Regular, Median and Huber Cross-Validation: A Computational Comparison

Chi-Wai Yu[1] and Bertrand Clarke[2*]

[1]*Department of Mathematics, The Hong Kong University of Science and Technology, Clearwater Bay, Kowloon, Hong Kong*

[2]*Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, USA*

Received 22 July 2014; revised 7 October 2014; accepted 12 October 2014
DOI:10.1002/sam.11254
Published online 7 January 2015 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: We present a new technique for comparing models using a median form of cross-validation and least median of squares estimation (MCV-LMS). Rather than minimizing the sums of squares of residual errors, we minimize the median of the squared residual errors. We compare this with a robustified form of cross-validation using the Huber loss function and robust coefficient estimators (HCV). Through extensive simulations we find that for linear models MCV-LMS outperforms HCV for data that is representative of the data generator when the tails of the noise distribution are heavy enough and asymmetric enough. We also find that MCV-LMS is often better able to detect the presence of small terms. Otherwise, HCV typically outperforms MCV-LMS for 'good' data. MCV-LMS also outperforms HCV in the presence of enough severe outliers.

One of MCV and HCV also generally gives better model selection for linear models than the conventional version of cross-validation with least squares estimators (CV-LS) when the tails of the noise distribution are heavy or asymmetric or when the coefficients are small and the data is representative. CV-LS only performs well when the tails of the error distribution are light and symmetric and the coefficients are large relative to the noise variance. Outside of these contexts and the contexts noted above, HCV outperforms CV-LS and MCV-LMS.

We illustrate CV-LS, HVC, and MCV-LMS via numerous simulations to map out when each does best on representative data and then apply all three to a real dataset from econometrics that includes outliers. © 2015 The Authors. *Statistical Analysis and Data Mining* published by Wiley Periodicals, Inc. Statistical Analysis and Data Mining 8: 14–33, 2015

Keywords: cross-validation; model selection; heavy-tailed errors; robustness; skewness; sparsity; outliers

1. INTRODUCTION

It is common for statistical researchers and subject matter users to assume normal noise in their linear models. To be more realistic, often this is phrased as saying the error distribution in a linear regression model is approximately symmetric, has light tails, and is unimodal leaving it aside how these criteria are to be interpreted precisely. To dramatize the pitfalls of the usual approach, consider the following example.

EXAMPLE 1: *Noise terms and variable selection.* Consider the model

$$Y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + E_i, \tag{1}$$

for $i = 1, \ldots, n$ where the Y_i is the response, E_i is the error term, and x_{1i}, x_{2i} are design points for two explanatory variables generically denoted by x_1 and x_2. Suppose we have data of the usual form, i.e. we have data (y_i, x_{1i}, x_{2i}) for $i = 1, \ldots, n$, and we assume the error terms E_i are independent and identical (IID) normal with zero mean or close enough that the approximation error can be ignored in comparison with other sources of error.

The usual estimator for the coefficients in Eq. (1) is found by least squares. Write this as $\hat{\beta}_{LS} = (\hat{\beta}_1, \hat{\beta}_2)^T$. Although Eq. (1) is very simple, it is standard to check if any submodel of Eq. (1) will fit as well as the full model. Obviously, this can be done by testing $\beta_1 = 0$ or $\beta_2 = 0$.

To set up our routine analysis, we generated 30 IID outcomes to use for x_1 by setting $X_1 \sim N(0, 1)$. We also generated 30 outcomes for x_2 by a different technique. So, an analyst cannot make any valid assumptions, at this stage, about the distributional properties of X_2. However, we ignore the randomness in the generation of (X_1, X_2) since we are regarding their values as deterministically chosen design point To find y_i's, a linear function of (x_1, x_2) was taken and perturbed by IID standard normal noise.

* *Correspondence to:* Bertrand Clarke (bclarke3@unl.edu)

Table 1. Inference for β_1 and β_2 in Eq. ((1)).

Parameter	Least squares estimate	Standard error	t-value	p-value
β_1	2.037	0.680	2.996	0.00567
β_2	1.003	0.008	128.854	$<2 \times 10^{-16}$

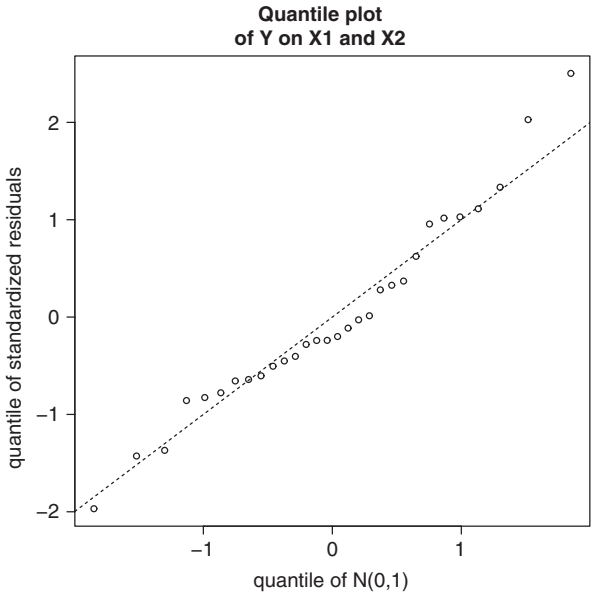

Fig. 1 Normal quantile plot of Y on the model with X_1 and X_2.

Now, consider fitting Eq. (1). The usual analysis is shown in Table 1. The standard errors indicate that $(\beta_1, \beta_2) = (0, 0)$ would be rejected at any reasonable level, meaning neither parameter can be taken as zero. This is reinforced by noting $R^2 = R^2_{adj} = 0.998$.

Likewise, a quantile plot, see Fig. 1, shows that Eq. (1) provides good fit. The residuals track the diagonal line reasonably closely apart from the right tail where two points look a little far from the regression line. This is a hint of non-normality in the right tail, but this is very slight as the deviations are within three standard errors. Overall, this appears to be a successful fit leading to the point predictor $\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$.

However, the model is far wrong and its predictions will quickly discredit it.

In fact, the data generator for Y was

$$Y_i = 2x_{1i} + E_i, \text{ for } i = 1, \ldots, n, \tag{2}$$

where $\{E_i : i = 1, \ldots, n\}$ were IID from the univariate skewed t-distribution with mean parameter zero 0, skewness parameter 5, variance parameter one, and 0.5 degrees of freedom. As stated before, X_1 was generated IID $N(0, 1)$. However, X_2 was artificially constructed by setting $X_2 = \hat{e}$ + N(0,25) noise, where $\hat{e}$ is the residual from the least squares estimators in Eq. (2) using only Y and X_1. Thus, there need not be any tipoff that the normality assumption is wrong and that downstream inferences will be poor. □

This example suggests more: In practice, analysts with large numbers of explanatory variables who assume normal errors may just be selecting those variables that happen to construct a normal noise rather than explaining the response, even when the correct noise term is far from normality. This is possible because only the sum of the linear function and error is identifiable, not each term individually. It is true that if $n \to \infty$ relative to the number of explanatory variables the problem seen in this example becomes less and less likely. However, it never disappears and can occur even in cases such as this example where there are 10 data points per parameter; here $n = 30$ and we have estimated β_1, β_2, and σ (even though σ does not exist for the error distribution used).

The family of skewed t-distributions that we use will be denoted by $skew - t(\nu, \gamma)$ where ν is the degrees of freedom and γ is the skewness parameter. That is, $Y \sim skew - t(\nu, \gamma)$ if and only if $Y = \gamma W + \sqrt{W} Z$ where $W \sim \mathsf{InverseGamma}(\nu/2, \nu/2)$ and $Z \sim N(0, 1)$. The $t(\nu, \gamma)$ class includes symmetric and asymmetric distributions controlled by γ as well as light- and heavy-tailed distributions controlled by ν. The $t(\nu, \gamma)$'s are a subset of more general skewed t-distributions that are multivariate, permit nonzero location parameters, and nonidentity variance matrices in the normal term. Skewed t-distributions have been used in a variety of linear models problems, see refs 1–5 (which have an extensive reference list) among others. As a separate class of distributions that permit control of skewness and heaviness of tails we have also used the Levy α stable distributions (see ref. 6 for a review) that regularly occur in the analysis of critical behavior in physics and in financial models. However, for appropriate choices of the two parameters in the Levy class, results qualitatively the same as we show in this paper for the $skew - t_(\nu, \gamma)$ distributions can be shown. Hence, we limit our focus to the $skew - t$ class because it is more familiar in statistics.

There are several subject matter disciplines where heavy-tailed distributions are common, such as climatology, see ref. 7, the study of dependence via copulas, see refs 8 and 9 (and the references therein), econometrics, see ref. 1, and computer traffic, see ref. 10. A recent overview can be found in ref. 11. The key feature that seems to unite these is that the response is the result of many small influences of substantial variability. This will be seen in the IMF data we analyze in Section 5.

As a separate issue, part of the popularity of the skewed t-distributions (or the Levy class) is that they permit

asymmetry, indeed strong asymmetry. There is a rich, if under-appreciated, literature on this as well. See, for instance, refs 12–14. In fact, one can argue that the whole burly field of quantile regression is devoted to the fact that knowing the quantiles of error terms is often essential—and this is most important in the asymmetric case.

Instead of positing a model and testing coefficients to find a submodel, one might compare linear models via cross-validation (CV). The central idea is to split the data into two sets. The first, called the training set, is used to fit a candidate model which is then evaluated on the second set, called the test set. Doing this repeatedly one can find the average cumulative discrepancy or CV error. The (leave-one-out) CV error is usually based on squared error loss and is given by

$$\mathrm{CV}(k) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_k^{-i}(\mathbf{x}_i))^2, \tag{3}$$

where $\hat{f}_k^{-i}(\mathbf{x}_i) = f_k(\mathbf{x}_i; \hat{\beta}^{-i})$ is an estimate of the regression function f_k at the point $\mathbf{x}_i$, $\hat{\beta}^{-i}$ is estimated without using $\mathbf{x}_i$ and $k = 1, \ldots K$ indexes the possible true models. As a generality, when using CV, the coefficients in the linear model are estimated by least squares to make the sense of distance the same for parameters and models. Since we are limiting attention to linear models, we write $f_k(\mathbf{x}_i; \beta) = \mathbf{x}_{k,i}\beta_k$ where k indicates a selection of the variables in the collection of all the explanatory variables $\mathbf{x}$ and their coefficients β_k in the full parameter vector β. The basic idea of CV is to choose the f_k with the smallest value of Eq. (3) since Eq. (3) estimates the prediction error of using f_k for Y. Note that Eq. (3) only makes sense when the error term has a finite variance.

Since the literature on CV is so vast we only note a few classic references [15–19]. In addition, for multifold CV, see refs 20, 21, and 22 and for the generalized CV, see 19. Arlot and Celisse [23] provide a recent review.

Because CV is based on squared error, one of its key problems is excessive sensitivity to influential data points, e.g. outliers. One way to reduce this is to use a robust form of the CV. For linear models one version of this was developed in ref. 24 and used the M-estimating approach of refs 25 and 26; see also refs 27 and 28. The idea is to choose a function ρ and then find the k that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\rho(y_i - \hat{f}_k^{-i}(\mathbf{x}_i)), \tag{4}$$

see refs 25 and 26. Note that Eq. (4) includes Eq. (3) by setting $\rho(t) = t^2$ but that when $\rho(t)$ increases slower than t^2, as $|t| \to \infty$, the minimum of Eq. (4) is less sensitive to extreme values of residuals than the minimum of Eq. (3) is. When ref. 24 uses a version of Eq. (4), they use a robust parameter estimator so the sense of distance used in the robust CV matches that for the parameters.

An obvious limitation of this methodology is one must choose ρ. In some nonparametric regression settings, Leung [28] shows that the minimum of Eq. (4) is asymptotically independent of the choice of ρ in large samples, this likely holds for linear regression as well. However, for small or moderate sample sizes, the minimum of Eq. (4) may depend nontrivially on the choice of ρ and many of the choices for ρ are essentially subjective. Nevertheless, Huber [25,26] suggests

$$\rho_c(t) = 2^{-1}t^2 I_{\{|t|\le c\}} + (c|t| - 2^{-1}c^2)I_{\{|t|>c\}}, \tag{5}$$

for some $c > 0$ for general usage. This approach has been extensively studied and recently has been extended to penalized regression settings, see ref. 29. A second obvious feature of this methodology is that while it prevents excessive reliance on outliers it does not discount them either. This is good if the outliers are 'for real' and hence influential data points, but may be bad if the outliers are genuinely unrepresentative of the data generator.

When we study robust forms of regression for linear models we use the R package rlm. One of the distances this package uses on models is Eq. (5) and we have chosen the default value (1.345) of c chosen internally to rlm. When providing parameter estimates, the default in rlm is to use Eq. (5) as well. Thus, the model selection and parameter estimation are based on the same sense of distance. Below, we refer to the results from this combination of parameter estimation and model selection using rlm as Huber CV (HCV). This is not identical to the method in ref. 24. However, the two are similar and the method in ref. 24 has only been coded in Fortran.

There are many reasonable choices for ρ in (4) that can be proposed. For instance, using $\rho(t) = |t|$ in Eq. (4) gives least absolute deviation CV (or regression), which is qualitatively different from regular CV under L^2 or HCV under Eq. (5). However, to avoid choosing ρ at all, we propose using the sample median in place of the mean in Eq. (3). That is, we find the model that minimizes

$$\operatorname*{med}_{1\le i\le n}(y_i - \hat{f}_k^{-i}(\mathbf{x}_i))^2, \tag{6}$$

over k. We call this median cross-validation (MCV). This sort of procedure was first proposed in ref. 30 and used in ref. 31 to choose the best number of neighbors to include in a nearest-neighbors approach to nonparametric curve fitting. Here, we advocate the idea more generally: Replace means by medians systematically to give alternative criteria with useful properties, in particular for model selection.

In addition to avoiding subjectivity in the choice of ρ, there are two immediate advantages of MCV over CV. First, MCV does not assume that any terms have any moments. So, MCV can be used with heavy-tailed errors terms such as in Example 1. Second, because the median does not depend on tail behavior of the residuals, the MCV is highly resistant to aberrant data points. A related fact is that loss functions have right-skewed distributions so the median of the squared residuals will be more representative of the distribution of the squared residuals than the mean is. We comment that, unlike the mean of the squared residuals, Eq. (6) is invariant to increasing transformations. So, it would be equivalent to use the median of any increasing function of the absolute error giving robustness to the choice of loss function.

MCV is the model selection analog of Rousseeuw's least median of squares estimator (LMSE) for parameters, see refs 32 and 33. The LMSE is

$$\hat{\beta}_{\mathrm{LMS}} = \arg\min_{\beta} \underset{1\le i\le n}{\mathrm{median}}[y_i - f(\mathbf{x}_i.\beta)]^2$$

Like MCV, LMSEs are defined in terms of a median, do not require any moments, and are as resistant to outliers as possible. So, to ensure that the same sense of distance was used for parameter estimation as model selection, we use LMSEs with MCV. However, while the MCV is unique this uniqueness is conditional on the uniqueness of the LMSE. As noted in ref. 32, the LMSE is only ill-defined when the data points used to find it are not in 'general position'. Since this happens only with probability zero (assuming a continuous probability measure) we have ignored this in our simulations below.

There are a variety of other techniques for robust model selection. For instance, for linear models, Muëller and Welsh [34] propose an objective function of three terms. The first is a familiar robust model selection term using a (weighted) form of Eq. (4) where $\rho(z) = \min(z^2, b^2)$ (for $b = 2$). The second is a compelxity penalty taken to be essentially the Bayes information criterion penalty of $k \log n$ where k is the number of parameters. The third term is a predictive term intended to ensure future predictions are close to future outcomes. This population-based term is estimated by a technique [34] called a stratified bootstrap. Minimizing the Mueller–Welsh objective function to find a model gives consistent model selection. However, the theorem establishing consistency uses the finiteness of the second moment of the error term in an essential way. Thus, the simulations for Table 5 in ref. 34 using a two-term model list and either the slash or Cauchy distribution show that their method never gave a probability of correct model selection higher than 0.5, even in very simple settings. Below some of our simulation results will show that Huber-based robust CV performed well with the (skewed) Cauchy error, but collapsed when the tails of skewed error become heavier. Hence, it is reasonable to conjecture that with error distributions even further from the normal than the Cauchy, the technique in ref. 34 would perform even more poorly, especially when the error is skewed. On the other hand, the change in performance from using a median in place of ρ remains unexplored. Regardless of this, when the error distributions are not far from mormal, the technique in ref. 34 was extended to generalized linear models in ref. 35.

Another technique by which to do robust model selection is in refs 36 and 37 and is based on variance inflation factors (VIFs) that are essentially scale factors on the $\hat{\beta}_k$'s. The key idea is to make the technique in ref. 38 for large, streaming datasets robust. The procedure is to start with a small model and sequentially test, as data accumulate, whether more explanatory variables are worth including without overfitting. The robustness is accomplished by using Tukey's or Huber's weights on residuals in an estimating equation. So far, simulations have only shown that this method is robust to outliers from a contaminated normal error. The VIF method could presumably be improved if a robust loss ρ were used in place of squared errors, but this extension does not seem to have been tested.

Despite the substantial literature on robust model selection, the bulk of the results in the sequel are comparisons among CV-LS, HCV, and MCV-LMS for a wide range of model lists and error terms that are typically heavy-tailed or skewed. Even for these settings, we make no claim that our comparisons below are exhaustive.

To be precise about our comparisons, let M_T denote the true model. We say that MCV-LMS works better than CV-LS if and only if

$$P_{M_T}(\text{MCV-LMS chooses } M_T) > P_{M_T}(\text{CV-LS chooses } M_T), \tag{7}$$

In this expression, it is understood that either P_{M_T} takes the nonuniqueness of the LMSE into account (for instance by averaging) or that the LMSE is unique. We use the analog of Eq. (7) for comparing CV-LS with HCV and HCV with MCV-LMS.

As a first comparison of CV-LS, HCV, and MCV-LMS consider the following toy problem.

EXAMPLE 2: *Three nested model classes.* Suppose the true model is

$$Y = 2 + .2x_1 + .8x_2 + E, \tag{8}$$

where x_1 and x_2 are explanatory variables and E is an error term and we are willing to consider three nested model

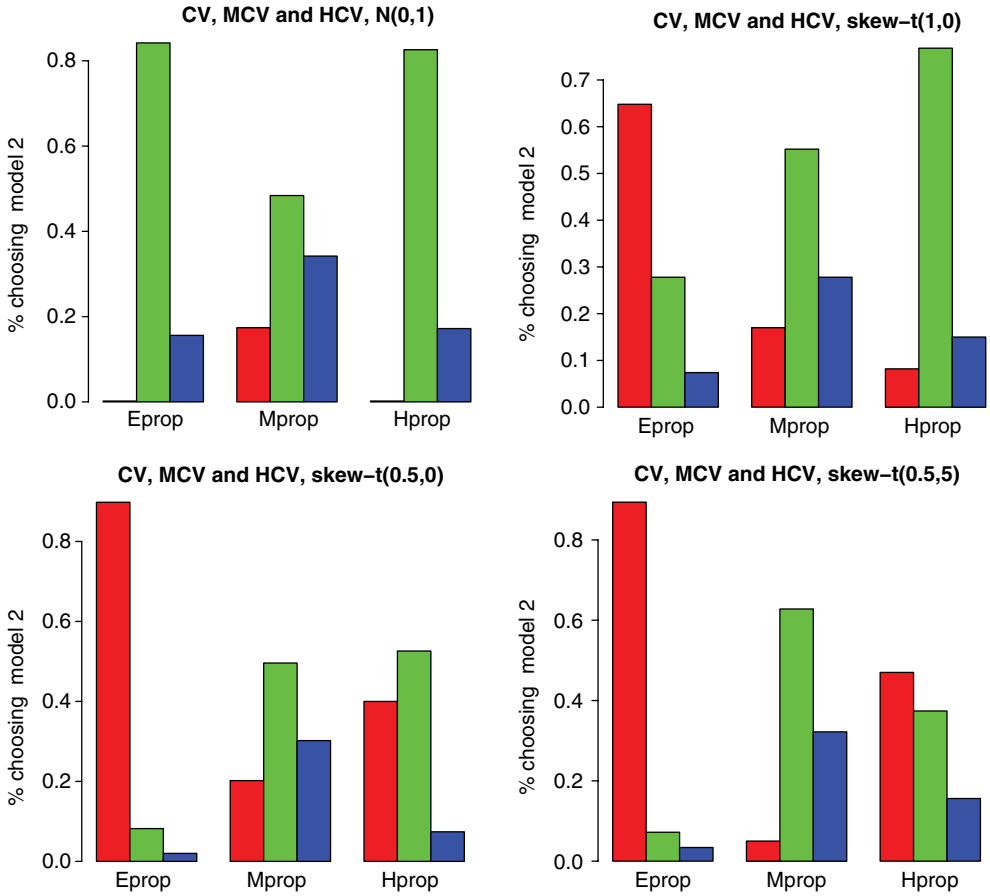

Fig. 2 Histograms of the sampling distributions for model class selection by tenfold CV-LS, HCV, and MCV-LMS. The proportions of selection for each of the three methods are denoted by Eprop (expectation proportion) or CV-LS, by Mprop (median proportion) or MCV-LMS, and by Hprop (Huber function proportion) or HCV. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

classes, namely

$$\begin{cases} \text{Model class 1: } Y \sim \beta_0 + \beta_1 x_1, \\ \text{Model class 2: } Y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2, \\ \text{Model class 3: } Y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3. \end{cases}$$

Suppose $N = 500$ IID samples of the form (x_{1i}, x_{2i}, x_{3i}), $i = 1, \ldots n = 50$ are drawn from the Unif[0,1] and in each iteration each of the explanatory variables is studentized. Next, consider four distributions for the error term, namely $N(0, 1)$, $skew - t(1, 0)$, $skew - t(0.5, 0)$ and $skew - t(0.5, 5)$ and for each case use all three cross-validatory methods.

The model selection results are shown in Fig. 2. The upper left panel shows that when the error term is $N(0, 1)$ CV-LS does best, but is only better than HCV by a very small amount, .016. This persists if the $N(0, 1)$ is replaced by a contamination distribution of teh form $(4/5)N(0, 1) + (1/5)N(a, 1)$ for $\mid a \mid \leq 3$. When $\mid a \mid > 3$ HCV performs best (figures not shown). MCV-LMS tends to do poorly because it chooses model class 3 too often. The upper right panel shows that when the tails of the error are $skew - t(1, 0)$, i.e. Cauchy and the error is symmetric, HCV performs much better than either of the other two methods. This persists even when the error distribution is asymmetric ($\gamma = 5, 10, 15$, figures not shown). The lower left panel shows that even when the tails are heavier than the Cauchy, HCV performs best when the error distribution is symmetric, although not by much. The lower right panel shows that when the tails of the error distribution are heavier than a Cauchy and asymmetric MCV-LMS performs best. Figure 2 also shows that when CV-LS does poorly it does so by choosing models that are too small. □

In Fig. 2 we could have used fivefold or leave-one-out CV, and the results would not have changed qualitatively. Indeed, in all our work we found that changing the k in k-fold CV did not affect our conclusions—although the results often had to be stabilized by bootstrapping

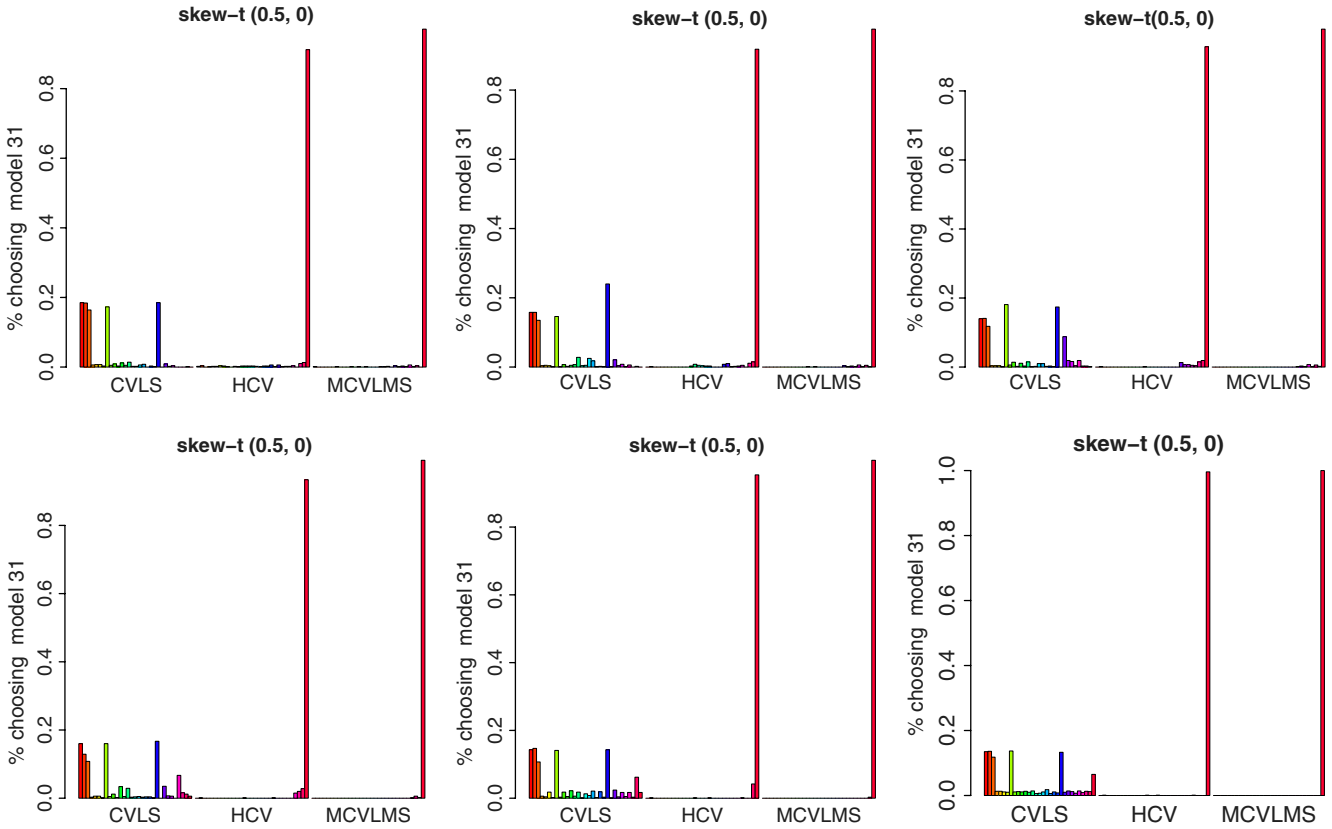

Fig. 3 Histograms of the sampling distributions for model class selection by tenfold CV-LS, HCV, and MCV-LMS. The panels are in the same order as the models in Eq. (9). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

as in Section 5. From a theoretical standpoint Shao [21] (Theorem 1) suggests that for CV-LS and error terms with a finite variance that asymptotically most of the data should be used for testing rather than training. Although not directly relevant here (since our focus is on cases where the variance does not exist), it is good intuition so we have followed it, making the k is k-fold CV as large as reasonably possible, even for HCV and MCV-LMS. More recently, Lund [39] in Section 2.1.4, implicitly suggests that reasonable K's range from 1 to 20. Moreover, Arlot and Celisse [23] state on p. 61: '...the best risk estimator is LOO, whereas 10-fold CV is more accurate for model selection'. Here, in using K as large as possible, the range was from four to ten—within the recommendations.

As a final introductory example, consider another toy problem with a bigger model list.

EXAMPLE 3: *31 non-nested model classes* Suppose the model list now consists of all $2^5 - 1 = 31$ nontrivial linear models using five explanatory variables $x_1, \ldots, x_5$ from $U(c(\alpha), c(1-\alpha))$ where $c(\alpha)$ is the 100α percentile of a standard Cauchy and the error distribution is $skew - t(0.5, 0)$. Consider six possible true models all from the same model class, namely,

$$
\begin{aligned}
Y &= 2 + 0.5x_1 + 0.5x_2 + 0.5x_3 + 0.5x_4 + 0.5x_5 + E \\
Y &= 2 + 5x_1 + 0.5x_2 + 0.5x_3 + 0.5x_4 + 0.5x_5 + E \\
Y &= 2 + 5x_1 + 5x_2 + 0.5x_3 + 0.5x_4 + 0.5x_5 + E \\
Y &= 2 + 5x_1 + 5x_2 + 5x_3 + 0.5x_4 + 0.5x_5 + E \\
Y &= 2 + 5x_1 + 5x_2 + 5x_3 + 5x_4 + 0.5x_5 + E \\
Y &= 2 + 5x_1 + 5x_2 + 5x_3 + 5x_4 + 5x_5 + E, \qquad (9)
\end{aligned}
$$

in which the number of terms with small coefficients is decreasing. From *Example 2*, we expect that HCV should do well because the error term is symmetric. In fact, the results are shown in Fig. 3 for a sample size of $n = 100$ and $N = 1000$.

In all panels MCV-LMS performs best—the reverse of what was seen in *Example 2*. Indeed, as the number of small terms decreases, HCV performs better compared with MCV-LMS until in the lower right panel, MCV-LMS and HCV are equivalent. □

The reason the performances of MCV-LMS and HCV reverse may be related to the fact that HCV tends to spread out over a larger class of possible models than MCV-LMS does when the data are spread out enough; this was not possible in Example 2 since the model list was so small. In effect, HCV may have a lower mode at the true model than MCV-LMS does. In addition, it is seen that the degree of *non*-sparsity is important. That is, MCV-LMS tends to give better performance than either HCV or CV-LS when the number of detectable small terms in the true model is high. Finally, as a generality, heavier tails and asymmetry tend to favor MCV-LMS more than they favor either HCV or CV-LS.

To date we have been unable to formalize our findings in theorems. Techniques from median regression seem not to apply. For instance, the Bahadur representation of the median has too large an error to identify lower order terms. Moreover, existing proofs for consistency of the CV or the optimality of HCV do not adapt to MCV. As a consequence we base our conclusions on extensive simulations and graphical analysis. On the other hand, we hold out hope that techniques from ref. 40 (or perhaps ref. 41) may be helpful for the future study of MCV.

The rest of this article is organized as follows. In Section 2, we state our procedure formally and then extend our findings from *Example 3* to a wider class of error distributions for 'good' data, i.e. data that is representative of the data generator, e.g. no outliers. In Section 3, we return to the nested case but use larger model lists to compare the sampling distributions of HCV and MCV-LMS for model selection, taking into account both error distributions and the size of the coefficients in the true model. In Section 4, we examine the effect of varying the model list and noise term taken as true, in particular, we show how HCV and MCV-LMS behave when the true model is on the model list, when it is not on the list, when its location on the list varies, and when the noise term varies. Sections 3 and 4 continue to use 'good' data. However, in Section 5, we analyze an econometrics dataset to show how MCV-LMS performs with complex data that includes outliers. In Section 6, we give recommendations for when each technique is appropriate.

2. NON-NESTED MODEL LISTS

For precision, we begin by stating the HCV and MCV-LMS procedures formally and stating the generic form of our simulations. The simulations do not include outliers; these are treated in Section 5. Then, we turn to the extension of *Example 2*.

2.1. Formal Method

Both the HCV and MCV-LMS methods are similar in structure to CV-LS. Here we present the MCV-LMS method because the key steps in HCV can be found in ref. 24 and CV-LS is well known.

To compare models $f_1, \ldots, f_K$, using (y_i, x_i), $i = 1, \ldots, n$ where x_i is a vector of explanatory variables, V-fold MCV-LMS is the following.

1. Split the sample of size n into V disjoint and exhaustive subsets $S_1, \ldots, S_V$.

2. For each k and $v = 1, \ldots, V$ use $S_v^c = \bigcup_{u \neq v} S_u$ to find LMSEs for β_k, the parameter in f_k.

3. For $v = 1, \ldots, V$, get a collection of

$$d_v(k) = \{ (y_i - \hat{f}_{k,LMS}^{-i}(\mathbf{x}_i))^2 : \forall i \in S_v \},$$

where the superscript $-i$ indicates that the ith data point was not used to form $\hat{f}_{k,LMS}$, the estimate of f_k using the LMSEs from S_v^c.

4. The V-fold MCV-LMS model is $f_{\hat{k}}$ for which

$$\hat{k} = \arg\min \left[\text{median}(\{d_1(k), \ldots, d_V(k)\}) \right].$$

This is the same as CV-LS but it uses the median in place of the mean, and LMSE in place of LSE. Likewise, HCV is the same except that it uses a truncated squared error criterion for robust estimation of the parameters and for robust evaluation of the CV error (see Eqs. ((4)) and ((5))).

Next, since we will be presenting numerous simulations, we describe their general form. We randomly generated design points X from a uniform distribution on the range $[c(\alpha), c(1-\alpha)]$ for $\alpha = .05$ where $c(\alpha)$ is the $\alpha 100\%$-th quantile of a Cauchy distribution. Given these values, we generated error terms E from various $skew - t$ distributions. Then we formed variables Y from $X\beta + E$ for various choices of parameter vector β. To compare MCV, HCV, and CV we only permitted the model selection procedures to use the design points X and the response values Y. Given that n data points were generated, V-fold CV of whichever form meant following the four steps above. Doing this for each candidate model gave the MCV-LMS, HCV, and CV-LS errors and we chose the models with the smallest MCV-LMS, HCV, or CV-LS errors.

2.2. Extending Example 3

Recall *Example 3* showed that for the $skew - t(.5, 0)$ error distribution MCV-LMS outperformed HCV and CV-LS in a non-nested model list setting. Here we extend this finding to a larger class of error distributions. Recall the model list consisted of models of the form

$$Y = \beta_0 + \gamma_1\beta_1 x_1 + \ldots + \gamma_k\beta_k x_k + E, \qquad (10)$$

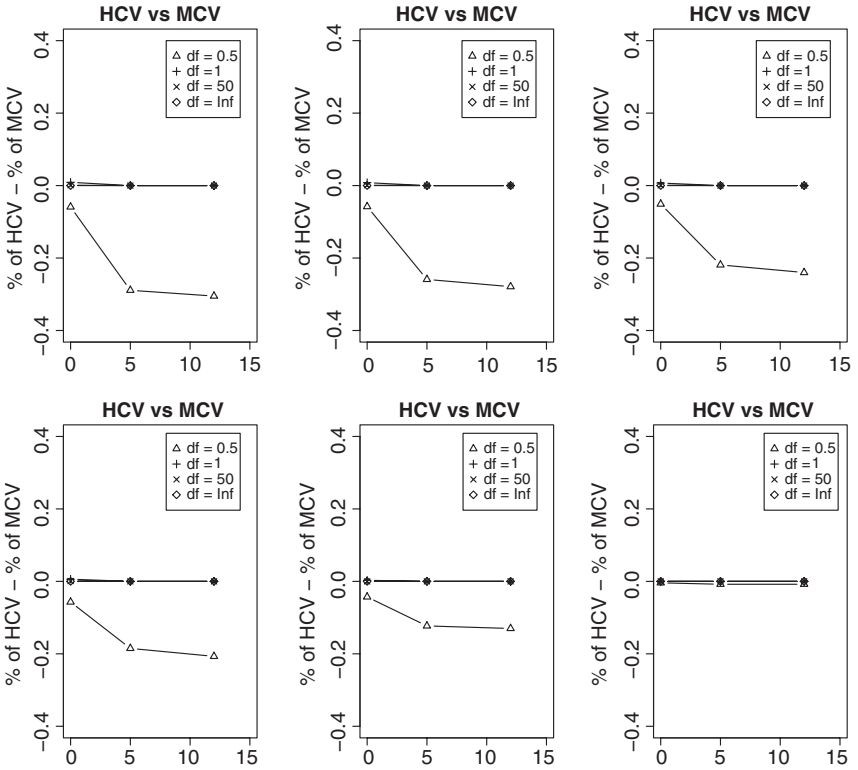

Fig. 4 The six panels in the figure show the percent decrease of using HCV instead of MCV-LMS for 12 error distributions in a non-nested model setting; the degree of asymmetry is indicated along the horizontal axis.

where γ_j = 0,1 according to whether x_j is or is not in the candidate model, for $j = 1, \ldots, k$ and we set $k = 5$ so there are 31 nontrivial models. Because all the x_j's were generated the same way, any two models with the same number of explanatory variables are equivalent. We continue to use $n = 100$, $N = 1000$, and tenfold CV.

To dramatize our results, rather than looking at the probability of correct model selection we look at the percent decrease in correct model selection when HCV is used instead of MCV-LMS. As before, the models taken as true for the six panels are as in Eq. (9). We chose the coefficients to be five and 0.5 on the grounds that (i) five was a simple number large enough that any method ought to be able to detect it given the range of the error distributions and explanatory variables and (ii) 0.5 was a simple number, large enough to matter if we were to use one of the models to make predictions but small enough that we would not expect the terms with coefficient 0.5 always to be found, i.e. it provided a nontrivial check on the model selection methods.

Here, we use 12 different error distributions: four choices for the heaviness of the tails ($\nu = .5, 1, 50, \infty$) and three choices for the degree of asymmetry ($\gamma = 0, 5, 12$). Our results are shown in the six panels of Fig. 4. (CV-LS was dropped since it performed so poorly in *Example 3*.)

It is seen that when the true model has all coefficients 0.5, HCV and MCV-LMS are essentially equivalent for $\nu = 1, 50, \infty$ for all choices of asymmetry parameter. However, when $\nu = 1/2$, MCV-LMS gives noticeably better performance over all values of the asymmetry parameter and improves relative to HCV as the asymmetry increases. This pattern is strongest when all five coefficients are 0.5 and decreases as the number of coefficients 0.5 decreases until, when all coefficients are 5, HCV and MCV-LMS are equivalent over the 12 error terms we used. This suggests that in model selection problems of a realistic size, MCV-LMS tends to perform better than HCV when the tails are heavy, the asymmetry is significant, or the true model is non-sparse. Results qualitatively the same as those presented in Fig. 4 are obtained if other sizes of true model are used provided the coefficients are similar or if other non-nested model lists are used.

3. NESTED MODEL LISTS

In this section we examine the effect of non-sparsity by presenting simulations that assume the true model is a sum of terms of diminishing influence on Y. We suggest that this mimics real scenarios where many small influence combine

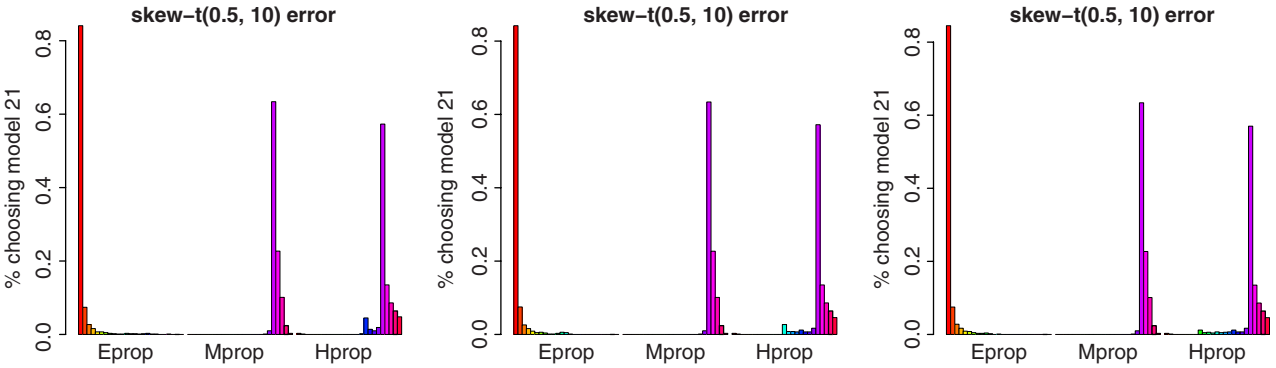

Fig. 5 Histograms of the sampling distributions for model class selection by tenfold CV-LS, HCV, and MCV-LMS assuming the three true models with coefficients 5, 0.5, and 0. The middle peak representing MCV-LMS is seen to be highest. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

to influence the behavior of Y and it is not *a priori* clear which ones should be regarded as part of an announced model. The model list therefore consists of models of the form (10) that have been put in equivalence classes based on the number of explanatory variables and then ordered by size so they can be assumed nested. While nesting reduces the number of models to be considered, it is a reasonable assumption when there are many terms that can be ordered for inclusion by, say, a shrinkage method such as LASSO or SCAD.

We consider two classes of models with non-sparsity. The first is an extension of Eq. (9) and Section 2.2 from five explanatory variables to 25 explanatory variables. The second permits the coefficients of the variables to decrease monotonically. All the results in this section assume a $skew-t(0.5, 10)$ error distribution and that the data have no aberrant points.

3.1. Coefficients 0, 0.5, 5

A model in this class is defined by having some leading terms with coefficient five, the next sequence of terms has coefficient 0.5, and the last terms have coefficient zero. So, write

$$Y = 2 + 5x_1 + 5x_2 + \cdots + 5x_{17} + .5x_{18} + \cdots + .5x_{21} + E,$$
$$Y = 2 + 5x_1 + 5x_2 + \cdots + 5x_{13} + .5x_{14} + \cdots + .5x_{21} + E,$$
$$Y = 2 + 5x_1 + 5x_2 + \cdots + 5x_9 + .5x_{10} + \cdots + .5x_{21} + E,$$

in which the last four explanatory variables $x_{22}, \ldots, x_{25}$ are decoys. Thus, the number of small terms is increasing even though the model size is fixed, i.e the model with 21 explanatory variables is always true.

In this case, Fig. 5 shows the results for CV-LS, MCV-LMS, and HCV. As might be expected, CV-LS does poorly—bailing out to the trivial model. Only MCV-LMS and HCV give useful results. Both have modes at the true model in all three cases, but MCV-LMS has a stronger mode indicating it puts more probability on or near the true model than HCV does.

The three panels all look the same meaning that MCV-LMS detects the smaller order terms better, no matter how many or fewer there are. A sidebar comment is that as the number of 0.5 terms increases, HCV spreads out a little while MCV does not and CV-LS concentrates ever more at the trivial model.

3.2. Decreasing Coefficients

The second model class consists of decreasing coefficient models of the form

$$Y = 2 + 5x_1 + 2x_2 + x_3 + (2/3)x_4 + \cdots + (2/(i-1))x_i + \ldots + (1/12)X_{25} + E, \tag{11}$$

where $i = 2, \ldots, 25$. For the sake of comparison we consider as true models the cases that $i = 9, 13, 17$, and 21. The results are in Fig. 6. They show that for $i = 9, 13$ MCV-LMS is much more likely to find the true model but that for $i = 17, 21$ as the coefficients shrink HCV is more likely to find the true model.

The results from Figs 5 and 6 are a little paradoxical but suggest that MCV-LMS imposes more sparsity than HCV (but much less than CV-LS that bails out to model 1, not shown). This may be the result of the same phenomenon seen in *Example 3*: HCV tends to give better performance than MCV-LMS with smaller model lists. That is, if Fig. 6 was regenerated without nesting the models MCV-LMS might outperform HCV. However, the computational burden increases rapidly with the number of explanatory variables making such comparisons difficult in practice.

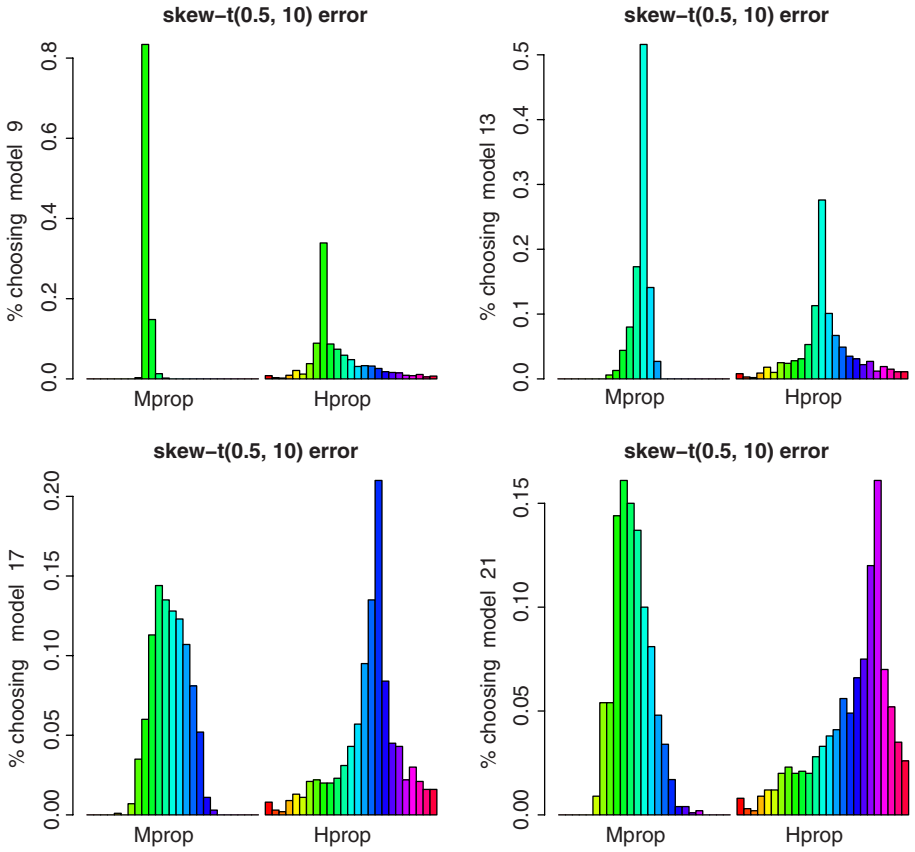

Fig. 6 Histograms of the sampling distributions for model class selection by tenfold HCV and MCV-LMS for the four models defined by taking the first 9, 13, 17, and 21 terms in Eq. (11). In the top two panels MCV-LMS does best; in the bottom two panels HCV does best. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

It is tempting to interpret the difference between Figs 5 and 6 as arising from the different treatment of sparseness by MCV-LMS and HCV: Below a threshold, MCV-LMS is unable to detect terms as effectively as HCV does. This is a little surprising since in most other settings the reverse is true—again tentatively suggesting it is the size of the model list that is the issue. Indeed, one expects a bias-variance tradeoff for model list selection: Too small a model list will give bias and too large a model list will give excess variance. For heavy-tailed or asymmetric errors MCV-LMS usually has a lower bias at the cost of a higher variance while HCV usually has a higher bias at the cost of a lower variance. On the other hand, this line of reasoning is limited because in the lower panels of Fig. 6 neither method achieves a probability of correct model selection above 0.25.

Figure 7 shows the probability of correct selection of models with decreasing coefficients for true models of size noted on the horizontal axis. MCV-LMS has a higher probability of correct model selection for all true models up to size 14, but this is only meaningful up to 13 because beyond model 12 MCV-LMS has probability of correct selection less than 0.5. By contrast, HCV always has probability less than 0.5.

Thus, the results of this subsection and the last are compatible: Regardless of the how the coefficients decrease, when HCV outperforms MCV-LMS, both methods are breaking down in the sense that the probability of correct model selection is < 0.5. Note that this is for $skew-t(.5, 10)$; the next section will look at how MCV-LMS compares to HCV over a range of error distributions.

4. EFFECT OF MODEL LIST SELECTION

In this section we continue to investigate the effect of the model list when the true model has attenuating coefficients. That is, we take true models to be of the form

$$\mathcal{M}_\mu : Y = 2 + 2X_1 + X_2 + (2/3)X_3 + \cdots + (2/i)X_i + \cdots + (2/\mu)X_\mu, \quad (12)$$

and the model class contains τ nested models, where the i-th model $\mathcal{M}_i$ consists of $X_1, \ldots, X_i$ in order. So,

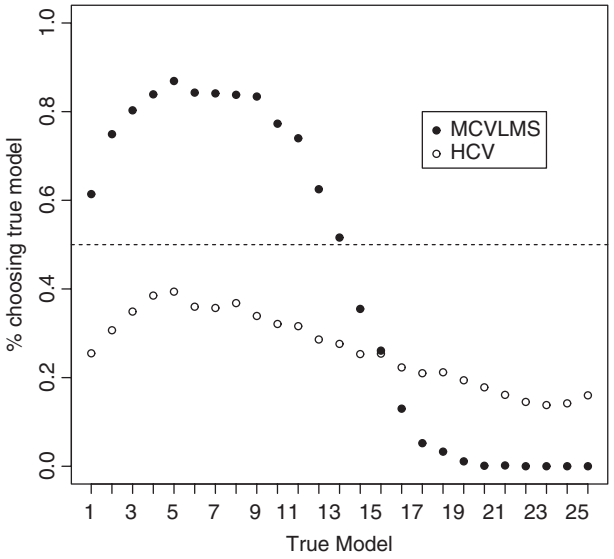

Fig. 7 Probabilities of correct selection when the error is $skew - t(.5, 10)$, the true model is of the form (11) with $i = 1, \ldots, 25$ using a nested model list of size 25. The dashed line indicates 0.5.

when $\mu \leq \tau$, the true model $\mathcal{M}_\mu$ is on the model list $\{\mathcal{M}_1, \ldots, \mathcal{M}_\tau\}$ but when $\mu > \tau$ the true model $\mathcal{M}_\mu$ is outside the class. Our goal is to see how the choice of μ and τ affects the model selection procedures when the error distributions are varied but outliers are not considered.

Note that this is the same class as used in Section 3.2. However, the results from Sections 3.2 and 3.1 are very similar so it seems safe to regard Eq. (12) as generally representative of the decreasing coefficient case for which CV-LS works very poorly—in spite of the data being 'good'. Hence, in this section, CV-LS is not included.

In Section 4.1 we present simulations assuming $\mu \leq \tau$, i.e. the model list contains models that are as big or bigger than the true model in the sense of number of terms. Thus, in some cases, the true model is at the right hand endpoint of the model list and in some case the model at the right hand endpoint of the list is larger than the true model. In Section 4.2 we present simulations for the case $\mu > \tau$, i.e. the true model is larger than any model on the model list; we regard this as a realistic scenario. In these cases, the true model is a submodel found by taking some coefficients zero. In Section 4.3, we present simulations for the case that the true model itself is not on the model list but the model list contains models that are too small and some that are too large.

4.1. Model List Contains Models Bigger than the True Model

Here, $\mu \leq \tau$ and we present results for HCV and MCV-LMS under four different error terms: $skew - t(1, 0)$, $skew - t(1, 10)$, $skew - t(.5, 0)$ and $skew - t(.5, 5)$. All these simulations assume a tenfold CV procedure with sample size $n = 150$ and number of replications $N = 1000$. In the tables of this subsection, we indicated the model list by ξ. That is, the model list is $\xi = \{\mathcal{M}_1, \ldots, \mathcal{M}_\xi\}$ where each of the models is of the form $\mathcal{M}_\mu$ in Eq. (12). (The double usage of ξ as the model list and the largest model on that list will not cause any confusion.) Then, the entry in a table for a given column ξ and a row μ is the probability of choosing the correct model μ when the model list is ξ.

One of the important qualitative features of all the tables in this section is that the probability of selection of the true model by MCV-LMS or HCV decreases as the model size and list increases. Sometimes this decrease is slow and sometimes it is relatively fast, but it always seems to happen.

Results for Cauchy noise, i.e. $skewt(1, 0)$, are given in Table 2. There are two points to note. First, all the entries in the top half of the table for MCV-LMS are much lower than the corresponding entries in the bottom half of the table for HCV. For instance, when model M_μ with $\mu = 8$ is true, the probability of selecting it when $\xi = 9$ is 0.62 for MCV-LMS and 0.84 for HCV. Second, there is a sudden drop as the list of candidate models goes even one step beyond the true model indicating both methods are parsimonious.

As a second example, results corresponding to Table 2 are given for asymmetric Cauchy noise $skew - t(1, 10)$ in Table 3. In this case, again, all the entries in the top half of the table for MCV-LMS are lower than the corresponding entries in the bottom half of the table for

Table 2. Model selection with the $skew - t(1, 0)$ noise term. Top: MCV-LMS. Bottom: HCV. Probabilities are rounded to two decimal places.

		Model List				
		ξ				
Percentage of choosing the true model M_μ		6	7	8	9	10
True model M_μ μ	6	1	0.67	0.59	0.55	0.53
	7		0.98	0.66	0.56	0.53
	8			0.94	0.62	0.57
	9				0.88	0.62
	10					0.79
		Model List				
		ξ				
Percentage of choosing the true model M_μ		6	7	8	9	10
True model M_μ μ	6	1	0.88	0.81	0.77	0.77
	7		1	0.84	0.82	0.78
	8			1	0.84	0.82
	9				1	0.86
	10					1

Table 3. Model selection with the $skew-t(1,10)$ noise term. Top: MCV-LMS. Bottom: HCV. Probabilities are rounded to two decimal places.

		Model List				
		ξ				
Percentage of choosing the true model M_μ		6	7	8	9	10
True model M_μ μ	6	1	0.72	0.64	0.61	0.63
	7		1	0.68	0.61	0.62
	8			1	0.70	0.63
	9				1	0.71
	10					0.99

		Model List				
		ξ				
Percentage of choosing the true model M_μ		6	7	8	9	10
True model M_μ μ	6	1	0.81	0.75	0.73	0.70
	7		1	0.81	0.76	0.74
	8			1	0.81	0.78
	9				1	0.81
	10					1

HCV. However, in comparison to Table 2, Table 3 shows that HCV outperforms MCV-LMS by a smaller margin. The effect of skewness is to improve the performance of MCV-LMS relative to HCV.

When the tails of the noise term are heavier than Cauchy but symmetric, for instance, $skew-t(.5,0)$, Table 4 suggests that for smaller model lists, MCV-LMS has a higher probability of choosing the correct model but that for larger model lists HCV still has a higher probability of choosing the correct model. This is reminiscent of Fig. 7 where larger models were unfavorable to MCV-LMS but here the reversal occurs at model size eight. Note also that the probabilities in a row do not strictly decrease. We regard this as an indication that more iterations would have to be done to get a finer resolution since in all other cases strictly decreasing probabilities were found. The values in Table 4 are satisfactory (especially in view of Fig. 8 in Section 4.3) since they suggest the two methods are performing more similarly to each other as the tails get heavier, the same as was observed for increasing asymmetry.

Finally for this subsection, suppose the noise term has heavier tails than a Cauchy and is asymmetric, for instance, a $skew-t(.5,10)$. It can be seen that the entries in the top half of Table 5 for MCV-LMS are higher than the corresponding entries in the bottom half of the table for HCV. That is, when the tails of the noise term are sufficiently heavy and asymmetric MCV-LMS outperforms HCV. (The value 0.77 for $\mu=6$ and $\xi=10$ is an anomaly of using simulations; sometimes the answer will not fit an established pattern.)

Table 4. Model selection with the $skew-t(.5,0)$ noise term. Top: MCV-LMS. Bottom: HCV. Probabilities are rounded to two decimal places.

		Model List				
		ξ				
Percentage of choosing the true model M_μ		6	7	8	9	10
True model M_μ μ	6	0.96	0.70	0.66	0.66	0.68
	7		0.90	0.68	0.61	0.62
	8			0.8	0.62	0.59
	9				0.64	0.50
	10					0.47

		Model List				
		ξ				
Percentage of choosing the true model M_μ		6	7	8	9	10
True model M_μ μ	6	0.91	0.82	0.81	0.80	0.80
	7		0.88	0.80	0.79	0.75
	8			0.86	0.74	0.71
	9				0.76	0.69
	10					0.70

Table 5. Model selection with the $skew-t(.5,10)$ noise term. Top: MCV-LMS. Bottom: HCV. Probabilities are rounded to two decimal places.

		Model List				
		ξ				
Percentage of choosing the true model M_μ		6	7	8	9	10
True model M_μ μ	6	1	0.79	0.72	0.73	0.77
	7		0.99	0.76	0.75	0.73
	8			0.98	0.74	0.71
	9				0.92	0.72
	10					0.79

		Model List				
		ξ				
Percentage of choosing the true model M_μ		6	7	8	9	10
True model M_μ μ	6	0.79	0.57	0.54	0.52	0.51
	7		0.69	0.55	0.50	0.48
	8			0.64	0.48	0.47
	9				0.54	0.44
	10					0.47

4.2. Model List Contains Models Strictly Smaller than the True Model

In this subsection we consider true models of the form $\mathcal{M}_\mu$, where $\mu>\tau$, i.e. where the model class $\{\mathcal{M}_1,\ldots,\mathcal{M}_\tau\}$ does not contain the true model as an interior or boundary point. Essentially, we allow $\mu>\tau$

Table 6. Sampling distributions for MCV-LMS (top) and HCV (bottom) for various model lists with the $skew-t(1,0)$ noise term. Probabilities are rounded to two decimal places.

			Candidate model $\mathcal{M}_\tau$								
			τ								
Proportions of choosing each candidate model			1	2	3	4	5	6	7	8	9
True model M_μ	μ	7	0	0	0	0.01	0.08	0.92			
		8	0	0	0.00	0.02	0.14	0.84			
		8	0	0	0	0.00	0.02	0.13	0.85		
		9	0	0.00	0.00	0.03	0.21	0.76			
		9	0	0	0.00	0.01	0.04	0.20	0.76		
		9	0	0	0	0	0.01	0.03	0.17	0.79	
		10	0	0.00	0.01	0.05	0.20	0.74			
	10	0	0	0.00	0.02	0.07	0.24	0.68			
		10	0	0	0	0.01	0.01	0.07	0.20	0.71	
	10	0	0	0	0.00	0.00	0.02	0.07	0.23	0.69	
			Candidate model $\mathcal{M}_\tau$								
			τ								
Proportions of choosing each candidate model			1	2	3	4	5	6	7	8	9
True model M_μ	μ	7	0	0.00	0.00	0.00	0.00	1			
		8	0	0	0	0	0.00	1			
		8	0	0	0	0	0	0	1		
		9	0	0	0	0	0	1			
		9	0	0	0	0	0	0	1		
		9	0	0	0	0	0	0	0	1	
		10	0	0	0	0	0.00	1			
		10	0	0	0	0	0	0.00	1		
		10	0	0	0	0	0	0	0.01	0.99	
		10	0	0	0	0	0	0	0	0.00	1

to mimic the case that the true model is bigger than any model on the model list. So, the best a model selection procedure can do is to give the largest model on the model list.

Here it is enough to show the results from MCV-LMS and HCV for $skew-t(1,0)$ and $skew-t(.5,10)$ noise terms since the other cases, $skew-t(1,10)$ and $skew-t(.5,0)$ examined in Section 4.1, give results similar to $skew-t(1,0)$ in that HCV outperforms MCV-LMS and the degree of outperformance decreases as the tails get heavier or more asymmetric. As in the last subsection, the probability of correct selection generally decreases as model size increases. The simulations here assume tenfold CV with $n=150$ and $N=1000$ replications and hence the data are representative of the data generator, i.e. there are no data points that can reasonably be regarded as outliers.

The tables here are of a different form from the tables in Section 4.1. Here, the rows are indexed by μ, the size of the true model. Each possible true model size may be associated with model lists of various sizes. Thus, in Table 6, when $\mu=9$ model lists $\xi=6,7,8$ may be used but not $\xi=9$ for then M_9 would be on it. Thus, each row gives the sampling distribution of a model selecction technique with the model list as its support. For instance, Table 6 is similar to Table 2 in that HCV outperforms MCV-LMS. The same holds here, as in Tables 3 and 4, when the noise is $skew-t(1,10)$ and $skew-t(.5,0)$, but to a smaller extent. Likewise, when the noise distribution is $skew-t(.5,10)$, Table 7, like Table 5, shows that MCV-LMS outperforms HCV.

4.3. True Model Not on a Two-Sided Model List

To conclude this section, we present simulations for model selection with a fixed model list while varying the noise term. The true model is $\mathcal{M}_6$ and the model list is $\{\mathcal{M}_1,\ldots,\mathcal{M}_5,\mathcal{M}_7,\ldots,\mathcal{M}_{10}\}$. In this case, the most appropriate model to choose is $\mathcal{M}_7$. Since there are many cases to handle, we used tenfold CV with $n=100$ data points and $N=500$ replications.

Figure 8 shows where MCV-LMS is better than HCV at identifying $\mathcal{M}_6$ (solid dots) and HCV is better than MCV-LMS (open dots) when the error term is a $skew-t$ with the indicated skewness and degrees of freedom. If the table is extended beyond asymmetry parameter 10, MCV-LMS continues to outperform HCV for degrees of freedom less than or equal to 0.6, i.e. for 0.7 or greater HCV outperforms MCV at choosing $\mathcal{M}_6$. When the degrees of freedom is too low and the skewness is small, MCV-LMS outperforms

Table 7. Sampling distributions for MCV-LMS (top) and HCV (bottom) for various model lists with the $skew - t(.5, 10)$ noise term. Probabilities are rounded to two decimal places.

		Candidate model $\mathcal{M}_\tau$								
		τ								
Proportions of choosing each candidate model		1	2	3	4	5	6	7	8	9
True model M_μ μ	7	0.00	0.00	0.00	0.00	0.06	0.94			
	8	0	0	0	0.02	0.14	0.84			
	8	0	0	0	0	0.01	0.10	0.89		
	9	0	0	0.00	0.02	0.18	0.80			
	9	0	0	0	0.00	0.05	0.19	0.75		
	9	0	0	0	0	0.00	0.03	0.17	0.80	
	10	0	0	0.00	0.06	0.21	0.73			
	10	0	0	0.00	0.01	0.05	0.23	0.70		
	10	0	0	0.00	0.00	0.01	0.08	0.25	0.66	
	10	0	0	0	0	0.00	0.02	0.07	0.21	0.698
		Candidate model $\mathcal{M}_\tau$								
		τ								
Proportions of choosing each candidate model		1	2	3	4	5	6	7	8	9
True model M_μ μ	7	0.01	0.03	0.03	0.06	0.13	0.75			
	8	0.01	0.02	0.04	0.06	0.14	0.74			
	8	0.01	0.01	0.02	0.03	0.06	0.16	0.71		
	9	0.01	0.02	0.03	0.06	0.13	0.76			
	9	0.00	0.01	0.02	0.03	0.07	0.16	0.70		
	9	0.01	0.01	0.02	0.04	0.05	0.07	0.19	0.61	
	10	0.01	0.02	0.04	0.06	0.16	0.71			
	10	0.01	0.02	0.03	0.05	0.08	0.15	0.68		
	10	0.01	0.01	0.03	0.03	0.05	0.08	0.20	0.59	
	10	0.01	0.02	0.02	0.02	0.04	0.07	0.10	0.18	0.542

HCV but neither does well; the probability of correct selection by MCV-LMS is below 0.5. This is indicated by solid triangles. Indeed, in many cases, even when MCV-LMS assigned a higher probability to $\mathcal{M}_7$ than HCV did, it also assigned a higher probability to $\mathcal{M}_5$ than $\mathcal{M}_7$. That is, MCV-LMS it is more likely to choose a a model that leaves out a term than to choose a model class for which a submodel would be right. This indicates that MCV-LMS has slightly more tendency toward sparsity than HCV does but nowhere near as much as CV-LS does. However, Fig. 8 indicates that below 0.3 degrees of freedom neither method gives good results and even at degrees of freedom 0.3, there must be a little skewness for MCV-LMS to choose the right model successfully. Of course, above degrees of freedom 0.7, HCV always does better than MCV-LMS and for a few cases with small degrees of freedom and small amounts of skewness HCV also does better than MCV-LMS.

Note that Fig. 8 is dual to Fig. 7. Specifically, Fig. 7 shows the collection of models for which MCV-LMS outperforms HCV for a fixed noise term while Fig. 8 indicates the collection of noise terms for which the most appropriate model is selected when a fixed model is true.

5. ECONOMETRIC DATA EXAMPLE

In this section we analyze a real dataset to express national gross domestic products (GDP) in terms of macroeconomic variables. When analysis ready, we will have 35 explanatory variables and sample size $n = 172$. The data exhibits both heavy tails and asymmetry, and has outliers. Overall, this is a relatively complicated dataset. Here, we apply CV-LS, HCV, and MCV-LMS to do model selection and parameter estimation.

To implement these methods we first sphere the data to transform the original explanatory variables so they will be approximately orthogonal. Then, we order the sphered variables by their (absolute) correlation with the response so we have a list of nested models. By bootstrapping, we obtain an approximation to the sampling distribution for CV-LS, MCV-LMS, and HCV. From these we choose the modal model. Then, we transform back to the original variables removing any terms for which the coefficients are too small. Apart from the bootstrapping, this procedure is much the same as in ref. 42 for selecting a useful number of principal components. We take the extra step of bootstrapping to find the sampling distribution to account for variability in the parameter estimation and model

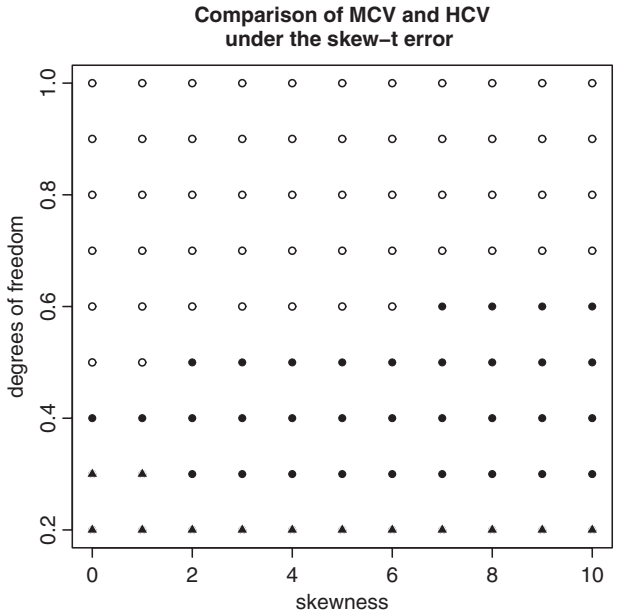

Fig. 8 Values of degrees of freedom and skewnesses for which HCV is better than MCV-LMS (open circles) and for which MCV-LMS is better than HCV (dark circles) at choosing model $\mathcal{M}_7$ when the model list is $\{\mathcal{M}_1 - \mathcal{M}_5\} \cup \{\mathcal{M}_7 - \mathcal{M}_{10}\}$, when the better of the two has probability at least 0.5. The dark triangles indicate degrees of freedom and skewnesses for which MCV-LMS is better than HCV but has probability of selecting $\mathcal{M}_7$ less than 0.5.

selection; this is important for complex data such as we are analyzing here.

To be precise, sphering means that the $n \times p$ design matrix $\mathbf{X}_n$ is transformed so that the arithmetic mean of each column is zero and the empirical covariance matrix is the identity. Let Σ_n be the empirical covariance matrix,

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_p)(X_i - \overline{X}_p)^T,$$

where X_i is the ith observation of the p variables, and $\overline{X}_p$ is the $p \times 1$ column vector of the arithmetic mean. The spectral representation of Σ_n gives

$$\Sigma_n = A_n \Lambda_n A_n^T,$$

where Λ_n is a diagonal matrix of the eigenvalue of the empirical covariance matrix, and A_n is the corresponding matrix of eigenvectors. The sphere designed matrix is

$$\mathbf{SX}_n = (\mathbf{X}_n - 1_{n\times 1}\overline{X}_p^T) \times A_n \Lambda_n^{-1/2}, \qquad (13)$$

where $1_{n\times 1}$ is the $n \times 1$ column vector of ones. Given this, we can apply CV-LS, MCV-LMS, and HCV to the dataset fairly since all variables have been located and scaled the same way.

Table 8. Definition of the variables

Variable	Definition
Y	Gross domestic product (GDP), purchasing-power-parity (PPP) share of world total
X1	Inflation, average consumer prices
X2	Inflation, end of period consumer prices
X3	Volume of imports of goods and services
X4	Volume of exports of goods and services
X5	Population
X6	General government revenue
X7	General government total expenditure
X8	General government net lending/borrowing
X9	Current account balance

5.1. IMF Data

Consider the 2009 financial data found at http://www.imf.org/external/pubs/ft/weo/2011/02/weodata/weoselgr.aspx. There were 46 variables measured on world economies; they were partitioned into six categories (national accounts, monetary, trade, people, government finance, and balance of payments). One was GDP (in purchasing power parity) which we try to explain using the other 45. In fact, we did not use all the other 45 variables, we used a selection of them since many had incomplete data. Out of 184 countries reporting we removed all those with more than ten missing variables. Then, we did a cull of the remaining variables by removing some that seemed to duplicate the information in other variables that we thought were more important to include. Thus, we reduced the 45 variables to nine. This still represents a relatively large number of variables because in many cases the products of these explanatory variables are important to include. Thus, the primary goal of our analysis is to determine the effects of nine financial variables X_1 - X_9, with their product terms, on the response variable Y, where Table 8 provides the short definitions of the variables. (The IMF webpage provides the formal definitions.)

It is easy to see the nine variables lead to 45 cross-terms. However, we retained only 26 of these, namely: X_5X_6, X_5X_7, X_4X_5 X_5X_8, X_5X_5, X_5X_9, X_2X_5, X_1X_5, X_7X_8, X_9X_9, X_4X_8, X_3X_8, X_7X_9, X_1X_6, X_1X_7, X_4X_7, X_3X_7, X_6X_9, X_4X_6, X_3X_6, X_1X_9, X_4X_9, X_6X_8, X_2X_9, X_7X_7, and X_1X_1. The statistical reasons are (i) these terms had the highest absolute magnitude of their Pearson correlations with Y and (ii) they led to covariance matrices that were not singular, a key issue for stability of results. Specifically, our selection process stopped when we get a negative eigenvalue of the covariance matrix of the sphered covariates. It is inconceivable that the model list contains the true model for GDP, however, we can hope that one

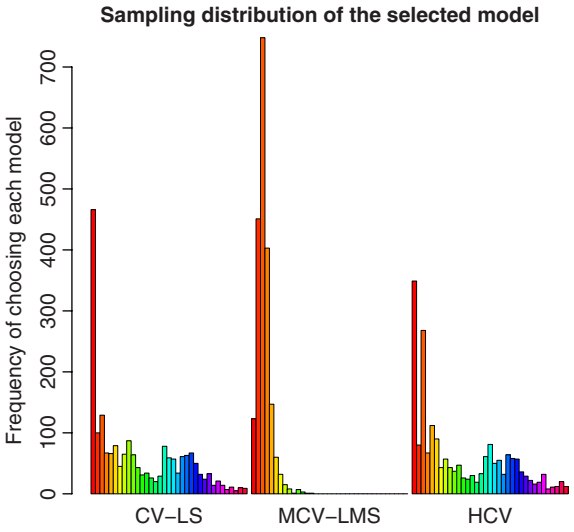

Fig. 9 Bootstrap estimates of the sampling distributions of CV-LS, MCV-LMS, and HCV. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

of the models is not hopelessly wrong. This scenario is somewhat like that in Section 4.3 except that for practical purposes we must reduce the full non-nested model list of size $2^{35} - 1$ to a nested model list.

A key feature of this dataset is that even if the dataset could be said to be representative of a single data generator, there are outliers, influential data points, and other aberrant data points that can affect the model selection. This is on top of the fact that there is reason to believe the error term is heavy-tailed and asymmetric. Moreover, it is plausible to regard the real economy as consisting of a large number of small influences that do not die away and are small relative to economic variability, somewhat like the models in Sections 3 and 4. Taken together these points suggest that CV-LS will do poorly, HCV will do better, and MCV-LMS will do best.

Proceeding, once sphering gave approximate orthogonality, we sorted them in order of the absolute magnitudes of their Pearson correlations with Y. This gave a class of 35 nested models. We then took 2000 bootstrap samples of size $n = 172$ from the sphered data to estimate the sampling distributions for the three methods CV-LS, HCV, and MCV-LMS. The use of the bootstrapping for CV-LS and HCV is to account for their instability under fourfold CV; the use of bootstrapping for MCV-LMS is to account for its instabilty under fourfold CV and the instability from the LMS estimation. The bootstrapping also accounts for the instability of parameter estimation in CV-LS and HCV, but it is not clear how important this is. (We comment that using the R function rlm gave more than 50 warning messages, even with 2000 iteration steps indicating some sort of problem with this implementation of HCV.)

The boostrapped sampling distributions are shown in Fig. 9. It is seen that the histogram for CV-LS bails out to one sphered variable. It is possible that the first sphered variable is the best model for the GDP, but the rest of the sampling distribution for CV-LS is essentially uniform over models of size two through 35 and the first model only gets approximate probability $466/2000 = 0.23$. (It is tempting to see a weakly bimodal sampling distribution but that is likely just the result of variability.) The sampling distribution for HCV is seen to break down similarly. There is an overall mode at the first sphered variable, a very low value for two sphered variables, and a higher value for three sphered variables. Assuming the low value at two sphered variables is an anomaly due to variability, HCV also bails out to one sphered variable but is concentrating at that single variable better than CV-LS is because HCV has a gradual decrease while CV-LS is sudden. Otherwise put, neither CV-LS nor HCV seems able to concentrate meaningfully anywhere credible. By contrast, MCV-LMS gives a well-defined mode at three sphered variables. Accordingly, we surmise that CV-LS and HCV both choose a one-sphered-variable model (but with different coefficients), while MCV-LMS chooses a three-sphered-variable model. The similarity between CV-LS and HCV may be due to the fact that HCV is a variation on CV-LS to stabilize it by using a loss function that is linear outside an interval. That is, HCV, like CV-LS, is still much more sensitive to large and small values than MCV is.

To examine the effectiveness of these three models for this dataset, we found the models explicitly and generated quantile plots for them; see Fig. 10. These plots are based on normal quantiles and so show departures from normality. Indeed, it is seen that the left panel for CV-LS strongly suggests that both the left and right tails of the error term are heavier than normal with the right hand tail being much heavier than the left tail. It also suggests a nearly normal shape on the mid-range. Overall, this panel suggests heavy tails and asymmetry. Unsurprisingly, the right panel for HCV suggests the same, likely because HCV and CV-LS are more similar to each other (in choosing the same number of sphered terms) than either is to MCV. The panel for MCV-LMS also suggests heavy tails and asymmetry but is smoother and hence more believable than the cusps on the right hand tails of the histograms for CV-LS and HCV. This is consistent with the fact that heavy tails with asymmetry are precisely the setting where MCV-LMS outperforms CV-LS and HCV as a model selection technique.

Let us now turn to an assessment of the effect of outliers and influential data points. One way to visualize this is to use the CV-LS model and generate partial residual plots for $X_1, \ldots, X_9$, the nine original variables. This is shown in Fig. 11. All the panels show that a significant proportion of the data points must be regarded as aberrant in the sense

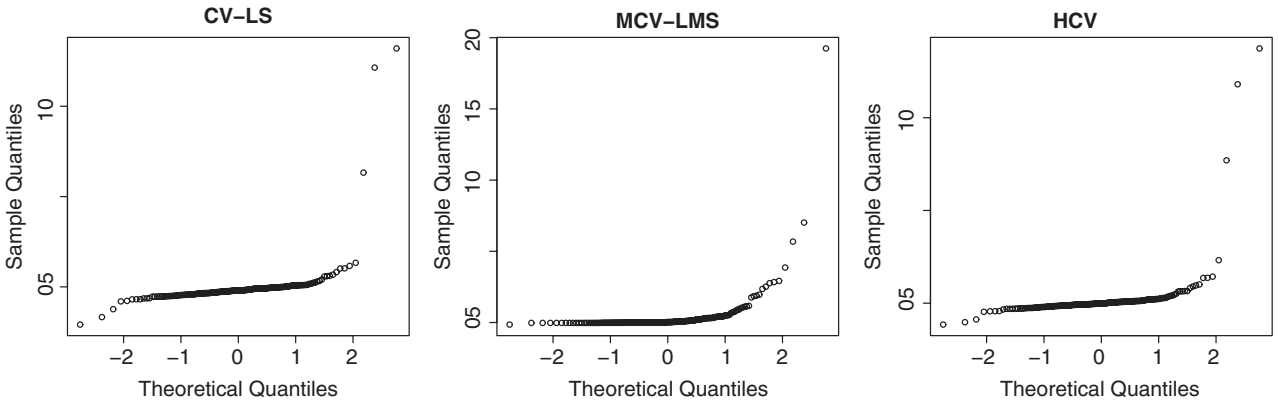

Fig. 10 Q–Q plots for the residuals from CV-LS, MCV-LMS, and HCV.

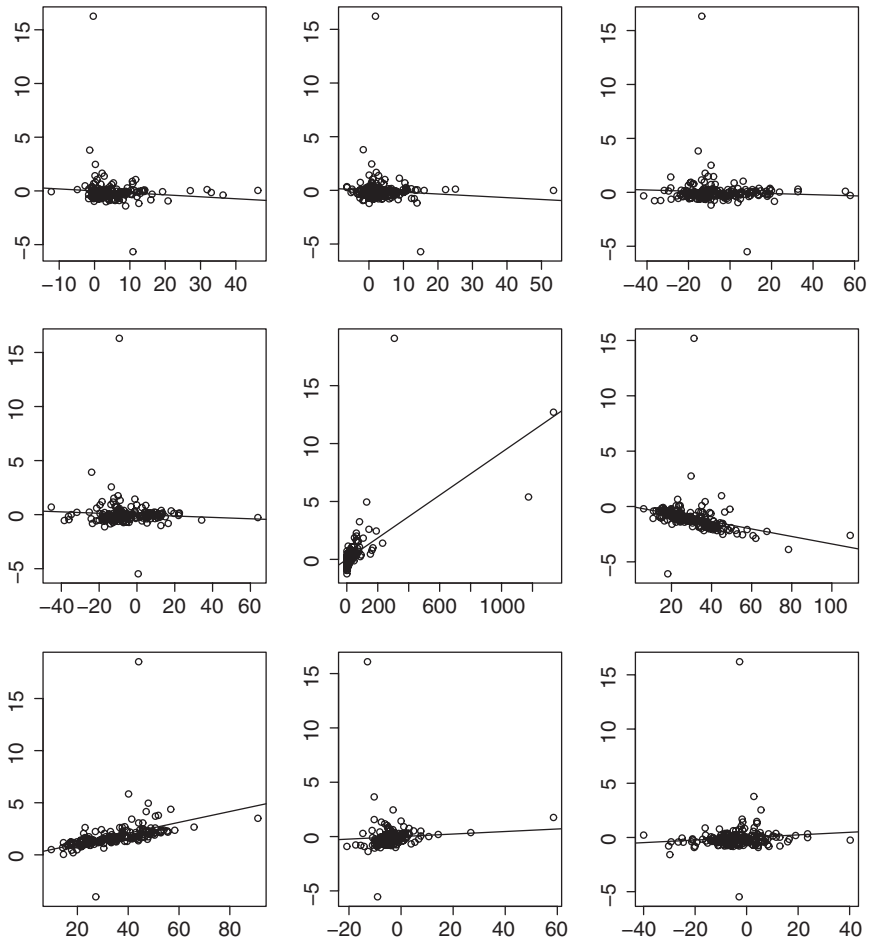

Fig. 11 Partial residual plots on the vertical axes for the first nine main effects terms; the horizontal exis represents X_j where the first row is X_1, X_2, X_3, the second row is X_4, X_5, X_6, and the third row is X_7, X_8, X_9.

that they are outliers or overly influential. Indeed, there seem to be enough of them of large enough magnitude that even reducing their effect to a linear ρ as with HCV will only reduce their influence not eliminate it. That is, Fig. 11 suggests outliers in general are a major problem with this dataset and merely reducing their influence somewhat may not be enough.

Taken together, this reasoning leads us to the view that the data generator has heavy, asymmetric tails and is prone to major problems with outliers—precisely the

setting where MCV-LMS, because it is based on medians, promises to give better results than CV-LS or HCV, as seen in Fig. 9. We comment that, as a generality, when a model selection procedure breaks down it often does so by choosing a trivial model—the smallest or the largest possible; see ref. 43 for examples of this. This may be what is happening with CV-LS and HCV; neither is capable of good inference with data as complex as this IMF data.

To complete this analysis, we find the three models based on the sphered data and transform back to the original variables for the sake of interpretability. First, for CV-LS, if we retain only those terms with coefficients at least 10^{-4}, we get a model based on six terms X_5X_6, X_5X_7, X_4X_5, X_5X_9, X_2X_5, and X_1X_5. For HCV, we get the same terms plus one more: X_7^2. That is, among the 35 covariates (9 variables, 26 cross-terms), we retain six and seven terms, respectively, for CV-LS and HCV. Retaining only the terms with coefficients at least 10^{-4} for MCV-LMS also gives seven terms but they are X_5X_6, X_5X_7, X_5X_8, X_2X_5,X_1X_5, X_4X_7, X_4X_6. So, the model selection methods agree on four product terms: X_5X_6, X_5X_7, X_2X_5, and X_1X_5, based on inflation, population, government revenue, and expenditure. Terms X_4X_5 and X_5X_9 that are in the CV-LS and HCV models, but not the MCV-LMS model represent exports, population, and current account balance. The extra term in the HCV model not in the CV-LS model is the square of expenditure. The extra terms in the MCV-LMS models not in the CV-LS or HCV models are X_5X_8, X_4X_7 and X_4X_6 depending on exports, revenue, and expenditure. Thus, the CV-LS and HCV models do not depend on X_3 or X_8 and the MCV-LMS model does not depend on X_9.

Aside from the argument that the data is heavy-tailed, asymmetric, and has many outliers so that MCV-LMS is more reasonable to use than CV-LS or HCV, one can also argue that the model MCV-LMS gives makes more sense than the models from either CV-LS or HCV. Going back to Table 8 it is seen that the CV-LS and HCV models for GDP do not depend on imports or government net borrowing while the MCV-LMS does not depend on the current account balance. If one recognizes that government net borrowing is often the most important a component in the current account balance then the MCV-LMS model is more reasonable because it includes imports and loses little from leaving out current account balance and so depends on more of the essential nine variables than CV-LS or HCV does. Thus, from an econometric standpoint, the fact that the MCV-LMS model includes imports while the CV-LS and HCV models do not means the MCV-LMS model seems more appropriate.

6. DISCUSSION

It is well known that conventional CV has limitations due to its sensitivity to outliers, its requirement that second moments exist, and its excessive sparsity in that it often does not detect terms that make small but detectable contributions to a response variable. Likewise, it is well know that HCV is a viable, if underused, alternative that, whatever its other flaws, seems to outperform CV-LS over a wide range of settings. These include noise distributions with tails that are heavy or asymmetric but not too heavy or asymmetric, especially for models that do not have a large number of small terms that contribute to the response. HCV is also resistant to outliers, although this resistance is only up to the point permitted by the Huber function.

However, HCV itself has limitations as well: When the tails are too heavy or asymmetric or there are too many outliers with enough severity, HCV does not give results that are much better than CV-LS. Therefore, we have proposed a different technique, MCV-LMS that seems to overcome the two main flaws of CV but not as well as HCV—except under more extreme conditions. In particular, MCV-LMS is insensitive to the tail behavior of the noise term because it does not rely on any moments existing. Second, despite being insensitive to tail behavior, MCV seems more sensitive to central behavior being more able in our simulations and data analysis to find non-sparse models, i.e. those with numerous small influences not just those with a few dominant terms. Overall, in contrast to MCV-LMS, CV-LS has a tendency to lump small terms incorrectly into the noise term. In contrast to HCV, MCV-LMS underperforms until the setting for model selection is complex enough. Even when MCV-LMS underperforms HCV, the degree of underperformance can be small (several percentage points).

One of the other benefits of MCV is that it has useful robustness properties. This is obvious if one recalls that outliers in X or in Y will not contribute to the MCV error unless over half the data points are outliers. This level of robustness is maintained, provided the least median of squares (or other similar estimators) are used to estimate the parameters in the model. By contrast, it is well known that the mean of the squared error, as commonly used in CV and LS estimators, is unstable in comparison to the median. Thus, even though we have not argued this in detail except in our data analysis, MCV should work better than CV in the presence of outliers. HCV has similar outlier robustness properties formalized in Theorem 1 of ref. 24, however, it likely has a breakdown point that is smaller than MCV-LMS.

Arguably, the most important pragmatic point is to be able to decide which of CV-LS, MCV-LMS, and HCV to use in a given setting, especially when the conclusions they give differ. Given our examples presented, and

literature search, we can state our main methodological recommendations for the use of MCV-LMS, HCV, and CV-LS as follows.

1. Outside of very narrow settings such as light tails, little asymmetry, and the true model consisting of relatively few terms with large coefficients, CV-LS should not be used. In fact, even when CV-LS outperforms HCV, the degree of outperformance is slight. CV-LS tends to choose models that are systematically too small by imposing too much sparsity.

2. HCV seems to be the preferred technique for scenarios that have tails that are not too heavy, do not exhibit high asymmetry, or do not have too many small terms. As a generality, HCV imposes the least sparsity—but its performance deteriorates as non-sparsity increases.

3. MCV-LMS seems to be the preferred technique when the conditions favoring HCV fail dramatically enough. For instance, when the tails are heavy enough and asymmetric enough, or there are many small terms. When these fall below detectability, HCV works better than MCV-LMS but for the detectable ones MCV-LMS works better. In this sense MCV-LMS imposes some sparsity. Moreover, MCV-LMS seems to outperform HCV when problems with outliers are sufficiently severe.

Note that these recommendations do not consider hybrid techniques, e.g. CV-LMS. In the simulation we did for these cases, the results turned out to be intermediate between the 'pure' cases where the distance for parameter estimation matched the distance model selection. Also, we have not included CV methods such as those based on least absolute deviation; we expect them to be similar to CV-LS.

In addition to these methodological recommendations, our data analysis leads us to the following pragmatic heuristics. First, if there is enough data that, say, ten quantile regressions for a few plausible models can be done to assess the deciles of a regression function (possibly chosen by CV-LS, HCV, or MCV-LMS) then the heaviness and asymmetry of the error term can be assessed directly. Otherwise, suppose we are trying to do model selection over a class of models using one of CV-LS, HCV, and MCV-LMS. Start by examining the quantile plot from the CV-LS model. If it looks 'nice' indicating a strongly unimodal, nearly symmetric noise distribution with light tails, then CV is a reasonable choice, provided that there is no extra knowledge to the effect that a sparse model is unlikely to hold. Then, we can examine the partial residual plot, or other plots, to ensure that any possible outliers are few and mild. If this is so, then it is hard to imagine the MCV-LMS model being better, although the HCV model might be comparable.

However, if the quantile plot from the CV-LS model suggests problems, e.g. the histogram of residuals is not 'nice' looking in the sense of asymmetry or heavy tails or we have reason to doubt the validity of a sparse model or we have evidence from residual plots of problems with outliers then we may be led to consider MCV or HCV. In particular, if the tails of the quantile plot seem moderately heavy and not more than slightly asymmetric and the outliers are not too many, and not too severe then HCV is indicated. However, if the error term seems to have very heavy tails and nontrivial asymmetry or the problems with outliers seem severe enough then MCV-LMS is indicated. Finally, if it is believed that the true model is sparse, then CV-LS or HCV might be preferable while sufficient non-sparsity seems to suggest the MCV-LMS model will be best.

REFERENCES

[1] I. Takeuchi, Y. Bengio, and T. Kanamori, Robust regression with asymmetric heavy-tail noise distributions, Neural Comp 14 (2002), 2469–2496.
[2] S. Demarta and A. MacNiel, The t-copula and related copulas, Int Stat Rev 73 (2005), 111–129.
[3] W. Hu and A. Kercheval, The skewed t-distribution for portfolio credit risk, 2006. http://www.math.fsu.edu/aluffi/archive/paper288.pdf. Accessed November 7, 2014.
[4] A. Azzalini and M. Genton, Robust likelihood methods based on the skew-t and related distributions, Int Stat Rev 76 (2008), 106–129.
[5] S. Lee and G. McLachlan, On mixtures of normal and skew -distributions, 2013. http://archiv.org/pdf/1211.3602.pdf. Accessed November 7, 2014.
[6] J. Nolan, Stable Distributions: Models for Heavy Tailed Data, 2014, http://academic2.american.edu/jpnolan/stable/chap1.pdf. Accessed November 7, 2014.
[7] R. Katz, Do weather or climate variables and their impacts have heavy-tailed distributions? 2002. http://www.isse.ucar.edu/staff/katx/docs/pdf/heavy.pdf. Accessed November 7, 2014.
[8] R. Zakaria, A. Metcalfe, P. Howlett, J. Piantadosi, and J. Boland, Using the skew -copula to model bivariate rainfal distribution, Aust Math Soc 51 (2010), 231–246.
[9] M. Smith, Q. Gan, and R. Kohn, Modeling dependence using skew copulas: Bayesian inference and applications, J Appl Econometr, 27 (2012), 500–522.
[10] S. Resnick, Heavy tail modeling and teletraffic data, Ann Stat 25 (1997), 1805–1869.
[11] R. M. Cooke and D. Nieboer, Heavy-tailed distributions: data, diagnostics and new developments, Resources for the Future Discussion Paper, 2011, 11–19.
[12] J. Collins, J. Sheahan, and Z. Zheng, Robust estimation in the linear model with asymmetric error distributions, J Mult Anal 20 (1986), 220–243.

[13] J. Lind, K. Mehra, and J. Sheahan, Asymmetric errors in linear models: estimation—theory and Monte Carlo, Statistics 23 (1992), 305–320.
[14] S. Kim, Inverse Circular Regression with Possibly Asymmetric Error Distribution. Ph.D. Thesis; Department of Statistics, UC Riverside, 2009.
[15] M. Stone, Cross-validatory choice and assessment of statistical predications, J R Stat Soc Ser B 36, (1974), 111–147.
[16] M. Stone, An asymptotic equivalence of choice of model by cross validation and Akaike's criterion, J R Stat Soc Ser B 39 (1977), 44–47.
[17] S. Geisser, The predictive sample reuse method with applications, J Am Stat Assoc 70 (1975), 320–328.
[18] B. Efron, How biased is the apparent error rate of a prediction rule? J Am Stat Assoc 81 (1986), 461–470.
[19] K. C. Li, Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set, Ann Stat 15 (1987), 958–975.
[20] P. Zhang, Model selection via multifold cross-validation, Ann Stat 21 (1993), 299–313.
[21] J. Shao, Linear model selection by cross-validation, J Am Stat Assoc 88 (1993), 486–495.
[22] J. Shao, An asymptotic theory for linear model selection, Stat Sin 7 (1997), 221–264.
[23] S. Arlot and A. Celisse, A survey of cross-validation procedures for model selection, Stat Surv 4 (2010), 40–79.
[24] E. Ronchetti, C. Field, and W. Blanchard, Robust linear model selection by cross validation, J Am Stat Assoc 92 (1997), 1017–1023.
[25] P. J. Huber, Robust estimation of a location parameter, Ann Math Stat 35 (1964), 73–101.
[26] P. J. Huber, Robust regression, Ann Stat 1 (1973), 799–821.
[27] E. Cantoni and E. Ronchetti, Resistant selection of the smoothing parameter for smoothing splines, Stat Comput 11 (2001), 141–146.
[28] H. Y. Leung, Cross-validation in nonparametric regression with outliers, Ann Stat 33(5) (2005), 2291–2310.
[29] S. Lambert-Lacroix and L. Zwald, Robust regression through Huber's criterion and the adaptive lasso penalty, Elec J Stat 5 (2011), 1015–1053.
[30] Y. Yang, Median cross-validation criterion Chinese Sci Bull 42 (1997), 1956–1959.
[31] Z. G. Zheng, and Y. Yang, Cross-validation and median criterion, Stat Sin 8 (1998), 907–921.
[32] P. J. Rousseeuw, Least median of squares regression, J Am Stat Assoc 79 (1984), 871–880.
[33] P. J. Rousseeuw and A. M. Leroy, Robust Regression and Outlier Detection, Hoboken, NJ, John Wiley and Sons, 2003.
[34] S. Muëller and A. Welsh, Outlier robust model selection in linear regression, J Am Stat Assoc 100 (2005), 1297–1310.
[35] S. Muëller and A. Welsh, Robust model selection in generalized linear models Stat, Sinica 19 (2005), 1155–1170.
[36] D. Dupuis and M. Vectoria-Feser, Fast robust mdoel selection in large data sets, J Am Stat Assoc, 106 (2011), 203–312.
[37] D. Dupuis and M. Victoria-Feser, Robust VIF regression with applications to variable selection in large data sets, Ann Appl Stat 7 (2013), 319–341.
[38] D. Lin, D. Foster, and L. Ungar, VIF regression: a fast regression for large data, J Am Stat Soc 106 (2011), 232–247.
[39] K. Lund, The Instability of Cross-Validated LASSO. Master's Thesis; Faculty of Mathematics and Natural Sciences, Univeristy of Oslo, 2013.
[40] W. Ip, Y. Yang, P. Kwan, and Y. Kwan, Strong convergence rate of the least median absolute estimator in linear regression models, Stat Pap 44 (2003), 183–201.
[41] A. Stromberg, Consistency of least median of squares estimators in nonlinear regression, Commun Stat - Theory Meth, 24 (1995), 1971–1984.
[42] H. Chipman and H. Gu, Interpretable dimension reduction, J Appl Stat 32 (2005), 969–987.
[43] J. Clarke, B. Clarke, and C. Yu, Prediction in -compelte problems with limited sample size, Bayes Anal 8 (2013), 647–690.