Inference about Clustering and Parametric Assumptions in Covariance Matrix Estimation[*]

Mikko Packalen[†] Tony Wirjanto[‡]

26 November 2010

Abstract

Selecting an estimator for the variance covariance matrix is an important step in hypothesis testing. From less robust to more robust, the available choices include: Eicker/White heteroskedasticity-robust standard errors, Newey and West heteroskedasticity-and-autocorrelation-robust standard errors, and cluster-robust standard errors. The rationale for using a less robust covariance matrix estimator is that tests conducted using a less robust covariance matrix estimator can have better power properties. This motivates tests that examine the appropriate level of robustness in covariance matrix estimation. We propose a new robustness testing strategy, and show that it can dramatically improve inference about the proper level of robustness in covariance matrix estimation. Our main focus is on inference about clustering although the proposed robustness testing strategy can also improve inference about parametric assumptions in covariance matrix estimation, which we demonstrate for the case of testing for heteroskedasticity. We also show why the existing clustering test and other applications of the White (1980) robustness testing approach perform poorly, which to our knowledge has not been well understood. The insight into why this existing testing approach performs poorly is also the basis for the proposed robustness testing strategy.

Keywords: variance covariance matrix; covariance estimator; robustness tests; cluster-robust; clustering; heteroskedasticity.

JEL Classification Codes: C10, C12, C13, C52.

[*] We thank Jay Bhattacharya for discussions. The usual disclaimer applies.

[†] Department of Economics, University of Waterloo, Canada. Email: packalen@uwaterloo.ca.

[‡] School of Accounting & Finance and Department of Statistics and Actuarial Science, University of Waterloo, Canada. Email: twirjant@uwaterloo.ca.

1 Introduction

Hypothesis testing about one or more parameters in a regression model is a central component of almost any empirical research project. In addition to the selection of an estimator for the parameters of the regression equation, hypothesis tests require that a researcher selects an estimator for the associated variance covariance matrix. The set of available covariance matrix estimators has grown considerably over the years, with each new estimator typically relaxing one or more of the assumptions of the older estimators. Newer covariance matrix estimators are therefore more robust than older estimators in the sense that while newer estimators are consistent estimators whenever older estimators are consistent estimators, older estimators are not always consistent estimators when newer estimators are consistent estimators. From less robust to more robust–and simultaneously from older to newer–the set of available covariance matrix estimators include:

1. Least Squares standard errors (under the assumption of homoskedasticity)
2. Eicker/White heteroskedasticity-robust standard errors
3. Newey and West heteroskedasticity-and-autocorrocorrelation-robust standard errors
4. Cluster-robust standard errors
5. Multi-way cluster-robust standard errors.

Robustness is of course a desirable feature of any estimation strategy. However, the selection of a covariance matrix estimator is complicated by the fact that a less robust covariance matrix estimator can often have better finite-sample properties than a more robust covariance matrix estimator when both are consistent estimators. For example, when the heteroskedasticity-robust estimator and the cluster-robust estimator are both consistent estimators and the number of clusters is small, the (less robust) heteroskedasticity-robust estimator has better size and power properties than the (more robust) cluster-robust estimator (see e.g. Hansen, 2007, Stock and Watson, 2008, and below). However, a less robust estimator may not have this potential advantage in all cases. For instance, the Least Squares covariance matrix estimator (calculated under the assumption of homoskedasticity) may not have such potential advantage over the heteteroskedasticity-robust estimator and thus the use of the Least Squares covariance matrix estimator may not be justified even in small samples (MacKinnon and White, 1985).

Nevertheless, in some cases–such as the selection between the cluster-robust estimator and the heteroskedasticity-robust estimator–the selection of a less robust covariance matrix estimator in favor of the more robust estimator can improve the quality of inference about regression equation parameters. This is the motivation for "robustness tests" that examine which of two covariance matrix estimators should be selected.

In this paper we propose a new robustness testing strategy, and show that when applied to inference about clustering the proposed approach has good finite-sample performance unlike the existing testing strategy. Moreover, we show why the existing approach performs poorly, which to our knowledge has not been well understood. The insight into why the existing approach performs poorly is the basis for proposed robustness testing strategy. The proposed robustness testing approach can be adapted to improve inference also about parametric assumptions in covariance matrix estimation. We demonstrate this for the case of testing for heteroskedasticity. Our main focus on testing for clustering is driven by the fact that selecting between the cluster-robust covariance estimator and a less robust covariance estimator is a central statistical issue in today's empirical research and by the fact that–as indicated above–inference about clustering is well-motivated as the choice of the less robust heteroskedasticity-robust covariance matrix estimator over the cluster-robust estimator can measurably improve the quality of inference about regression parameters.

The existing approach to testing whether the use of the cluster-robust covariance matrix estimator is necessary was presented in Kezdi (2003) and Hansen (2007) and is a modification of the White (1980) heteroskedasticity test. In this robustness test–as in the original White (1980) heteroskedasticity test–the null hypothesis is that also the less robust covariance matrix estimator is a consistent estimator. These tests are constructed as a Wald test statistic from the contrast between the more robust and the less robust covariance matrix estimates. While this existing testing approach is firmly grounded on an asymptotic theory, the finite-sample performance of asymptotic and bootstrapped versions of this approach is poor both when applied to inference about clustering (see Hansen, 2007, and below) and when applied to inference about heteroskedasticity (see MacKinnon and White, 1985, and below). Specifically, this approach tends to have low power; it often fails to reject the consistency of the less robust covariance matrix estimator when the less robust estimator is an inconsistent estimator. However, to our knowledge the reason for this result has not been previously well understood.

The underlying reason for why the existing approach to testing for clustering performs poorly is that the construction of the Wald test statistic converts the tails of an asymmetric distribution into one tail. How this occurs is most accessibly demonstrated in the single regressor model. In this case the existing robustness test statistic is constructed as the square of a ratio. In the numerator in this ratio is the contrast between the more robust cluster-robust covariance matrix estimate and the less robust heteroskedasticity-robust estimate. In the denominator in this ratio is an estimate of the variance of the contrast in the numerator. The two tails of the distribution the ratio are different because the contrast in the numerator and the variance of the contrast in the denominator are correlated. And because the existing robustness test is constructed as the square of the ratio, the test converts the two very different tails of a distribution into one tail. As a result, this robustness test has poor small sample properties even when bootstrap is used.

The correlation between the contrast and the estimator of its variance in turn arises because the distribution of the variable which average forms the contrast has an asymmetric distribution. When the more robust estimator is the cluster-robust estimator and the less robust estimator is the heteroskedasticity-robust estimator, the distribution of the individual terms in the average is asymmetric because–as we show in this paper–the sum of all cross-products of T independent random variables has an asymmetric distribution even asymptotically.

In regression models with $K > 1$ regressors the existing robustness testing approach is based on the construction of a one-dimensional Wald test statistic from the $K + K(K-1)/2$ unique elements of the contrast between a more robust and a less robust covariance matrix estimator. Again, this existing approach has poor power properties because it converts the tails of an asymmetric distribution into one tail.

This insight into why the existing robustness tests perform poorly is the basis for our proposed alternative robustness testing strategy. In models with a single regressor the proposed testing strategy is to again first calculate the ratio of the contrast and the square root of an estimate of the variance of the contrast and then use the ratio itself–rather than the square of the ratio–as the test statistic. In models with multiple regressors our proposed robustness testing strategy is to first partial out the effect of all other explanatory variables except the variable associated with the parameter of interest, and then calculate the ratio of the contrast and the square root of an estimate of the variance of the contrast for this parameter and again use the ratio as the robustness test statistic. While some information is obviously lost when this proposed dimension-reduction approach is applied, our analysis shows that when applied to clustering the proposed approach can still be expected to dramatically outperform the existing approach. With sufficient computational resources, the proposed alternative robustness testing strategy can also be based on two (or more) parameters of the regression model. In this case a three-dimensional test statistic and the associated bootstrapped three (or higher) -dimensional rejection region are constructed.

Unlike the existing approach, the proposed approach does not convert the tails of an asymmetrically distributed variable into one tail. Consequently, as our analysis shows, the proposed robustness testing strategy has much better finite-sample performance than the existing approach even when the regression model has multiple regressors and the proposed robustness testing strategy is based on the ratio of the contrast and the square root of an estimate of its variance only for the main parameter of interest.

Our analysis has two main contributions. First, our analysis shows why applications of the White (1980) robustness testing strategy often perform poorly. This analysis shares some features with Altonji and Segal (1996) who examine the small-sample bias in Generalized Method of Moments (GMM) estimation of covariance structures. Second, we propose an alternative robustness testing strategy that performs well in small samples when applied to clustering. Our analysis also demonstrates that the proposed robustness testing strategy can improve inference about heteroskedasticity in comparison to the White (1980) heteroskedasticity test and the Wooldridge (1991) heteroskedasticity test, which is a heterokurtosis-robust version of the auxiliary regression based second heteroskedasticity test in White (1980). The literature on Lagrange multipliers (LM) type heteroskedasticity and autocorrelation tests is large, and we refer the reader to the recent contributions to this literature by Baltagi et al. (2010) and Montes-Rojas and Sosa-Escudero (2010) for the relevant references.

The good finite-sample performance of our proposed robustness testing strategy is important for two reasons. First, finite-sample performance is an especially important factor in the context of robustness tests because typically a small sample size is a necessary condition for a less robust covariance matrix estimator to have measurably better power properties than a more robust estimator. Thus, robustness tests are typically well-motivated only if the sample size is small. Second, use of robustness tests with better power properties in applied work would decrease the rate of false rejections of null hypotheses about the regression equation parameters. Accordingly, in addition to measuring the power of each robustness test, we measure the impact that applying each robustness test has on the probability of false rejections of a null hypothesis about a regression equation parameter. The latter statistics show that use of the proposed testing strategy can decrease the probability of erroneous inference about regression equation parameters quite dramatically compared to when the existing robustness testing approach is employed.

In the next section we first present the linear regression model and the cluster-robust and heteroskedasticity-robust covariance matrix estimators, and then demonstrate the potential advantage of using the less robust heteroskedasticity-robust estimator. In the third section we present the existing clustering test and examine why it performs poorly. The proposed alternative testing strategy and its finite-sample performance are examined in the fourth section. In the fifth section we apply the proposed testing strategy to testing for heteroskedasticity. The sixth section concludes.

2 The Linear Regression Model

In the next two subsections we present the linear regression model and two covariance matrix estimators. In the third subsection we compare the power properties of the associated hypothesis tests to demonstrate that there can be a downside to choosing the more robust estimator.

2.1 The Linear Regression Model

We examine the linear regression model

$$y_{it} = x_{it}'\beta + \varepsilon_{it}, \tag{1}$$

where observations are indexed on two dimensions $i \in \{1, ..., G\}$ and $t \in \{1, ..., T\}$, the variable x_{it} is a vector of K observable explanatory variables, the variable y_{it} is the dependent variable, and the variable ε_{it} is the error term. Let $\hat{\beta}$ denote the Least Squares estimator of the parameter vector β, and let $\hat{\varepsilon}_{it}$ denote the associated Least Squares residual $\hat{\varepsilon}_{it} = y_{it} - \hat{\beta}x_{it}$. Throughout the analysis we assume that $E\left[\varepsilon_{it}|\, x_{it}\right] = 0$ and that the usual conditions on the fourth moments of the observed variables y_{it} and x_{it} hold, so that the Least Squares estimator $\hat{\beta}$ is a consistent estimator. This enables us to focus solely on issues surrounding the estimation of the covariance matrix of the parameter estimates.

It is well-known that properties of different estimators of the covariance matrix of the parameter estimator $\hat{\beta}$ depend on the structure of the covariance matrix of the unobserved error terms ε_{it}. For the sake of expositional convenience and analytical tractability, we assume that the error terms ε_{it} are independent across the G-dimension and examine only the potential impacts of dependence among the error terms in the T-dimension. Hence, the relationships between the error terms ε_{it} are captured by the matrices

$$\Omega_i \equiv E\left[\varepsilon_i\varepsilon_i'|\, x_i\right], \text{ for all } i \in \{1, ..., G\}, \tag{2}$$

where the elements of the matrix Ω_i may depend on x_i.

2.2 The Heteroskedasticity-Robust and Cluster-Robust Estimators

While the matrix

$$W \equiv \lim_{G\to\infty} \sum_{i=1}^{G} E\left[x_i'\Omega_i x_i\right] \tag{3}$$

is *not* the actual covariance matrix of the estimator $\hat{\beta}$, different estimators of the covariance matrix of the estimator $\hat{\beta}$ mainly differ in terms of how this matrix W is estimated (see e.g. Hansen, 2007 and Kezdi, 2003). Accordingly, for expositional brevity, we use the term "covariance matrix" interchangeably in reference to the matrix W and the actual covariance matrix of the estimator $\hat{\beta}$.

The heteroskedasticity-robust estimator of the covariance matrix of the parameter $\hat{\beta}$ is based on the assumption that the matrix Ω_i is a diagonal matrix. When this property holds, the heteroskedasticity-robust estimator

$$\hat{W}_{HS} \equiv \frac{1}{GT} \sum_{i=1}^{G} \sum_{t=1}^{T} \hat{\varepsilon}_{it}^2 x_{it} x_{it}', \tag{4}$$

is a consistent estimator of the covariance matrix W.[1]

The cluster-robust estimator, in contrast, is motivated by potential within-group correlation (in the T-dimension) in the error terms ε_{it}, and is derived without any assumptions on the matrix Ω_i. The cluster-robust estimator of the covariance matrix W is given by

$$\hat{W}_{CLUSTER} \equiv \frac{1}{GT} \sum_{i=1}^{G} x_i'\hat{\varepsilon}_i\hat{\varepsilon}_i'x_i. \tag{5}$$

Unlike the heteroskedasticity-robust estimator $\hat{W}_{HS}$, the cluster-robust estimator $\hat{W}_{CLUSTER}$ is a consistent estimator of the covariance matrix W regardless of whether the matrix Ω_i is a diagonal matrix.[2]

[1] As Kezdi (2003) notes, the heteroskedasticity-robust estimator is consistent also if Ω_i are not diagonal but each explanatory variable x_{it} is uncorrelated within clusters. Stock and Watson (2008) show that in fixed effects models this heteroskedasticity-robust covariance matrix estimator is inconsistent under fixed T asymptotics (with $G \to \infty$), and offer an alternative heteroskedasticity-robust estimator which is consistent for fixed T.

[2] As is well-known, the least squares residuals $\hat{\varepsilon}_{it}$ are typically smaller than the error terms ε_{it} due to overfitting. Moreover, within-cluster correlation between least squares residuals is typically smaller than the within-cluster cor-

2.3 Robustness and the Power of Hypothesis Tests

The cluster-robust estimator $\hat{W}_{CLUSTER}$ relaxes one of the assumptions of the heteroskedasticity-robust estimator $\hat{W}_{HS}$. Hence, the cluster-robust estimator is in a sense a more robust estimator than the heteroskedaticity-robust estimator. However, the selection between the more robust and the less robust covariance matrix estimator can involve a genuine trade-off as the power of hypothesis tests about regression equation parameters can be higher when constructed using the less robust covariance matrix estimator. We next demonstrate this argument using Monte Carlo analyses conducted in the fixed effects and random effects specifications employed previously by Hansen (2007). The Hansen (2007) fixed effects (FE) specification can be expressed as

$$\begin{aligned} y_{it} &= x_{it}\beta + \alpha_i + \varepsilon_{it}, \qquad \alpha_i \sim N\left(0, .5\right) \\ x_{it} &= .5x_{it-1} + v_{it}, \qquad v_{it} \sim N\left(0, .75\right) \\ \varepsilon_{it} &= \rho\varepsilon_{it} + u_{it}\sqrt{(1-a) + a x_{it}^2}, \qquad u_{it} \sim N\left(0, 1-\rho\right), \end{aligned} \tag{6}$$

whereas the Hansen (2007) random effects (RE) specification is given by

$$\begin{aligned} y_{it} &= x_{it}\beta + \varepsilon_{it} \\ x_{it} &= z_i + v_{it}, \qquad z_i \sim N\left(0, .8\right),\ v_{it} \sim N\left(0, 1-.8\right) \\ \varepsilon_{it} &= \alpha_i + u_{it}, \qquad \alpha_i \sim N\left(0, \rho\right),\ u_{it} \sim N\left(0, 1-\rho\right). \end{aligned} \tag{7}$$

Either model can be written as (1) although for the fixed effects model the variables y_{it}, x_{it}, and ε_{it} then represent the fixed effect demeaned versions of the corresponding original variables. When

relation between the error terms (Bell and McCaffrey, 2002). Consequently, the heteroskedasticity-robust estimator $\hat{W}_{HS}$ and the cluster-robust estimator $\hat{W}_{CLUSTER}$ are biased estimators in finite samples. For presentational convenience, and because each of the many available bias-correction methods is only guaranteed to work under a very restrictive set of circumstances (MacKinnon and White, 1985, and Bell and McCaffrey, 2002), we do not employ a bias-correction in the analysis in the text. However, in constructing the heteroskedasticity-robust estimator and the cluster-robust estimator in our Monte Carlo simulations we apply the adjusted residuals $\hat{\varepsilon}\sqrt{\frac{GT}{GT-K}}$ and $\hat{\varepsilon}\sqrt{\frac{G}{G-1}}$, respectively, instead of the original residuals $\hat{\varepsilon}$. These bias-corrections are implemented in commonly used software packages such as Stata and SAS.

$\rho = 0$ in either model, both the heteroskedasticity-robust estimator and cluster-robust covariance estimator are consistent variance covariance matrix estimators.[3]

The two panels of Table 1 show the power functions with nominal size 0.05 against the null hypothesis H_0: $\beta = 0$ for $\beta \geq 0$ when $G = 10$, $T = 20$ and $a = 0.5$ in the FE model and when $G = 10$ and $T = 10$ in the RE model. In both the power functions for the heteroskedasticity-robust estimator are shown in columns 1 and 3 and the power functions for the cluster-robust estimator are shown in columns 2 and 4. Power functions calculated using bootstrapped distributions of the test statistic are indicated in bold (columns 3-4), and power functions calculated using asymptotic distributions of the test statistic are shown in columns 1-2.[4]

Comparison of row 1 of column 1 and row 1 of column 2 for the RE model demonstrates that when tests are conducted using asymptotic distributions, use of the less robust covariance matrix estimator can yield better size properties than use of the more robust variance covariance estimator. When tests are conducted using bootstrapped distributions this size advantage is eliminated but power can increase by up to 25% in the FE model (from .40 to .50 when $\beta = 0.2$) and by up to 47% in the RE model (from .30 to .44 when $\beta = 0.2$) when the heteroskedasticity-robust variance covariance matrix is employed instead of the cluster-robust estimator. These results do not represent the upper bound for the potential advantage of the less robust estimator. For example, in the RE model the difference in the power of tests is increasing in within-group correlation in the observed explanatory variable x_{it}. When the explanatory variable is generated with $x_{it} = z_i$, where $z_i \sim N(0,1)$, and $\beta = 0.2$ power of bootstrapped tests increases by 78% (from .27 to .48) if the heteroskedasticity-robust covariance estimator is employed instead of the cluster-robust estimator.

These results demonstrate that the power of hypothesis tests can be considerably higher when the less robust heteroskedasticity-robust estimator of the covariance matrix is applied, which implies

[3] Consistency of the heteroskedasticity-robust estimator in the FE model requires that $T \to \infty$ (see footnote 1).

[4] Each cell in Tables 1-2 is calculated using 2000 simulated samples. In constructing the bootstrapped power functions we impose the null hypothesis, employ the wild-bootstrap method, and obtain the distribution of the t-statistic (see Cameron et al., 2008). For each of the 2000 simulated samples we obtain the bootstrapped estimate of the distribution of the t-statistic using 799 samples bootstrapped from the simulated sample.

that the selection of a more robust covariance matrix estimator can have a considerable negative impact on the probability of correct statistical inference. This motivates the use and analysis of "robustness tests" that examine the appropriate level of robustness in covariance matrix estimation.

FE Model	Power of Test		Power of Test		RE Model	Power of Test		Power of Test	
	t_{HS}^{asym}	t_{CL}^{asym}	t_{HS}^{boot}	t_{CL}^{boot}		t_{HS}^{asym}	t_{CL}^{asym}	t_{HS}^{boot}	t_{CL}^{boot}
$\beta = 0$	.07	.07	**.06**	**.06**	$\beta = 0$	.06	.09	**.05**	**.05**
$\beta = .1$	.20	.17	**.19**	**.14**	$\beta = .1$	.17	.19	**.15**	**.11**
$\beta = .2$	.52	.43	**.50**	**.40**	$\beta = .2$	.47	.47	**.44**	**.30**
$\beta = .3$	.85	.75	**.82**	**.72**	$\beta = .3$	.77	.75	**.75**	**.55**
$\beta = .4$	.97	.94	**.97**	**.92**	$\beta = .4$	.94	.91	**.92**	**.74**
$\beta = .5$	1.00	.98	**1.00**	**.98**	$\beta = .5$	.98	.97	**.98**	**.88**

Table 1: Power function in the FE and RE models.

It is also important to note that this advantage of the less robust covariance matrix estimator decreases quickly as the number of clusters G increases. Monte Carlo simulations reported in Table 2 illustrate this result. In these simulations we set $\beta = 0.2$, $a = 0.5$ and $T = 20$ in the FE model and $\beta = 0.2$ and $T = 10$ in the RE model. The original error terms are multiplied by $\sqrt{G}/\sqrt{10}$ so that the variance of the Least Squares estimator $\hat{\beta}$ is approximately the same for all G. This finding, that the potential advantage of the less robust heteroskedasticity-robust covariance matrix estimator decreases considerably or even disappears as G increases, implies that in selecting between different robustness tests a strong emphasis should be based on the robustness tests' small sample properties.

FE Model	Power of Test		Power of Test		RE Model	Power of Test		Power of Test	
	t_{HS}^{asym}	t_{CL}^{asym}	t_{HS}^{boot}	t_{CL}^{boot}		t_{HS}^{asym}	t_{CL}^{asym}	t_{HS}^{boot}	t_{CL}^{boot}
$G = 10$	.54	.44	**.52**	**.40**	$G = 10$	.49	.48	**.46**	**.31**
$G = 15$	.53	.46	**.51**	**.43**	$G = 15$	.47	.49	**.45**	**.37**
$G = 20$	.50	.46	**.49**	**.44**	$G = 20$	.49	.50	**.48**	**.40**
$G = 50$	.52	.47	**.52**	**.46**	$G = 50$	.50	.52	**.50**	**.47**
$G = 100$	.50	.47	**.50**	**.46**	$G = 100$	.52	.51	**.51**	**.48**

Table 2: Power as a function of the number of clusters G in the FE and RE models.

3 The White (1980) Robustness Testing Approach

In this section we first show how the White (1980) robustness testing approach is applied to test for clustering, and then explain why this test performs poorly in small samples.

3.1 The White (1980) Test Adapted to Clustering

Kezdi (2003) and Hansen (2007) have previously presented how the White (1980) heteroskedasticity test is adapted to testing whether clustering in the error terms has an impact on the covariance matrix. In this robustness test the null and alternative hypotheses are

$$\begin{array}{lll} H_0 & \text{(no clustering):} & \text{plim}_{G\to\infty}\left[\hat{W}_{CLUSTER}-\hat{W}_{HS}\right]=0 \\ H_1 & \text{(clustering):} & \text{plim}_{G\to\infty}\left[\hat{W}_{CLUSTER}-\hat{W}_{HS}\right]\neq 0 \end{array}$$

and the test statistic is based on the contrast

$$\hat{W}_{CLUSTER}-\hat{W}_{HS} \tag{8}$$

between the values of the two covariance matrix estimators $\hat{W}_{CLUSTER}$ and $\hat{W}_{HS}$, which are defined above in expressions (5) and (4), respectively. When the null hypothesis of no clustering holds (does not hold), the $K+K\left(K-1\right)/2$ unique individual elements of this contrast matrix will be relatively small (large) in absolute value. The existing robustness test statistic, which we denote by S^*, is constructed from the contrast (8) in the form of the Wald test statistic as

$$S^* = GT \times vec\left(\hat{W}_{CLUSTER}-\hat{W}_{HS}\right) D^{-} vec\left(\hat{W}_{CLUSTER}-\hat{W}_{HS}\right)', \tag{9}$$

where the matrix D is an estimator of the variance of $vec\left(\hat{W}_{CLUSTER}-\hat{W}_{HS}\right)$ and D^- denotes the generalized inverse of D. We estimate the parameter matrix D using the estimator

$$\hat{D} \equiv \left(\frac{1}{GT}\sum_{i=1}^{G}\left(vec\left(x_i'\hat{\varepsilon}_i\hat{\varepsilon}_i'x_i-\sum_{t=1}^{T}\hat{\varepsilon}_{it}^2 x_{it}x_{it}'\right)\right)\left(vec\left(x_i'\hat{\varepsilon}_i\hat{\varepsilon}_i'x_i-\sum_{t=1}^{T}\hat{\varepsilon}_{it}^2 x_{it}x_{it}'\right)\right)'\right). \tag{10}$$

The estimator $\hat{D}$ is one of two estimators of D mentioned in Hansen (2007). The existing clustering test performs better with estimator $\hat{D}$ than with the alternative variance estimator.

Hansen (2007) provides sufficient conditions for the result that under the null hypothesis of "no clustering" the test statistic S^* has the asymptotic distribution $\chi^2_{k(k+1)/2}$ both when $G \to \infty$ and T is fixed and when $G \to \infty$ and $T \to \infty$ jointly. However, as Monte Carlo results in Hansen (2007) show, the finite-sample performance of this test is poor when the number of clusters G is small.

3.2 Why Does the White (1980) Robustness Test Perform Poorly?

For expositional convenience we now focus on the model with just one regressor, x_{it}. The two covariance matrix estimators $\hat{W}_{HS}$ and $\hat{W}_{CLUSTER}$ can then be written simply as

$$\hat{W}_{HS} = \frac{1}{GT} \sum_{i=1}^{G} \sum_{t=1}^{T} x_{it}^2 \hat{\varepsilon}_{it}^2 \tag{11}$$

and

$$\hat{W}_{CLUSTER} = \frac{1}{GT} \sum_{i=1}^{G} \sum_{t=1}^{T} \sum_{s=1}^{T} x_{it} \hat{\varepsilon}_{it} \hat{\varepsilon}_{is} x_{is}. \tag{12}$$

Moreover, the cluster-robust estimator can be rewritten as

$$\hat{W}_{CLUSTER} = \frac{1}{GT} \sum_{i=1}^{G} \left[\sum_{t=1}^{T} x_{it}^2 \hat{\varepsilon}_{it}^2 + 2 \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} x_{it} \hat{\varepsilon}_{it} \hat{\varepsilon}_{is} x_{is} \right]. \tag{13}$$

Substituting expressions (11) and (13) to the contrast $\hat{W}_{CLUSTER} - \hat{W}_{HS}$ yields

$$\hat{W}_{CLUSTER} - \hat{W}_{HS} = \frac{1}{G} \sum_{i=1}^{G} \frac{2}{T} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} x_{it} \hat{\varepsilon}_{it} \hat{\varepsilon}_{is} x_{is}. \tag{14}$$

The contrast $\hat{W}_{CLUSTER} - \hat{W}_{HS}$ is thus calculated as the average of G observations on the sum

$$\frac{2}{T} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} x_{it} \hat{\varepsilon}_{it} \hat{\varepsilon}_{is} x_{is}. \tag{15}$$

Expression (15) consists of the summation of all $T\,(T-1)\,/2$ cross-products of the T random variables $x_{i1}\hat{\varepsilon}_{i1}$ through $x_{iT}\hat{\varepsilon}_{iT}$. In what follows we first show that the distribution of the sum of all cross-products of T random variables is asymmetric even when the T random variables are jointly independent as the variables $x_{i1}\hat{\varepsilon}_{i1}$ through $x_{iT}\hat{\varepsilon}_{iT}$ are (asymptotically) under the null hypothesis of "no clustering" To facilitate both finite-sample and asymptotic analysis we conduct this analysis in terms of the normalized sum of cross-products expressed as

$$\sqrt{T\,(T-1)\,/2} \times \frac{\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} x_{it}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}x_{is}}{T\,(T-1)\,/2}. \tag{16}$$

We then show that the asymmetric distribution of (16) implies that the contrast $\hat{W}_{CLUSTER} - \hat{W}_{HS}$ is positively correlated with the estimator of the variance of this contrast. In the third step we show that, due to this positive correlation, the ratio of the contrast $\hat{W}_{CLUSTER} - \hat{W}_{HS}$ and the square root of an estimate of the variance of this contrast has an asymmetric distribution. In the final step we show that because the existing robustness test statistic is calculated as the square of this asymmetrically distributed ratio, the finite-sample properties of the test are destined to be poor.

3.2.1 Distribution of All Cross-Products of Independent Variables is Asymmetric

We now show that when the null hypothesis of "no clustering" holds, both the finite-sample and asymptotic distributions of the sum of cross-products (16) are asymmetric.

Consider first the sign of the individual terms in the sum of cross-products (16). When $x_{it}\hat{\varepsilon}_{it}$ is positive for all t, the sum of cross-products (16) is of course positive because then all terms in the sum are positive. In contrast, not all of the terms in the sum of cross-products in (16) can be simultaneously negative. For example, when $T = 3$, only two of the three terms can be negative at once.[5] This result arises because the individual terms in the sum of cross-products (16) are *not* jointly independent even though they are pairwise uncorrelated asymptotically (as $\text{plim}_{G\to\infty}\hat{\beta} \to \beta$

[5] If the terms $x_{i1}\hat{\varepsilon}_{i1}\hat{\varepsilon}_{i2}x_{i2}$ and $x_{i1}\hat{\varepsilon}_{i1}\hat{\varepsilon}_{i3}x_{i3}$ are negative, the term $x_{i3}\hat{\varepsilon}_{i3}\hat{\varepsilon}_{i2}x_{i2}$ cannot be negative, as either $x_{i1}\hat{\varepsilon}_{i1} < 0$ and $x_{i2}\hat{\varepsilon}_{i2} > 0$ and $x_{i3}\hat{\varepsilon}_{i3} > 0$ hold or $x_{i1}\hat{\varepsilon}_{i1} > 0$ and $x_{i2}\hat{\varepsilon}_{i2} > 0$ and $x_{i3}\hat{\varepsilon}_{i3} < 0$ hold.

and thereby $\hat{\varepsilon}_{it} \approx \varepsilon_{it}$) when the null hypothesis of "no clustering" holds. That all terms in the sum of cross-products (16) can be simultaneously positive but not simultaneously negative immediately implies that the finite-sample distribution of the sum of cross-products (16) is asymmetric.

Consider next the sum of cross-products (16) for large T. In Appendix 1.1 we show that in the limit, as $G \to \infty$ (so that $P(\hat{\varepsilon}_{it} < 0)$ is arbitrarily close to $P(\varepsilon_{it} < 0)$ independent of the cluster size T) and $T \to \infty$, the probability that the sum of $T(T-1)/2$ cross-products (16) has more negative terms than positive terms is more than two thirds, provided that for the error term ε_{it} negative and positive values are equally likely. Formally, we show that

$$\lim_{G,T\to\infty} P\left(\frac{\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} I_{x_{it}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}x_{is}>0} + \sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\left(I_{x_{it}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}x_{is}>0} - 1\right)}{\sqrt{T(T-1)/2}} < 0\right) \approx 0.6823, \qquad (17)$$

where $I_{x_{it}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}x_{is}>0} = 1$ if $x_{it}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}x_{is} > 0$ and zero otherwise.

If x_{it} and ε_{it} have symmetric discrete distributions with $P(x_{it} = 1) = 0.5$ and $P(\varepsilon_{it} = 1) = 0.5$, so that $\lim_{G\to\infty} P(\hat{\varepsilon}_{it} < 0) = P(\varepsilon_{it} < 0)$ and $\text{plim}_{G\to\infty} |x_{it}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}x_{is}| = 1$, the above result (17) immediately implies that in the limit (as $G \to \infty$ and $T \to \infty$) also the probability that the value of the sum-of cross-products (16) itself is positive is less than one third. Obtaining analytical asymptotic results for other distributions is complicated by the dependence between terms in the sum of cross-products. However, that similar results apply to other distributions for the variables x_{it} and ε_{it} is easily verified using Monte Carlo simulations.

Table 3 reports such Monte Carlo results for the FE and RE models (6) and (7), with the sum of cross-products (16) denoted by Δw_i. Column 1 reports the empirical probability that the sum of cross-products (16) is negative.[6] Columns 2 and 3, respectively, report the empirical probability that the sum of cross-products (16) is less than $c_{0.75}$ and $c_{.975}$, where c_p is defined as the empirical critical value that satisfies $P(\Delta w_i > c_p) = 1 - p$. Thus, for a symmetric distribution

[6] Each entry in Table 3 is constructed using the observation on the sum of cross-products (16) for one cluster in 1000 simulated samples. In these simulations the number of clusters G is set relatively high so that the bias of the estimators $\hat{W}_{HS}$ and $\hat{W}_{CLUSTER}$ is small and, consequently, $E(\Delta w_i)$ is close to zero. In all other reported simulations we apply the bias-corrections (see the end of Section 2.2) in estimating $\hat{W}_{HS}$ and $\hat{W}_{CLUSTER}$.

$P\left(\Delta w_i < -c_{0.75}\right) = 0.25$ and $P\left(\Delta w_i < -c_{0.975}\right) = 0.025$. The results show that the distribution of the sum of cross-products (16) remains highly asymmetric as T increases. Moreover, for large T the probability that the sum of cross-products (16) is negative is again around two thirds.

		$P\left(\Delta w_i < 0\right)$	$P\left(\Delta w_i < -c_{0.75}\right)$	$P\left(\Delta w_i < -c_{0.975}\right)$
FE Model, $\rho = 0,\ a = 0,\ G = 100$	$T = 20$	.61	.37	.00
	$T = 100$	.67	.55	.00
	$T = 200$	.69	.63	.00
FE Model, $\rho = 0,\ a = .5,\ G = 100$	$T = 20$	.61	.31	.03
	$T = 100$	.66	.47	.04
	$T = 200$	.70	.65	.00
RE Model, $\rho = 0,\ G = 100$	$T = 20$	.68	.56	.04
	$T = 100$	.67	.53	.00
	$T = 200$	.67	.50	.06

Table 3: Distribution of the sum of cross-products (16).

3.2.2 Variables $\hat{W}_{CLUSTER} - \hat{W}_{HS}$ and $\widehat{var}(\hat{W}_{CLUSTER} - \hat{W}_{HS})$ are Correlated

The right-skewed distribution of the sum of cross-products (15) implies that the contrast $\hat{W}_{CLUSTER} - \hat{W}_{HS}$ is positively correlated with the estimator of the variance of this contrast. We prove this result in Appendix 1.2. The proof is similar to the proof of a related result in Altonji and Segal (1996) who examine bias in the GMM estimation of covariance structures and in which the variable for which the average is calculated has an asymmetric distribution because it is the second sample moment of a random variable.

Figure 1 illustrates the correlation between the contrast $\hat{W}_{CLUSTER} - \hat{W}_{HS}$ and the estimator of its variance. Here each sub-figure depicts 5000 observations on $\sqrt{GT}(\hat{W}_{CLUSTER} - \hat{W}_{HS})/\sqrt{2(T-1)}$ (vertical axis) and $\sqrt{\widehat{var}(\hat{W}_{CLUSTER} - \hat{W}_{HS})}/\sqrt{2\,(T-1)}$ (horizontal axis) in the FE model when $T = 100$, $a = 0$ and $\rho = 0$.[7] Depicting $\sqrt{GT}(\hat{W}_{CLUSTER} - \hat{W}_{HS})/\sqrt{2(T-1)}$ rather than $(\hat{W}_{CLUSTER} -$

[7] Estimates shown in Figures 1-3 are calculated using the bias-corrections mentioned in footnote 2. In these simulations we set $a = 0$ and $T = 100$ so that the two conditions–homoskedastic error terms and $\left(x'x\right)^{-1}\left(x_i'x_i\right)$ constant across clusters–under which the bias-correction $\sqrt{\frac{G}{G-1}}$ eliminates the bias of $\hat{W}_{CLUSTER}$ approximately hold (see Theorem 1 in Bell and McCaffrey, 2002).

$\hat{W}_{HS}$) or $\sqrt{GT}(\hat{W}_{CLUSTER} - \hat{W}_{HS})$ on the vertical axis in Figure 1 keeps the variance of the measured variable roughly constant across different G as the contrast $\hat{W}_{CLUSTER} - \hat{W}_{HS}$ consists of the summation of $2GT(T-1)$ terms. Figure 1 shows that the contrast and the estimator of its variance are correlated and that this result is not limited to the case when the number of clusters G is small. However, because the variance estimator $\widehat{var}(\hat{W}_{CLUSTER} - \hat{W}_{HS})$ approaches $var(\hat{W}_{CLUSTER} - \hat{W}_{HS})$ in probability as the number of clusters G increases, the variance of the variance estimator decreases as G increases. Consequently, when the number of clusters G is large, the impact of correlation between the contrast and the estimator of its variance on the ratio of the contrast and the square root of an estimate of its variance–which we examine next–is small.

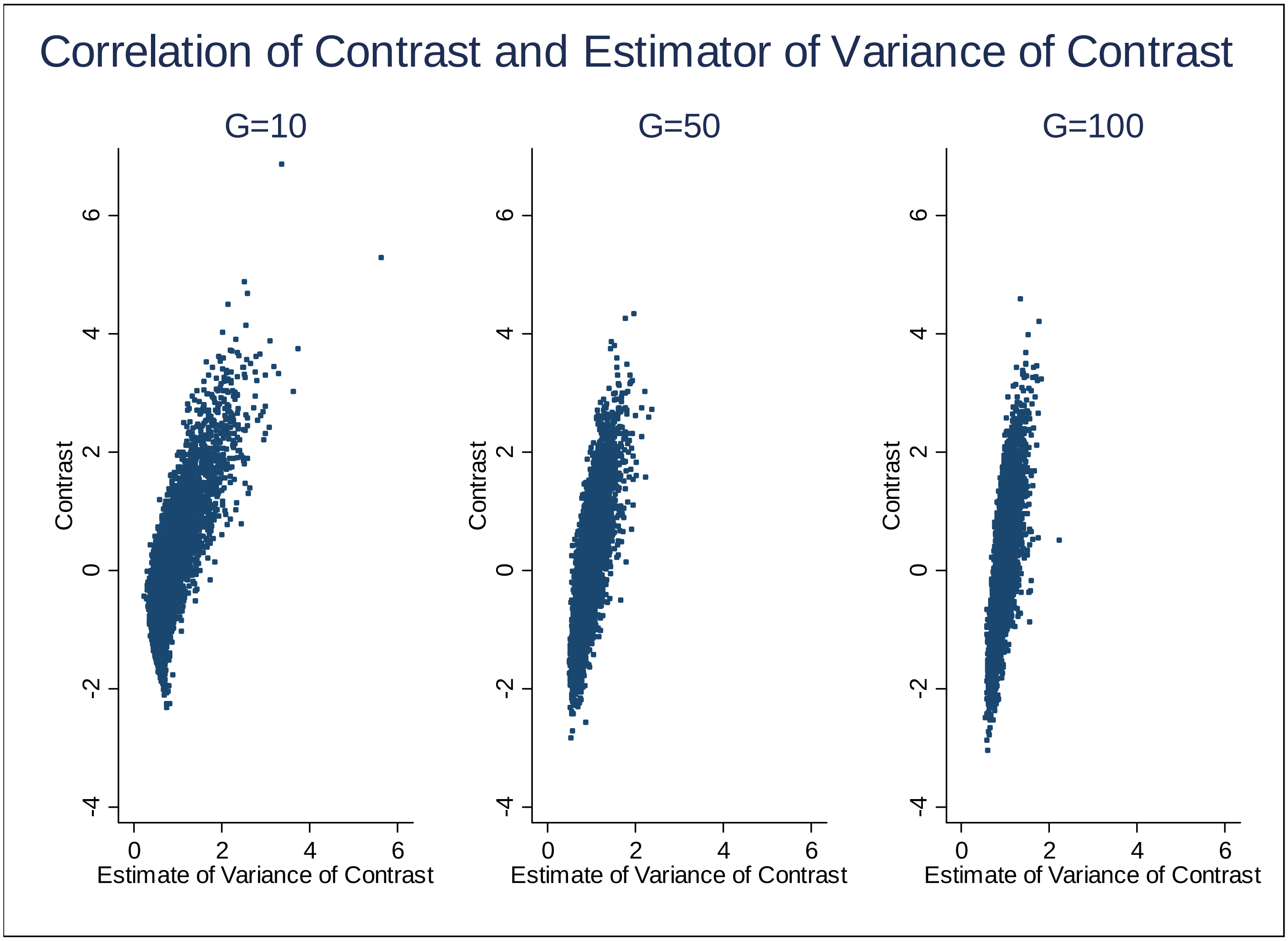

Figure 1: Scatterplot of $\hat{W}_{CLUSTER} - \hat{W}_{HS}$ and $\sqrt{\widehat{var}(\hat{W}_{CLUSTER} - \hat{W}_{HS})}$ in the FE model.

3.2.3 Distribution of the Ratio of $\hat{W}_{CLUSTER} - \hat{W}_{HS}$ and $\sqrt{\widehat{var}(\hat{W}_{CLUSTER} - \hat{W}_{HS})}$ is Asymmetric in Small Samples

We now examine the implications of the positive correlation between the contrast $\hat{W}_{CLUSTER} - \hat{W}_{HS}$ and the estimator of its variance, $\widehat{var}(\hat{W}_{CLUSTER} - \hat{W}_{HS})$, on the ratio

$$t_R \equiv \sqrt{GT} \frac{\hat{W}_{CLUSTER} - \hat{W}_{HS}}{\sqrt{\widehat{var}(\hat{W}_{CLUSTER} - \hat{W}_{HS})}} \tag{18}$$

when the null hypothesis of "no clustering" holds.

The positive correlation between the contrast $\hat{W}_{CLUSTER} - \hat{W}_{HS}$ and $\widehat{var}\left(\hat{W}_{CLUSTER} - \hat{W}_{HS}\right)$ implies that in constructing the ratio t_R a positive estimate of the contrast is generally divided by a larger estimate of its variance than the corresponding negative estimate of the contrast. When the expected value of the contrast itself is zero, $E(\hat{W}_{CLUSTER} - \hat{W}_{HS}) = 0$, this in turn implies that the expected value of the the ratio t_R is generally negative (Altonji and Segal, 1996, present a related informal argument). Formally,

$$E\left(\sqrt{GT} \frac{\hat{W}_{CLUSTER} - \hat{W}_{HS}}{\sqrt{\widehat{var}(\hat{W}_{CLUSTER} - \hat{W}_{HS})}}\right) < 0. \tag{19}$$

When the result (19) holds, then $P\left(t_R < -c\right) > P\left(t_R > c\right)$ at least for some critical values c. In other words, the two tails of the distribution of the ratio t_R are different and, more specifically, the distribution of the ratio t_R has more probability mass in the left tail than in the right tail. Even if the condition $E(\hat{W}_{CLUSTER} - \hat{W}_{HS}) = 0$ holds, this property does not necessarily hold for all critical values c because–due to the asymmetric distribution of the sum of cross-products (16)–for the contrast $\hat{W}_{CLUSTER} - \hat{W}_{HS}$ itself large positive values can be more likely than corresponding negative values. Moreover, the condition $E(\hat{W}_{CLUSTER} - \hat{W}_{HS}) = 0$ does not usually hold exactly as the estimators $\hat{W}_{HS}$ and $\hat{W}_{CLUSTER}$ are generally biased (see footnote 2). However, the result (19) generally continues to hold if this bias is small enough for both estimators. And even when these biases are large enough for the result (19) not to hold, the ratio t_R still has an asymmetric distribution because larger values of the contrast are divided by larger estimates of the variance, and because the distribution of the contrast is asymmetric.

Figure 2 depicts the distribution of the ratio t_R in the FE model for different numbers of clusters G, with each distribution estimated using 5000 simulations, when $\rho = 0$, $a = 0$, and $T = 100$. As expected, the left tail of the distribution of t_R has more probability mass than the left tail of the distribution when G is small. And, also as expected, this asymmetry disappears as G becomes large.

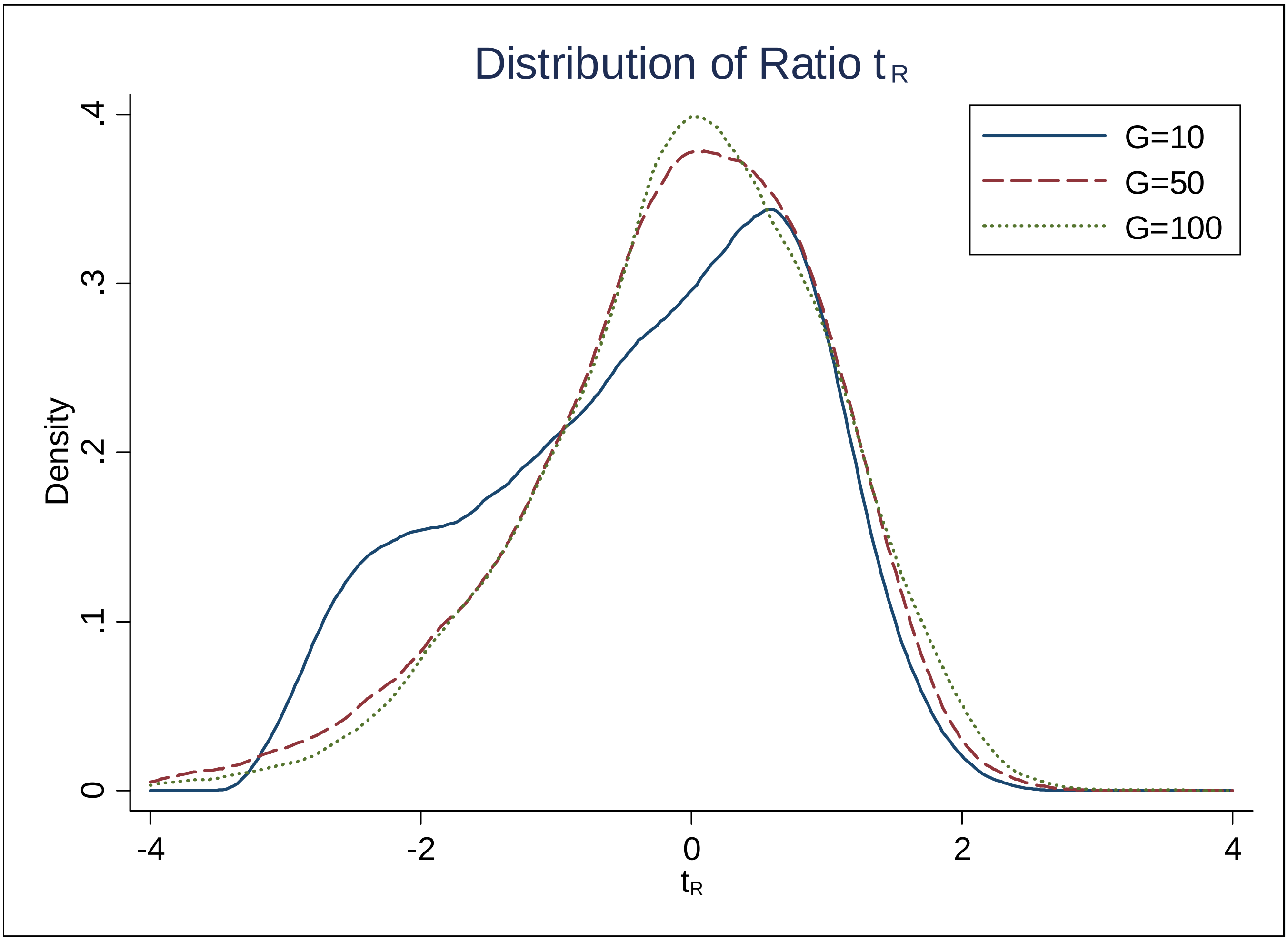

Figure 2: Distribution of the ratio t_R in the FE model as a function of the number of clusters G.

3.2.4 Power of the White (1980) Robustness Test

We now consider the implications of the asymmetry in the distribution of the ratio t_R on using the existing clustering test statistic S^* as written in (9). Because the model under consideration has only one regressor, the test statistic S^* can be rewritten simply as the square of the ratio t_R,

$$S^* = \left[\sqrt{GT} \frac{\hat{W}_{CLUSTER} - \hat{W}_{HS}}{\sqrt{\widehat{var}(\hat{W}_{CLUSTER} - \hat{W}_{HS})}} \right]^2 . \tag{20}$$

Suppose first that the null hypothesis "no clustering" holds. As was discussed in Section 3.2.3, under the null hypothesis the left tail of the distribution of the ratio t_R tends to have more probability mass than the right tail of the distribution. The construction of the test statistic S^* as the square of the ratio t_R ignores this asymmetry in the distribution of the ratio t_R by converting the two tails of its distribution into one tail. Consequently, when the null hypothesis holds, observations on the test statistic S^* that fall into the constructed rejection region mostly correspond to the negative values of the ratio t_R. Importantly, this occurs even if the bootstrapped distribution of S^* is used to construct the rejection region for the test statistic S^*.

Suppose next that the null hypothesis does not hold. And, more specifically, consider the case of positive clustering, which is arguably the empirically more relevant part of the alternative hypothesis. Positive clustering shifts the distribution of the ratio t_R to the right and thus increases its expected value. Yet, because the expected value of the ratio t_R is negative under the null hypothesis, positive clustering initially decreases the *absolute* value of the expected value of the ratio t_R. And because both the asymptotic and bootstrapped rejection regions of the test statistic S^* mostly correspond to the negative values of the ratio t_R under the null hypothesis, a shift in the distribution of the ratio t_R to the right initially does not necessarily increase the probability that an observation on S^* is in the constructed rejection region for S^*.

With weak positive clustering the distribution of the test statistic S^* therefore does not necessarily overlap much with the relevant rejection region; the power of the test may even be lower than its size, as can be seen from some of our Monte Carlo simulations below. The power of the test will only exceed its size when positive clustering is strong enough to shift also the distribution of the test statistic S^* to the right in comparison to its distribution under the null hypothesis.

Figure 3 illustrates the distribution of the ratio t_R and the test statistic S^* under null and alternative hypotheses in the FE model, with each distribution estimated using 5000 simulations, when $G = 10$, $T = 100$ and $a = 0$. Under the null hypothesis ($\rho = 0$) the distribution of the ratio t_R is right-skewed. While positive clustering ($\rho > 0$) shifts the distribution of the ratio t_R to the right, it does not initially increase the probability mass in the right tail of the distribution of the test statistic S^*. Consequently, the power of the existing clustering test can be lower than its size.

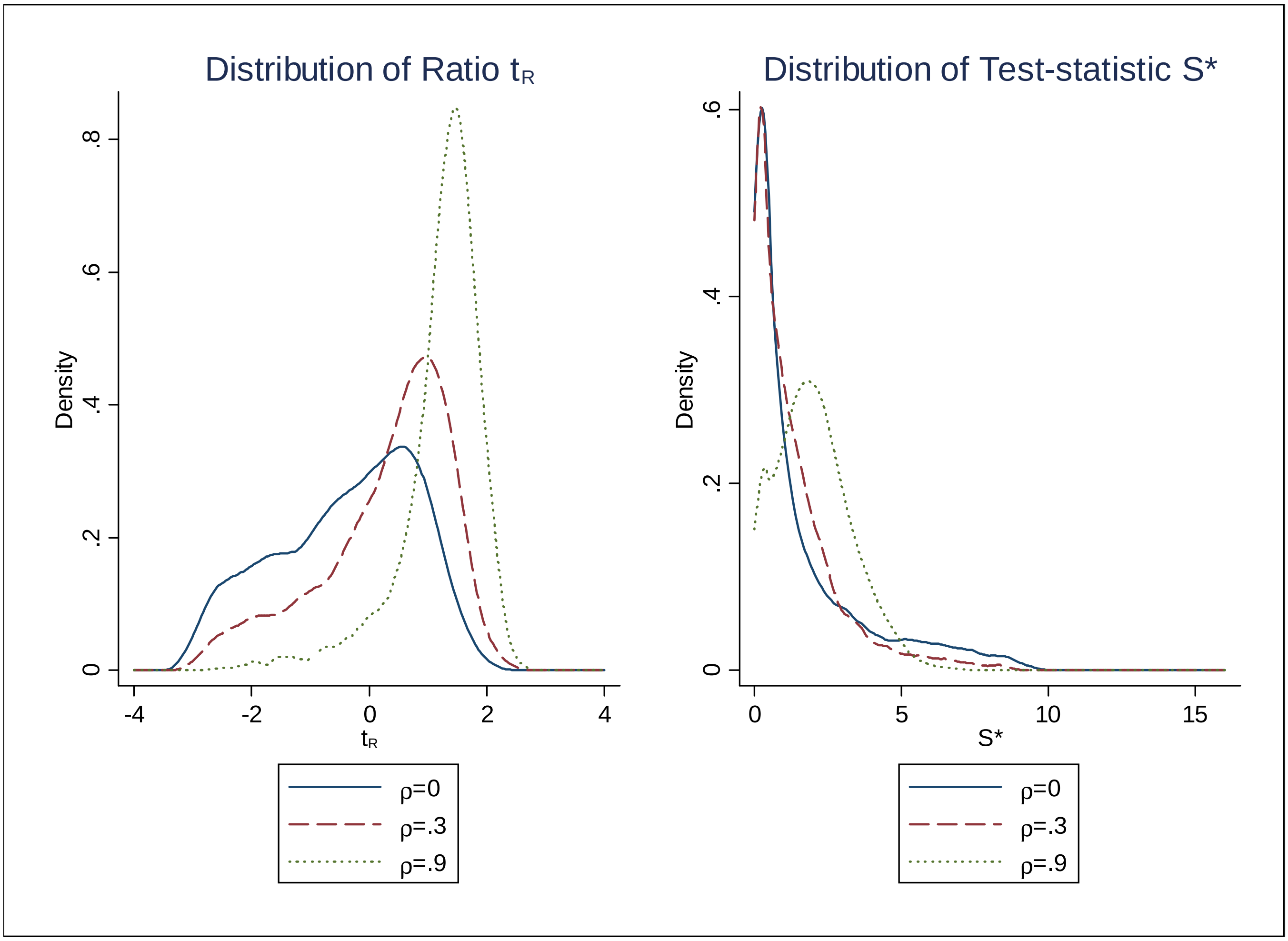

Figure 3: Distribution of t_R and S^* under null and alternative hypotheses in the FE model.

The two panels of Table 4 demonstrate the size and power properties of the asymptotic and bootstrapped versions of this existing clustering test in the FE and RE models (6) and (7). In these Monte Carlo simulations we set $a = 0.5$ for the FE model, $G = 10$ for both models, and vary the parameters ρ and T. In both panels of Table 4 columns 1 and 3 (columns 2 and 4) show the power function for the asymptotic (bootstrapped) version of the existing robustness test. Results for the asymptotic version of the test are based on 1000 replications. Results for the bootstrapped version of this robustness test are generated as in Table 5 below (see footnote 8 for the details).

As expected, the results show that for small values of ρ the power of the existing robustness test is smaller than the size of the test. This occurs both for the asymptotic and bootstrapped versions of the test. When $T = 50$ the size of the asymptotic version of the test exceeds the nominal size. The bootstrapped version of the test eliminates this size distortion but has only little power in the RE model and no power in the FE model.

FE Model

	Power of Robustness Test S^*			
	$T = 20$		$T = 50$	
	Asym.	Boot.	Asym.	Boot.
ρ= 0	.05	.06	.10	.04
ρ= .1	.04	.02	.08	.04
ρ= .2	.04	.02	.06	.04
ρ= .3	.03	.03	.05	.01
ρ= .6	.05	.03	.05	.01
ρ= .9	.04	.03	.06	.03

RE Model

	Power of Robustness Test S^*			
	$T = 10$		$T = 50$	
	Asym.	Boot.	Asym.	Boot.
ρ= 0	.06	.04	.11	.05
ρ= .1	.01	.03	.08	.01
ρ= .2	.02	.03	.10	.02
ρ= .3	.06	.02	.18	.05
ρ= .6	.12	.05	.26	.04
ρ= .9	.18	.14	.26	.12

Table 4: Asymptotic and bootstrapped power functions for the White (1980) robustness testing strategy when applied to testing for clustering.

In summary, in this section we have shown why even the bootstrapped versions of the existing clustering test perform poorly against positive clustering when the number of clusters G is small. This finding is important since clustering tests are only well-motivated when the number of clusters G is small and positive clustering arguably forms the important part of the alternative hypothesis.

4 An Alternative Robustness Testing Strategy

The alternative robustness testing strategy that we propose in this paper is based on the insight in the previous section on why the White (1980) robustness testing approach performs poorly when applied to inference about clustering. In the case of one explanatory variable, the problems arise because construction of the existing test statistic S^* as the square of the ratio (18) converts the two very different tails of the distribution of the ratio t_R into one tail. As a solution, in the case of one explanatory variable, we propose using this ratio t_R itself as a test statistic. Hence, the proposed robustness test statistic, which we denote by $\tilde{S}$, is

$$\tilde{S} \equiv \sqrt{GT} \frac{\hat{W}_{CLUSTER} - \hat{W}_{HS}}{\sqrt{\widehat{var}(\hat{W}_{CLUSTER} - \hat{W}_{HS})}}. \tag{21}$$

In principle, in models with K multiple regressors, this same approach can be applied through the construction of the corresponding $K + K(K-1)/2$-dimensional test statistic, which is constructed by dividing each element of $\sqrt{GT} vec(\hat{W}_{CLUSTER} - \hat{W}_{HS})$ by the square root of the corresponding element of $\widehat{var}(vec(\hat{W}_{CLUSTER} - \hat{W}_{HS}))$, and the construction of the associated bootstrapped $K + K(K-1)/2$-dimensional rejection region. In practice, however, calculating the multi-dimensional bootstrapped rejection region is likely to be computationally too burdensome except in the case of $K = 2$, in which case the test statistic and the rejection region are three-dimensional. Consequently, in models with multiple regressors a more practical version of the proposed robustness testing approach consists of first partialling out the impact of all other regressors except the variables corresponding to the one or two main parameters of interest, and then constructing a one or three dimensional test statistic $\tilde{S}$ and the associated bootstrapped rejection region. This dimension reduction approach obviously has its disadvantages. Thus, it is important to show that even when the proposed robustness testing strategy involves such a dimension reduction it can be expected to outperform the existing approach.

In the next two subsections we present results on Monte Carlo simulations that examine the performance of the proposed robustness testing strategy and how it compares with the performance of the White (1980) robustness testing approach when applied to inference about clustering. We first focus on the case of one regressor and then examine the case of multiple regressors. We only report results for the bootstrapped version of the existing test because it performs better than the asymptotic version of this test when the number of clusters G is small (see Section 3.2.4).

The main purpose of robustness tests is to enable researchers to avoid false rejections of null hypotheses about regression equation parameters. In most cases such null hypothesis is H_0: $\beta_1 = 0$, and thus the performance of robustness tests in the case $\beta_1 = 0$ is particularly important. Accordingly, for the parameter of interest β_1 we set $\beta_1 = 0$. Moreover, in addition to calculating the size and power of each robustness test, we calculate for each robustness test the associated probability that the null hypothesis H_0: $\beta_1 = 0$ about the parameter of interest β_1 is rejected when the researcher follows the following three-step hypothesis testing strategy:

Step 1. Test if the null hypothesis H_0: $\beta = 0$ is rejected using the more robust cluster-robust estimator of $var(\hat{\beta})$.

Step 2. If the null hypoothesis H_0: $\beta = 0$ was not rejected in Step 1, use a robustness test to test whether also the less robust heteroskedasticity-robust estimator of $var(\hat{\beta})$ is a consistent estimator.

Step 3. If the consistency of the less robust heteroskedasticity-robust estimator was not rejected in Step 2, test the null hypothesis H_0: $\beta = 0$ using the less robust heteroskedasticity-robust estimator of $var(\hat{\beta})$.

This three-step testing strategy is common in applied work although it is not always explicitly stated and in many contexts its use has been limited by the poor power of existing robustness tests. We refer to the probability that the null hypothesis H_0: $\beta_1 = 0$ is rejected (either in Step 1 or in Step 3) as a "*conditional β-size*" of a robustness test when the null hypothesis H_0: $\beta_1 = 0$ holds and this three-step research strategy is followed. Comparison of the conditional β-size of the existing robustness testing strategy and the conditional β-size of the proposed robustness testing strategy yields an indication of the how much the proposed testing strategy can improve the quality of inference about regression parameters.

4.1 Comparisons of Robustness Tests: One Explanatory Variable

Tables 5 and 6 report Monte Carlo results comparing the properties of the White (1980) robustness tests and the properties of the proposed alternative robustness testing strategy in the FE and RE models (6) and (7) with one explanatory variable.[8] In the FE model we set $a = 0.5$. We vary the parameter ρ, which captures the within-cluster dependence in the error terms, the number of clusters G and the number of observations T within each cluster. Properties of the two robustness tests are reported in columns 1-4. The properties of the proposed alternative testing strategy are indicated in bold (columns 2 and 4). Columns 1 and 2 report the power function for each robustness test, and columns 3 and 4 report the associated probability of false rejections of the null hypothesis H_0: $\beta = 0$ when researchers follow the three-step research strategy discussed above. Columns 5 and 6 report the probability of false rejections of the null hypothesis H_0: $\beta_1 = 0$ when the test is conducted using the cluster-robust and the heteroskedasticity-robust variance estimator, respectively. Comparison of entries in columns 5 and 6 thus give a rough measure of the importance of using the cluster-robust variance estimator in cases with within-cluster dependence in the error terms (i.e. cases with $\rho > 0$).[9]

[8] The Monte Carlo results reported in each cell of Tables 5-8 are based on 400 simulated samples. The rejection frequencies for the test statistics S^* and $\tilde{S}$ are constructed from the bootstrapped distribution of each test statistic. The bootstrapped distribution of the test statistic is constructed using 399 bootstrapped samples from the original sample. We use the wild bootstrap method with the relevant weight (-1 or 1) independently drawn for each observation. This approach imposes the null hypothesis that the errors are not clustered.

In constructing the test statistics for the RE model we first de-mean the data so that the constant is not treated as a regressor when the relevant contrast for the existing clustering test is constructed. The results for the existing test are worse when the constant is included. Moreover, when the constant is omitted from the construction of the contrast for the existing test statistic (and thus the contrast is one-dimensional) the difference in the performance of the two robustness testing approaches can be solely attributed to the asymmetric nature of the distribution of the ratio of the contrast and the estimator of its variance.

[9] To limit the computational burden we rely on the asymptotic rejection region to calculate the probability of false rejections of the null hypothesis H_0: $\beta_1 = 0$ for each covariance matrix estimator, although in applied work the asymptotic rejection region should not be employed when the number of clusters is as small as 10 (see Cameron et al., 2008, and Table 1 above).

	Power of Robustness Test		Conditional β-Size		β-Size	β-Size
	S^*	$\tilde{S}$	S^*	$\tilde{S}$	Cluster-Robust	HS-Robust
$T = 20,\ G = 10$						
$\rho = .0$	.06	**.06**	.10	**.10**	.07	.08
$\rho = .3$	.03	**.13**	.13	**.11**	.08	.11
$\rho = .6$	.03	**.19**	.16	**.14**	.04	.16
$\rho = .9$	.03	**.25**	.23	**.19**	.08	.23
$T = 50,\ G = 10$						
$\rho = .0$	.04	**.07**	.07	**.07**	.05	.05
$\rho = .3$	.01	**.10**	.08	**.07**	.04	.08
$\rho = .6$	.01	**.25**	.16	**.13**	.06	.16
$\rho = .9$	.03	**.36**	.18	**.13**	.06	.19
$T = 20,\ G = 50$						
$\rho = .0$	.04	**.07**	.05	**.05**	.04	.04
$\rho = .3$	.14	**.41**	.07	**.05**	.04	.08
$\rho = .6$	.40	**.76**	.10	**.06**	.04	.17
$\rho = .9$	.56	**.92**	.09	**.04**	.03	.18

Table 5: Power and the associated conditional β-Size of robustness tests in the FE Model.

	Power of Robustness Test		Conditional β-Size		β-Size	β-Size
	S^*	$\tilde{S}$	S^*	$\tilde{S}$	Cluster-Robust	HS-Robust
$T = 10, G = 10$						
$\rho = .0$	.04	**.04**	.12	**.12**	.11	.07
$\rho = .3$	.02	**.41**	.29	**.19**	.10	.28
$\rho = .6$	.05	**.71**	.40	**.19**	.09	.41
$\rho = .9$	.14	**.75**	.45	**.20**	.12	.51
$T = 50, G = 10$						
$\rho = .0$	.05	**.05**	.11	**.11**	.08	.05
$\rho = .3$	.05	**.86**	.54	**.16**	.12	.57
$\rho = .6$	.04	**.90**	.66	**.16**	.09	.69
$\rho = .9$	.12	**.87**	.69	**.17**	.08	.76
$T = 10, G = 50$						
$\rho = .0$	.05	**.06**	.09	**.09**	.08	.06
$\rho = .3$	.65	**1.00**	.17	**.09**	.09	.30
$\rho = .6$	.91	**1.00**	.09	**.07**	.07	.39
$\rho = .9$	1.00	**1.00**	.07	**.07**	.07	.45

Table 6: Power and the associated conditional β-Size of robustness tests in the RE Model.

The first two columns in Tables 5 and 6 reveal that in both models the proposed alternative robustness testing strategy performs much better than the existing robustness testing strategy. When the number of clusters is small, $G = 10$, the existing approach has no power. In stark contrast, when $G = 10$, the power of the proposed approach is as high as 0.36 in the FE model and as high as 0.90 in the RE model. When the number of clusters is larger, $G = 50$, the both approaches have power. However, as was demonstrated in Section 2.2, robustness tests are only well-motivated when the number of clusters G is small. Hence, the selection between two robustness testing approaches should be mainly based on their small-sample performance.

The good performance of the proposed robustness testing strategy is also reflected in columns 3 and 4. For example, when $G = 10$ and $T = 50$ and researchers follow the three-step hypothesis testing strategy, the application of the proposed robustness testing strategy instead of the existing robustness testing can decrease the probability of false rejections of the null hypothesis H_0: $\beta = 0$ by 30% in the FE model (from 0.18 to 0.13) and by 75% (from 0.69 to 0.17) in the RE model.

4.2 Comparisons of Robustness Tests: Multiple Explanatory Variables

The multiple regressor models that we examine are modifications of the single regressor FE and RE models (6) and (7). Let $x_{it} = (x_{1,it}, x_{2,it}, ..., x_{K,it})$ denote the vector of K regressors, and let $\beta = (\beta_1, \beta_2, ..., \beta_K)$ denote the associated parameter vector. We introduce a new parameter ρ_x that governs correlation between regressors. Denoting the sign function by $\text{sgn}(\cdot)$, the modified FE model with K regressors plus a constant is

$$\begin{aligned}
y_{it} &= x'_{it}\beta + \alpha_i + \varepsilon_{it}, \qquad \alpha_i \sim N(0, .5) \\
x_{k,it} &= .5x_{k,it-1} + \text{sgn}\left[(\rho_x)^k\right] \times \sqrt{|\rho_x|} \times \omega_{it} + \sqrt{1 - |\rho_x|} \times v_{k,it}, \\
\omega_{it} &\sim N(0, .75) \qquad v_{k,it} \sim N(0, .75), \\
\varepsilon_{it} &= \rho\varepsilon_{it} + u_{it}\sqrt{(1-a) + a\frac{1}{K}\sum_{k=1}^{K} x_{k,t}^2}, \qquad u_{it} \sim N(0, 1-\rho),
\end{aligned} \tag{22}$$

and the modified RE model with K regressors plus a constant is

$$\begin{aligned}
y_{it} &= \beta_0 + x_{it}'\beta + \varepsilon_{it} \qquad (23)\\
x_{k,it} &= \operatorname{sgn}\left[(\rho_x)^k\right] \times \sqrt{|\rho_x|} \times (\eta_i + \omega_{it}) + \sqrt{1-|\rho_x|} \times (z_{k,i} + v_{k,it}), \qquad (24)\\
\eta_i &\sim N(0,.8), \qquad \omega_{it} \sim N(0,1-.8), \qquad z_{k,i} \sim N(0,.8), \qquad v_{k,it} \sim N(0,1-.8),\\
\varepsilon_{it} &= \alpha_i + u_{it}, \qquad \alpha_i \sim N(0,\rho), \qquad u_{it} \sim N(0,1-\rho).
\end{aligned}$$

If $\rho_x = 0$, the K regressors are uncorrelated in both models. If $\rho_x = 0$ and $K = 1$, models (22) and (24) correspond to the single regressor FE and RE models (6) and (7). If $\rho_x > 0$, all K regressors are positively correlated. If $\rho_x < 0$, each odd-numbered regressor is positively (negatively) correlated with each odd-numbered (even-numbered) regressor.

Monte Carlo results are reported in Tables 7 and 8. In these simulations we again set $\beta_0 = 0$ and $\beta_1 = 0$, and we set $\beta_i = 1$ for all $i > 1$. Moreover, we set $G = 10$, $T = 20$ and $a = 0.5$ in the FE model and $G = 10$ and $T = 10$ in the RE model. In both models we set $\rho = 0.6$. We vary the number of regressors K and the parameter ρ_x which governs correlation between regressors.[10] The proposed alternative robustness test statistic $\tilde{S}$ is constructed by first partialling out the effect all variables except the variable $x_{1,it}$ associated with the parameter of interest β_1.

Results in Tables 7 and 8 demonstrate that the superior performance of the proposed robustness testing strategy in comparison with the existing robustness testing approach extends to the case of multiple regressors. This occurs in spite of the fact that the proposed alternative robustness test statistic $\tilde{S}$ is constructed using the partialling out approach.

The only case in which the existing clustering test should be expected to outperform the proposed clustering test is the case in which within-cluster dependence has only a small impact on the variance of the estimator of the parameter of interest β_1 but a large impact on the variance of the estimators of

[10] The rank of the variance estimator $\hat{D}$ applied in the construction of the existing test statistic is limited by the number of clusters G. This limits the number of regressors K to those that satisfy $K(K+1)/2 < G$ (without constant as regressor) and $K(K+1)/2 - 1 < G$ (with constant as regressor).

other regression parameters β_0 and β_2 through β_K. However, whenever within-cluster dependence has only a small impact on the variance of the estimator of the parameter of interest β_1, the (incorrect) use of the less robust heteroskedasticity-robust estimator has only a small impact on the probability false rejections of the null hypothesis H_0: $\beta_1 = 0$. Thus, while the potential advantage from using the proposed alternative testing strategy instead of the existing approach is often quite large–as the Monte Carlo analyses in this section have shown–the potential advantage from using the existing robustness testing approach instead of the proposed approach appears relatively small.

	Power of Robustness Test		Conditional β-Size		β-Size	β-Size
	S^*	$\tilde{S}$	S^*	$\tilde{S}$	Cluster-Robust	HS-Robust
$K = 2$						
$\rho_x = -.9$	.01	**.22**	.17	**.13**	.06	.16
$\rho_x = -.5$	.02	**.19**	.15	**.14**	.06	.15
$\rho_x = 0$	.00	**.19**	.17	**.13**	.06	.17
$\rho_x = .5$	.01	**.18**	.11	**.14**	.05	.13
$\rho_x = .9$	.03	**.22**	.17	**.14**	.05	.17
$K = 3$						
$\rho_x = -.9$	.02	**.19**	.15	**.13**	.06	.15
$\rho_x = -.5$	.02	**.16**	.14	**.13**	.06	.14
$\rho_x = 0$	.02	**.20**	.18	**.14**	.07	.17
$\rho_x = .5$	.01	**.25**	.17	**.15**	.07	.16
$\rho_x = .9$	.02	**.20**	.12	**.11**	.04	.11

Table 7: Properties of robustness tests in the FE Model with multiple regressors.

	Power of Robustness Test		Conditional β-Size		β-Size	β-Size
	S^*	$\tilde{S}$	S^*	$\tilde{S}$	Cluster-Robust	HS-Robust
$K = 2$						
$\rho_x = -.9$	.02	**.65**	.41	**.19**	.11	.41
$\rho_x = -.5$	.02	**.64**	.45	**.23**	.13	.46
$\rho_x = 0$	.03	**.61**	.41	**.22**	.12	.42
$\rho_x = .5$	.02	**.64**	.42	**.20**	.12	.42
$\rho_x = .9$	.02	**.63**	.44	**.24**	.14	.45
$K = 3$						
$\rho_x = -.9$	.05	**.54**	.42	**.28**	.17	.44
$\rho_x = -.5$	.05	**.57**	.41	**.22**	.13	.42
$\rho_x = 0$	.04	**.58**	.44	**.24**	.14	.46
$\rho_x = .5$	.03	**.56**	.42	**.23**	.14	.42
$\rho_x = .9$	.04	**.56**	.45	**.28**	.16	.46

Table 8: Properties of robustness tests in the RE Model with multiple regressors.

5 Application to Testing for Heteroskedasticity

We now examine how the proposed robustness testing strategy performs when applied to testing for heteroskedasticity, and compare the results with the corresponding results for the White (1980) heteroskedasticity test.[11] We employ the cross-sectional version of the linear regression model (1). The N independent observations are indexed by i. We assume that the matrix $\Omega \equiv E\left(\varepsilon\varepsilon'|x\right)$ is a diagonal matrix so that the heteroskedasticity-robust covariance matrix estimator is a consistent estimator.

The White (1980) heteroskedasticity test is constructed from the contrast

$$\hat{W}_{HS} - \hat{W}_{LS}, \tag{25}$$

where $\hat{W}_{HS}$ and $\hat{W}_{LS}$, respectively, are the heteroskedasticity-robust estimator (4) and the "Least Squares" estimator (calculated under the assumption of homoskedasticity) of the covariance matrix $W \equiv E\left[x'\Omega x\right]$. With $K+1$ regressors the Least Squares covariance estimator $\hat{W}_{LS}$ is defined as

$$\hat{W}_{LS} = \hat{\sigma}_{\varepsilon}^{2} \frac{1}{N} \sum_{t=1}^{N} x_i' x_i, \tag{26}$$

where $\hat{\sigma}_{\varepsilon}^{2} \equiv \frac{1}{N-K-1} \sum_{i=1}^{N} \hat{\varepsilon}_i^2$. Thus, for the model with a single regressor x_i the contrast (25) is

$$\hat{W}_{HS} - \hat{W}_{LS} = \frac{1}{N} \sum_{t=1}^{N} \left(\hat{\varepsilon}_i^2 - \hat{\sigma}_{\varepsilon}^2\right) x_i^2. \tag{27}$$

The presence of the factor $\left(\hat{\varepsilon}_i^2 - \hat{\sigma}_{\varepsilon}^2\right)$ in this contrast (27) implies that the contrast is again–as in Section 3.2–constructed as the average of asymmetrically distributed random variables. In the analysis of the clustering test in Section 3.2, the asymmetric distribution of the variables that enter the average in the contrast was discovered as the cause for the poor performance of the existing

[11] Results for the bootstrapped version of the Wooldridge (1991) heteroskedasticity test, which is a homokurtosis-robust version of the second heteroskedasticity test offered in White (1980), were similar to the results for the bootstrapped version of the White (1980) heteroskedasticity reported in this section.

clustering test which is an application of the same robustness testing strategy. Accordingly, the same reasoning reveals why the White (1980) heteroskedasticity test has poor small-sample performance. The asymmetric distribution of the variables that enter the average in the contrast (27) implies that the contrast and the estimator of the variance of this contrast are correlated (see White (1980) for the estimator of the variance of the contrast). This in turn implies that the ratio of the contrast and the square root of the estimate of the variance of the contrast is asymmetrically distributed. In the single regressor model, the White (1980) heteroskedasticity test is constructed as the square of this ratio. The construction of the test thus converts two very different tails of a distribution into one tail which explains the poor small-sample performance of the test.

Application of the proposed alternative testing strategy for inference about hetereroskedasticity is straightforward. Impact of all other regressors except the variable associated with the parameter of interest is first partialled out. The contrast $\hat{W}_{HS} - \hat{W}_{LS}$ and the estimator of the variance of the contrast are then constructed for the parameter of interest and their ratio is used as the test statistic. This test statistic is analogous to the proposed test statistic in expression (21) constructed for the case of testing for clustering. The only difference is that $\hat{W}_{CLUSTER} - \hat{W}_{HS}$ is replaced with $\hat{W}_{HS} - \hat{W}_{LS}$.

We now compare the performance of these heteroskedasticity tests in the regression model

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \qquad \varepsilon_t = u_t\sqrt{(1-a) + a x_t^2},\ x_t \sim N(0,1) \text{ and } \varepsilon_t \sim N(0,1) \tag{28}$$

with only one regressor plus a constant. The parameter a governs the extent of heteroskedasticity. When $a = 0$ (when $a > 0$) the error terms are homoskedastic (heteroskedastic). Again, we set $\beta_1 = 0$ for the parameter of interest β_1, and we set $\beta_0 = 0$. We vary the parameter a and the number of observations N. Monte Carlo results are reported in Table 9.[12] Results for the proposed

[12] In constructing both robustness test statistics, we first de-mean the data. Omission of the constant in the construction of the White (1980) heteroskedasticity test improves the performance of the test. The results reported in each cell of Tables 9-10 are calculated using 400 simulated samples. The rejection frequencies for the test statistics S^* and $\tilde{S}$ are constructed from the bootstrapped distribution of each test statistic. The bootstrapped distribution of the test statistic is constructed using *399* bootstrapped samples from the original sample. We use the standard un-

testing strategy are indicated in bold in columns 2 and 4. Results for the bootstrapped version of the White (1980) heteroskedasticity test are reported in columns 1 and 3.

	Power of Robustness Test		Conditional β-Size		β-Size	β-Size
	S^*	$\tilde{S}$	S^*	$\tilde{S}$	HS-Robust	OLS
$N = 20$						
$a = 0$	.03	**.06**	.09	**.09**	.09	.06
$a = .5$	.06	**.17**	.14	**.13**	.11	.13
$a = 1$	.21	**.40**	.25	**.23**	.15	.28
$N = 50$						
$a = 0$	.06	**.06**	.05	**.04**	.04	.03
$a = .5$	.13	**.41**	.15	**.11**	.07	.16
$a = 1$	.51	**.82**	.18	**.13**	.08	.24
$N = 100$						
$a = 0$	.05	**.04**	.08	**.08**	.08	.07
$a = .5$	.37	**.76**	.11	**.07**	.05	.15
$a = 1$	.70	**.93**	.12	**.08**	.06	.23

Table 9: Properties of heteroskedasticity tests with a single regressor.

Results in Table 9 show that when the number of observations is small, $N = 20$, the proposed testing strategy performs better than the White (1980) heteroskedasticity test but neither approach performs well. When the number of observations is larger, $N = 50$ or $N = 100$, the proposed testing approach outperforms the White (1990) robustness testing approach in terms of power as well as the impact that the robustness tests have on the quality of inference about the regression parameter of interest β_1 as measured by the reported conditional β-size for these robustness tests.

We next examine the performance of these robustness tests in the multiple regression model

$$\begin{aligned} y_t &= \beta_0 + x_t'\beta + \varepsilon_t, \qquad \varepsilon_t = u_t\sqrt{(1-a) + a\frac{1}{K}\textstyle\sum_{k=1}^{K} x_{k,t}^2},\ u_t \sim N(0,1), \\ x_{k,t} &= \text{sign}\left[(\rho_x)^k\right] \times \sqrt{|\rho_x|} \times \omega_t + \sqrt{1-|\rho_x|} \times v_{k,t}, \omega_t \sim N(0,1), v_{k,t} \sim N(0,1) \end{aligned}$$

with $K > 1$ regressors plus a constant. The parameter ρ_x reflects the dependence between the K regressors. If $\rho_x = 0$, the regressors are uncorrelated. If $\rho_x > 0$, the regressors are positively

weighted non-parametric bootstrap (the residuals are randomly sampled with replacement (see Hodoshima and Ando (2008) for an application of this approach to heteroskedasticity tests). This approach imposes the null hypothesis that the errors are homoskedastic.

correlated. If $\rho_x < 0$, odd-numbered regressors are negatively (positively) correlated with even-numbered (other odd-numbered) regressors. We set $\beta_0 = 0$, $\beta_1 = 0$, and $\beta_i = 1$ for all $i > 1$, and we set $a = 0.5$, and $N = 50$. We vary the parameter ρ_x and the number of regressors K.

Monte Carlo results are reported in Table 10. A comparison of the results for the case of uncorrelated regressors ($\rho_x = 0$) with the results in Table 9 for the single regressor model shows that power of both robustness tests decreases as the number of regressors increases, and the proposed robustness testing strategy continues to maintain its advantage over the existing approach. However, additional simulations not reported here in detail show that when $K = 3$, $\rho_x = 0$ and $a = 1$ so that the extent of heteroskedasticity is more severe than in the case for which results are shown in Table 10, the existing robustness testing approach performs better than the proposed robustness testing approach. Results in Table 10 also indicate that in this regression model both robustness tests have poor power when the regressors are correlated ($\rho_x \neq 0$).

	Power of Robustness Test		Conditional β-Size		β-Size	β-Size
	S^*	$\tilde{S}$	S^*	$\tilde{S}$	HS-Robust	OLS
$k = 2$						
$\rho_x = -.9$	.10	**.08**	.11	**.11**	.09	.09
$\rho_x = -.5$	.06	**.14**	.11	**.10**	.09	.11
$\rho_x = 0$	.08	**.18**	.08	**.08**	.06	.07
$\rho_x = .5$	.07	**.12**	.08	**.08**	.06	.07
$\rho_x = .9$	.08	**.05**	.07	**.07**	.06	.06
$k = 3$						
$\rho_x = -.9$	.07	**.05**	.09	**.09**	.08	.07
$\rho_x = -.5$	.07	**.05**	.10	**.10**	.07	.09
$\rho_x = 0$	.05	**.08**	.11	**.10**	.08	.10
$\rho_x = .5$	.07	**.07**	.11	**.10**	.08	.09
$\rho_x = .9$	.07	**.05**	.07	**.07**	.05	.06

Table 10: Power of heteroskedasticity tests with multiple regressors.

We reiterate (from the introduction) that unlike clustering tests, heteroskedasticity tests may not be well-motivated because the use of the less robust Least Squares covariance matrix $\hat{W}_{LS}$ estimator in favor of the more robust heteroskedasticity-robust estimator $\hat{W}_{HS}$ may not be justified. Results in this section therefore merely serve as a demonstration that in principle our insight into why the existing clustering test performs poorly allso explains why other applications of the White (1980) robustness testing approach perform poorly and that, accordingly, application of the proposed alternative robustness testing strategy can improve also inference about parametric assumptions in covariance matrix estimation.

6 Conclusion

In this paper we have examined “robustness tests” that help researchers select between different covariance matrix estimators in the ever-expanding set of available covariance matrix estimators. Our main focus has been on clustering tests that examine the choice between the cluster-robust covariance matrix estimator and a less robust covariance matrix estimator.

We have proposed a new robustness testing strategy, which implementation is straightforward. We have also shown why the existing clustering test and other applications of the White (1980) robustness testing strategy perform poorly in small samples. Moreover, we have shown that when applied to inference about clustering the proposed robustness testing strategy performs well in small samples. As we have argued in this paper, the small-sample performance is especially important in the context of robustness tests: these tests are well-motivated only when the less robust estimator has a potential advantage over the more robust estimator for which a small number of observations is typically a necessary condition. For the clustering tests examined here–in which the alternative covariance matrix estimator is the heteroskedasticity-robust estimator–this principle implies that the tests are only well-motivated when the number of clusters is small.

An important topic for future research is the performance of the proposed robustness testing strategy in relation to more complex covariance matrix estimators such as the multi-way cluster-robust estimator. Another worthy topic for future research is whether and to what extent the construction of a multi-dimensional bootstrapped rejection region or the application of the Bonferroni method can improve the performance of the proposed robustness testing strategy in models with multiple explanatory variables relative to the dimension reduction approach applied here.

References

Altonji, J. G. and L. M. Segal, 1996, "Small-Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business & Economic Statistics*, vol. 14, pp. 353-66.

Baltagi, B. H., Jung, B. C. and S. H. Song, 2010, "Testing for Heteroskedasticity and Serial Correlation in a Random Effects Panel Data Model," *Journal of Econometrics*, vol. 154, pp. 122-4.

Bell, R. M. and D. F. McCaffrey, 2002, "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples," *Survey Methodology*, vol. 28, pp. 169-81.

Cameron, A. C., Gelbach, J. B. and D. L. Miller, 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, vol. 90, 414-27.

Hansen, C. B., 2007, "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T is Large," *Journal of Econometrics*, vol. 141, pp. 597-620.

Hodoshima, J. and M. Ando, 2008, "The Finite-Sample Performance of White's Test for Heteroskedasticity Under Stochastic Regressors," *Communications in Statistics–Simulation and Computation*, vol. 36, pp. 1201-15.

Kezdi, G., 2003, "Robust Standard Error Estimation In Fixed-Effects Panel Models," Mimeo

MacKinnon, J. G, and H. White, 1985, "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite-Sample Properties," *Journal of Econometrics*, vol. 29, pp. 305-25.

Monte-Rojas, G. and W. Sosa-Escudero, 2010, "Robust Tests for Heteroskedasticity in the One-Way Error Components Model," *Journal of Econometrics*, forthcoming.

Stock, J. H. and M. W. Watson, 2008, "Heteroskedasticity–Robust Standard Errors for Fixed Effects Panel Data Regression," *Econometrica*, vol. 76. pp. 155-74.

White, H., 1980, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, vol. 48, pp. 817-38.

Wooldridge, J. M., 1991, "On the Application of Robust, Regression-based Diagnostics to Models of Conditional Means and Conditional Variances." *Journal of Econometrics*, vol. 47, pp. 5-46.

APPENDIX 1: Proofs

Appendix 1.1. Sign of Terms in the Sum of Cross-Products

In this appendix we prove the result (17). Let P denote the number positive factors among the T factors $\{x_{i1}\hat{\varepsilon}_{i1}, x_{i1}\hat{\varepsilon}_{i1}, ..., x_{iT}\hat{\varepsilon}_{iT}\}$. The number of negative cross-products among the $T(T-1)/2$ cross-products in the sum (16) is $P(T-1-(P-1))/2 + (T-1-(P-1))P/2$, which can be rewritten as $-PP + PT$. Hence, the share of negative terms in the sum of cross-products (16) is greater than c if $(PP - PT)/(T(T-1)/2) > c$, which can be rewritten as

$$-\left(\frac{P}{T}\right)^2 + \frac{P}{T} - \frac{c(T-1)}{2T} > 0 \tag{29}$$

Solving for the roots of this quadratic equation yields

$$\left(\frac{P}{T}\right)^* = \frac{1 \pm \sqrt{1 - \frac{2c(T-1)}{T}}}{2}, \tag{30}$$

and substituting $c = \frac{1}{2}$ yields

$$\left(\frac{P}{T}\right)^* = \frac{1 \pm \sqrt{\frac{1}{T}}}{2}. \tag{31}$$

Hence, the sum of cross products (16) has more negative terms than positive terms if the share $\frac{P}{T}$ of positive factors among the T factors $\{x_{i1}\hat{\varepsilon}_{i1}, x_{i1}\hat{\varepsilon}_{i1}, ..., x_{iT}\hat{\varepsilon}_{iT}\}$ is in the interval $\left(\frac{1}{2} - \frac{1}{2}\sqrt{\frac{1}{T}}, \frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{T}}\right)$. By definition $P = \sum_t I_{x_{it}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}x_{is}>0}$, where $I_{x_{it}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}x_{is}>0}$ is the indicator function in the text. Provided that positive and negative values are equally likely for the error terms ε_{it}, so that $P(\varepsilon_{it} > 0) = P(\varepsilon_{it} < 0) = 0.5$ (modification of the proof to the $P(\varepsilon_{it} > 0) = P(\varepsilon_{it} < 0) < 0.5$ is straightforward and omitted), the variable I_{it} has the Bernoulli distribution with parameter p arbitrarily close to $\frac{1}{2}$.[13] By the central limit theorem the asymptotic distribution of

$$\frac{\sqrt{T}\frac{\sum_t I_{x_{it}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}x_{is}>0} - \mu_I}{T}}{\sigma_I}, \tag{32}$$

where $\mu_I = \frac{1}{2}$ and $\sigma_I = \frac{1}{2}$, is the standard normal distribution. Thus, in the limit (as $G, T \to \infty$), the probability that the share of positive factors, $\frac{P}{T}$, is within the distance $\frac{1}{2}\sqrt{\frac{1}{T}}$ from $\frac{1}{2}$ is

[13] As the number of clusters $G \to \infty$, the least squares estimate $\hat{\beta}$ and the associated least squares residuals $\hat{\varepsilon}_{it}$ approach β and ε_{it}, respectively. Thus, positive and negative values are also equally likely for the least squares residual $\hat{\varepsilon}_{it}$ so that $P(\hat{\varepsilon}_{it} > 0)$ and $P(\hat{\varepsilon}_{it} < 0)$ are arbitrarily close to 0.5 for large enough G.

$2 \times (\Phi(1) - 0.5) \approx 0.6823$. Consequently, in the limit (as $G, T \to \infty$), also the probability that the sum of cross-products (16) has more negative terms than positive terms is approximately 0.6823.

Appendix 1.2: $\hat{W}_{CLUSTER} - \hat{W}_{HS}$ and $\widehat{var}\left(\hat{W}_{CLUSTER} - \hat{W}_{HS}\right)$ are Correlated

This appendix shows that $Cov(\hat{W}_{CLUSTER} - \hat{W}_{HS}, \hat{Var}(\hat{W}_{CLUSTER} - \hat{W}_{HS}))$ is positive provided that the distribution of $\sum_{t=1}^{T-1} \sum_{s=t+1}^{T} x_{it}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}x_{is}$ is skewed to the right.

Using expression (14) the contrast can be written as

$$\hat{W}_{CLUSTER} - \hat{W}_{HS} = \frac{1}{G}\sum_{i=1}^{G}\frac{2}{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} x_{it}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}x_{is}. \tag{33}$$

Applying the formula (10) a consistent estimate of the variance of the contrast $(\hat{W}_{CLUSTER} - \hat{W}_{HS})$ is given by

$$\hat{Var}(\hat{W}_{CLUSTER} - \hat{W}_{HS}) = \frac{1}{G^2}\sum_{i=1}^{G}\frac{2}{[T(T-1)/2]^2}\left[\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} x_{it}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}x_{is}\right]^2, \tag{34}$$

which can be rewritten simply as

$$\hat{Var}(\hat{W}_{CLUSTER} - \hat{W}_{HS}) = \frac{8}{[GT(T-1)]^2}\sum_{i=1}^{G}\left[\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} x_{it}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}x_{is}\right]^2. \tag{35}$$

To prove the claim, we introduce several convenient notations. Let $d = \hat{W}_{CLUSTER} - \hat{W}_{HS}$. Let $\omega = \hat{Var}(d)$. Then we want to show that $Cov(d, \omega)$ depends on the skewness of $\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} x_{it}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}x_{is}$. which we denote as D_i. Next we write d and ω as $d = \frac{2}{GT}\sum_{i=1}^{G} D_i$ and $\omega = \frac{8}{[GT(T-1)]^2}\sum_{i=1}^{G} D_i^2$ respectively.

Now, from the covariance formula, we have that

$$Cov(d, \omega) = E(d\omega) - E(d)E(\omega), \tag{36}$$

where

$$E(d\omega) = E\left[\left(\frac{1}{G}\frac{2}{T}\sum_{i=1}^{G} D_i\right)\left(\frac{1}{G^2}\frac{8}{T^2(T-1)^2}\sum_{i=1}^{G} D_i^2\right)\right] \tag{37}$$

and

$$E(d)E(\omega) = E\left[\frac{1}{G}\frac{2}{T}\sum_{i=1}^{G} D_i\right]E\left[\frac{1}{G^2}\frac{8}{T^2(T-1)^2}\sum_{i=1}^{G} D_i^2\right]. \tag{38}$$

But we note that

$$E\left[\sum_{i=1}^{G} D_i \sum_{i=1}^{G} D_i^2\right] = E\left[\sum_{i=1}^{G} D_i^3 + \sum_{i=1}^{G} \sum_{j=1, j\neq i}^{G} D_i^2 D_j\right]. \tag{39}$$

Next let us denote $m_1 = \frac{1}{T}\sum_{i=1}^{G} D_i (\equiv D)$, $m_2 = \frac{1}{[T(T-1)]^2}\sum_{i=1}^{G} D_i^2$, and $m_3 = \frac{1}{T^3(T-1)^2}\sum_{i=1}^{G} D_i^3$, where $E(m_j) = \mu_j$ for $j = 1, 2, 3$. Then we have

$$Cov(d, \omega) = \frac{16}{G^3}\left[E(m_3) + E(m_1 m_2) - E(m_1)E(m_2)\right], \tag{40}$$

which can be rewritten as

$$Cov(d, \omega) = \frac{16}{G^3}[\mu_3 + E(m_1 m_2) - \mu_1 \mu_2]. \tag{41}$$

A little bit of algebras then show that

$$Cov(d, \omega) = \frac{16}{G^3}[\mu_3 + 2\mu_1(\mu_2 - \mu_1) - \mu_1 \mu_3]. \tag{42}$$

In the remaining steps, we work with the right–hand side bracketed expression of the above equation

$$\begin{aligned} \mu_3 + 3\mu_2\mu_1 + 2\mu_1^3 &= \mu_3 + 3\mu_2\mu_1 + 2\mu_1\mu_1^2 - \mu_1^2\mu_1 + \mu_1^3 \\ &= E(m_3) + 3E(m_2\mu_1) + 2E(m_1)\mu_1^2 - \mu_1^2 E(m_1) + \mu_1^3 \\ &= E(m_3 - 2m_2\mu_1 - m_2\mu_1 + 2m_1\mu_1^2) - E(\mu_1^2 m_1) + E(\mu_1^3) \\ &= E(m_3 - 2m_1\mu_1)(m_1 - \mu_1) - E(\mu_1^2(m_1 - \mu_1)) \\ &= E[(m_2 - 2m_1\mu_1 + \mu_1^2)(m_1 - \mu_1)] \\ &= E[(m_1 - \mu_1)^2(m_1 - \mu_1)] \\ &= E[(m_1 - \mu_1)^3. \end{aligned} \tag{43}$$

Thus,

$$Cov(d, \omega) = \frac{16}{G^3} E\left(D - \mu_1\right)^3. \tag{44}$$

In other words, $Cov(d, \omega)$ is positive when the distribution of $D_i \equiv \sum_{t=1}^{T-1}\sum_{s=t+1}^{T} x_{it}\hat{\varepsilon}_{it}\hat{\varepsilon}_{is}x_{is}$ is skewed to the right.