00:20:44.100 --> 00:20:46.590 and we know that there is a fertile complication

361 00:20:46.590 --> 00:20:50.310 between X and Y, which is unknown link function F.

362 00:20:50.310 --> 00:20:54.330 So how do we utilize the information from the column

363 00:20:54.330 --> 00:20:57.300 from the rho and also the anomaly function

364 00:20:57.300 --> 00:21:01.405 to generate a same method

365 00:21:01.405 --> 00:21:06.405 that can help us to the variable selection

366 00:21:06.515 --> 00:21:09.483 that is to select influential columns for X.

367 00:21:12.930 --> 00:21:15.750 Okay so, let us get back to the matrix

368 00:21:15.750 --> 00:21:18.060 we derived in the previous slides.

369 00:21:18.060 --> 00:21:21.990 We have the conditional variance in the denominator

370 00:21:21.990 --> 00:21:24.780 and we have variance X in the numerator

371 00:21:24.780 --> 00:21:26.730 and the variance of X in the denominator.

372 00:21:26.730 --> 00:21:29.310 And with simple statistics,

373 00:21:29.310 --> 00:21:32.820 this can be simplified to variants of expectation

374 00:21:32.820 --> 00:21:37.600 of Z given Y where Z is a standardized X.

375 00:21:37.600 --> 00:21:41.292 Okay, further we used a singular variety competition,

376 00:21:41.292 --> 00:21:45.600 Z can be simplified to UV transpose

377 00:21:45.600 --> 00:21:47.820 and then it's IJ element basically

378 00:21:47.820 --> 00:21:51.363 is in the product of Ui and Vj,

379 00:21:52.332 --> 00:21:55.557 so Vi basically contains both raw information

380 00:21:55.557 --> 00:21:57.690 and column information.

381 00:21:57.690 --> 00:22:00.870 And then we proposed the weighted leverage score,

382 00:22:00.870 --> 00:22:05.820 which is defined in this equation.

383 00:22:05.820 --> 00:22:08.730 And the interpretation of the Wj,

384 00:22:08.730 --> 00:22:12.480 which is the weight leverage score for J's predictor

385 00:22:12.480 --> 00:22:13.770 is threefold.

386 00:22:13.770 --> 00:22:15.570 So first of all you can see it contains

387 00:22:15.570 --> 00:22:19.593 both the column information and the raw information of X.

388 00:22:20.535 --> 00:22:25.535 And we know and thus, in the second fold,

389 00:22:27.330 --> 00:22:30.390 you can see in the middle, basically contains

390 00:22:30.390 --> 00:22:32.760 the information from the unknown function F

391 00:22:32.760 --> 00:22:35.670 because we have the conditional expectation here,

392 00:22:35.670 --> 00:22:38.190 expectation of Ui given Y,

393 00:22:38.190 --> 00:22:41.160 basically it's a kind of a reflection

394 00:22:41.160 --> 00:22:44.493 of the anomaly function F.

395 00:22:46.414 --> 00:22:49.440 And third, this method is viewed

396 00:22:49.440 --> 00:22:51.210 under the general index model

397 00:22:51.210 --> 00:22:52.517 and it is model three,

398 00:22:52.517 --> 00:22:55.320 in the case that the general index model

399 00:22:55.320 --> 00:22:57.933 encompasses many different model types.

400 00:22:59.011 --> 00:23:03.570 Okay so this is kind of a population version

401 00:23:03.570 --> 00:23:05.610 of the weighted leverage score.

402 00:23:05.610 --> 00:23:07.467 In terms of estimation,

403 00:23:07.467 --> 00:23:09.294 you will see we only need to estimate

404 00:23:09.294 --> 00:23:11.460 in the matrix in the middle,

405 00:23:11.460 --> 00:23:14.733 which is the variance of the expectation of Ui given Y.

406 00:23:17.149 --> 00:23:19.413 To estimate this matrix,

407 00:23:20.400 --> 00:23:24.030 there's a, we can see that Ui is actually three dimensional

408 00:23:24.030 --> 00:23:27.354 because this is a directly single value composition.

409 00:23:27.354 --> 00:23:28.920 Ui is a three dimensional vector.

410 00:23:28.920 --> 00:23:30.780 Y is only one dimensional.

411 00:23:30.780 --> 00:23:33.480 This is a function of one dimensional variable.

412 00:23:33.480 --> 00:23:38.400 So it can be easily approximated by dividing, okay,

413 00:23:38.400 --> 00:23:42.310 the range of Y into h slices as much as h

414 00:23:43.200 --> 00:23:45.570 and within each slice, okay?

415 00:23:45.570 --> 00:23:48.600 Within each slice we calculate the slice mean

416 00:23:48.600 --> 00:23:51.363 for all roles of U.

417 00:23:53.663 --> 00:23:54.900 Lastly, illustrated inn this graph,

418 00:23:54.900 --> 00:23:58.830 we can first the slice into Y into edge slices

419 00:23:58.830 --> 00:24:01.920 and then within each slice, if those Yi

420 00:24:01.920 --> 00:24:03.330 fall into the same slice,

421 00:24:03.330 --> 00:24:05.536 we find out their corresponding use

422 00:24:05.536 --> 00:24:10.536 and then do, calculate its mean for each U.

423 00:24:13.890 --> 00:24:16.424 And in that way, we can simplify,

424 00:24:16.424 --> 00:24:20.610 we can simply estimate expectation of Ui given Y.

425 00:24:20.610 --> 00:24:23.300 And then further we can estimate the variance

426 00:24:23.300 --> 00:24:25.533 of those averages.

427 00:24:27.390 --> 00:24:30.480 Basically just taking that the variance of U one bar

428 00:24:30.480 --> 00:24:32.283 to U edge bar.

429 00:24:37.170 --> 00:24:40.440 So this is the way how we estimate the variance

430 00:24:40.440 --> 00:24:42.693 of the expectation of Ui given Y.

431 00:24:45.180 --> 00:24:48.277 Okay so in that way we actually generate our estimate

432 00:24:48.277 --> 00:24:49.909 weighted leverage score

433 00:24:49.909 --> 00:24:53.300 we define as the right leverage score

434 00:24:53.300 --> 00:24:55.803 weighted by the matrix in the middle.

435 00:24:58.260 --> 00:25:01.534 Okay so first of all, this weighted leverage score

436 00:25:01.534 --> 00:25:06.534 is built, say, upon the general index model,

437 00:25:08.030 --> 00:25:10.020 it is considered as model free

438 00:25:10.020 --> 00:25:14.474 because Yi, the response is connected with

439 00:25:14.474 --> 00:25:18.660 hitting combination of X through anomaly function F.

440 00:25:18.660 --> 00:25:21.210 And this model is general to encompass

441 00:25:21.210 --> 00:25:22.920 many different model types.

442 00:25:22.920 --> 00:25:25.023 So we can consider it as model free.

443 00:25:26.040 --> 00:25:31.040 And second, this to generate this weighted leverage score

444 00:25:31.350 --> 00:25:36.171 where there is no need to estimate the covariance matrix

445 00:25:36.171 --> 00:25:40.260 and there is no need to estimate anomaly function F.

446 00:25:40.260 --> 00:25:43.000 So we can bypass all those procedures

447 00:25:43.860 --> 00:25:45.960 to calculate this weighted leverage score.

448 00:25:51.810 --> 00:25:53.312 So this weighted leverage score

449 00:25:53.312 --> 00:25:55.680 actually encompass a very good feature

450 00:25:55.680 --> 00:25:57.450 that is it is an indicator

451 00:25:57.450 --> 00:26:00.530 of how influential of the columns are

452 00:26:00.530 --> 00:26:03.510 and we can basically run our predictors

453 00:26:03.510 --> 00:26:05.760 according to the weighted leverage score.

454 00:26:05.760 --> 00:26:07.260 The higher the score is,

455 00:26:07.260 --> 00:26:10.043 the more influential the predictor will be.

456 00:26:10.043 --> 00:26:14.160 And later I will show why this ranking properties

457 00:26:14.160 --> 00:26:17.307 would help or even with the leverage score.

458 00:26:17.307 --> 00:26:20.340 So this is a basic procedures

459 00:26:20.340 --> 00:26:21.720 for giving weighted leverage score

460 00:26:21.720 --> 00:26:24.600 to a variable selection or variable screening.

461 00:26:24.600 --> 00:26:27.630 So given that we have this matrix,

462 00:26:27.630 --> 00:26:30.750 which is an impart matrix and we have the responses,

463 00:26:30.750 --> 00:26:33.600 the labels, for each of the location,

464 00:26:33.600 --> 00:26:38.600 we only need one time singular value composition, okay?

465 00:26:39.120 --> 00:26:42.810 This is a rank D singular value composition of X.

466 00:26:42.810 --> 00:26:45.210 And then we can just calculate the weighted leverage score

467 00:26:45.210 --> 00:26:48.240 according to the equations,

468 00:26:48.240 --> 00:26:49.920 rank those weighted leverage score

469 00:26:49.920 --> 00:26:51.870 from the highest to the lowest.

470 00:26:51.870 --> 00:26:53.850 Select the predictor that we use,

471 00:26:53.850 --> 00:26:55.700 the highest weighted leverage scores.

472 00:26:56.640 --> 00:26:58.770 This is the basic screening procedure

473 00:26:58.770 --> 00:27:00.370 using the weighted average score

474 00:27:01.470 --> 00:27:03.480 and there is still implementation issue

475 00:27:03.480 --> 00:27:05.250 that we will later address.

476 00:27:05.250 --> 00:27:08.070 First one, how can we determine the number of D?

477 00:27:08.070 --> 00:27:11.133 So given the data which is an IP, how can we determine,

478 00:27:12.480 --> 00:27:17.340 say, how many, say, spiked or how many singular values

479 00:27:17.340 --> 00:27:18.740 to be included in the model?

480 00:27:20.280 --> 00:27:22.680 And then the second implementation issue

481 00:27:22.680 --> 00:27:25.290 is determine the number of variables

482 00:27:25.290 --> 00:27:26.853 to be selected in the model.

483 00:27:29.070 --> 00:27:31.740 So you can see that the weight leverage score procedure

484 00:27:31.740 --> 00:27:34.020 is screening procedure only include

485 00:27:34.020 --> 00:27:37.263 one type of singularity, competition is quite efficient.

486 00:27:41.130 --> 00:27:44.160 Okay so, in the next, let us using two slides

487 00:27:44.160 --> 00:27:47.070 to discuss a little but, basically just one slides

488 00:27:47.070 --> 00:27:48.930 to discuss the ranking properties

489 00:27:48.930 --> 00:27:50.700 of the weighted leverage score.

490 00:27:50.700 --> 00:27:52.680 So as I mentioned, the weighted leverage score

491 00:27:52.680 --> 00:27:54.570 has a very nice property.

492 00:27:54.570 --> 00:27:58.290 So it is guaranteed by the theorem here.

493 00:27:58.290 --> 00:28:01.770 We show that, given certain conditions are satisfied,

494 00:28:01.770 --> 00:28:06.770 we have the minimum value of the weighted leverage score

495 00:28:06.930 --> 00:28:08.747 from the true predictors

496 00:28:08.747 --> 00:28:12.180 will always rank higher than the maximum value,

497 00:28:12.180 --> 00:28:13.830 maximum weighted leverage score

498 00:28:13.830 --> 00:28:15.603 of those redundant predictors.

499 00:28:17.550 --> 00:28:22.550 And this holds for the population with leverage score.

500 00:28:24.150 --> 00:28:26.790 In terms of the estimation weighted leverage score,

501 00:28:26.790 --> 00:28:29.280 we utilize the two step procedure.

502 00:28:29.280 --> 00:28:30.120 So first of all,

503 00:28:30.120 --> 00:28:33.660 we show that the estimate weighted leverage score

504 00:28:33.660 --> 00:28:36.900 is very close to the population version

505 00:28:36.900 --> 00:28:38.671 of the weighted leverage.

506 00:28:38.671 --> 00:28:43.671 Okay and then the estimated weighted leverage score

507 00:28:43.980 --> 00:28:46.090 will also have the ranking property

508 00:28:47.640 --> 00:28:49.955 that is the estimated weighted leverage score

509 00:28:49.955 --> 00:28:53.773 of the active predictors or important predictors

510 00:28:53.773 --> 00:28:57.570 ranks higher than the estimated weighted leverage score

511 00:28:57.570 --> 00:29:01.653 of the down predictors with probability turning to one.

512 00:29:04.740 --> 00:29:07.560 Okay so this, the ranking properties

513 00:29:07.560 --> 00:29:09.090 of the weighted leverage score

514 00:29:09.090 --> 00:29:12.453 basically is guaranteed by these two properties.

515 00:29:16.440 --> 00:29:18.300 And then further, we also know

516 00:29:18.300 --> 00:29:20.670 that there are two implementation issue.

517 00:29:20.670 --> 00:29:22.020 The first one is determine

518 00:29:22.020 --> 00:29:25.380 the number of spiked singular values d.

519 00:29:25.380 --> 00:29:29.190 So how many single values we need to include in our model?

520 00:29:29.190 --> 00:29:32.496 This is a question and it's quite crucial

521 00:29:32.496 --> 00:29:35.890 because we need to know how many of the signals

522 00:29:37.230 --> 00:29:39.930 contain the data and we need to remove

523 00:29:39.930 --> 00:29:43.140 all those redundant or noise information.

524 00:29:43.140 --> 00:29:45.960 Okay so here I develop a criterion

525 00:29:45.960 --> 00:29:50.793 based on the properties of those aken values.

526 00:29:52.367 --> 00:29:56.070 DR is a function of, R is the number of values

527 00:29:56.070 --> 00:29:57.813 to be included in a model.

528 00:29:58.964 --> 00:30:02.280 DR is a function of R and the theta I hat

529 00:30:02.280 --> 00:30:06.790 is the ratio between the highest aken value

530 00:30:07.950 --> 00:30:11.523 and the largest aken, value, number one hat.

531 00:30:12.360 --> 00:30:17.165 And then, you can see that as we include more, say,

532 00:30:17.165 --> 00:30:18.870 single values in the model,

533 00:30:18.870 --> 00:30:20.770 this, the first term will decrease

534 00:30:22.080 --> 00:30:24.030 and then tier to some point.

535 00:30:24.030 --> 00:30:27.690 The first, the decreasing of the first term

536 00:30:27.690 --> 00:30:30.120 is smaller than the increasing of the second term.

537 00:30:30.120 --> 00:30:31.803 Then DR starts to increase.

538 00:30:33.810 --> 00:30:37.710 And then we can use the criterion to find D hat,

539 00:30:37.710 --> 00:30:40.173 we show that we had is very close to a true D.

540 00:30:41.260 --> 00:30:42.780 Okay, you meet this criterion,

541 00:30:42.780 --> 00:30:46.953 we can select the true number of signals in the model.

542 00:30:51.180 --> 00:30:53.400 And the second implementation issue

543 00:30:53.400 --> 00:30:58.230 is about how many predictors, how many true predictors

544 00:30:58.230 --> 00:31:00.570 we need to include in our model.

545 00:31:00.570 --> 00:31:04.330 Okay again, we're ranking our weighted leverage scores

546 00:31:05.460 --> 00:31:09.630 and here we utilize the criterion here

547 00:31:09.630 --> 00:31:12.240 based on the properties of the weighted leverage score.

548 00:31:12.240 --> 00:31:17.174 Okay, as we include more predictors into the active set,

549 00:31:17.174 --> 00:31:19.680 the first term will decrease, okay?

550 00:31:19.680 --> 00:31:22.020 The summation of the weighted leverage score will increase,

551 00:31:22.020 --> 00:31:24.580 but the increment will decrease

552 00:31:24.580 --> 00:31:26.880 and then the second term will increase, okay?

553 00:31:26.880 --> 00:31:31.563 So, as we include more predictors in the model,

554 00:31:32.610 --> 00:31:35.070 there's some changing point

555 00:31:35.070 --> 00:31:40.070 when the increment is smaller than the increment

556 00:31:42.547 --> 00:31:45.840 of the second penalty term.

557 00:31:45.840 --> 00:31:48.090 We show that, using this criteria,

558 00:31:48.090 --> 00:31:51.720 the set we selected using this criteria,

559 00:31:51.720 --> 00:31:56.160 which is A, will always we'll say can include

560 00:31:56.160 --> 00:32:00.543 all the true predictors with probability pending to one.

561 00:32:03.420 --> 00:32:05.400 Okay so that's how we using this criterion

562 00:32:05.400 --> 00:32:07.050 to determine the number of predictors

563 00:32:07.050 --> 00:32:08.450 to be selected in the model.

564 00:32:10.290 --> 00:32:11.190 Okay in the next step,

565 00:32:11.190 --> 00:32:14.099 let me show some empirical study results

566 00:32:14.099 --> 00:32:16.797 of using the weighted leverage score

567 00:32:16.797 --> 00:32:19.803 to do the variable selection in the model.

568 00:32:21.521 --> 00:32:23.880 In the example one, as I mentioned,

569 00:32:23.880 --> 00:32:27.510 we are utilizing, we are proposing our method

570 00:32:27.510 --> 00:32:31.080 under the general index model framework.

571 00:32:31.080 --> 00:32:33.783 So the first model is the general index model.

572 00:32:34.620 --> 00:32:37.080 Well why there's two directions?

573 00:32:37.080 --> 00:32:38.780 The first one is in the numerator,

574 00:32:39.660 --> 00:32:42.870 so this is so called a beta one transpose X.

575 00:32:42.870 --> 00:32:45.240 The second direction beta two transpose X

576 00:32:45.240 --> 00:32:47.777 is in the variable system,

577 00:32:48.900 --> 00:32:50.070 the term within the variant,

578 00:32:50.070 --> 00:32:52.323 this is beta two transpose X.

579 00:32:53.700 --> 00:32:55.971 Okay so this is called a general index model

580 00:32:55.971 --> 00:33:00.190 and we assume that X is generated from

581 00:33:03.000 --> 00:33:05.250 a very normal distribution

582 00:33:05.250 --> 00:33:08.050 and we submit our zero and covariant structure

583 00:33:09.300 --> 00:33:10.530 back in this way, okay?

584 00:33:10.530 --> 00:33:12.810 We let rho equal to 0.5

585 00:33:12.810 --> 00:33:17.283 which will generate a matrix with moderate correlations.

586 00:33:20.820 --> 00:33:23.920 So in this way, we generate both X and Y,

587 00:33:23.920 --> 00:33:28.860 let's see how the performance of variable selection

588 00:33:28.860 --> 00:33:31.277 give you the weighted leverage score.

589 00:33:31.277 --> 00:33:34.200 In our scenarios, we let N equal to 1000

590 00:33:34.200 --> 00:33:35.300 and the rho equal 0.5.

591 00:33:37.020 --> 00:33:40.743 In example one, there are four different scenarios.

592 00:33:41.910 --> 00:33:46.080 For scenario one, we let P as 200

593 00:33:46.080 --> 00:33:48.780 and then we increase P to 200, sorry 22,000 and 2,500.

594 00:33:54.120 --> 00:33:58.260 We also increase the variance of the error turn

595 00:33:58.260 --> 00:33:59.977 as 1.5 this now to 1.3.

596 00:34:02.010 --> 00:34:04.378 Okay, there are three criteria we used

597 00:34:04.378 --> 00:34:08.160 to evaluate the performance of the method.

598 00:34:08.160 --> 00:34:10.380 The first one is the false positive.

599 00:34:10.380 --> 00:34:13.623 So it means how many of the variables are falsely selected?

600 00:34:14.580 --> 00:34:17.940 False negative which shows how many variables

601 00:34:17.940 --> 00:34:20.880 falsely excluded, how many true predictors

602 00:34:20.880 --> 00:34:22.713 are falsely excluded.

603 00:34:24.090 --> 00:34:26.011 The last one, the last criterion

604 00:34:26.011 --> 00:34:29.670 is because, is basically is our model size

605 00:34:29.670 --> 00:34:33.090 because we have this ranking properties

606 00:34:33.090 --> 00:34:36.450 of all the methods here, okay?

607 00:34:36.450 --> 00:34:38.340 All those methods have ranking properties.

608 00:34:38.340 --> 00:34:42.780 So I want to know how many variables I need to include

609 00:34:42.780 --> 00:34:45.132 in the model so that all true predictors

610 00:34:45.132 --> 00:34:49.650 are included because we have this ranking property, okay?

611 00:34:49.650 --> 00:34:51.790 You will see that weighted leverage score

612 00:34:53.670 --> 00:34:55.560 basically have a better performance

613 00:34:55.560 --> 00:34:58.210 in terms of the false positive and the false negative

614 00:35:00.600 --> 00:35:02.493 and also the model size.

615 00:35:05.490 --> 00:35:09.080 I want to say say a bit more about model size.

616 00:35:09.080 --> 00:35:13.440 We can see that when N is 1000, P is 200,

617 00:35:13.440 --> 00:35:17.520 the minimum model size is 6.14, meaning that we, in total,

618 00:35:17.520 --> 00:35:21.060 we only have six true predictors, okay?

619 00:35:21.060 --> 00:35:23.040 We only need to include the six variables

620 00:35:23.040 --> 00:35:28.040 to 6.14 variables to encompass all true predictors.

621 00:35:28.740 --> 00:35:31.980 Basically all those six variables are rank higher

622 00:35:31.980 --> 00:35:35.430 than all our other novel variables, right?

623 00:35:35.430 --> 00:35:38.883 In a second model, when P increases to 2,200,

624 00:35:40.200 --> 00:35:42.600 we only need seven predictors

625 00:35:42.600 --> 00:35:45.292 in order to, on average, in order to include

626 00:35:45.292 --> 00:35:46.140 all the true predictors.

627 00:35:46.140 --> 00:35:50.670 Meaning that overall, all the true predictors

628 00:35:50.670 --> 00:35:53.103 ranks higher than those redundant predictors.

629 00:35:54.570 --> 00:35:57.900 And then when sigma increases and when P increases,

630 00:35:57.900 --> 00:36:01.000 we still need only the minimum number of variables

631 00:36:02.880 --> 00:36:04.743 just to include all true predictors.

632 00:36:06.510 --> 00:36:08.880 Okay so this is for the first example

633 00:36:08.880 --> 00:36:12.060 of the model index model.

634 00:36:12.060 --> 00:36:13.800 The second is more challenging,

635 00:36:13.800 --> 00:36:17.010 it's called a heteroscedastic model.

636 00:36:17.010 --> 00:36:19.271 You will find that, in the first model,

637 00:36:19.271 --> 00:36:23.580 the X, those active predictors only influence

638 00:36:23.580 --> 00:36:25.590 the main response, okay?

639 00:36:25.590 --> 00:36:28.260 Only influence Y in its means.

640 00:36:28.260 --> 00:36:30.870 But here, you can see four different types

641 00:36:30.870 --> 00:36:32.700 for those variables, the average of Y,

642 00:36:32.700 --> 00:36:34.050 the mean of Y is zero

643 00:36:34.050 --> 00:36:36.130 because error term is in the numerator

644 00:36:37.119 --> 00:36:41.640 and those X, those active predictors will influence Y

645 00:36:41.640 --> 00:36:43.083 in its variance.

646 00:36:44.730 --> 00:36:46.953 So it's a much challenging case.

647 00:36:49.170 --> 00:36:52.260 In this case, we also assume that X follows

648 00:36:52.260 --> 00:36:53.970 a very normal distribution,

649 00:36:53.970 --> 00:36:58.970 given mean is zero and a covariant structure like this.

650 00:36:59.580 --> 00:37:01.680 In our scenarios, we let N equal to 1,000.

651 00:37:03.030 --> 00:37:05.880 So let's see, in a heteroscedastic model,

652 00:37:05.880 --> 00:37:10.383 what are the behaviors of our methods?

653 00:37:11.397 --> 00:37:14.430 Okay, sorry, I forget to introduce the method

654 00:37:14.430 --> 00:37:18.060 that we are compare these with true independence ranking

655 00:37:18.060 --> 00:37:20.130 and screening and the distance correlation.

656 00:37:20.130 --> 00:37:22.740 So both of these methods can be utilized

657 00:37:22.740 --> 00:37:27.740 to measure, say, the association between the response

658 00:37:28.077 --> 00:37:30.090 and the predictors.

659 00:37:30.090 --> 00:37:33.780 And both of the methods have the ranking properties.

660 00:37:33.780 --> 00:37:36.330 So we can compare the minimum model size

661 00:37:36.330 --> 00:37:37.383 for all the methods.

662 00:37:39.030 --> 00:37:41.670 Regards the false positive of these two methods,

663 00:37:41.670 --> 00:37:43.380 there's no criteria proposed

664 00:37:43.380 --> 00:37:45.930 to select the number of predictors.

665 00:37:45.930 --> 00:37:47.670 And in all those methods,

666 00:37:47.670 --> 00:37:49.860 I just use a harder stretch holding

667 00:37:49.860 --> 00:37:52.320 to select the number of predictors

668 00:37:52.320 --> 00:37:54.990 in order to be included in the model.

669 00:37:54.990 --> 00:37:58.200 Okay let's see, in a heteroscedastic model,

670 00:37:58.200 --> 00:38:00.543 what are the behaviors of those three methods?

671 00:38:02.910 --> 00:38:07.910 Okay, when we have N greater the number of predictors,

672 00:38:08.010 --> 00:38:12.480 P here, you will find that there are slightly larger,

673 00:38:12.480 --> 00:38:15.540 with false negative for the weighted leverage score.

674 00:38:15.540 --> 00:38:19.920 This is because both the methods within the our threshold,

675 00:38:19.920 --> 00:38:24.570 they select around 140, more than 140,

676 00:38:24.570 --> 00:38:29.010 true predictors out of 200 from the model.

677 00:38:29.010 --> 00:38:32.970 So that's why they have very small false negative,

678 00:38:32.970 --> 00:38:35.771 but if you look at the minimum model size,

679 00:38:35.771 --> 00:38:38.880 you will find that our weighted leverage score

680 00:38:38.880 --> 00:38:41.610 still maintains a very good performance.

681 00:38:41.610 --> 00:38:46.080 Okay, it has a smaller value of the minimum model size.

682 00:38:46.080 --> 00:38:48.690 Okay in general, we only need 46 variables

683 00:38:48.690 --> 00:38:53.013 in order to include our two predictors in the model.

684 00:38:55.025 --> 00:39:00.025 And then as P diverges, as CSP increased to 2,500,

685 00:39:00.668 --> 00:39:02.970 basically the weighted leverage score

686 00:39:02.970 --> 00:39:06.852 will measure 1.3 variable true predictors from the model

687 00:39:06.852 --> 00:39:11.040 and every method have a really hard time

688 00:39:11.040 --> 00:39:14.010 to identify all the true predictors.

689 00:39:14.010 --> 00:39:16.713 They have really large minimum model size.

690 00:39:20.400 --> 00:39:25.167 So this is basic a performance of the using with

691 00:39:25.167 --> 00:39:29.640 the leverage score to perform a variable screening

692 00:39:29.640 --> 00:39:31.290 under general index model.

693 00:39:31.290 --> 00:39:33.540 So I only present two examples here

694 00:39:33.540 --> 00:39:35.730 for interest in odd scenarios.

695 00:39:35.730 --> 00:39:37.923 We can talk about that at the top.

696 00:39:42.303 --> 00:39:45.483 Okay so let's get back to our real data example.

697 00:39:46.650 --> 00:39:48.750 So in a motivating example, as I mentioned,

698 00:39:48.750 --> 00:39:52.620 we utilize this spatial transcriptomics data.

699 00:39:52.620 --> 00:39:57.120 We are sequencing the grid point within each section, okay?

700 00:39:57.120 --> 00:40:01.650 Basically these locations are invasive cancer areas,

701 00:40:01.650 --> 00:40:05.550 the other areas are the noninvasive cancer areas,

702 00:40:05.550 --> 00:40:08.040 and then these are the normal areas.

703 00:40:08.040 --> 00:40:11.400 So how to determine the invasive, non-invasive

704 00:40:11.400 --> 00:40:13.137 and the normal area?

705 00:40:13.137 --> 00:40:16.785 These are determined by qualified doctors

706 00:40:16.785 --> 00:40:20.160 and they're utilizing some logical information

707 00:40:20.160 --> 00:40:21.750 of these locations.

708 00:40:21.750 --> 00:40:24.390 Okay in general, for these two sections,

709 00:40:24.390 --> 00:40:28.320 we have identified 518 locations,

710 00:40:28.320 --> 00:40:33.320 64 invasive areas and 73 are noninvasive areas.

711 00:40:33.720 --> 00:40:37.533 And there are rest of the areas, 381, they are normal.

712 00:40:39.600 --> 00:40:44.600 And we have our gene, about 3,572 expressions,

713 00:40:46.350 --> 00:40:50.580 gene expressions across the section.

714 00:40:50.580 --> 00:40:54.380 Okay so in general, basically we have our data matrix

715 00:40:54.380 --> 00:40:59.380 it is about 518 times 3,572.

716 00:40:59.730 --> 00:41:04.170 So we trying to identify biomarkers, okay,

717 00:41:04.170 --> 00:41:06.060 within those three genes

718 00:41:06.060 --> 00:41:10.440 that can help us discriminate between invasive cancer,

719 00:41:10.440 --> 00:41:12.423 non-invasive cancer and normal areas.

720 00:41:14.430 --> 00:41:17.676 So we utilize the weighted leverage score,

721 00:41:17.676 --> 00:41:20.310 we apply the weight leverage score screening procedure

722 00:41:20.310 --> 00:41:21.810 for this data set.

723 00:41:21.810 --> 00:41:26.810 And we identified around 225 genes

724 00:41:27.150 --> 00:41:30.570 among all those P genes.

725 00:41:30.570 --> 00:41:33.706 In the plot, a heat map here show the results

726 00:41:33.706 --> 00:41:37.260 because just for the ease of presentation,

727 00:41:37.260 --> 00:41:40.110 I only printed around 20 genes here

728 00:41:40.110 --> 00:41:45.110 and with the top, say, weighted leverage scores,

729 00:41:45.120 --> 00:41:47.640 you can see that there are certain patterns here.

730 00:41:47.640 --> 00:41:51.096 This group of genes are more highly expressed

731 00:41:51.096 --> 00:41:54.210 for the non-invasive cancer area.

732 00:41:54.210 --> 00:41:57.360 There are certain group of genes right here.

733 00:41:57.360 --> 00:41:59.070 They are more highly expressed

734 00:41:59.070 --> 00:42:02.940 in the invasive cancer areas.

735 00:42:02.940 --> 00:42:06.540 Okay so this is the gene expression patterns

736 00:42:06.540 --> 00:42:09.123 of those top 20 genes.

737 00:42:13.273 --> 00:42:18.273 And then we also plot the expressions of those genes

738 00:42:20.220 --> 00:42:22.380 in these sections.

739 00:42:22.380 --> 00:42:25.020 Again, these are invasive areas,

740 00:42:25.020 --> 00:42:27.300 noninvasive and the normal areas.

741 00:42:27.300 --> 00:42:29.730 So we plot a group of genes,

742 00:42:29.730 --> 00:42:32.880 I can’t remember exactly what our genes are,

743 00:42:32.880 --> 00:42:36.600 but these genes have, you can see,

744 00:42:36.600 --> 00:42:41.600 have a higher expression on those noninvasive cancer areas.

745 00:42:41.610 --> 00:42:43.080 And we plug another group of genes,

746 00:42:43.080 --> 00:42:45.300 basically are these three genes,

747 00:42:45.300 --> 00:42:48.390 in the section, the expression shows a higher,

748 00:42:48.390 --> 00:42:51.776 this means that these three genes have higher expression,

749 00:42:51.776 --> 00:42:56.776 okay, in the invasive cancer areas, okay?

750 00:42:56.820 --> 00:42:59.458 Basically this means that the genes that we selected

751 00:42:59.458 --> 00:43:04.458 show a remarkable spatially differential expressed patterns

752 00:43:05.699 --> 00:43:07.623 across the tissue sections.

753 00:43:11.070 --> 00:43:16.070 And later we do a, say, pathway analysis

754 00:43:19.290 --> 00:43:24.290 and see that there are 47 functional classes

755 00:43:24.420 --> 00:43:28.290 for those all those gene 225 gene that we have identified.

756 00:43:28.290 --> 00:43:31.590 And there are several cancer hallmarks, for example,

757 00:43:31.590 --> 00:43:34.565 38 of the genes that we identified

758 00:43:34.565 --> 00:43:39.565 enriched in the regulation of apoptotic process.

759 00:43:41.190 --> 00:43:44.130 This is a kind of cancer hallmark.

760 00:43:44.130 --> 00:43:47.070 And then another 41 gene that we have identified

761 00:43:47.070 --> 00:43:49.983 are involved in the regulation of cell death.

762 00:43:51.630 --> 00:43:54.420 More specifically, because we are really interested

763 00:43:54.420 --> 00:43:57.120 in the invasive cancer,

764 00:43:57.120 --> 00:43:59.490 so we identified these three,

765 00:43:59.490 --> 00:44:01.620 there are like three genes for example,

766 00:44:01.620 --> 00:44:05.353 in the regulation of brain process,

767 00:44:05.353 --> 00:44:10.353 they have many relations with the breast cancer, okay?

768 00:44:11.370 --> 00:44:16.370 And later we can investigate or, say,

769 00:44:16.920 --> 00:44:20.293 even adaptation of those, those genes

770 00:44:20.293 --> 00:44:24.303 that are enriched in the revelation of apoptotic process.

771 00:44:31.890 --> 00:44:36.150 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 the two 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 predicts 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 --> 00:47:05.596 that is based on the slicing versus regression.

819 00:47:05.596 --> 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,

824 00:47:22.513 --> 00:47:27.390 they all just utilize the partial and partial correlation.

825 00:47:27.390 --> 00:47:28.743 So between X and Y.

826 00:47:30.000 --> 00:47:33.303 Okay so it does not utilize any of the information within X.

827 00:47:34.280 --> 00:47:37.590 It's kind of a marginal correlation

828 00:47:37.590 --> 00:47:41.130 between each variable X and then one, okay?

829 00:47:41.130 --> 00:47:42.713 However, the weighted leverage score

830 00:47:42.713 --> 00:47:46.290 will utilize both the raw information

831 00:47:46.290 --> 00:47:48.570 and also the variance information,

832 00:47:48.570 --> 00:47:52.143 the correlation structure within the model,

833 00:47:53.880 --> 00:47:56.285 which is the V matrix, as I mentioned,,

834 00:47:56.285 --> 00:48:01.230 is derived from the covariance structure of X.

835 00:48:01.230 --> 00:48:03.107 The V matrix basically is the vector

836 00:48:03.107 --> 00:48:06.420 of the covariance structure, covariance of X.

837 00:48:06.420 --> 00:48:10.500 So it utilize the, say, kind of a correlation

838 00:48:10.500 --> 00:48:15.116 between all the X variables and a variable screening.

839 00:48:15.116 --> 00:48:17.880 So I'm not sure if this answers your question.

840 00:48:17.880 --> 00:48:18.947 <v Student>Yeah, thank you.</v>

841 00:48:18.947 --> 00:48:21.690 I think it is, you answered my question.

842 00:48:21.690 --> 00:48:25.153 Essentially, I'm thinking like if those important features,

843 00:48:25.153 --> 00:48:28.440 they are actually not the top contributors

844 00:48:28.440 --> 00:48:30.180 to the top other lectures,

845 00:48:30.180 --> 00:48:33.030 then we wouldn't expect the weighted leverage score

846 00:48:33.030 --> 00:48:33.863 to aim true way.

847 00:48:35.850 --> 00:48:36.863 <v Professor Liu>Thank you.</v>

848 00:48:42.240 --> 00:48:43.320 <v Robert>Any other questions</v>

849 00:48:43.320 --> 00:48:45.687 anyone wants to bring up right now?

850 00:48:50.800 --> 00:48:53.800 (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)