Cross-Validation for Selecting the Best Model/Procedure

Yuhong Yang

School of Statistics

University of Minnesota

Collaborators: Yongli Zhang and Zishu Zhan

Outline

- The need of selection of a tuning parameter or a model/method/algorithm
- We may also want to do selection[2]
- Cross-validation is a general solution, but also a tricky business!
- Consistent selection of a model selection/regression procedure
- Stop the war between AIC and BIC by CV
- Numerical results for insight and illustration
- Profile Electoral College CV

The need of selection of a model selector or a modeling procedure

- There are many high-dimensional model selection methods. Conditions for their optimality (if any) are typically uncheckable. So which one to choose for the data at hand?
- One may also need to choose among parametric and nonparametric procedures.
- There are many non-traditional leaning methods (e.g., random forest, deep learning).
- Cross-validation provides a possible solution.

CV is a tricky business!

- The current theory does not provide enough guidance on how to implement CV for high-dimensional regression
- The center piece is the data splitting ratio
- There are various recommendations/folklores

Some popular ones

- "The best method to use for model selection is 10-fold CV"
- Better estimation (e.g., in bias and variance) of the prediction error by CV means better model selection
- Leave-one-out(LOO) CV has smaller bias but larger variance than leave-more-out CV

Are these statements mostly right?

CV Paradox

- Suppose a statistician's original data splitting scheme works for consistency in selection.
- The same amount of (or more) independent and identically distributed data is given to the statistician.
- He decides to add half of the new data to the estimation part and the remaining half to the evaluation part.
- He naturally thinks that with improvement in both the training and evaluation components, the comparison of the candidate procedures becomes more reliable.

Is that the case?

A simulation

We compare two different uses of Fisher's LDA method.

- $n = 100$
- For 40 observations with $Y = 1$, we generate three independent random variables X_1, X_2, X_3, all standard-normally distributed
- For the remaining 60 observations with $Y = 0$, we generate the three predictors with $N(0.4, 1)$, $N(0.3, 1)$ and $N(0, 1)$ distributions
- We compare LDA based on only X_1 and X_2 with LDA based on all of the three predictors.

Is MORE automatically helpful for selecting the better procedure? We evenly split the additional observations. The initial data splitting ratio is 30/70.

$n = 100$	300	500	700	900
0.835	0.825	0.803	0.768	0.772

How about maintaining the ratio of 30/70 in data splitting?

$n = 100$	300	500	700	900
0.835	0.892	0.868	0.882	0.880

How about an increasing ratio in favor of evaluation size?

Say, 70%, 75%, 80%, 85%, and 90%, respectively.

$n = 100$	300	500	700	900
0.835	0.912	0.922	0.936	0.976

When the estimation size is increased by e.g. half of the original sample size, since the estimation accuracy is improved for both of the classifiers, their difference may no longer be distinguishable with the same order of evaluation size (albeit increased).

The surprising requirement of the evaluation part in CV to be dominating in size (i.e., $n_2/n_1 \to \infty$) for differentiating nested parametric models was first noted by Shao (1993) in the context of linear regression. Yang (2007) handled general comparisons of regression procedures.

Why is CV so tricky?

- There are three main goals of CV
- They are obviously related, but not the same
- Unfortunately, recommendations were often made without distinguishing the goals
- For high-dimensional regression, there are new issues related to proper use of CV. For instance, when the true model size increases, the true model and a model with one more term may be much closer to each other than that when the true model is fixed.

Three main goals in the use of CV

- The first is for estimating the prediction performance of a model or a modeling procedure.
- The second and third goals are often both under the same name of CV for model selection. However, there are different objectives of model selection, one as an internal step in the process of producing the final estimator, the other as for identifying the best candidate model or modeling procedure.
- The second use of CV is to choose a tuning parameter of a procedure or a model/modeling procedure among a number of possibilities with the end goal of producing asymptotically the best performance.

- The third use of CV is to figure out which model/modeling procedure works the best for the data.

The optimal data splitting ratio for the different goals are usually different!

- The second and third goals are closely related.
- The third use of CV may be applied for the second goal, i.e., the declared best model/modeling procedure can be then used for estimation, which can result in asymptotically optimal performance in estimation.
- A caveat is that this asymptotic optimality may not always be satisfactory. For instance, when selecting among parametric models in a practically nonparametric situation with the fixed true model being one of the candidates, a model selection method built for the third goal (such as BIC) may perform very poorly for the second goal. (*Conflict between point-wise optimality and minimax optimality. See e.g., Yang, 2005*)
- In the reverse direction, the best CV for the second goal does not necessarily imply the achievement of the third goal.

Our objectives

- Understand how data should be split for consistent selection of the best model selection method or a modeling procedure in high-dimensional regression
- As an application, we use CV to stop the "war" between AIC and BIC
- Conduct a serious numerical work to gain insight on the related matters.

Regression problem:

- Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be iid observations
- Data generating model

$$Y = \mu(\mathbf{X}) + \varepsilon,$$

  - μ is the unknown true regression function to be estimated

- Multiple general learning methods are considered.

Definitions

For a general result, we need some preparation.

Definition 1: Let $0 < \xi_n \leq 1$ be a sequence of positive numbers. Procedure δ_1 (or $\{\widehat{\mu}_{n,\delta_1}\}_{n=1}^{\infty}$, or simply $\widehat{\mu}_{n,\delta_1}$) is asymptotically ξ_n-better than δ_2 (or $\{\widehat{\mu}_{n,\delta_2}\}_{n=1}^{\infty}$, or $\widehat{\mu}_{n,\delta_2}$) under the L_2 loss if for every $0 < \epsilon < 1$, there exists a constant $c_\epsilon > 0$ such that when n is large enough,

$$P\big(\, \|\mu - \widehat{\mu}_{n,\delta_2}\|_2^2 \geq (1 + c_\epsilon \xi_n^2)\, \|\mu - \widehat{\mu}_{n,\delta_1}\|_2^2 \,\big) \geq 1 - \epsilon. \tag{1}$$

Definition 2: A procedure δ (or $\{\widehat{\mu}_{n,\delta}\}_{n=1}^{\infty}$) is said to converge exactly at rate $\{a_n\}$ in probability under the loss L_2 if $\|\mu - \widehat{\mu}_{n,\delta}\|_2 = O_p(a_n)$, and for every $0 < \epsilon < 1$, there exists $c'_\epsilon > 0$ such that when n is large enough, $P\big(\, \|\mu - \widehat{\mu}_{n,\delta}\|_2 \geq c'_\epsilon a_n \big) \geq 1 - \epsilon$.

Main theorem for consist selection by CV

Suppose there are a finite number of candidate procedures in Λ. Consider a procedure $\delta \in \Lambda$ that produces $\widehat{\mu}_{n,\delta}$ at each sample size n. Let $\widehat{\mu}_{n,\widehat{\delta}}$ be the estimator of μ based on the procedure $\widehat{\delta}$ selected by CV among the $|\Lambda|$ candidates.

- The error variances $E(\varepsilon_i^2|\mathbf{x})$ are upper bounded by a constant $\overline{\sigma}^2 > 0$ almost surely for all $i \geq 1$.
- Under the L_2 loss, for some $\xi_n > 0$, one of the procedures is asymptotically ξ_n-better than any other procedure considered.
- For each $\delta \in \Delta$, the estimator $\widehat{\mu}_{n,\delta}$ converges exactly at rate $a_{n,\delta}$.

Let CV_v and CV_a be CV based on voting and averaging, respectively.

Let $\underline{a}_n$ denote the minimum of $a_{n,\delta}$ over $\delta \in \Lambda$, except that the best procedure is excluded.

Let $\mathcal{S}$ be a collection of data splittings at the same ratio of training verse evaluation.

Theorem: If the data splitting ratio satisfies

- i. $n_v \to \infty$ and $n_t \to \infty$;
- ii. $\sqrt{n_v}\xi_{n_t}\underline{a}_{n_t} \to \infty$,

then the delete-n_v CV_v is consistent for any set $\mathcal{S}$, i.e., the best procedure is selected with probability approaching 1. It follows that the CV_v selection is asymptotically optimal:

$$\frac{\|\mu - \widehat{\mu}_{n,\widehat{\delta}}\|_2}{\inf_{\delta\in\Lambda} \|\mu - \widehat{\mu}_{n,\delta}\|_2} \to 1 \text{ in probability.}$$

If the size of $\mathcal{S}$ is uniformly bounded, then CV_a has the same asymptotic properties as CV_v above.

Conflicting properties of AIC and BIC

Consider a traditional situation of using a sequence of nested models for regression modeling.

- When one of the candidate models holds, BIC is consistent in selection, but AIC is not. BIC is also asymptotically efficient, but AIC is not.
- When none of the candidate models holds, AIC is asymptotically efficient, but BIC is not.
- In the parametric case, BIC is pointwise-risk adaptive, but AIC is minimax-rate adaptive. The price of being consistent is the inflation of worse-case risk by a factor of

$$\log(1/\text{prob. of over fitting}).$$

AIC or BIC?

- Some researchers rule out BIC based on the argument that parametric models cannot be right.
- A popular saying is that: Use AIC for a nonparametric situation and use BIC for a parametric situation.
- A more accurate statement is: Use AIC for a *practically* nonparametric situation and use BIC for a *practically* parametric situation.
- A parametricness index can help us decide which scenario we are in (Liu and Yang, 2012; Ding, Tarokh, and Yang, 2017+).

Let δ be a model selection rule to choose between $\widehat{f}_{n,1}(x)$ and $\widehat{f}_{n,2}(x)$. Let A_δ be the event that model 2 is selected.

The risk of the estimator based on δ at a given x_0 is

$$E\left(f(x_0) - \left(\widehat{f}_{n,1}(x_0)I_{A_\delta^c} + \widehat{f}_{n,2}(x_0)I_{A_\delta}\right)\right)^2$$

A commonly told story is: BIC should be used for a parametric case and AIC should be used for a nonparametric case.

This is far from being an accurate statement!

A simple demonstration

Consider

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, 2, ..., n,$$

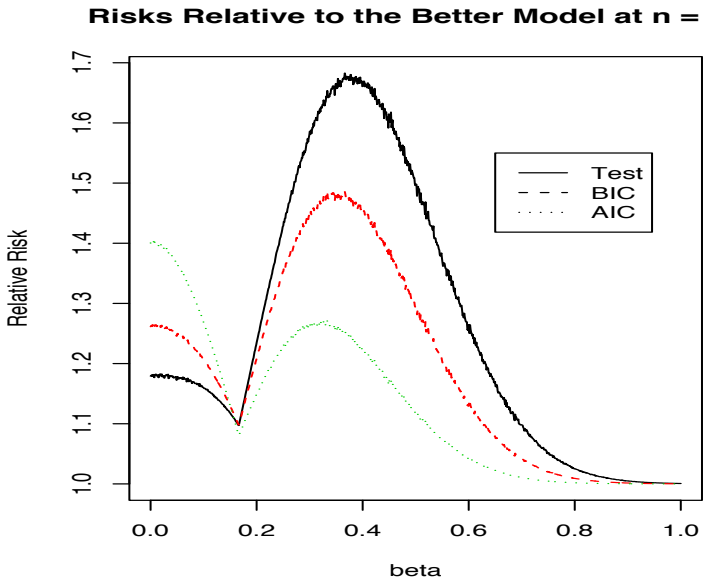
- $x \in [-1, 1]$ is the design variable with $\overline{x}_n = 0$
- $\{\varepsilon_i\}$ are Gaussian errors

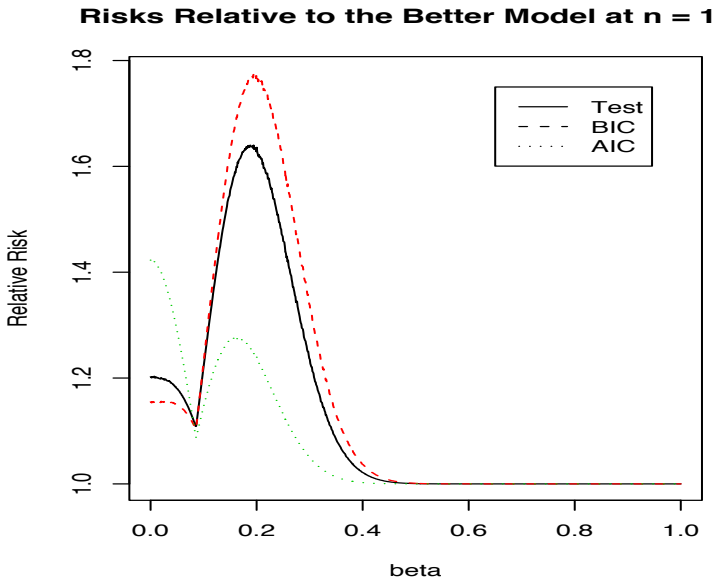
Our interest: point prediction of Y at a new value x_0 under the squared error loss

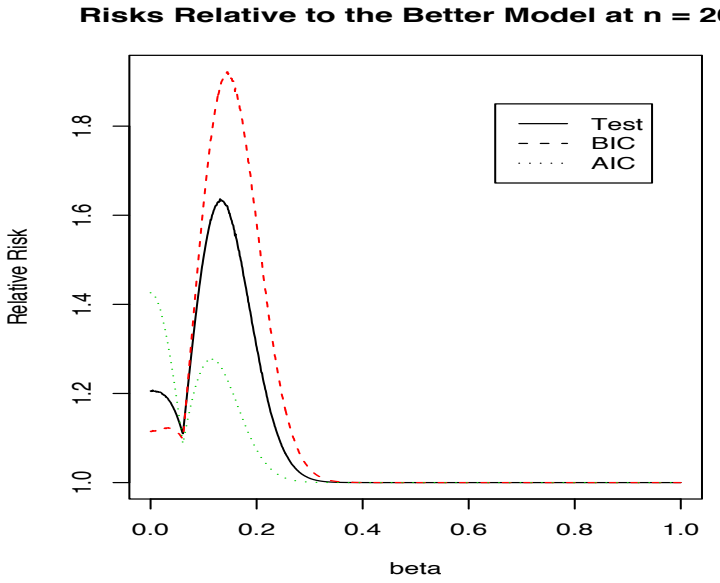
Model 0: $Y_i = \alpha + \varepsilon_i$

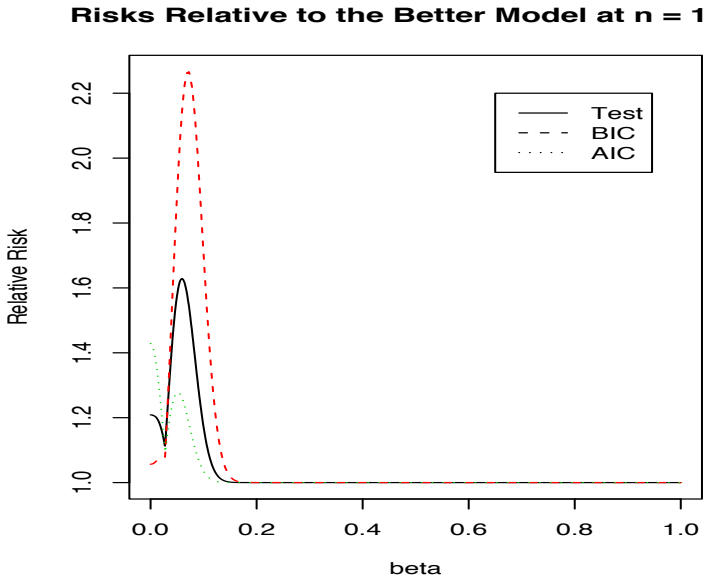
Model 1: $Y_i = \alpha + \beta x_i + \varepsilon_i$

- $n = 25, 100, 200, 1000$
- $x_0 = 0.5$
- $\sigma = 0.5$


Risks Relative to the Better Model at n = 25
Relative Risk
beta
Test
BIC
AIC
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
0.0
0.2
0.4
0.6
0.8
1.0


Risks Relative to the Better Model at n = 100
Relative Risk
beta
Test
BIC
AIC
1.0
1.2
1.4
1.6
1.8
0.0
0.2
0.4
0.6
0.8
1.0


Risks Relative to the Better Model at n = 200
Relative Risk
beta
Test
BIC
AIC
1.0
1.2
1.4
1.6
1.8
0.0
0.2
0.4
0.6
0.8
1.0


Risks Relative to the Better Model at n = 1000
Relative Risk
beta
Test
BIC
AIC
1.0
1.2
1.4
1.6
1.8
2.0
2.2
0.0
0.2
0.4
0.6
0.8
1.0

The nice property of BIC being asymptotically optimal in this setting is not quite in line with the simulation results.

Stopping the war between AIC and BIC

Can we design a super criterion to resolve the conflicts?

- Asymptotic efficiency of BIC and AIC shown for different situations can be shared (Yang, 2007; Ing, 2007; Liu and Yang, 2012; Zhang and Yang, 2015; Ding, Tarokh, and Yang, 2017+)
  - Based on choosing between AIC and BIC;
  - Or using a parametricness index
- Consistency of BIC and minimax-rate optimality of AIC cannot be resolved (Yang, 2005). See also Erven, Grunwald and de Rooij (2012).

Many issues are subtle. A bottom line is: You may not hit two hoops in one shoot!

Stop the war between AIC and BIC by CV

- For illustration, consider estimating a regression function on [0,1] based on series expansion. A model selection method, e.g., AIC and BIC, is used to select the order of the expansion.
- It is well-known that AIC and BIC have optimality for typical nonparametric and parametric situations respectively.
- But one does not know which scenario describes the data well.
- CV, at the second-level of model selection, comes to rescue.
- We use CV to choose between AIC and BIC.

Assumption 1: In the nonparametric case, we suppose AIC is asymptotically efficient in the sense that $\|\mu - \widehat{\mu}_{n,AIC}\|_2 / \inf_{M \in \mathcal{M}} \|\mu - \widehat{\mu}_{n,M}\|_2 \to 1$ in probability. BIC is suboptimal in the sense that there exists a constant $c > 1$ such that with probability going to 1, we have

$$\|\mu - \widehat{\mu}_{n,BIC}\|_2 / \inf_{M \in \mathcal{M}} \|\mu - \widehat{\mu}_{n,M}\|_2 \geq c$$

In the parametric case, BIC is consistent in selection.

The assumption holds for various nonparametric and parametric situations.

Proposition: Consider the delete-n_v CV with

- $n_t \to \infty$
- $n_t = o(n_v)$

to choose between AIC and BIC. Suppose Assumption 1 is satisfied. Then the CV method is consistent for selection between AIC and BIC in the sense that when the true model is among the candidates, the probability of BIC being selected goes to 1; and when the true regression function is infinite-dimensional, then with probability going to 1 AIC is selected.

So the second level CV achieve adaptivity between AIC and BIC. Note that the first level CV on the series expansion models cannot do the same.

Numerical work

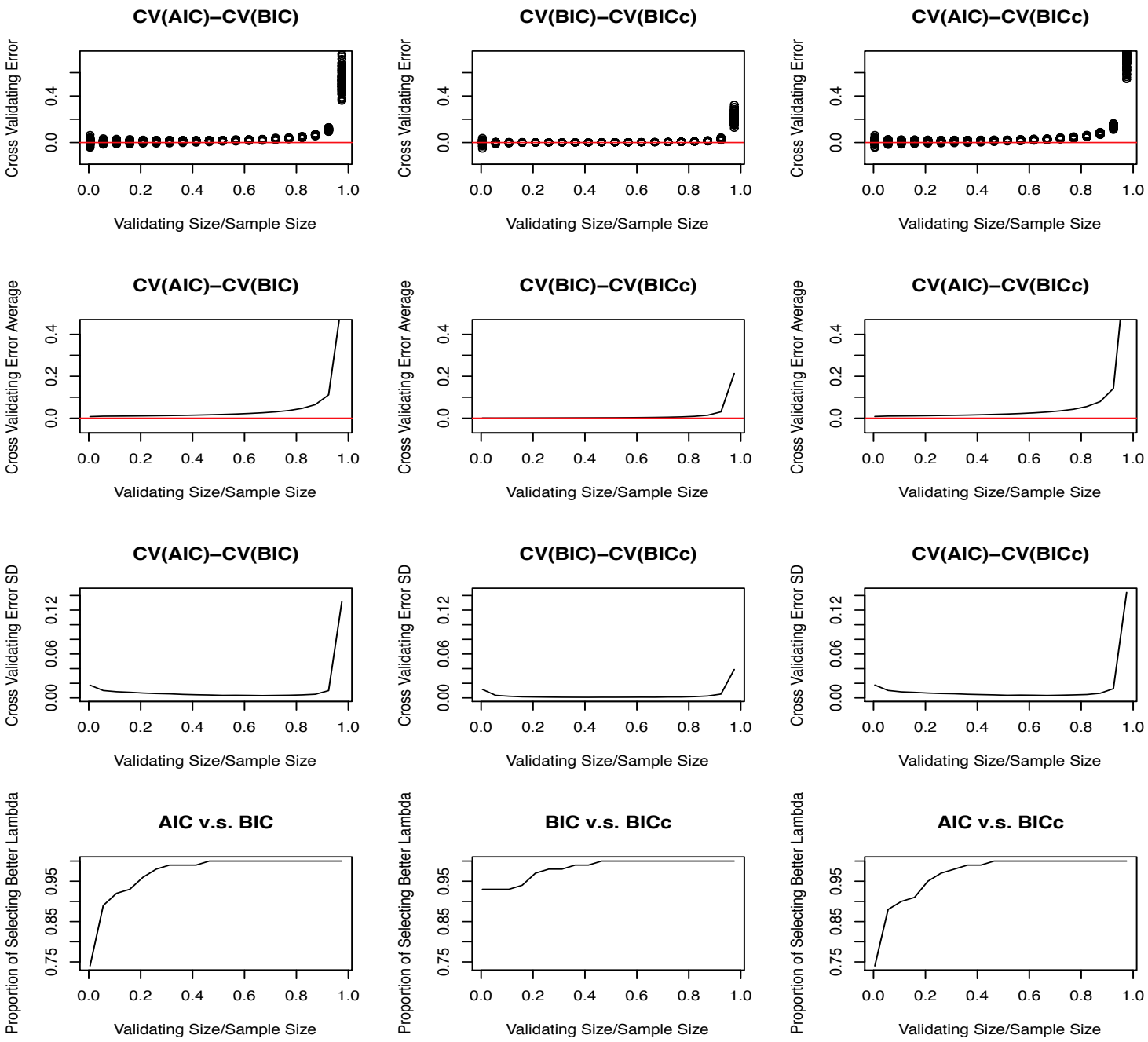
Consider linear models. The design matrix $X = (X_{i,j})$ $(i = 1, \cdots, n;$ $j = 1, \cdots, p_n)$ is $n \times p_n$ and each row of X is generated from the multivariate normal distribution with mean $\mathbf{0}$ and an AR(1) covariance matrix with marginal variance 1 and autocorrelation coefficient ρ, $\rho = -0.5$ and 0.5. The responses are generated from the model

$$Y_i = \sum_{j=1}^{p_n} \beta_j X_{i,j} + \varepsilon_i \tag{2}$$

where $\varepsilon_i' s$ $(i = 1, \cdots, n)$ are iid $N(0, 1)$ and $\beta = (\beta_1, .., \beta_{p_n})^T$ is a p_n-dimensional vector including q_n nonzero coefficients and $(p_n - q_n)$ zeros.

- A representative parametric scenario: $(\beta_1, \beta_2) = (2, 2)$ and $\beta_j = 0$ $(3 \leq j \leq 20)$. BIC is the star.

- A "practically nonparametric" scenario: $\beta_j = 1/j$ $(j = 1, \cdots, 20)$, where, with p_n fixed at 20 and n not very large (e.g., around 1000), AIC is the star.


CV(AIC)−CV(BIC)
CV(BIC)−CV(BICc)
CV(AIC)−CV(BICc)
Cross Validating Error
Cross Validating Error Average
Cross Validating Error SD
Proportion of Selecting Better Lambda
Validating Size/Sample Size
AIC v.s. BIC
BIC v.s. BICc
AIC v.s. BICc

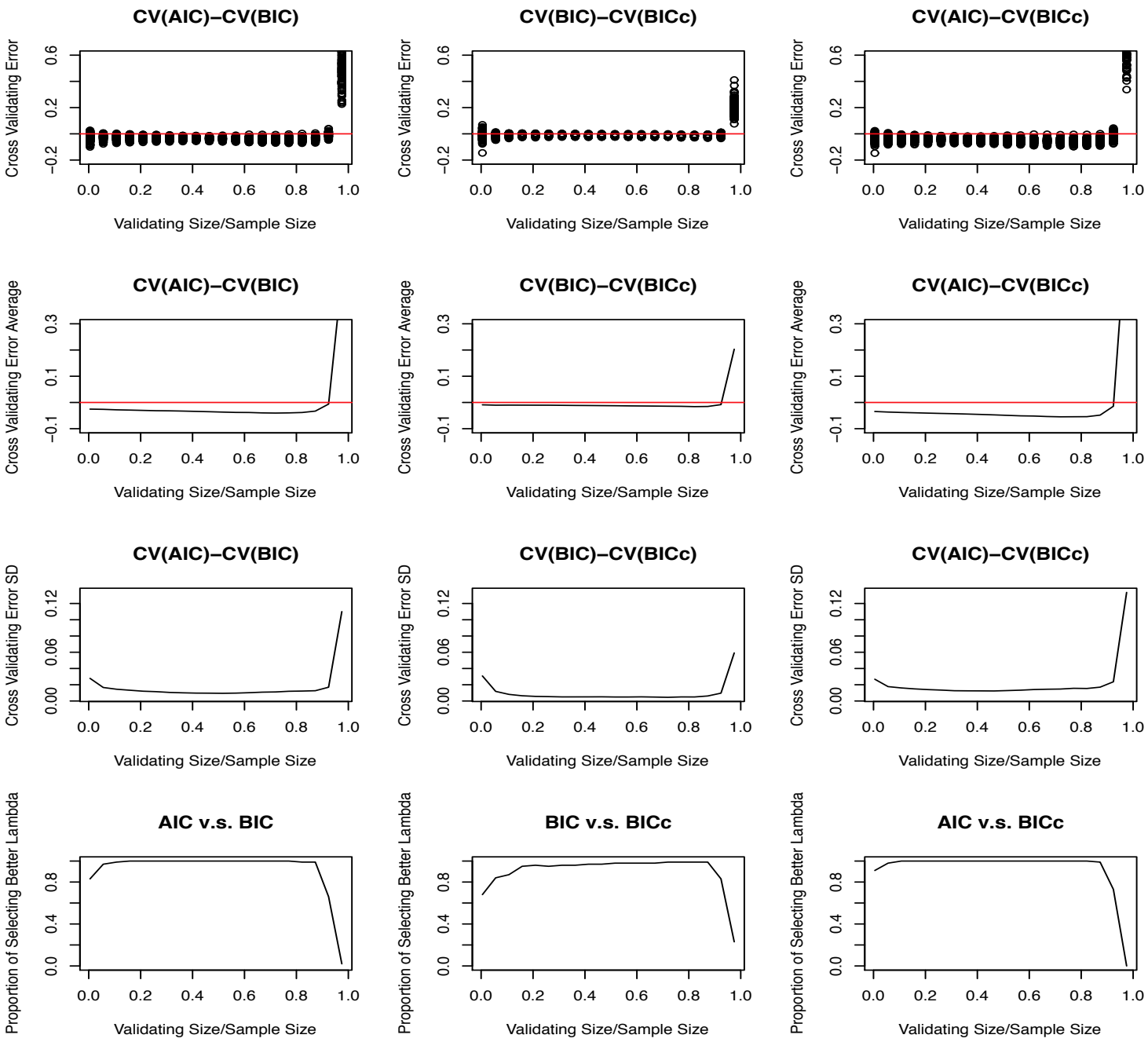

CV(AIC)−CV(BIC)
CV(BIC)−CV(BICc)
CV(AIC)−CV(BICc)
Cross Validating Error
Cross Validating Error Average
Cross Validating Error SD
Proportion of Selecting Better Lambda
AIC v.s. BIC
BIC v.s. BICc
AIC v.s. BICc
Validating Size/Sample Size

Table 1: Comparison of AIC, BIC and CV (with 400 data splittings) in terms of MSE (in the unit of $1/n$) based on 500 replications with $\sigma = 1$, $p_n = 15$, $\beta_j = 0.25/j$ $(1 \le j \le 10)$ and $\beta_j = 0$ $(11 \le j \le 15)$. The standard errors (in the unit of $1/n$) are shown in the parentheses.

n	AIC	BIC	del-0.5n	del-0.8n	del-0.2n	10-fold
$\rho = -0.5$						
100	13.23 (0.33)	8.92 (0.27)	8.01 (0.25)	7.99 (0.23)	8.94 (0.30)	9.65 (0.31)
10000	17.01 (0.32)	32.68 (0.45)	18.26 (0.41)	20.96 (0.54)	18.43 (0.40)	18.82 (0.40)
500000	13.53 (0.25)	10.93 (0.21)	11.01 (0.22)	10.93 (0.21)	11.41 (0.24)	11.87 (0.25)
$\rho = 0.5$						
100	14.64 (0.32)	12.63 (0.26)	12.21 (0.24)	12.18 (0.23)	13.07 (0.27)	13.27 (0.28)
10000	16.33 (0.31)	28.28 (0.43)	16.33 (0.31)	19.33 (0.41)	17.10 (0.37)	18.01 (0.37)
500000	13.76 (0.25)	10.83 (0.21)	10.89 (0.21)	10.83 (0.21)	11.34 (0.23)	11.68 (0.23)

Table 2: Comparison of SCAD, MCP, LASSO, STRIC (Stepwise plus RIC) and CV (with 400 data splittings) in terms of MSE (in the unit of $1/n$) based on 500 replications with $\sigma = 1$, $n = 500$, $p_n = 500$, $\beta_j = 6/j$ for $j \leq q_n$ and $\beta_j = 0$, otherwise. The standard errors (in the unit of $1/n$) are shown in the parentheses.

q_n	SCAD	MCP	LASSO	STRIC	del-0.5n	10-fold
$\rho = -0.5$						
1	1.01 (0.06)	1.03 (0.06)	45.36 (0.57)	4.84 (0.31)	1.02 (0.07)	1.95 (0.2)
5	5.13 (0.14)	5.13 (0.14)	411.50 (3.12)	10.34 (0.44)	5.14 (0.15)	6.65 (0.31)
10	255.02 (3.46)	105.37 (3.79)	834.40 (4.89)	14.79 (0.38)	14.79 (0.38)	14.51 (0.38)
$\rho = 0.5$						
1	1.07 (0.07)	1.07 (0.06)	46.53 (0.58)	4.62 (0.29)	1.10 (0.10)	2.11 (0.21)
5	18.26 (1.14)	7.57 (0.29)	205.53 (1.39)	9.14 (0.36)	6.40 (0.20)	7.39 (0.27)
10	425.97 (4.51)	238.29 (4.71)	369.36 (2.34)	14.20 (0.38)	14.20 (0.38)	14.20 (0.38)

Profile Electoral College Cross-Validation

Zishu Zhan and Yuhong Yang (2022). *Information Sciences*, 586, 24-40.

- It is hard to digest the differences of the CV errors of competing methods. Are the differences statistically meaningful?
- It is useful to repeat k-fold CV a number of times and then record the number of times a candidate wins. This is called *electoral college CV*, or EC-CV. The winning frequencies of the candidate models/procedures give information on how comfortably the best candidate beats the others.
- No single data splitting ratio works the best generally.
- How about considering multiple data splitting ratios? The profile of the winning frequencies presents a better and often much more clear understanding on the relative strengths of the candidates.

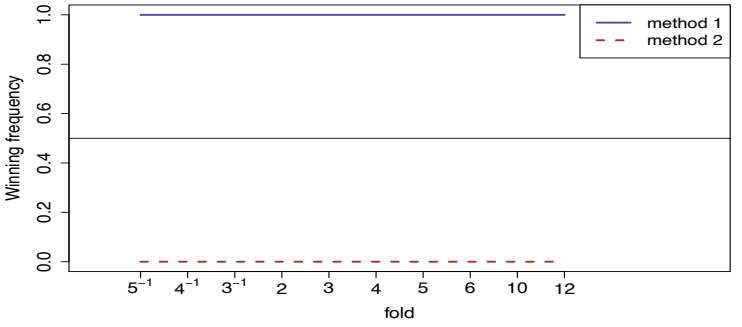


Figure 1: Common pattern 1 of PEC-CV: Dominating

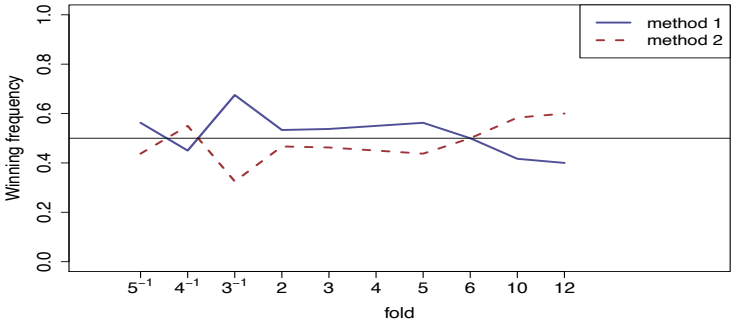


Figure 2: Common pattern 2 of PEC-CV: Indifferent

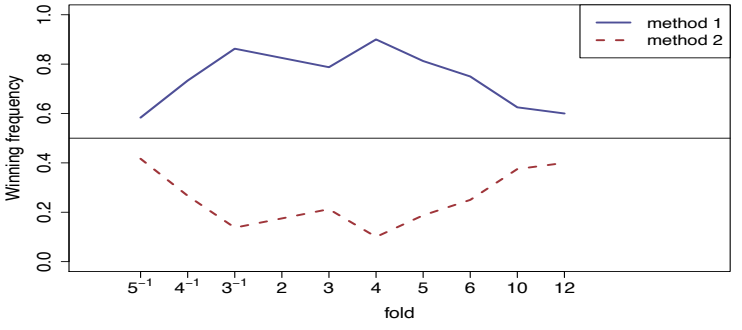


Figure 3: Common pattern 3 of PEC-CV: Marginally better

- Pattern 3 (*Marginally better*). As shown in Fig. 3, consistently at all the different DSRs, one method is clearly above the 50% line, but to a limited degree, especially towards the end. This profile suggests that we are quite confident that one method is generally better, but only marginally so.

- Pattern 4 (*Strong switching*). This is illustrated in Fig. 4. Here, the winning frequency of one method (denoted as method 1) is below 50% in the beginning but increases substantially (passing 50% line eventually) as the

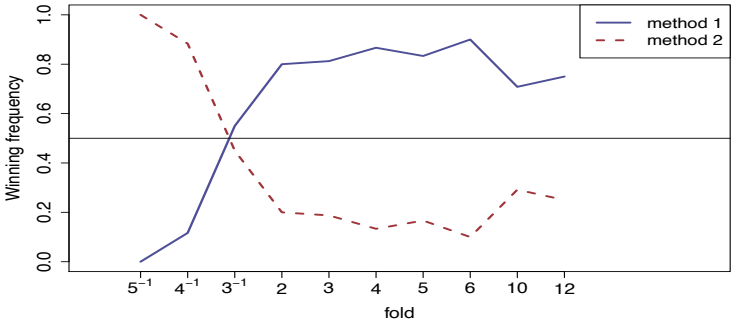


Figure 4: Common pattern 4 of PEC-CV: Strong switching

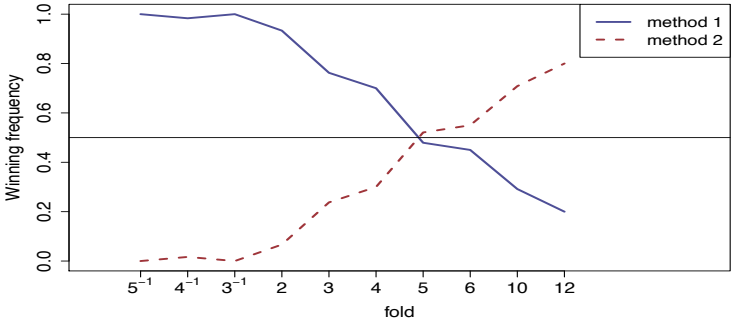


Figure 5: Common pattern 5 of PEC-CV: Marginal switching

training sample size increases. The switching can happen early or late, but the winning frequency of method 1 is close to 1 towards the end. This situation may happen when the estimation process of method 1 is more complex than method 2, and therefore, method 1 suffers from the low sample size more than method 2. When the training size gets large enough, method 1 becomes clearly more competitive than method 2. This profile shows that the selection of method 1 at the full sample size is a confident one (but method 2 may be preferred at a smaller sample size).

- Pattern 5 (*Marginal switching*). As depicted in Fig. 5, method 2 has transitioned to be better as the number of folds increases, but its advantage over method 1 is not quite certain. Here when k gets larger, the number of observations used to assess the predictive performance becomes smaller, making the comparison of the candidates less certain for a tough situation (which is the reason behind the cross-validation paradox). For this delicate case, if we just use the regular 10-fold CV, the selection outcome is quite random and unreliable, and we have to handle it with kid gloves. The profile CV provides a bigger picture and it suggests that method 2 might be better than method 1 with more training samples, but the confidence level on this is low. Since method 1 is actually better, the decision to choose method 2 by focusing only on 10-fold or 12-fold would be wrong. The PEC-CV profile can clearly warn against simply trusting the selection outcome by the popular 10-fold. Furthermore, the PEC-CV may actually prefer method 1 as the winner when integrating together the performances at the different DSRs (see Section 4.3).

- Pattern 6 (*V-shape*). As shown in Fig. 6, method 1 has winning frequencies above 50%, but it roughly has a

V-shape: the winning frequency of method 1 is high at both ends but lower in the middle. The profile suggests that the choice of method 1 is most likely safe at the full sample size. We note that sometimes it may be possible that the winning frequency of method 1 can drop close to or even slightly below 0.5 in the middle. Note that such a profile can occur when comparing model selection methods. For example, when using BIC [23] or Least Absolute Shrinkage and Selection Operator (LASSO) [28] to select among linear regression models, when the true coefficients have different magnitudes, it has been observed in the literature that when the training sample size is small, LASSO performs better, but when the sample size gets larger, BIC performs better, and when the sample size gets much larger, LASSO wins back the competition [18].

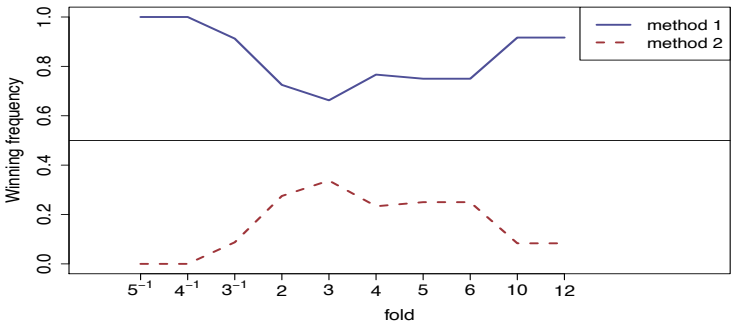


Figure 6: Common pattern 6 of PEC-CV: V-shape

- Pattern 7 (*Asymmetric*). This pattern is unique in some sense. Sometimes, the two competing methods may give identical regression estimates. For instance, in the comparison of BIC and BICc [15, 35], they may actually select the same model, in which case no one wins the competition. Thus their winning frequencies do not necessarily add to 1 and it is no longer the case that one wanes while the other waxes. As shown in Fig. 7, the winning frequency curves are asymmetrical about the line of 0.5. The profile plot suggests that the two methods agree frequently, but it is quite clear in this case that method 1 should be selected.

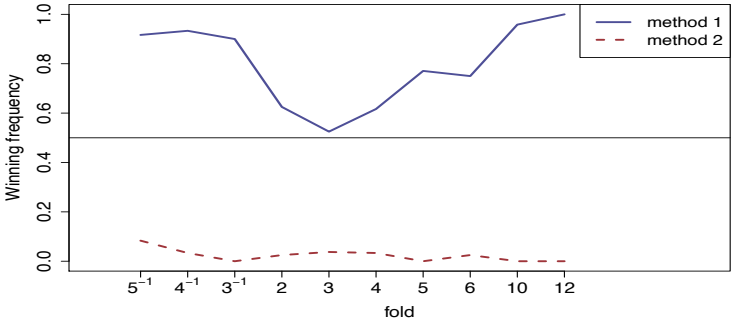


Figure 7: Common pattern 7 of PEC-CV: Asymmetric

As explained above, the PEC-CV plots provide much richer and more reliable info on relative performances of the competitors compared to the regular CV. We have highlighted several common patterns, based on which more robust and accurate decisions can be made. Of course, due to randomness of the data, sometimes the PEC-CV plot may falsely

Summary on CV

- We should always keep the objective in mind when applying CV!
- The 2nd level CV really helps and works broadly for regression with substantial modeling uncertainty.
- 10-fold CV may not perform as well as perceived, regardless of the goal.
- For consistently identifying the best candidate model or modeling procedure by CV, the evaluation part has to be sufficiently large, the larger the better as long as the ranking of the candidates in terms of risk at the reduced sample size of the training part stays the same as that under the full sample size.

- The benefits of having a large portion for evaluation are two-fold:
  - more observations for evaluation provide better capability to distinguish the close competitors;
  - the fewer observations in the training part make the accuracy difference between the close competitors magnified and the difference becomes easier to detect even with the same amount of evaluation data.
- Try different data splitting ratios to have a better understanding.
- The pre-determined training and testing data in machine learning may leave a space for unreliable conclusions or manipulated results!