Submitted to the Annals of Statistics
arXiv: math.PR/0000000

SUPPLEMENT TO "PARAMETRIC OR NONPARAMETRIC? A PARAMETRICNESS INDEX FOR MODEL SELECTION"

BY WEI LIU* AND YUHONG YANG*

University of Minnesota

In this supplement, we provide complete descriptions of our numerical work and also give additional results that support our conclusions. The first section deals with simulation and the second section presents real data examples.

1. Simulation Results. We consider single-predictor and multiple-predictor cases, aiming at a serious understanding of the practical utility of PI. In all the numerical examples in this paper, we choose $\lambda_n = 1$ and $d = 0$.

1.1. *Single predictor.*

Example 1. Compare two different situations:

Case 1: $Y = 3\sin(2\pi x) + \sigma_1\epsilon$,
Case 2: $Y = 3 - 5x + 2x^2 + 1.5x^3 + 0.8x^4 + \sigma_2\epsilon$, where $\epsilon \sim N(0, 1)$ and $x \sim N(0, 1)$.

BIC is used to select the order of polynomial regression between 1 and 30. The estimated σ from the selected model is used to calculate the PI. Representative scatterplots at $n = 200$ with $\sigma_1 = 3$, $\sigma_2 = 7$ can be found in Figure 1. Note that the function estimate based on the selected model by BIC is visually more different from that based on the smaller model with one fewer term for the parametric scenario than the nonparametric one.

Quantiles for the PIs in both scenarios based on 300 replications are presented in Table 1.

Example 2. Compare the following two situations:

Case 1: $Y = 1 - 2x + 1.6x^2 + 0.5x^3 + 3\sin(2\pi x) + \sigma\epsilon$

*Supported by NSF grant DMS-0706850.

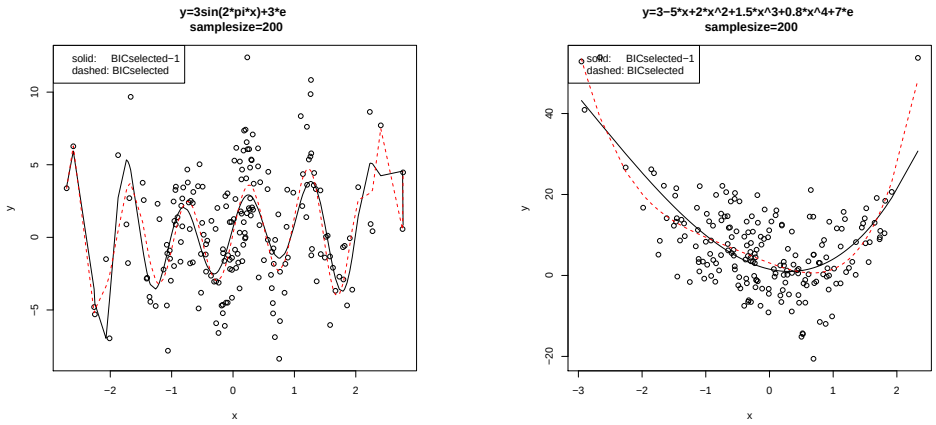

FIG 1. *Scatterplots For Example 1*

TABLE 1
Percentiles of PI for Example 1

	case 1			**case 2**		
percentile	order selected	PI	$\hat{\sigma}$	order selected	PI	$\hat{\sigma}$
10%	1	0.47	2.78	4	1.14	6.53
20%	13	1.02	2.89	4	1.35	6.67
50%	15	1.12	3.03	4	1.89	6.96
80%	16	1.34	3.21	4	3.15	7.31
90%	17	1.54	3.52	4	4.21	7.49

Case 2: $Y = 1 - 2x + 1.6x^2 + 0.5x^3 + \sin(2\pi x) + \sigma\epsilon$.

The two mean functions are the same except the coefficient of the $\sin(2\pi x)$ term. Scatterplots and table similar to those for Example 1 are in Figure 2 and Table 2, respectively. As we can see from Table 2, although both cases are of a nonparametric nature, they have different behaviors in terms of model selection uncertainty and PI values. Case 2 can be called 'practically' parametric and the large PI values provide information in this regard.

1.2. *Factors that influence PI.* As we know, most model selection problems, if not all, are affected by many factors like the regression function itself, the noise level, and the sample size. We expect that these factors influence the behavior of PI as well. We investigate the effects of these factors on PI and report some representative results below. As we will see, PI, as a diagnostic measure, can tell us whether a problem is 'practically' paramet-

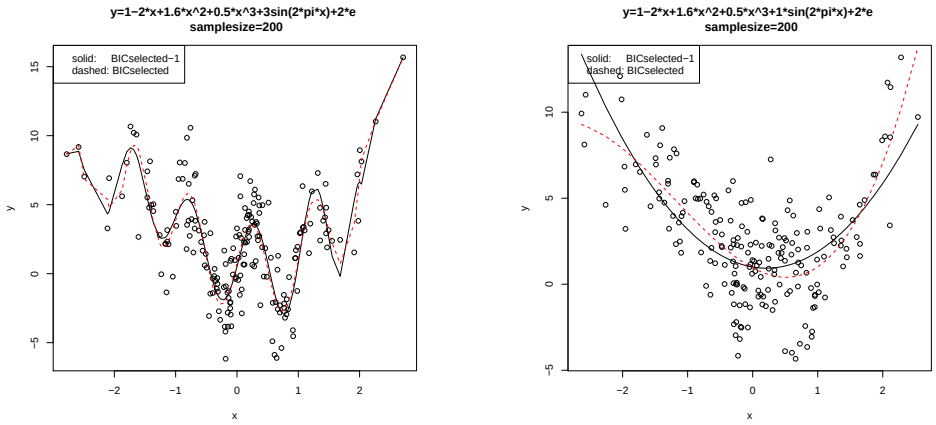

FIG 2. *Scatterplots For Example 2*

TABLE 2
Percentiles of PI for Example 2

	case 1			**case 2**		
percentile	order selected	PI	$\hat{\sigma}$	order selected	PI	$\hat{\sigma}$
10%	15	1.01	1.87	3	1.75	1.99
20%	15	1.05	1.92	3	2.25	2.03
50%	16	1.14	2.00	3	3.51	2.12
80%	17	1.4	2.11	3	5.33	2.22
90%	18	1.63	2.17	3	6.62	2.26

ric/nonparametric due to the influences of all the factors that affect model selection.

1.2.1. *The effect of sample size.* We calculated the PI at different sample sizes with 300 replications for each and report the results for Examples 1 and 2 in Figures 3 and 4.

For Example 1, from Figure 3, we see the PIs in case 1 basically fall in between 1 and 1.6, whereas the PIs in case 2 become larger as sample size increases.

From Figure 4, in case 2 of Example 2, the PIs first increase and then drop down as the sample size increases. This is due to the fact that in the beginning, the sine term is better to be ignored due to lack of information, and when the sample size is bigger, say 300-400, the PI indicates a strong parametric scenario. With a sample size in this range, the problem is 'practically' parametric. With more and more data we are then gradually able

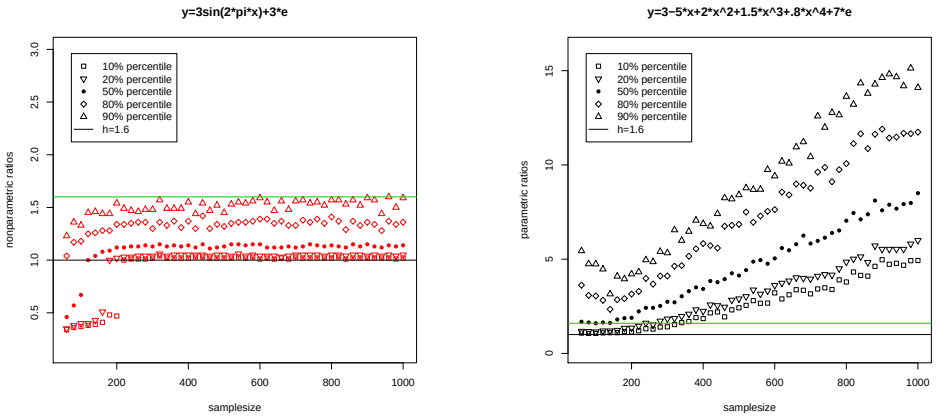

FIG 3. *Sample size effect for Example 1*

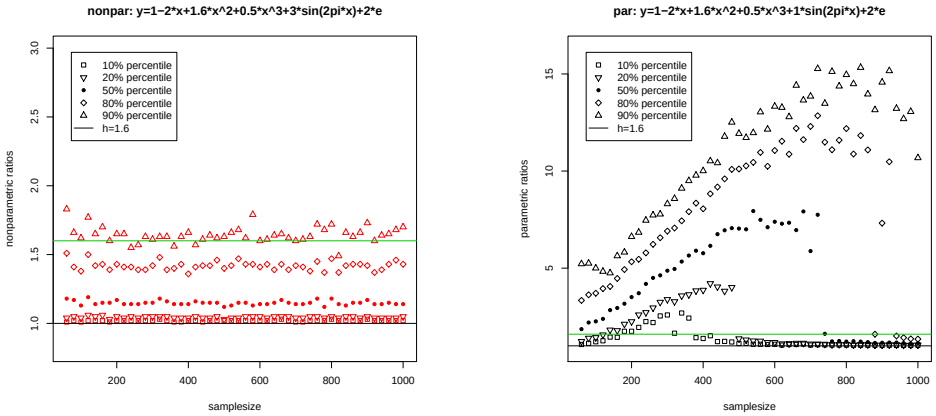

FIG 4. *Sample size effect for Example 2*

to detect the signal of the $\sin(2\pi x)$ term, thus capturing the nonparametric nature of the mean function.

In case 1 of Example 2 we have the 90% percentiles slightly exceeding 1.6. Also notice that the percentiles in case 2 drop at different levels of sample sizes. For example, the 10% drops below 1.6 when the sample size is bigger than 400, while the 50% drops below 1.6 when the sample size is bigger than 800.

The examples show that given the regression function and the noise

level, the value of PI indicates whether the problem is 'practically' parametric/nonparametric at the current sample size.

1.2.2. *The effect of coefficient.* We study the PIs for different values of the coefficient of the last term in case 2 of Example 1 and Example 2, respectively. The results are reported in Figure 5.

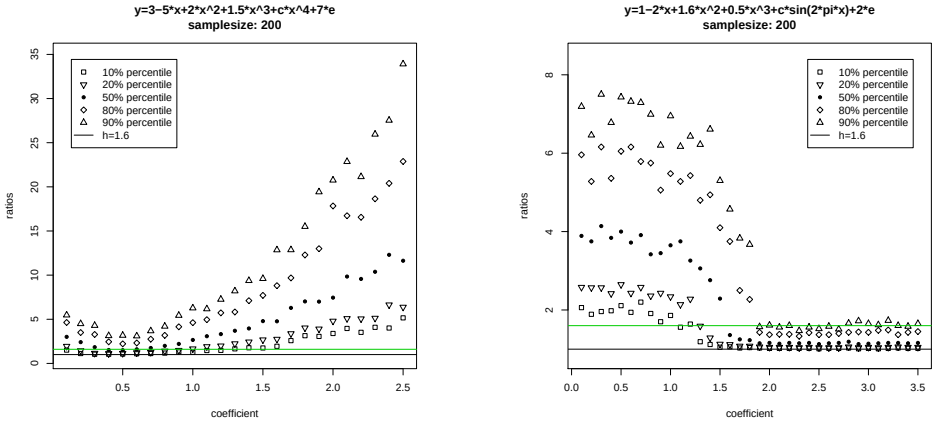

FIG 5. *Effect of coefficient*

For Example 1, the values of PI first decrease and then increase as the coefficient of x^4 increases. This is because when the coefficient for the term of x^4 is small (less than .5), the true mean function behaves just like a polynomial of order 3 at the current sample size. As the coefficient gets slightly larger, there is no clear distinction between a polynomial of order 3 and a polynomial of order 4 at the current sample size. That is why we see the PIs drop a little in the beginning. However, when the coefficient gets bigger than .5 or .6, then we can detect the term of x^4 and the PIs increase with the coefficient. Overall, the PI values are mostly larger than 1.6 in this example.

For Example 2, the PIs drop as the coefficient of $\sin(2\pi x)$ increases. This is because as the coefficient gets larger, the nonparametric signal becomes stronger. When the coefficient is small (less than 1.3), most of the PIs are bigger than 1.6 and the problem is 'practically' parametric. When the coefficient is bigger than 1.9, most of the PIs fall in between 1 and 1.6 and the problem is 'practically' nonparametric.

The examples show that given the noise level and the sample size, when the nonparametric part is very weak, PI has a large value, which properly

indicates that the nonparametric part is negligible; but as the nonparametric part gets strong enough, PI will drop close to 1, indicating a clear nonparametric scenario. For a parametric scenario, the stronger the signal, the larger PI as is expected.

1.3. *Multiple predictors.* Now we study several examples with multiple predictors. The first two examples were used in the original lasso paper [10].

Unlike what we did in the single predictor cases, in these multiple-predictor examples we are going to do all subset selection. We generate data from a linear model (except example 7):

$$Y = \beta^T \mathbf{x} + \sigma\epsilon,$$

where $\mathbf{x}$ is generated from a multivariate normal distribution with mean 0, variance 1, and correlation structure given in each example. For each generated data set, we apply the Branch and Bound algorithm [5] to do all subset selection by BIC and then calculate the PI value (part of our code is modified from the aster package of Geyer [4]). Unless otherwise stated, in these examples, the sample size is 200 and we replicate 300 times.

Example 3. In this example, $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. The correlation between x_i and x_j is $\rho^{|i-j|}$ with $\rho = 0.5$. We set $\sigma = 5$.

Example 4. This example is the same as example 3, but with $\beta_j = .85, \forall j$ and $\sigma = 3$.

Example 5. In this example, $\beta = (0.9, 0.9, 0, 0, 2, 0, 0, 1.6, 2.2, 0, 0, 0, 0)^T$. There are 13 predictors and the pairwise correlation between x_i and x_j is $\rho = 0.6$ and $\sigma = 3$.

Example 6. This example is the same as example 5 except that $\beta = (0.85, 0.85, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0)^T$ and $\rho = 0.5$.

Example 7. This example is the same as example 3 except that we add a nonlinear component in the mean function and $\sigma = 3$, i.e., $Y = \beta^T \mathbf{x} + \phi(u) + \sigma\epsilon$, where $u \sim uniform(-4, 4)$ and $\phi(u) = 3(1 - 0.5u + 2u^2)e^{-u^2/4}$. All subset selection is carried out with predictors $x_1, \cdots, x_8, u, \cdots, u^8$ which are coded as 1-8 and A-G in the Table 3.

The selection behaviors and PI values are reported in Table 3 and Table 4, respectively. From those results, we see that the PIs are large for Example 3 and small for Example 4. Note that in Example 3 we have 82% chance

TABLE 3
Proportion of selecting true model

Example	true model	proportion
3	125	0.82
4	12345678	0.12
5	12589	0.43
6	125	0.51
7	1259ABCEG*	0.21

TABLE 4
Quartiles of PIs

example	Q1	Q2	Q3
3	1.26	1.51	1.81
4	1.02	1.05	1.10
5	1.05	1.15	1.35
6	1.09	1.23	1.56
7	1.02	1.07	1.16

selecting the true model, while in Example 4 the chance is only 12%. Although both Example 3 and Example 4 are of parametric nature, we would call Example 4 'practically nonparametric' in the sense that at the given sample size many models are equally likely and the issue is to balance the approximation error and estimation error. For Examples 5 and 6, the PI values are in-between, so are the chances of selecting the true models. Note that the median PI values in Examples 5 and 6 are around 1.2. These examples together show that the values of PI provide sensible information on how strong the parametric message is and that information is consistent with stability in selection. More discussions about these examples in terms of PI and statistical risks will be given later in this section. (In the lasso paper σ was chosen to be 3 for Example 3. But even with a higher noise level $\sigma = 5$, the parametric nature of this example is still obvious.)

Example 7 is quite interesting. Previously, without the $\phi(u)$ component, even at $\sigma = 5$, large values of PI are seen. Now with the nonparametric component present, the PI values are close to 1. (The asterisk mark (*) in Table 3 indicates the model is the most frequently selected one instead of being the true model.)

An illuminating example. We now look at a special example. We still generate data from a linear model with $\beta = (2, 2, 0.3, 0.3, 0.1, 0.1, 0, 0, 0, 0)^T$ and $\sigma = 2$. The pairwise correlation among the predictors is 0.5. For this example we do all-subset selection by BIC at different sample sizes. Our thinking is that since some of the coefficients are large and others are small, BIC is going to pick up the significant predictors gradually as the sample size increases. We expected to see both big and small PI values alternating to some degree when the sample size changes. In this example, we replicate 500 times for each sample size.

The results of median PIs at different sample sizes are shown in figure 6. From the plot we see PI first increases with the sample size, then decreases, then increases and decreases again, and finally increases. This is because when the sample size is small, most of the time BIC only picks up

x_1 and x_2 and the PI increases with the sample size. As the sample size further increases, BIC finds the predictors x_3 and x_4 relevant and the PI then decreases since the coefficients for x_3 and x_4 are small (but not too small) so that BIC is not quite sure about the best model. When the sample size gets big enough so that most of the times BIC chooses x_1, x_2, x_3, and x_4, the PI increases again with sample size. A similar story repeats for the predictors x_5, and x_6. If we choose 1.2 as a cutoff point, we would see (practically) parametric and (practically) nonparametric scenarios alternating as the sample size changes.

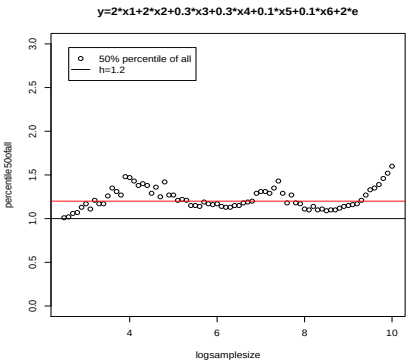

FIG 6. *Behavior of PI for the special example*

Inference after model selection (PI as practical identifiability). With Examples 3 and 4, we assess the accuracy of statistical inferences after model selection. We first generate an evaluation set of predictors from the same distribution as that for observations. Then for each replication, we generate data and do subset selection by BIC. After selecting a model, we get the resulting predicted value (estimated regression function at the given point), the standard error and the 95% confidence interval for the regression estimate at each of the new design points in the evaluation set. We replicate 500 times. Then at each new design point we calculate the actual standard deviation (which is called se.fit.t in the output table) of the 500 regression estimates and compare that to quantiles of the 500 standard errors that are obtained based on the selected model. We also take a look at the actual coverage of the 95% CIs for the true mean of the regression function at each new design point. Results at 10 randomly chosen new design points are reported in Table 5 and Table 6 for the two examples.

From the results we can see the actual coverages in Example 3 are reasonably close to 95% while the ones in Example 4 are much worse than the nominal 95% level. Also the (simulated) actual standard errors of the regression estimation are quite close to the ones from the selected model in Example 3, and in contrast, in Example 4, the reported uncertainty of regression estimation is grossly under-estimated. We tried several evaluation data sets with different sizes, the results are similar.

TABLE 5
Reliability of inference for Example 3

new design point	quantiles of standard errors of fit 5%	25%	50%	75%	95%	se.fit.t	coverage
1	0.439	0.474	0.496	0.525	0.583	0.564	0.930
2	0.351	0.375	0.396	0.416	0.480	0.457	0.934
3	0.333	0.356	0.375	0.395	0.446	0.471	0.932
4	0.359	0.386	0.405	0.428	0.464	0.453	0.932
5	0.360	0.385	0.405	0.426	0.493	0.498	0.926
6	0.229	0.247	0.258	0.271	0.349	0.350	0.920
7	0.391	0.420	0.443	0.468	0.516	0.582	0.906
8	0.398	0.426	0.447	0.473	0.509	0.502	0.926
9	0.631	0.679	0.712	0.747	0.809	0.738	0.960
10	0.238	0.254	0.265	0.277	0.296	0.252	0.972

TABLE 6
Reliability of inference for Example 4

new design point	quantiles of standard errors of fit 5%	25%	50%	75%	95%	se.fit.t	coverage
1	0.448	0.551	0.637	0.716	0.795	1.340	0.626
2	0.368	0.537	0.667	0.721	0.783	1.250	0.656
3	0.727	0.920	1.097	1.200	1.304	2.122	0.660
4	0.298	0.458	0.516	0.551	0.600	0.941	0.662
5	0.618	0.728	0.798	0.856	0.946	1.365	0.758
6	0.353	0.411	0.438	0.463	0.501	0.654	0.796
7	0.543	0.683	0.773	0.830	0.914	1.471	0.672
8	0.393	0.457	0.507	0.560	0.610	0.991	0.684
9	0.537	0.624	0.676	0.727	0.795	1.130	0.740
10	0.566	0.688	0.786	0.867	0.959	1.720	0.634

It is now well known that model selection has an impact on subsequent statistical inferences (see, e.g., [11, 8, 3, 9]). For observational data, typically one cannot avoid making various modeling choices (such as which type of statistical analysis to pursue, which kind of models to consider) after seeing the data, but their effects are very difficult to quantify. Thus it can be very helpful to know when the choices have limited impact on the final results. The above results together with Tables 3 and 4 show that the value of PI can provide valuable information on the parametricness of the underlying regression function and hence on how confident we are on the accuracy of subsequent inferences.

Combining strengths of AIC and BIC based on PI. Still with Examples 3-7, we investigate the performance of an adaptive choice between AIC and BIC based on the PI value. Again we first generate an evaluation data set

with 500 new design points from the same distribution as the one for observations. Then for each replication, we use both AIC and BIC to select a model. The combined procedure is BIC if the PI value is larger than a cutoff point (chosen as 1.2 in these examples) and AIC otherwise. Then for each procedure (AIC, BIC, and the combined) in each replication, we calculate the average squared error (which is the average squared difference between the true regression mean and the fitted value based on the selected model) at the new design points in the evaluation data. We replicate 500 times and the statistical risk is estimated to be the average of the 500 average squared errors. The risk ratios are reported in Table 7 with BIC as the reference.

TABLE 7
Statistical risks of AIC, BIC, and the Combined procedure

	Statistical Risk			Risk Ratio		
Example	AIC	BIC	Combined	AIC	BIC	Combined
3	0.335	0.227	0.230	1.474	1.000	1.014
4	0.543	1.045	0.680	0.520	1.000	0.651
5	0.562	0.513	0.564	1.096	1.000	1.098
6	0.502	0.402	0.459	1.250	1.000	1.142
7	0.835	0.927	0.899	0.901	1.000	0.969

From the results we see in all these examples the combined procedure shows capability of adaptation between AIC and BIC in terms of the statistical risk. We also see from Tables 7, 3 and 4 that in Examples 5 and 7, the PIs are roughly around 1.2 and AIC and BIC have similar performance in terms of statistical risks, while in the other examples the PIs are either large or small and correspondingly, either BIC or AIC has a smaller statistical risk. These results show that PI provides helpful information regarding whether AIC or BIC works better or they have similar performances in statistical risks. Therefore, PI can be viewed as a **P**erformance **I**ndicator of AIC versus BIC.

1.4. *A summary.* From our simulation outcomes (some are not presented due to space limitation), we summarize a few points here.

1. Factors other than the nature of the regression function also influence the value of PI, including the sample size and the noise level. From a practical point of view, PI, as a diagnostic measure, indicates whether a specific problem is 'practically' parametric/nonparametric with the influences of all those factors.
2. Model selection effect on subsequent statistical inferences may or may not be reasonably ignored, and the value of PI provides useful information in that regard.

3. For Examples 3 and 4, both being parametric, one is practically parametric and the other practically nonparametric for $n = 200$. Correspondingly, BIC works better for the former and AIC for the latter in terms of risk for estimating the regression function. This phenomenon will be seen again in the next section.
4. Combining AIC and BIC based on the PI value shows adaptation capability in terms of statistical risk. That is, the composite rule yields a risk close to the better one of AIC and BIC.
5. In nested model problems (like order selection of series expansion), a cutoff point of $c = 1.6$ seems to be good. In subset selection problems, we expect the cutoff point to be smaller since the infimum is taken over many models. The choice of 1.2 seems to be reasonably good based on our numerical investigations, which is also supported by the observation that when PI is around 1.2, AIC and BIC perform similarly.

2. Real Data Examples. In this section, we study three data sets: the Ozone data (e.g. [1]), the Boston housing data (e.g. [6]), and the Diabetes data (e.g. [2]).

In these examples, we conduct all subset selection by BIC using the Branch and Bound algorithm. Besides finding the PI values for the full data, we also do the same with sub-samples from the original data at different sample sizes. In addition, we carry out a parametric bootstrap from the model selected by BIC based on the original data to assess the stability of model selection. (The design points of the predictors are randomly selected with replacement from the original data.) Like in the multiple-predictor simulation study, we also combine AIC and BIC based on the PI value when doing parametric bootstrap. Unless otherwise stated, the subsampling and the bootstrap are both replicated 500 times at each sample size. (In the results, the predictors are coded to be a single digit between 1 and 9 and then a single capital letter between 'A' and 'Z', i.e, letter 'A' stands for the 10th predictor, 'B' for the 11th, and so on.)

Ozone Data. There are 9 variables with 8 predictors and 330 observations. We followed the transformations of the predictors and the response suggested by Hawkins [7]. (In that paper a ninth predictor, day of the year, was also included. We left this predictor out as many others did. See [1].) After the transformations, we have 10 predictors with quadratic terms of two predictors added.

Boston Housing Data. The data consists of 14 variables (1 response and 13 predictors). There are 506 observations. We followed the transformations of the variables in Harrison and Rubinfeld's paper [6].

Diabetes Data. There are 11 variables with 10 predictors and 442 observations.

The PIs from the original data for these three examples are: 1.277 (ozone), 1.028 (Boston housing), and 1.298 (diabetes). The results of subsampling and bootstrap are reported in Tables 8-9 and Tables 10-11, respectively.

TABLE 8
Quartiles of PIs from subsamples of size 400

Data	Q1	Q2	Q3
Ozone	-	-	-
Boston	1.02	1.04	1.1
Diabetes	1.17	1.23	1.28

TABLE 9
Quartiles of PIs from subsamples of size 200

Data	Q1	Q2	Q3
Ozone	1.08	1.21	1.47
Boston	1.02	1.05	1.11
Diabetes	1.06	1.13	1.24

TABLE 10
The 6 most frequently selected models and their frequencies with a sample size of 400

	Ozone		Boston Housing		Diabetes	
	model	proportion	model	proportion	model	proportion
1	-	-	145689ABCD	0.28	23479	0.732
2	-	-	15689ABCD	0.238	3479	0.078
3	-	-	15689ABD	0.092	349	0.058
4	-	-	145689ABD	0.084	23489	0.016
5	-	-	145689BCD	0.062	2349	0.012
6	-	-	14568BCD	0.046	3459	0.012

TABLE 11
The 6 most frequently selected models and their frequencies with a sample size of 200

	Ozone		Boston Housing		Diabetes	
	model	proportion	model	proportion	model	proportion
1	1269	0.474	15689ABCD	0.088	23479	0.318
2	126	0.248	15689ABD	0.088	349	0.17
3	1236	0.06	1568BD	0.07	39	0.128
4	1239	0.046	1589ABD	0.062	3479	0.102
5	167	0.028	14568BD	0.05	2349	0.042
6	12	0.012	1568BCD	0.044	379	0.042

From the tables, we see the PIs for the ozone data are mostly larger than 1.2, while those for the Boston housing data are smaller than 1.2.

Moreover, the parametric bootstrap suggests that for the Ozone data, the model selected from the full data still reasonably stands out even when the sample size is reduced to about 200 and noises are added (not all shown due to space limitation). For the Boston housing data, however, even at a sample size of 400 we only have 28% chance selecting the same model as the one selected with the full data. Interestingly, the diabetes data exhibit a parametric behavior when $n = 400$, but with the sample size reduced by half, it looks more like a nonparametric scenario.

TABLE 12
Combining AIC and BIC based on PI with full sample size

	Statistical Risk			Risk Ratio		
Data	AIC	BIC	Combined	AIC	BIC	Combined
Ozone	7.66e-4	6.44e-4	6.82e-4	1.189	1.000	1.060
Boston Housing	8.18e-4	1.05e-3	8.65e-4	0.779	1.000	0.824
Diabetes	63.05	57.42	58.19	1.098	1.000	1.014

Combining AIC and BIC based on PI. Similar to the simulation results in the previous section, by parametric bootstrap at the original sample size from the selected model, in these data examples, combining AIC and BIC based on PI shows good overall performance in terms of statistical risk (Table 12). The combined procedure has a statistical risk close to the better one of AIC and BIC in each case.

REFERENCES

[1] Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.*, **80**, 580-598.

[2] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R., (2004). Least angle regression. *Ann. Statist.*, **32**, 407-451.

[3] Faraway, J.J. (1992). On the Cost of Data Analysis. *J. Computational and Graphical Statist.*, **1**, 213229.

[4] Geyer, C. and Shaw, R. (2008). Model selection in estimation of fitness landscapes. *Technical Report.*, University of Minnesota.

[5] Hand, D. J. (1981). Branch and bound in statistical data analysis. *The Statistician.*, **30**, 1-13.

[6] Harrison, D. and Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air. *J. Environmental Economics Management.*, **5**, 81-102.

[7] Hawkins, D. (1989). Flexible parsimonious smoothing and additive modeling: discussion. *Technometrics*, **31**, 31-34.

[8] Hurvich, C. M., and Tsai, C.-L. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, **44**, 214217.

[9] Kabaila P., Leeb H. (2006). On the large-sample minimal coverage probability of confidence intervals after model selection. *Journal of the American Statistical Association*, **101**, 619-629.

[10] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B*, **58**, 267-288.

[11] Zhang, P. (1990). Inference after variable selection in linear regression models. *Biometrika*, **79**, 741-746.

Wei Liu
School of Statistics
University of Minnesota
313 Ford Hall
224 church street S.E.
Minneapolis, MN 55455, US
E-mail: william050@stat.umn.edu

Yuhong Yang
School of Statistics
University of Minnesota
313 Ford Hall
224 church street S.E.
Minneapolis, MN 55455, US
E-mail: yyang@stat.umn.edu
http://www.stat.umn.edu/~yyang