$t-$statistic based correlation and heterogeneity robust inference*

Rustam Ibragimov
Harvard University
Economics Department
1875 Cambridge Street
Cambridge, MA 02138

Ulrich K. Müller
Princeton University
Economics Department
Fisher Hall
Princeton, NJ 08544

Abstract

We develop a general approach to robust inference about a scalar parameter when the data is potentially heterogeneous and correlated in a largely unknown way. The key ingredient is the following result of Bakirov and Székely (2005) concerning the small sample properties of the standard $t-$test: For a significance level of 5% or lower, the $t-$test remains conservative for underlying observations that are independent and Gaussian with heterogenous variances. One might thus conduct robust large sample inference as follows: partition the data into $q \geq 2$ groups, estimate the model for each group and conduct a standard $t-$test with the resulting q parameter estimators. This results in valid and in some sense efficient inference when the groups are chosen in a way that ensures the parameter estimators to be asymptotically independent, unbiased and Gaussian of possibly different variances. We provide examples of how to apply this approach to time series, panel, clustered and spatially correlated data.

JEL classification: C32

Keywords: $t-$test, dependence, least favorable distribution, variance estimation, Fama-MacBeth method

*The authors would like to thank seminar and conference participants at Berkeley, Brown, Chicago GSB, Columbia, Harvard/MIT, Ohio State, Penn State, Princeton, Rice, Rochester, Stanford, Vanderbilt, the CIREQ Time Series Conference 2006, the European Econometric Society Meeting 2007, the Greater New York Metropolitan Area Econometrics Colloquium 2006 and the NBER Summer Institute 2006 for valuable comments and discussions. We also thank T. McElroy and D. Politis for sharing their Monte Carlo simulation codes.

1 Introduction

Empirical analyses in economics often face the difficulty that the data is correlated and heterogeneous in some unknown fashion. Many estimators of parameters of interest remain valid and interesting even under the presence of correlation and heterogeneity, but it becomes considerably more challenging to correctly estimate their sampling variability.

The typical approach is to invoke a law of large numbers to justify inference based on consistent variance estimators: For an OLS regression with independent but not identically distributed disturbances, see White (1980). In the context of time series, popular heteroskedasticity and autocorrelation consistent ("long-run") variance estimators were derived by Newey and West (1987) and Andrews (1991). For clustered data, which includes panel data as a special case, Rogers' (1993) clustered standard errors provide a consistent variance estimator. Conley (1999) derives consistent non-parametric standard errors for data sets that exhibit spatial correlations. Other important references for related approaches include Liang and Zeger (1986) and Arellano (1987); also see Wooldridge (2002) for a textbook treatment. While quite general, the consistency of the variance estimator is obtained through an assumption that asymptotically, an infinite number of observable entities are essentially uncorrelated: heteroskedasticity robust estimators achieve consistency by averaging over an infinite number of uncorrelated disturbances; clustered standard errors achieve consistency by averaging over an infinite number of uncorrelated clusters; long-run variance estimators achieve consistency by averaging over an infinite number of (essentially uncorrelated) low frequency periodogram ordinates; and so forth. Inference based on such nonparametric consistent variance estimators is therefore inapplicable or yields poor results when correlations are pervasive and pronounced enough.[1]

More recently, a number of inference procedures have been developed that do not rely on consistency of the variance estimator. In a time series context, Kiefer, Vogelsang, and Bunzel (2000) show that it is possible to conduct asymptotically justified inference in a linear time series regression based on long-run variance estimators with a nondegenerate limiting distribution. These results were extended and scrutinized by Kiefer and Vogelsang

[1] Also block bootstrap and subsampling techniques derive their asymptotic validity from averaging over an infinite number of essentially uncorrelated blocks.

(2002, 2005), Jansson (2004) and Sun, Phillips, and Jin (2006). Müller (2007b) shows that all consistent long-run variance estimators lack robustness in a certain sense, and determines a class of inconsistent long run variance estimators with some optimal trade off between robustness and efficiency. Donald and Lang (2004) point out that linear regression inference in a setting with clusters may be based on Student$-t$ distributions with a finite number of degrees of freedom under an assumption that both the random effects and cluster averages of the individual disturbances are approximately i.i.d. Gaussian across clusters. Hansen (2007) finds that the asymptotic null distribution of test statistics based on the standard clustered error formula for a panel with one fixed dimension and one dimension tending to infinity become that of a Student$-t$ with a finite number of degrees of freedom (suitably scaled), as long as the fixed dimension is "asymptotically homogeneous".

This paper develops a general strategy for conducting inference about a scalar parameter with potentially heterogenous and correlated data, when relatively little is known about the precise property of the correlations. The key ingredient to the strategy is a result by Bakirov and Székely (2005) concerning the small sample properties of the usual $t-$test used for inference on the mean of independent normal variables: For a significance levels of 8.3 percent or lower, the usual $t-$test remains conservative[2] when the variances of the underlying independent Gaussian observations are not identical. This insight allows the construction of asymptotically valid test statistics for general correlated and heterogenous data in the following way: Assume that the data can be classified in a finite number q of groups that allow asymptotically independent normal inference about the scalar parameter of interest β. This means that the parameter estimator $\hat{\beta}_j$ from each group j is approximately $\hat{\beta}_j \sim \mathcal{N}(\beta, v_j^2)$, and $\hat{\beta}_j$ is approximately independent of $\hat{\beta}_i$ for $j \neq i$. The observations $\hat{\beta}_1, \cdots, \hat{\beta}_q$ can thus be treated as independent normal observations with common mean β (but not necessarily equal variance), and the usual $t-$test concerning β constructed from $\hat{\beta}_1, \cdots, \hat{\beta}_q$ (with $q-1$ degrees of freedom) is conservative. If the number

[2] When the variances are viewed as unknown nuisance parameters, the small sample $t-$test is, of course, overall exact for underlying independent Gaussian variates. We nevertheless refer to the property of a smaller than nominal rejection probability under the null hypothesis for unequal variances as 'conservativeness' in the remainder.

of observations is reasonably large in all groups, the approximate normality $\hat{\beta}_j \sim \mathcal{N}(\beta, v_j^2)$ is of course a standard result for most models and estimators, linear or nonlinear.

The knowledge about the correlation structure in the data is embodied in the assumption that $\hat{\beta}_1, \cdots, \hat{\beta}_q$ are (approximately) independent. Compared to consistent variance estimators, this approach reduces the requirement of uncorrelatedness in the data to a finite amount. What is more, by invoking the results of Müller (2007a), we show that the $t-$statistic approach efficiently exploits this assumption in some sense. Of course, a stronger (correct) assumption, i.e. larger q, will typically lead to more powerful inference, so that one faces the usual trade-off between robustness and power in choosing the number of groups. In the benchmark case of underlying i.i.d. observations, a 5% level $t-$statistic based test with $q = 16$ equal sized groups loses at most a 5.8 percentage points of asymptotic local power compared to inference with known (or correctly consistently estimated) asymptotic variance in an exactly identified GMM problem. The robustness vs. power trade-off is thus especially acute only for data sets where an even coarser partition is required to yield independent information about the parameter of interest.

This $t-$statistic approach to inference in the presence of pervasive correlations advances the literature on inference based on inconsistent variance estimators in several ways: First, it is generic strategy that can be employed in different contexts, such as in time series data, panel data or spatially correlated data. Second, due to the small sample conservativeness result, the approach allows for unknown and unmodelled heterogeneity. In a time series context, for instance, this means that unlike Kiefer and Vogelsang (2005), we can allow for low frequency variability in the second moment of the moment condition, and in a panel context, we do not require the asymptotic homogeneity as in Hansen (2007). Third, the $t-$statistic approach is very easy to implement, and does not require any new tables of critical values. Fourth, the crucial regularity condition—the assumption that $\hat{\beta}_1, \cdots, \hat{\beta}_q$ are approximately independent and distributed $\mathcal{N}(\beta, v_j^2)$—is more explicit and maybe easier to interpret than, say, the primitive conditions underlying consistent long-run variance estimators, or the value of the bandwidth as a fraction of the sample size in Kiefer and Vogelsang (2005). Finally, and perhaps most importantly from an econometric theory perspective, the $t-$statistic approach in some sense efficiently exploits the information contained in this regularity condition; to the best of our knowledge, this is the first general large sample efficiency claim about the test of a parameter value that does not involve consistent estimation of the asymptotic variance.

The approximate normality of $\hat{\beta}_j$ of group j data stems from an appeal to a central limit theorem. If within group correlations become very strong, this argument might become less plausible, as $\hat{\beta}_j$ can no longer be thought of as an average of many approximately independent quantities.[3] As long as the independence of the $\hat{\beta}_j$'s remains a compelling assumption, one could resort to non-parametric location tests such as sign and rank tests and perform asymptotically valid inference in that way. We do not treat such an extension in this paper.

The $t-$statistic approach to inference has an important precursor in the work of Fama and MacBeth (1973). Their work on empirical tests of the CAPM has motivated the following widespread approach to inference in panel regressions with firms or stocks as individuals: Estimate the regression separately for each year, and then test hypotheses about the coefficient of interest by the $t-$statistic of the resulting yearly coefficient estimates. The Fama-MacBeth approach is thus a special case of the method described above, where observations of the same year are collected in a group. While this approach is routinely applied, we are not aware of a formal justification. One contribution of this paper is to provide such a justification, and we find that as long as year coefficient estimators are approximately normal (or scale mixtures of normals) and independent, the Fama-MacBeth method results in valid inference even for a short panel that is heterogenous over time.

The rest of the paper is organized as follows: Section 2 discusses properties of the small sample $t-$statistic for independent Gaussian observations of potentially heterogeneous

[3] In such circumstances, of course, also the usual way of conducting inference based on consistent variance estimator fails, as the full sample estimator $\hat{\beta}$ is typically a weighted average of the $\hat{\beta}_j$'s, $j = 1, \cdots, q$, and thus not normally distributed. Under the assumption of independent $\hat{\beta}_j$'s, one might argue that the approximate normality of $\hat{\beta}$ remains a more convincing approximation; but the $t-$statistic also remains approximately distributed Student$-t$ even when the underlying observations are not exactly Gaussian. In fact, as a further consequence of the small sample conservativeness result, the usual $t-$test remains exactly valid as long as the $\hat{\beta}_j$'s are independent and the distribution of each $\hat{\beta}_j$ can be written as a scale mixture of normals, which is the case for a rather large class of non-normal distributions. The $t-$statistic approach is thus formally more robust against non-normal $\hat{\beta}_j$'s than standard inference based on a consistent variance estimator.

variance. Section 3 lays out in detail how this result can be exploited to obtain robust large sample inference, formalizes the large sample efficiency claim and derives some other theoretical properties of this approach. In Section 4, we discuss applications to time series data, panel data, clustered data and spatially correlated data, and provide some Monte Carlo evidence. Section 5 concludes.

2 The Small Sample $t-$test

Let $X_j, j = 1, \cdots, q$, with $q \geq 2$, be independent Gaussian random variables with common mean $E[X_j] = \mu$ and variances $V[X_j] = \sigma_j^2$. The usual $t-$statistic for the hypothesis test

$$H_0 : \mu = 0 \quad \text{against } H_1 : \mu \neq 0 \tag{1}$$

is given by

$$t = \sqrt{q}\frac{\bar{X}}{s_X} \tag{2}$$

where $\bar{X} = q^{-1}\sum_{j=1}^{q} X_j$ and $s_X^2 = (q-1)^{-1}\sum_{j=1}^{q}(X_j - \bar{X})^2$, and the null hypothesis is rejected for large values of $|t|$.[4] Note that $|t|$ is a scale invariant statistic, that is a replacement of $\{X_j\}_{j=1}^q$ by $\{cX_j\}_{j=1}^q$ for any $c \neq 0$ leaves $|t|$ unchanged. If $\sigma_j^2 = \sigma^2$ for all j, by definition, the critical value cv of $|t|$ is given by the appropriate percentile of the distribution of a Student$-t$ distributed random variable T_{q-1} with $q-1$ degrees of freedom.

In a recent paper, Bakirov and Székely (2005) show that for a given critical value, the rejection probability under the null hypothesis of a test based on $|t|$ is maximized when $\sigma_1^2 = \cdots = \sigma_k^2$ and $\sigma_{k+1}^2 = \cdots = \sigma_q^2 = 0$ for some $1 \leq k \leq q$. Their results imply the

[4] To be precise, we define the $t-$statistic (2) to be equal to zero if $X_i = X_j$ for all i, j, which happens with probability one when $\sigma_j^2 = 0$ for all j.

following Theorem.[5,6]

Theorem 1 (Bakirov and Székely, 2005) *Let* $\mathrm{cv}_q(\alpha)$ *be the critical value of the usual two-sided* $t-$*test based on (2) of level* α, *i.e.* $P(|T_{q-1}| > \mathrm{cv}_q(\alpha)) = \alpha$, *and let* Φ *denote the cumulative density function of a standard normal random variable.*

(i) If $\alpha \leq 2\Phi(-\sqrt{3}) = 0.08326...$, *then for all* $q \geq 2$,

$$\sup_{\{\sigma_1^2,\cdots,\sigma_q^2\}} P(|t| > \mathrm{cv}_q(\alpha)|H_0) = P(|T_{q-1}| > \mathrm{cv}_q(\alpha)) = \alpha. \tag{3}$$

(ii) Equation (3) also holds true for $2 \leq q \leq 14$ *if* $\alpha \leq \alpha_1 = 0.1$, *and for* $q \in \{2,3\}$ *if* $\alpha \leq \alpha_2 = 0.2$. *Moreover, define* $\widetilde{\mathrm{cv}}_q(\alpha_i) = \sqrt{k_i(q-1)\,\mathrm{cv}_{k_i}(\alpha_i)^2}/\sqrt{q(k_i-1)+(q-k_i)\,\mathrm{cv}_{k_i}(\alpha_i)^2}$, $i \in \{1,2\}$, *where* $k_1 = 14$ *and* $k_2 = 3$. *Then for* $q \geq k_i + 1$, $\sup_{\{\sigma_1^2,\cdots,\sigma_q^2\}} P(|t| > \widetilde{\mathrm{cv}}_q(\alpha_i)|H_0) = \alpha_i$.

The usual 5% level two sided tests of (1) based on the usual $t-$test thus remains valid for all values of $\{\sigma_1^2, \cdots, \sigma_q^2\}$, and all $q \geq 2$. Also, by symmetry of the $t-$statistic under the null hypothesis, Theorem 1 (ii) implies conservativeness of the usual one-sided $t-$test of significance level 5% or lower as long as $q \leq 14$. For $q \geq 15$, however, the rejection probability of a 10% level two-sided test (or a 5% level one-sided test) under the null hypothesis is maximized at $\sigma_1^2 = \cdots = \sigma_{14}^2$ and $\sigma_{15}^2 = \cdots = \sigma_q^2 = 0$. So usual two-sided $t-$tests of level 10% are not automatically conservative for large q, and the appropriate critical value of a robust test is a function of the critical value of the usual $t-$test when $q = 14$. In the following, our focus is on the empirically most relevant case of two-sided tests of level 5% or lower.

[5] Part (ii) of Corollary 1 in Bakirov and Székely (2005) implies that (3) holds for all $\alpha \leq P(|T_{q-1}| > \sqrt{\frac{(q-1)r(q-1)}{q-r(q-1)}}) = \alpha_0(q)$, where, for $k \geq 2$, $r(k)$ is the unique point in the interval $(1,k)$ such that the difference $\Delta_k(R) = P(|T_k| > \sqrt{\frac{Rk}{k+1-R}}) - P(|T_{k-1}| > \sqrt{\frac{R(k-1)}{k-R}})$ is negative for $0 < R < r(k)$ and is positive for $r(k) < R < k$ (existence and uniqueness of $r(k)$ is established in part (i) of Proposition in Bakirov and Székely (2005)). Since $r(q-1) < r(q)$ by part (ii) of their Proposition, one has $\Delta_{q-1}(r(q-1)) < 0$ and, thus, $\alpha_0(q) > P(|T_q| > \sqrt{\frac{qr(q-1)}{q+1-r(q-1)}}) > P(|T_q| > \sqrt{\frac{qr(q)}{q+1-r(q)}}) = \alpha_0(q+1)$. Iterating on the above arguments, we thus have that, for $k \geq 1$, $\alpha_0(q) > \alpha_0(q+k)$ and, since $r(q+k) \to 3$ by part (iii) of their Proposition, we have $\alpha_0(q) \geq 2\Phi(-\sqrt{3})$.

[6] Before becoming aware of the paper by Bakirov and Székely (2005), we have proven (3) for $\alpha \leq 0.05$ and $q \geq 2$ by refining earlier results by Bakirov (1989). Our proof differs from the approach in Bakirov and Székely (2005), and is available on our websites.

One immediate application of Theorem 1 concerns the construction of confidence intervals for μ: a confidence interval for μ of level $C \geq 95\%$ based on the usual formulas for i.i.d. Gaussian observations has effective coverage level of at least C for all values of $\{\sigma_1^2, \cdots, \sigma_q^2\}$. As long as the realized value of $|t|$ is larger than the smallest $\text{cv}_q(\alpha)$ for which (3) holds, also $p-$values constructed from the cumulative distribution function of the Student$-t$ distribution maintain their usual interpretation as the lowest significance level at which the test still rejects.[7] As stressed by Bakirov and Székely (2005), a further implication of Theorem 1 is the conservativeness of the usual $t-$test against i.i.d. observations that are scale mixtures of Gaussian variates: Let $Y_j = Z_j V_j$ where $Z_j \sim i.i.d.\ \mathcal{N}(\mu, 1)$ and V_j is i.i.d. and independent of $\{Z_j\}_{j=1}^q$. Then by Theorem 1, the usual $t-$test based on $\{Y_j\}_{j=1}^q$ of the null hypothesis (1) of level 5% or lower is conservative conditional on $\{V_j\}_{j=1}^q$, and hence also unconditionally. The usual $t-$test of level 5% or lower thus yields a valid test for the median (which is equal to mean, if it exists) of i.i.d. observations with a distribution that can be written as a scale mixture of normals. This is a rather large class of distributions: it includes, for instance, the Student$-t$ distribution with arbitrary degrees of freedom (including the Cauchy distribution), the double exponential distribution, the logistic distribution and all symmetric stable distributions. For $q \leq 4$, this result was already established by Benjamini (1983), who also provides a heuristic argument for the conservativeness of the usual two-sided $t-$test against scale mixtures of normals for $q \geq 5$.

More generally, as long as $\{V_j\}_{j=1}^q$ is independent of $\{Z_j\}_{j=1}^q$, Theorem 1 and the conditioning argument above imply conservativeness of the usual $t-$test of significance level 5% or lower, with an arbitrary joint distribution of $\{V_j\}_{j=1}^q$. We discuss an application of this property in Section 4.2 below.

We now explore how conservative the usual $t-$test becomes when the underlying observations are independent Gaussian of unequal variance. For large q, as long as none of the σ_j^2 dominates the average $\bar{\sigma}_q^2 = q^{-1}\sum_{j=1}^q \sigma_j^2$ (more precisely, if $\lim_{q\to\infty} q^{-2}\sum_{j=1}^q \sigma_j^4 = 0$), a law of large numbers applied to s_X^2 yields $s_X^2 - \bar{\sigma}_q^2 \xrightarrow{p} 0$, and the $t-$test is asymptotically

[7] Based on additional results of Bakirov and Székely (2005), it is also possible to compute the appropriate $p-$values for smaller realizations of $|t|$. These $p-$values are always larger than 8.3% and thus typically not of primary interest.

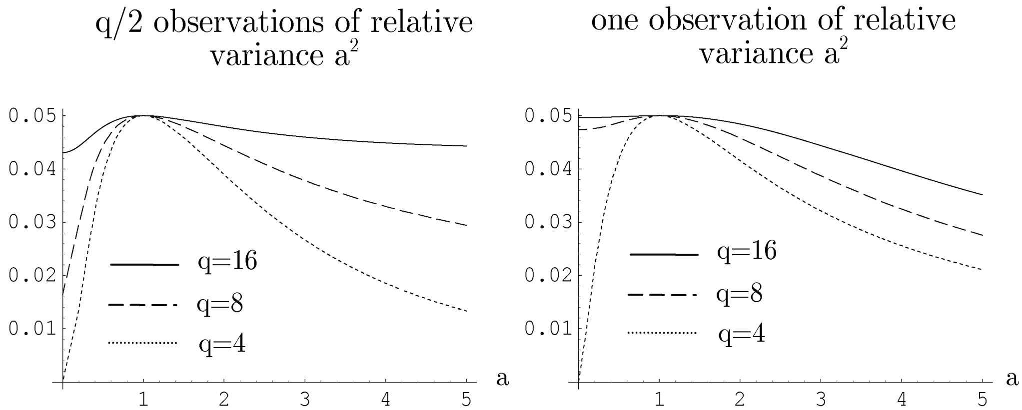

Figure 1: Effective Rejection Probabilities of a 5% Level $t-$test

of correct size as $q \to \infty$. Theorem 1 shows that for a nominal level of 5% or smaller, this convergence to the nominal level is from below for any sequence $\{\sigma_j^2\}$. For small q, Figure 1 depicts the effective size of the 5% level two-sided $t-$test for $q = 4, 8$ and 16 when (i) there are two equal sized groups of i.i.d. Gaussian observations, and the ratio of their variances is equal to a^2: for $i, j \leq q/2$, $\sigma_i^2 = \sigma_j^2$, $\sigma_{q+1-i}^2 = \sigma_{q+1-j}^2$ and $\sigma_1^2/\sigma_q^2 = a^2$ and (ii) all observations excepts one are of the same variance, that is for $i, j \geq 2$, $\sigma_i^2 = \sigma_j^2$, and $\sigma_1^2/\sigma_q^2 = a^2$. Due to the scale invariance, the description in terms of the ratio of variances is without loss of generality. Rejection probabilities in Figure 1 (and Figures 2 and 3 below) were computed by numeric inversion of the characteristic function of the appropriate Gaussian quadratic form—see Imhof (1961). As can be seen from Figure 1, for small q, the effective size can be much lower than the nominal level, but for $q = 16$, the effective size does not drop much below 4% in either scenario.

Theorem 1 provides conditions under which the usual $t-$test remains a valid test. We now turn to a discussion of the optimality of the $t-$statistic when the underlying Gaussian variates X_j are not necessarily of equal variance. Recall that if the variances are identical, then the usual two-sided $t-$test is not only the uniformly most powerful unbiased test of (1), but also the uniformly most powerful scale invariant test (see Ferguson (1967), p. 246, for instance). For a significance level of 5% or lower, Theorem 1 shows that the effective level for the $t-$test never exceeds the nominal level if the variances σ_j^2 are not identical.

So if we consider the hypothesis test

$$H_0 : \mu = 0 \text{ and } \{\sigma_j^2\}_{j=1}^q \text{ arbitrary} \qquad \text{against} \qquad H_1 : \mu \neq 0 \text{ and } \sigma_j^2 = \sigma^2 \text{ for all } j \quad (4)$$

and restrict attention to scale invariant tests, then the least favorable distribution for the q dimensional nuisance parameter $\{\sigma_j^2\}_{j=1}^q$ is the case of equal variances. In other words, the usual $t-$test is the optimal scale invariant test of (4) for any given alternative $\mu \neq 0$ when the level constraint is most difficult to satisfy. By the generalized Neyman-Pearson Lemma (Theorem 7 of Lehmann (1986), p. 104-105), we thus have the following result.

Theorem 2 *Let α and q be such that (3) holds. A test that rejects the null hypothesis for $|t| > \mathrm{cv}_q(\alpha)$ is the uniformly most powerful scale invariant level α test of (4).*

If one is uncertain about the actual variances of X_j, and considers the case of equal variances a plausible benchmark, then the usual 5% level $t-$test maximizes power against such benchmark alternatives in the class of all scale invariant tests. Since the one-sided $t-$test is also known to be the uniformly most powerful invariant test under the (sign-preserving) scale transformations $\{X_j\}_{j=1}^q \to \{cX_j\}_{j=1}^q$ for $c > 0$ (Ferguson (1967), p. 246), the analogous result also holds for the one-sided $t-$test of small enough level.

Note that this optimality result is driven by the conservativeness of the *usual* $t-$test. For $\alpha = 10\%$ and $q = 20$, say, according to Theorem 1, the critical value of the $t-$statistic must be amended to induce conservativeness. The resulting test is thus not optimal when $\sigma_j^2 = \sigma^2$ for all j under both H_0 and H_1. It is also not optimal against the worst case alternative with 14 variances identical and 6 variances zero—the optimal test against such an alternative would certainly exploit that if 6 equal realizations of X_j are observed, they are known to be equal to μ.

In some applications, one might have some *a priori* information about the variances σ_j^2. In that case, it might not be attractive to base inference on a test that maximizes power against alternatives with equal variances. For $\{\sigma_j^2\}_{j=1}^q$ known, the uniformly most powerful test of (1) is based on $\tilde{z} = (\sum_{j=1}^q X_j/\sigma_j^2)/\sum_{j=1}^q 1/\sigma_j^2$, which is standard normal under the null hypothesis. Let v_j^2 be a (nonrandom) guess of σ_j^2, and define $\tilde{X}_j = X_j/v_j^2$. Then a test based on the statistic $\tilde{t} = \sqrt{q}\overline{\tilde{X}}/\tilde{s}_X$ with $\overline{\tilde{X}} = q^{-1}\sum_{j=1}^q \tilde{X}_j$ and $\tilde{s}_X^2 = (q - 1)^{-1}\sum_{j=1}^q(\tilde{X}_j - \overline{\tilde{X}})^2$ is typically an attractive choice: If $v_j^2 = C\sigma_j^2$ for some $C > 0$ for all

j and $\lim_{q\to\infty} q^{-2}\sum_{j=1}^{q} 1/\sigma_j^4 = 0$, then $\tilde{t} - \tilde{z}$ converges in probability to zero as $q \to \infty$ under the null and local alternatives, making inference based on $\tilde{t}$ large sample efficient. At the same time, since $\tilde{X}_j \sim i.i.d.\ \mathcal{N}(0, \sigma_j^2/v_j^4)$ whether or not $v_j^2 = C\sigma_j^2$ for all j, a two-sided test based on $\tilde{t}$ of level 5% or below with the usual critical value is small sample conservative by Theorem 1. Note, however, that no small sample optimality claim akin to Theorem 2 can be made for inference based on $\tilde{t}$: If indeed $v_j^2 = C\sigma_j^2$, then the appropriate critical value for a test based on $\tilde{t}$ would be the $(1-\alpha/2)$ percentile of the (nonstandard) null distribution of $\tilde{t}$, rather than the $(1-\alpha/2)$ percentile of T_{q-1}. Also, if $\{v_j^2\}_{j=1}^q$ are too heterogenous, a test based on $\tilde{t}$ can have low power even for very distant alternatives, especially for small q (see the end of Section 3.1 below for a related point).

3 Large Sample $t-$statistic Based Inference

3.1 Asymptotic Validity and Consistency

Our main interest in the small sample results on the $t-$statistic stems from the following application: Suppose we want to do inference on a scalar parameter β of an econometric model in a large data set with n observations. For a wide range of models and estimators $\hat{\beta}$, it is known that $\sqrt{n}(\hat{\beta} - \beta) \Rightarrow \mathcal{N}(0, \sigma^2)$ as $n \to \infty$, where "$\Rightarrow$" denotes convergence in distribution. Suppose further that the observations exhibit correlations of largely unknown form. If such correlations are pervasive and pronounced enough, then it will be very challenging to consistently estimate σ^2, and inference procedures for β that ignore the sampling variability of a candidate consistent estimator $\hat{\sigma}^2$ will have poor small sample properties.

Now consider a partition the original data set into $q \geq 2$ groups, with n_j observations in group j, and $\sum_{j=1}^q n_j = n$. Denote by $\hat{\beta}_j$ the estimator of β using observations in group j only. Suppose the groups are chosen such that $\sqrt{n}(\hat{\beta}_j - \beta) \Rightarrow \mathcal{N}(0, \sigma_j^2)$ for all j, and, crucially, such that $\sqrt{n}(\hat{\beta}_j - \beta)$ and $\sqrt{n}(\hat{\beta}_i - \beta)$ are asymptotically independent for $i \neq j$—this amounts to the convergence in distribution

$$\sqrt{n}(\hat{\beta}_1 - \beta, \cdots, \hat{\beta}_q - \beta)' \Rightarrow \mathcal{N}(0, \mathrm{diag}(\sigma_1^2, \cdots, \sigma_q^2)), \quad \max_{1\leq j\leq q} \sigma_j^2 > 0 \tag{5}$$

and $\{\sigma_j^2\}_{j=1}^q$ are, of course, unknown. The asymptotic Gaussianity of $\sqrt{n}(\hat{\beta}_j - \beta)$, $j =$

$1, \cdots, q$, typically follows from the same reasoning as the asymptotic Gaussianity of the full sample estimator $\hat{\beta}$. The argument for an asymptotic independence of $\hat{\beta}_j$ and $\hat{\beta}_i$ for $i \neq j$, on the other hand, depends on the choice of groups and the details of the application. We discuss such arguments in more detail for some common econometric models in Section 4 below.

Under (5), for large n, the q estimators $\hat{\beta}_j$, $j = 1, \cdots, q$, are approximately independent Gaussian random variables with common mean β and variances σ_j^2. Thus, by Theorem 1 above, one can perform an asymptotically valid test of level α, $\alpha \leq 0.05$ of $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$ by rejecting H_0 when $|t_\beta|$ exceeds the $(1 - \alpha/2)$ percentile of the Student$-t$ distribution with $q - 1$ degrees of freedom, where t_β is the usual $t-$statistic

$$t_\beta = \sqrt{q}\frac{\overline{\hat{\beta}} - \beta_0}{s_{\hat{\beta}}} \tag{6}$$

with $\overline{\hat{\beta}} = q^{-1}\sum_{j=1}^{q}\hat{\beta}_j$ and $s_{\hat{\beta}}^2 = (q-1)^{-1}\sum_{j=1}^{q}(\hat{\beta}_j - \overline{\hat{\beta}})^2$. By Theorem 1 (and the Continuous Mapping Theorem), this inference is asymptotically valid whenever (5) holds, irrespective of the values of σ_j^2, $j = 1, \cdots, q$. Also, by implication, the confidence interval $\overline{\hat{\beta}} \pm \text{cv}\, s_{\hat{\beta}}$ where cv is the usual $(1 + C)/2$ percentile of the Student$-t$ distribution with $q - 1$ degrees of freedom has asymptotic coverage of at least C for all $C \geq 0.95$.[8]

For some applications, a slightly more general regularity condition than (5) is useful: Suppose

$$\{m_n(\hat{\beta}_j - \beta)\}_{j=1}^{q} \Rightarrow \{Z_j V_j\}_{j=1}^{q} \tag{7}$$

for some real sequence m_n, where $Z_j \sim i.i.d.\ \mathcal{N}(0, 1)$, the random vector $\{V_j\}_{j=1}^{q}$ is independent of the vector $\{Z_j\}_{j=1}^{q}$ and $\max_j |V_j| > 0$ almost surely. In analogy to the

[8] On first sight, the $t-$statistic based approach described may seem to be similar to subsampling. However, the similarities do not go beyond the use of blocking the data in both the $t-$statistic approach and subsampling: First, the $t-$statistic approach works for a fixed number q of groups, while subsampling requires the number of subsamples to grow with the sample size. Second, in the $t-$statistic based approach one calculates the *estimators* $\hat{\beta}_i$ for each block and then forms the $t-$statistic t_β in (6) to conduct asymptotically valid inference. Subsampling a $t-$statistic, in contrast, would approximate the distribution of the full sample $t-$statistic with the empirical cdf of $t-$*statistics* calculated from each subsample. Third, subsampling typically requires additional weak dependence assumptions, such as strong mixing, to result in asymptotically valid inference (see, for instance, Chapters 3 and 4 in Politis, Romano, and Wolf (1999)).

discussion of Theorem 1 in Section 2 above, condition (7) accommodates convergences (at an arbitrary rate) to independent but potentially heterogeneous mixed normal distributions, such as the family of stable symmetric distributions, and also convergences to conditionally normal variates which are unconditionally dependent through their second moments. Under (7), inference based on t_β remains asymptotically valid conditionally on $\{V_j\}_{j=1}^q$ by the Continuous Mapping Theorem and an application of Theorem 1, and thus also also unconditionally.

Under the fixed alternative $\beta \neq \beta_0$, (5) or (7) imply that $s_{\hat{\beta}} = O_p(n^{-1/2})$ and $\bar{\hat{\beta}} - \beta = O_p(n^{-1/2})$, so that $P(|t_\beta| > K) \to 1$ for all K and a test based on $|t_\beta|$ is consistent at any level of significance. Under fixed heterogeneous alternatives of the null hypothesis $\beta_0 = 0$ with the true value of β in group j given by β_j (and $\beta_j \neq \beta_i$ for some j and i), $\hat{\beta}_j \xrightarrow{p} \beta_j$ for $j = 1, \cdots, q$, and a test based on $|t_\beta|$ with critical value cv rejects asymptotically with probability one if

$$\frac{\left(q^{-1}\sum_{j=1}^q \beta_j\right)^2}{q^{-1}\sum_{j=1}^q \beta_j^2} > \frac{\mathrm{cv}^2}{\mathrm{cv}^2 + q - 1}. \tag{8}$$

Especially for small q and large cv, (8) might not be satisfied when $\{\beta_j\}_{j=1}^q$ are very heterogenous, even when all β_j are of the same sign. On the other hand, a calculation shows that for $q \geq 7$, a 5% level test is consistent for all alternatives $\{\beta_j\}_{j=1}^q$ of equal sign that are less heterogeneous (in the majorization sense, see Marshall and Olkin (1979)) than $\beta_1 = \cdots = \beta_{\lfloor q/2 \rfloor} = 0$ and $\beta_{\lfloor q/2 \rfloor + 1} = \cdots = \beta_q \neq 0$, where $\lfloor \cdot \rfloor$ denotes the greatest lesser integer function.

3.2 Asymptotic Efficiency

The arguments so far imply that the $t-$statistic approach to inference in large samples yields an asymptotically valid and consistent test statistic under the weak convergence (5) or (7). Furthermore, Theorem 2 provides a small sample efficiency claim about the $t-$test under heterogeneous variances. This section draws on the recent results in Müller (2007a) and shows in which sense a corresponding large sample efficiency claim can be made about the large sample $t-$statistic approach under (5).

Suppose the n observations in the potentially correlated large data set are of dimension $r \times 1$, so that the overall data $\mathbf{Y}_n$ is an element of $\mathbb{R}^{rn}$. In general, tests φ_n of $H_0 : \beta = \beta_0$

are sequences of (measurable) functions from $\mathbb{R}^{rn}$ to the unit interval, where $\varphi_n(y_n) \in [0,1]$ indicates the probability of rejection conditional on observing $\mathbf{Y}_n = \mathbf{y}_n$. If φ_n takes on values strictly between zero and one then φ_n is a randomized test. As usual in large sample testing problems, consider the sequence of local alternatives $\beta = \beta_n = \beta_0 + \mu/\sqrt{n}$, so that the null null hypothesis becomes $H_0 : \mu = 0$. Under such local alternatives, (5) implies

$$\{\sqrt{n}(\hat{\beta}_j - \beta_0)\}_{j=1}^q \Rightarrow \{X_j\}_{j=1}^q \tag{9}$$

where X_j, $j = 1, \cdots, q$ are as in the small sample case of Section 2 above, that is X_j are independent and distributed $\mathcal{N}(\mu, \sigma_j^2)$. Furthermore, by the Continuous Mapping Theorem, it also holds that

$$\left\{ \frac{\hat{\beta}_j - \beta_0}{\bar{\hat{\beta}} - \beta_0} \right\}_{j=1}^q \Rightarrow \{R_j\}_{j=1}^q = \left\{ \frac{X_j}{\bar{X}} \right\}_{j=1}^q \tag{10}$$

where $\bar{X} = q^{-1} \sum_{j=1}^q X_j$. The interest of (10) over (9) is that $\{R_j\}_{j=1}^q$ is a maximal invariant of the group of transformations $\{X_j\}_{j=1}^q \to \{cX_j\}_{j=1}^q$ for $c \neq 0$, so that in the "limiting problem" with $\{R_j\}_{j=1}^q$ observed, Theorem 2 implies that a level-α test based on $|t| = \sqrt{q}\bar{R}/s_R = \sqrt{q}\bar{X}/s_X$ with $\bar{R} = q^{-1}\sum_{j=1}^q R_j$ and $s_R^2 = (q-1)^{-1}\sum_{j=1}^q (R_j - \bar{R})^2$ maximizes power against alternatives with $\mu \neq 0$ and $\sigma_j^2 = \sigma$ for $j = 1, \cdots, q$ as long as $\alpha \leq 0.05$.

Let $F_n(m, \mu, \{\sigma_j^2\}_{j=1}^q)$ be the distribution of $\mathbf{Y}_n$ in a specific model m with parameter $\beta = \beta_0 + \mu/\sqrt{n}$ and asymptotic variance of $\hat{\beta}_j$ equal to σ_j^2; think of m as describing all aspects of the data generation mechanism beyond the parameters μ and $\{\sigma_j^2\}_{j=1}^q$, such as the correlation structure of $\mathbf{Y}_n$. The unconditional rejection probability of a test φ_n then is $\int \varphi_n dF_n(m, \mu, \{\sigma_j^2\}_{j=1}^q)$, and the asymptotic null rejection probability is $\limsup_{n\to\infty} \int \varphi_n dF_n(m, 0, \{\sigma_j^2\}_{j=1}^q)$.

The weak convergences (9) and (10) obviously only hold for some sequences of underlying distributions $F_n(m, \mu, \{\sigma_j^2\}_{j=1}^q)$ of $\mathbf{Y}_n$, that is some models m. The assumption (9) is an asymptotic regularity condition that restricts the dependence in $\mathbf{Y}_n$ in a way that in large samples, each $\hat{\beta}_j$ provides independent and Gaussian information about the parameter of interest β. The convergence (10) is a very similar, but slightly weaker regularity condition. Denote by $\mathcal{M}_0^X$ and $\mathcal{M}_0^R$ the set of models m for which (9) and (10) hold under the null hypothesis of $\mu = 0$, respectively (so that $\mathcal{M}_0^X \subset \mathcal{M}_0^R$), and analogously, denote by $\mathcal{M}_1^X$ and $\mathcal{M}_1^R$ the set of models m for which (9) and (10) hold pointwise for $\mu \neq 0$. A concern about strong and pervasive correlations in $\mathbf{Y}_n$ of largely unknown form means that little is known about properties of $F_n(m, \mu, \{\sigma_j^2\}_{j=1}^q)$. In an effort to obtain robust inference for a large set of possible data generating processes m, one might want to impose that level-α tests φ_n are asymptotically valid for all $m \in \mathcal{M}_0^X$ or $m \in \mathcal{M}_0^R$, that is

$$\limsup_{n\to\infty} \int \varphi_n dF_n(m, 0, \{\sigma_j^2\}_{j=1}^q) \leq \alpha \quad \text{for all } m \in \mathcal{M}_0 \text{ and } \{\sigma_j^2\}_{j=1}^q \text{ with } \max_j \sigma_j^2 > 0 \tag{11}$$

for $\mathcal{M}_0 = \mathcal{M}_0^X$ or $\mathcal{M}_0 = \mathcal{M}_0^R$.

Denote by $\varphi_n^*(\alpha) = \mathbf{1}[|t_\beta| > \mathrm{cv}_q(\alpha)]$ the $\mathbb{R}^{rn} \mapsto \{0, 1\}$ test of asymptotic size $\alpha \leq 0.05$ that rejects for large values of $|t_\beta|$ as defined in (6), and note that, by scale invariance, t_β can also be computed from the observations $\{(\hat{\beta}_j - \beta_0)/(\overline{\hat{\beta}} - \beta_0)\}_{j=1}^q$. As discussed in Section 3.1, the test $\varphi_n^*(\alpha)$ satisfies (11) for $\mathcal{M}_0 = \mathcal{M}_0^X$ as long as $\alpha \leq 0.05$, and scale invariance implies that (11) also holds for $\mathcal{M}_0 = \mathcal{M}_0^R$. What is more, for any data generating process satisfying (10) under the alternative with $\mu \neq 0$, i.e. for any $m \in \mathcal{M}_1^R$ or $m \in \mathcal{M}_1^X$, $\varphi_n^*(\alpha)$ has the same local power $\lim_{n\to\infty} \int \varphi_n^*(\alpha) dF_n(m, \mu, \{\sigma_j^2\}_{j=1}^q)$ for $\mu \neq 0$ as the small sample $t-$test in the "limiting problem" with $\{X_j\}_{j=1}^q$ (or $\{R_j\}_{j=1}^q$) observed. An asymptotic efficiency claim about the $t-$statistic approach now amounts to the statement that no other test φ_n satisfying (11) has higher local asymptotic power. The following Theorem, which follows straightforwardly from the general results in Müller (2007a) and Theorem 2 above, provides such a statement for the case of equal asymptotic variances.

Theorem 3 *(i) For any test φ_n that satisfies (11) for $\mathcal{M}_0 = \mathcal{M}_0^R$ and $\alpha \leq 0.083$, $\limsup_{n\to\infty} \int \varphi_n dF_n(m, \mu, \{\sigma^2\}_{j=1}^q) \leq \lim_{n\to\infty} \int \varphi_n^*(\alpha) dF_n(m, \mu, \{\sigma^2\}_{j=1}^q)$ for all $\mu \neq 0$ and $m \in \mathcal{M}_1^R$.*

(ii) Suppose there exists a group of transformations $G_n(c)$ of $\mathbf{Y}_n$ that induces the transformations $\{\hat{\beta}_j - \beta\}_{j=1}^q \to \{c(\hat{\beta}_j - \beta)\}_{j=1}^q$ for $c \neq 0$. For any test φ_n that is invariant to G_n and that satisfies (11) for $\mathcal{M}_0 = \mathcal{M}_0^X$ and $\alpha \leq 0.083$, $\limsup_{n\to\infty} \int \varphi_n dF_n(m, \mu, \{\sigma^2\}_{j=1}^q) \leq \lim_{n\to\infty} \int \varphi_n^*(\alpha) dF_n(m, \mu, \{\sigma^2\}_{j=1}^q)$ for all $\mu \neq 0$ and $m \in \mathcal{M}_1^X$.

Part (i) of Theorem 3 shows the $t-$statistic approach to be the asymptotically most powerful against the benchmark alternative of equal asymptotic variances among all tests that provide asymptotically valid inference under the regularity condition (10). Part (ii) contains the same claim under the slightly more natural condition (9) when attention is restricted to tests that are appropriately invariant. For example, in a regression context, an adequate underlying group of transformations is the multiplication of the dependent variable by c. Also, the analogous asymptotic efficiency statements hold for one-sided tests based on t_β of asymptotic level smaller than 4.1%.

Note that this asymptotic optimality of the $t-$statistic approach holds for *all* models in $\mathcal{M}_1^X$ and $\mathcal{M}_1^R$, that is whenever (9) and (10) holds with $\mu \neq 0$. In other words, for any test that has higher asymptotic power for some data generating process for which (9) and (10) holds with $\mu \neq 0$ and equal asymptotic variances, there exists a data generating process satisfying (9) and (10) with $\mu = 0$ for which the test has asymptotic rejection probability larger than α.

In particular, this implies that it is not possible to use data dependent methods to determine an appropriate q: Suppose one is conservatively only willing to assume (10) to hold for some small $q = q_0$, but the actual data is much more regular in the sense that (10) also holds for $q = 2q_0$, with each group divided into two subgroups. Then any data dependent method that leads to higher asymptotic local power for this more regular data necessarily lacks robustness in the sense that there exists some data generating process for which (10) holds with $q = q_0$, and this method overrejects asymptotically. Thus, with (10) viewed as a regularity condition on the underlying large data set, the $t-$statistic approach efficiently exploits the available information, with highest possible power in the benchmark case of equal asymptotic variances.

If maximizing power against this benchmark alternative is obviously inappropriate because the asymptotic variances are (at least approximately) known and very different, then it might be preferable to base inference on the analogue of the weighted least squares $t-$statistic $\tilde{t}$ discussed at the end of Section 2 above.

3.3 Size Control under Dependence

Tests of level 5% or lower based on t_β are asymptotically valid whenever (5) holds. As usual, when applying this result in small samples, one will incur an approximation error, as the sampling distribution of $\{\hat{\beta}_j\}_{j=1}^q$ will not be exactly that of a sequence of independent normals with common mean β. In particular, depending on the applications, the estimators from different groups $\hat{\beta}_j$ might not be exactly independent. We now briefly investigate what kind of correlations are necessary to grossly distort the size of tests based on t_β, while maintaining the assumption of multivariate Gaussianity.

Specifically, we consider two correlation structures for $\{\hat{\beta}_j\}_{j=1}^q$: (i) $\hat{\beta}_j$ are a strictly stationary autoregressive process of order one (AR(1)), i.e. the correlation between $\hat{\beta}_i$ and $\hat{\beta}_j$ is $\rho^{|i-j|}$; (ii) $\{\hat{\beta}_j\}_{j=1}^q$ has the correlation structure of a random effects model, i.e. the correlation between $\hat{\beta}_i$ and $\hat{\beta}_j$ is ρ for $i \neq j$. For both cases, we consider the two types of variance heterogeneity discussed above, with either two equal-sized identical variance groups of relative variance a^2, or all observations of equal variance except for one of relative variance a^2. Figure 2 depicts the effective size of a 5% level two-sided $t-$tests under these four scenarios. As might be expected, negative ρ lead to underrejections throughout. More interestingly, $t-$tests for q small are somewhat robust against correlations in the underlying observations. This effect becomes especially pronounced if combined with strong heterogeneity in the variances: with $a = 5$, ρ needs to be larger than 0.4 before effective size of a $t-$test based on $q = 4$ observations exceeds the nominal level in both the AR(1) and the random effects model for both types of variance heterogeneity. But even in the case of equal variances, the size of a test based on $q = 4$ observations exceeds 7.5% only when ρ is larger than 0.18 in the AR(1) model. So while (5) is the essential assumption of the approach suggested here, inference based on t_β continues to have reasonable properties as long as the dependence in $\{\hat{\beta}_j\}_{j=1}^q$ is weak, especially when q is small.

3.4 Comparison with Inference under Known Variance

We now turn to a discussion of the relative performance of this robust approach and inference based on the full sample estimator $\hat{\beta}$ with σ^2 known. When (5) summarizes the amount of regularity that one is willing to impose, then this is a purely theoretical exercise. On the other hand, one might be willing to consider stronger assumptions that enable consistent estimation of σ^2, and it is interesting to explore the relative gain in power.

With $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$, the standard approach to inference is, of course, to reject when $|z_\beta|$

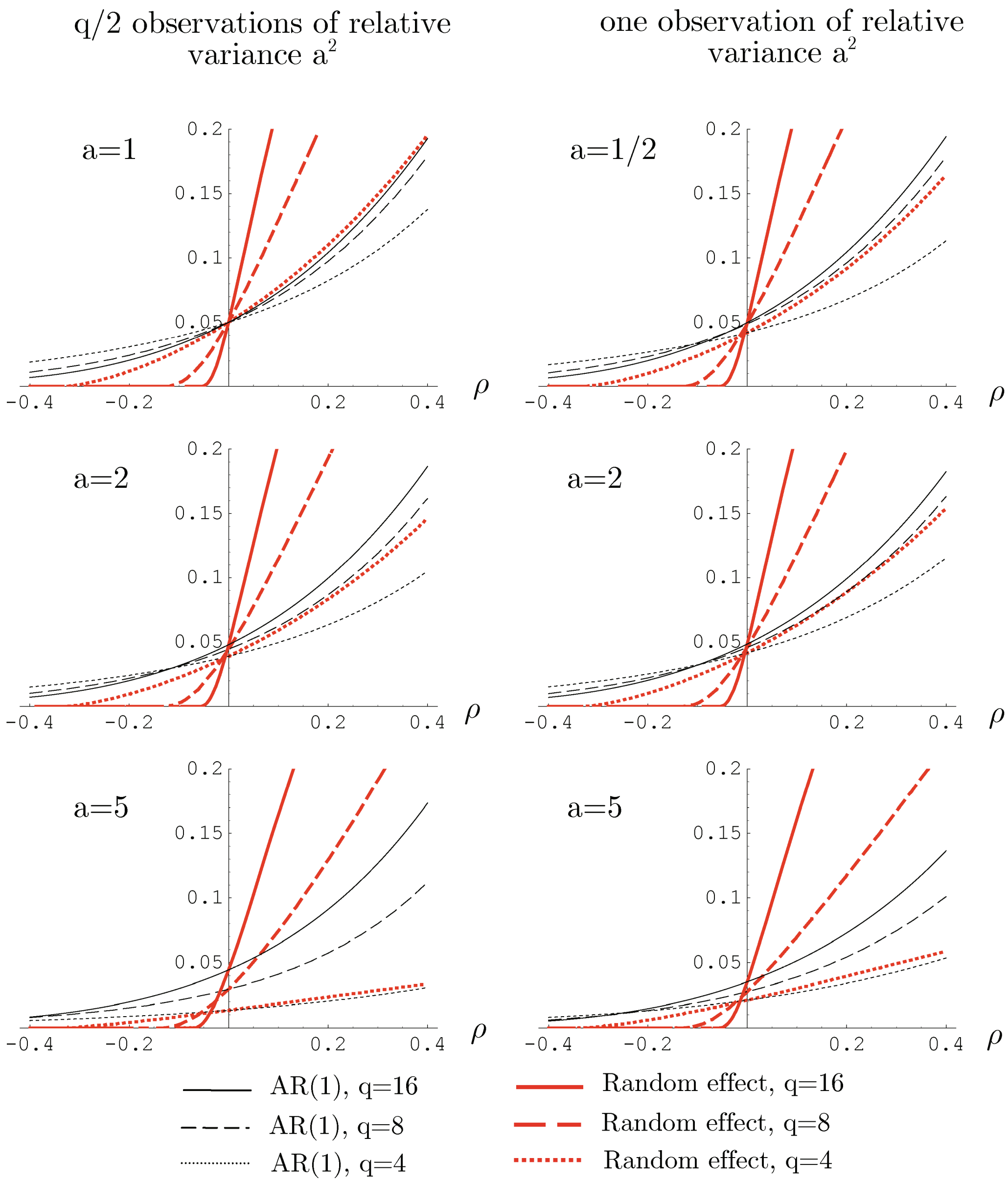

Figure 2: Effective rejection probabilities of 5% level $t-$statistics under correlation

exceeds the critical value for a standard normal, where z_β is given by

$$z_\beta = \sqrt{n}\frac{\hat{\beta} - \beta_0}{\hat{\sigma}} = \sqrt{n}\frac{\hat{\beta} - \beta_0}{\sigma} + o_p(1) \tag{12}$$

under the null and local alternatives. In this case, a comparison of the asymptotic power of a test based on t_β with the asymptotic power of a test based on z_β approximates the efficiency cost of the higher robustness of inference based on t_β.

To investigate this issue, we impose more structure on the econometric model. Specifically, suppose the model is in the class of exactly identified Generalized Method of Moments (GMM) models (cf. Hansen (1982)) with moment condition $E[g(\theta, y_i)] = 0$, where g is a known $k \times 1$ vector valued function, θ is a $k \times 1$ vector of parameters and y_i, $i = 1, \cdots, n$, are possibly vector-valued observations. Without loss of generality, we assume that the first element of θ is the parameter of interest β, so that the last $k - 1$ elements of θ are nuisance parameters. Denote by $\mathcal{G}_j$ the set of indices of group j observations, such that y_i is in group j if and only if $i \in \mathcal{G}_j$. Assume that the GMM estimator $\hat{\theta}_j$ based on group j observations satisfies

$$\sqrt{n}(\hat{\theta}_j - \theta) = \Gamma_j^{-1} Q_j + o_p(1)$$

where $n^{-1} \sum_{i \in \mathcal{G}_j} \frac{\partial g(a, y_i)}{\partial a}|_{a = \hat{\theta}_j} \xrightarrow{p} \Gamma_j$ (of full rank for all j), and $Q_j = n^{-1/2} \sum_{i \in \mathcal{G}_j} g(\theta, y_i) \Rightarrow \mathcal{N}(0, \Omega_j)$. In addition, in analogy to (5), we assume the GMM group estimators to be asymptotically independent, which requires $(Q_1', \cdots, Q_q') \Rightarrow \mathcal{N}(0, \mathrm{diag}(\Omega_1, \cdots, \Omega_q))$. Under these assumption, the simple average of the q group estimators $\bar{\hat{\theta}} = q^{-1} \sum_{j=1}^q \hat{\theta}_j$ satisfies

$$\sqrt{n}(\bar{\hat{\theta}} - \theta) = q^{-1} \sum_{j=1}^q \Gamma_j^{-1} Q_j + o_p(1) \Rightarrow \mathcal{N}(0, \bar{\Sigma}_q), \tag{13}$$

where $\bar{\Sigma}_q = q^{-2} \sum_{j=1}^q \Gamma_j^{-1} \Omega_j (\Gamma_j')^{-1}$. In contrast, the full sample GMM estimator $\hat{\theta}$ which solves $n^{-1} \sum_{i=1}^n g(\hat{\theta}, y_i)' g(\hat{\theta}, y_i) = 0$, satisfies under the same assumptions

$$\sqrt{n}(\hat{\theta} - \theta) = \left(\sum_{j=1}^q \Gamma_j \right)^{-1} \sum_{j=1}^q Q_j + o_p(1) \Rightarrow \mathcal{N}(0, \Sigma_q) \tag{14}$$

where $\Sigma_q = \left(\sum_{j=1}^q \Gamma_j \right)^{-1} \left(\sum_{j=1}^q \Omega_j \right) \left(\sum_{j=1}^q \Gamma_j' \right)^{-1}$. In general, this full sample GMM estimator is not efficient: with heterogeneous groups, it would be more efficient to compute the optimal GMM estimator of the q conditions $E[g(\theta, y_i)] = 0$ for $i \in \mathcal{G}_j$, $j = 1, \cdots, q$. But this efficient full sample estimator requires the consistent estimation of the optimal weighting matrix, which involves Ω_j, $j = 1, \cdots, q$. This is unlikely to be feasible or appropriate in applications with pronounced correlations and heterogeneity, so that the relevant comparison for $\bar{\hat{\theta}}$ is with $\hat{\theta}$ as characterized in (14).

Comparing $\bar{\Sigma}_q$ with Σ_q, we find that while $\sqrt{n}$-consistent and asymptotically Gaussian, the estimators $\hat{\theta}$ and $\bar{\hat{\theta}}$ (and thus $\hat{\beta}$ and $\bar{\hat{\beta}}$) are not asymptotically equivalent. The asymptotic power of tests based on t_β and z_β thus not only differ through differences in the denominator, but also through their numerator. The relationship between $\bar{\Sigma}_q$ and Σ_q is summarized in the following Theorem, whose proof is given in the appendix.

Theorem 4 *Let $\mathcal{Q}_k$ be the set of full rank $k \times k$ matrices, and let $\mathcal{P}_k \subset \mathcal{Q}_k$ denote the set of symmetric and positive definite $k \times k$ matrices. For any $q \geq 2$:*

(i) Let ι be the $k \times 1$ vector with 1 in the first row and zeros elsewhere. Then $\inf_{\{\Gamma_j\}_{j=1}^q \in \mathcal{Q}_1^q, \{\Omega_j\}_{j=1}^q \in \mathcal{P}_1^q} \bar{\Sigma}_q / \Sigma_q = 0$,

$$\inf_{\{\Gamma_j\}_{j=1}^q \in \mathcal{P}_k^q, \{\Omega_j\}_{j=1}^q \in \mathcal{P}_k^q} \frac{\iota' \Sigma_q \iota}{\iota' \bar{\Sigma}_q \iota} = 0 \quad \text{and} \quad \inf_{\{\Gamma_j\}_{j=1}^q \in \mathcal{P}_k^q, \{\Omega_j\}_{j=1}^q \in \mathcal{P}_k^q} \frac{\iota' \bar{\Sigma}_q \iota}{\iota' \Sigma_q \iota} = \begin{cases} 1/q^2 & \text{if } k = 1 \\ 0 & \text{if } k \geq 2 \end{cases}.$$

(ii) For any sequence $\{\Gamma_j\}_{j=1}^q \in \mathcal{Q}_k^q$ there exists $\{\bar{\Omega}_j\}_{j=1}^q \in \mathcal{P}_k^q$ so that $\Sigma_q - \bar{\Sigma}_q$ is positive semidefinite for $\{\Omega_j\}_{j=1}^q = \{\bar{\Omega}_j\}_{j=1}^q$, and for any sequence $\{\Gamma_j\}_{j=1}^q \in \mathcal{P}_k^q$ there exists $\{\underline{\Omega}_j\}_{j=1}^q \in \mathcal{P}_k^q$ so that $\Sigma_q - \bar{\Sigma}_q$ is negative semidefinite for $\{\Omega_j\}_{j=1}^q = \{\underline{\Omega}_j\}_{j=1}^q$.

(iii) If $\Gamma_j = \Gamma$ for $j = 1, \cdots, q$, then $\bar{\Sigma}_q = \Sigma_q$ for all $\{\Omega_j\}_{j=1}^q$.

Part (i) of Theorem 4 shows that very little can be said in general about the relative magnitudes of the asymptotic variances of $\bar{\hat{\beta}}$ and $\hat{\beta}$. Only for $k = 1$ and Γ_j restricted to positive numbers there exists a bound on the relative asymptotic variances, and this bound is so weak that even for q as small as $q = 4$, one can construct an example where the local asymptotic power of a two-sided 5% level test based on $|t_\beta|$ greatly exceeds the local asymptotic power of a test based on $|z_\beta|$ for almost all alternatives, despite the much larger critical value for $|t_\beta|$ (which is equal to 3.18 for $q = 4$ compared to 1.96 for $|z_\beta|$). What is more, as shown in part (ii), it is not possible to determine whether $\hat{\theta}$ is more efficient than $\bar{\hat{\theta}}$ without knowledge of $\{\Omega_j\}_{j=1}^q$, and vice versa in the important special case where Γ_j are symmetric and positive definite.[9]

When $\Gamma_j = \Gamma$ for all j, however, the two estimators become asymptotically equivalent. This special case naturally arises when the groups have an equal number of observations n/q, and the average of the derivative of the moment condition is homogenous across groups. One important set-up with this feature is the case of underlying i.i.d. observations. With $\Gamma_j = \Gamma$, $\sqrt{n}(\bar{\hat{\theta}} - \theta) = \sqrt{n}(\hat{\theta} - \theta) + o_p(1)$ and $\hat{\beta}$ and $\bar{\hat{\beta}}$ are asymptotically equivalent (up to order $\sqrt{n}$) under the null and local alternatives. There is thus no asymptotic efficiency cost for basing inference about β on $\bar{\hat{\beta}}$ associated with the re-estimation of the last $k-1$ elements of θ in each of the q groups. The asymptotic local power of tests based on t_β and z_β simply reduces to the small sample power of the $t-$statistic (2) as discussed in Section 2 and the $z-$statistic $z = \sqrt{q}\bar{X}/\bar{\sigma}_q$ in the hypothesis test (1), where σ_j^2 is the (1,1) element of $\Gamma^{-1}\Omega_j\Gamma^{-1}$ and $\bar{\sigma}_q^2 = q^{-1}\sum_{j=1}^q \sigma_j^2$. Figure 3 depicts the power of such 5% level tests for various q and the two scenarios for the variances considered in Figures 1 and 2 above. The scale of the variances is normalized to ensure $\bar{\sigma}_q^2 = 1$, and the magnitude of the alternative μ is the value on the abscissa divided by $\sqrt{q}$, so that the power of the $z-$statistic is the same for all q.

When all variances are identical ($a = 1$), the differences in power between the $t-$statistic and $z-$statistic are substantial for small q, but become quite small for moderate q: The largest difference in power is 32 percentage points for $q = 4$, is 13 for $q = 8$ and is 5.8 for $q = 16$. In both scenarios and all considered values of $a \neq 1$, the maximal difference in power between the $z-$statistic and $t-$statistic is smaller than this equal variance benchmark, despite the fact that the $t-$statistic underrejects under the null hypothesis when variances are unequal. When $\Gamma_j = \Gamma$ for all j, the loss in local asymptotic power of inference based on t_β compared to z_β is thus approximately bounded above by the largest loss of power of a small sample $t-$statistic over the $z-$statistic in an i.i.d. Gaussian set-up. Interestingly, for very unequal variances with $a = 5$, the $t-$statistic is sometimes even more powerful than the $z-$statistic. This is possible because the $z-$statistic is not optimal in the case of unequal variances. Intuitively, for small realizations of the high variance

[9] There exist $\{\Gamma_j\}_{j=1}^q \notin \mathcal{P}_k^q$ that make $\bar{\hat{\theta}}$ the more efficient estimator for all possible values of $\{\Omega_j\}_{j=1}^q$; for instance, for $k = 1$ and $q = 2$, let $\Gamma_1 = 1$ and $\Gamma_2 = -1/2$.

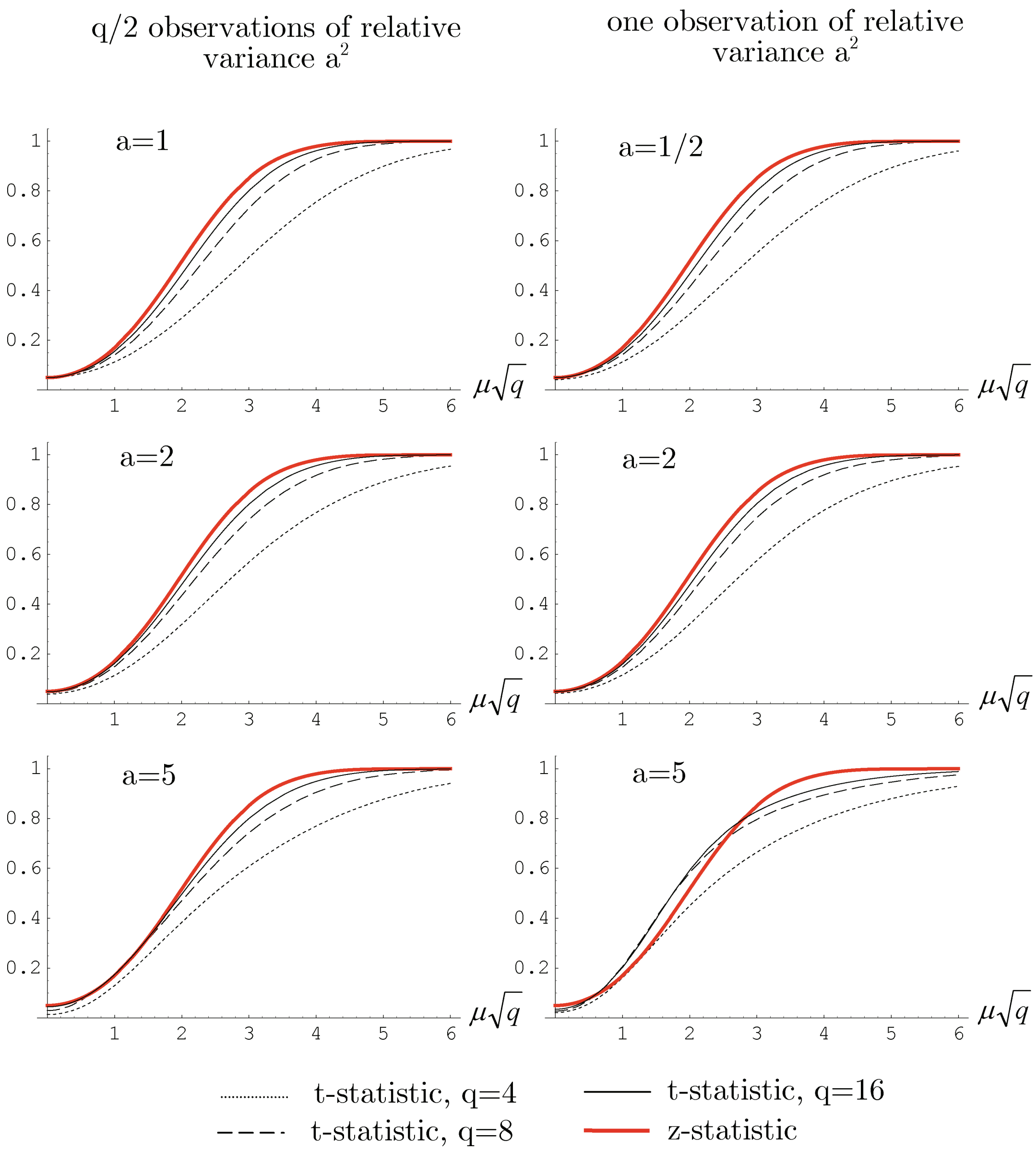

Figure 3: Power of 5% level $t-$statistics and z-statistic

observation, s_X^2 is much smaller than $\bar{\sigma}_q^2$, and the $t-$statistic exceeds the (larger) critical value more often under moderate alternatives.

To sum up, in an exactly identified GMM framework, tests based on t_β and z_β compare as follows: Both tests are consistent and have power against the same local alternatives. Without additional assumptions on Γ_j—the sample average of the derivative of the moment condition in group j—little can be said about their local asymptotic power, as either procedure may be the more powerful one, depending on the values of Ω_j, the group j asymptotic covariance. In the important special case where $\Gamma_j = \Gamma$ for all j, the largest gain in power of inference based on 5% level two-sided z_β over t_β is typically no larger than the largest difference in power between a small sample $z-$statistic over a $t-$statistic for i.i.d. Gaussian observations. By implication, as soon as q is moderately large (say, $q = 16$) there exist only modest gains in terms of local asymptotic power (less than 6 percentage points for 5% level tests) of efforts to consistently estimate the asymptotic variance σ^2.

3.5 A Simple Test of Potentially Consistent Variance Estimators

In many applications, there will be uncertainty whether the additional assumptions required for consistent variance estimation hold in the data at hand. We now discuss a simple test whether such assumptions are rejected by the data, maintaining throughout that (5) holds.

Typically, assumptions that allow for consistent estimation of σ^2 also allow for consistent estimation of σ_j^2, $j = 1, \cdots, q$, in (5). For example, under the assumption of no intra-group correlation, one can consistently estimate σ_j^2 by applying the usual White (1980) formula to the observations of each group. Denote by $\hat{\sigma}_j^2$, $j = 1, \cdots, q$, a set of such estimators. If indeed $\hat{\sigma}_j^2 \xrightarrow{p} \sigma^2$, then under (5), we have the approximately Gaussian least squares regression

$$\sqrt{n}\frac{\hat{\beta}_j}{\hat{\sigma}_j} = \sqrt{n}\frac{\beta}{\hat{\sigma}_j} + \varepsilon_j, \quad \varepsilon_j \Rightarrow i.i.d.\, \mathcal{N}(0,1), \quad j = 1, \cdots, q, \tag{15}$$

so that the sum of squared residuals $\sum_{j=1}^q \hat{\varepsilon}_j^2$ is approximately distributed χ_{q-1}^2. If, in contrast, $\hat{\sigma}_j^2$ systematically underestimates σ_j^2, say because of some ignored positive correlation in the data, then $\sum_{j=1}^q \hat{\varepsilon}_j^2$ tends to be larger. A simple test whether the data is consistent with the assumptions required to obtain consistent estimators of $\{\sigma_j^2\}_{j=1}^q$ (and thus σ^2) is therefore to use $\sum_{j=1}^q \hat{\varepsilon}_j^2$ as a test statistic, which is asymptotically distributed χ^2_{q-1}under the null hypothesis. If such a test rejects, one might abandon the attempt of consistently estimating σ^2 and instead rely on the $t-$statistic approach to inference based on (5).

4 Applications

We now discuss applications of the $t-$statistic approach, and provide some Monte Carlo evidence of its performance compared to alternative approaches. Specifically, we consider time series data, panel data, data where observations are categorized in clusters and spatially correlated data. The Monte Carlo evidence focusses on inference about OLS linear regression coefficients. This is for convenience and comparability to other simulation studies in the literature, since the $t-$statistic approach is also applicable to instrumental variable regressions and nonlinear models, as noted above. Also, we mostly consider data generating processes where the variances of the $\hat{\beta}_j$ are similar. This is again to ensure comparability with other simulation studies[10], and it also represents the case where the theoretical results above predict size control to be most difficult for the $t-$statistic approach.

4.1 Time Series Data

With observations ordered in time, the default assumption driving most of time series inference is that the further apart the observations, the weaker their potential correlation. For the $t-$statistic approach, in absence of more specific information regarding the potential time series correlation, this suggests dividing the sample of size T into q (approximately) equal sized groups of consecutive observations: the observation indexed by t, $t = 1, \cdots, T$, is element of group j if $t \in \mathcal{G}_j = \{t : (j-1)T/q < t \leq jT/q\}$ for $j = 1, \cdots, q$. The smaller q, the less approximate independence in time is imposed.

[10] Given the theoretical results provided in Theorem 4 above, it is quite evident that under asymptotic heterogeneity, one can construct examples where $\bar{\hat{\beta}}$ is a vastly inferior estimator than $\hat{\beta}$, and vice versa.

Under a wide range of assumptions on the underlying model and observations, exactly identified GMM with k dimensional moment condition $E[g(\theta, y_t)] = 0$ satisfies

$$\sup_{0\leq s\leq 1} ||T^{-1}\sum_{t=1}^{\lfloor sT\rfloor}\frac{\partial g(a, y_t)}{\partial a}|_{a=\theta} - \int_0^s \Gamma(\lambda)d\lambda|| \xrightarrow{p} 0 \quad \text{and} \quad T^{-1/2}\sum_{t=1}^{\lfloor \cdot T\rfloor} g(\theta, y_t) \Rightarrow \int_0^{\cdot} h(\lambda)dW(\lambda) \tag{16}$$

for some nonstochastic, positive definite $k \times k$ matrix function $\Gamma(\cdot)$ and nonstochastic nonzero $k\times 1$ function $h(\cdot)$. For the groups chosen as above, we thus have by the Continuous Mapping Theorem

$$\sqrt{T}\begin{pmatrix} \hat\theta_1 - \theta \\ \hat\theta_2 - \theta \\ \vdots \\ \hat\theta_q - \theta \end{pmatrix} \Rightarrow \begin{pmatrix} (\int_0^{1/q}\Gamma(\lambda)d\lambda)^{-1}\int_0^{1/q} h(\lambda)dW(\lambda) \\ (\int_{1/q}^{2/q}\Gamma(\lambda)d\lambda)^{-1}\int_{1/q}^{2/q} h(\lambda)dW(\lambda) \\ \vdots \\ (\int_{(q-1)/q}^{1}\Gamma(\lambda)d\lambda)^{-1}\int_{(q-1)/q}^{1} h(\lambda)dW(\lambda) \end{pmatrix}$$

so that $\{\sqrt{T}(\hat\beta_j - \beta)\}_{j=1}^q$ are asymptotically independent and Gaussian. Therefore, whenever (16) holds, $t-$statistic based inference is asymptotically valid for any $q \geq 2$. The $t-$statistic approach can hence allow for asymptotically time varying information (nonconstant $\Gamma(\cdot)$) and pronounced stochastic volatility (nonconstant $h(\cdot)$). In contrast, the approach of Kiefer and Vogelsang (2002, 2005) requires $\Gamma(\cdot)$ and $h(\cdot)$ to be constant. There is substantial empirical evidence for persistent instabilities in the second moment of macroeconomic and financial time series: see, for instance, Bollerslev, Engle, and Nelson (1994), Kim and Nelson (1999), McConnell and Perez-Quiros (2000), Andersen, Bollerslev, Christoffersen, and Diebold (2007), and Müller and Watson (2007). The additional robustness of the $t-$statistic approach is thus arguably of practical relevance.

In fact, the $t-$statistic based approach suggested here is, to the best knowledge of the authors, the only known way of conducting asymptotically valid inference whenever (16) holds, as least under double-array asymptotics: Müller (2007b) demonstrates that in the scalar location model, for any equivariant variance estimator that is consistent for the variance of Gaussian white noise, there exists a double array that satisfies a functional central limit theorem which induces the "consistent" variance estimator to converge in probability to an arbitrary positive value. Since all usual consistent long-run variance estimators are both scale equivariant and consistent for the variance of Gaussian white noise, none of these estimators yields generally valid inference under (16). The general validity of the $t-$statistic approach under (16) is thus a further analytical reflection of its increased robustness, beyond the results of Sections 3.1 and 3.2.

Table 1 reports small sample properties of various approaches to inference. The small sample experiment is the one considered in Andrews (1991), Andrews and Monahan (1992) and Kiefer, Vogelsang, and Bunzel (2000) and concerns inference in a linear regression with 5 regressors. In addition to $t-$statistic based inference described above with $q = 2, 4, 8$ and 16 and groups $\mathcal{G}_j = \{t : (j-1)T/q < t \leq jT/q\}$, we include in our study the approach developed by Kiefer and Vogelsang (2005) and usual inference (as in (12)) based on two standard consistent long-run variance estimators. Specifically, we follow Kiefer and Vogelsang (2005) and focus on the quadratic spectral kernel estimator $\hat{\omega}^2_{QS}(b)$ and Bartlett kernel estimator $\hat{\omega}^2_{BT}(b)$ with bandwidths equal to a fixed fraction $b \leq 1$ of the sample size, with asymptotic critical values as provided by Kiefer and Vogelsang (2005) in their Table 1. For standard inference based on consistent long-run variance estimators, we include the quadratic spectral estimator $\hat{\omega}^2_{QA}$ with an automatic bandwidth selection using an AR(1) model for the bandwidth determination as suggested by Andrews (1991), and an AR(1) prewhitened long-run variance estimator $\hat{\omega}^2_{PW}$ with a second stage automatic bandwidth quadratic spectral kernel estimator as described in Andrews and Monahan (1992), where the critical values are those from a standard normal distribution.

As can be seen from Table 1, the $t-$statistic approach is remarkably successful at controlling size, the only instance of a moderate size distortion occurs in the AR(1) model with $\rho \geq 0.9$ and $q \geq 8$. This performance may be understood by observing that strong autocorrelations (which induce overrejection) co-occur with strong heterogeneity in the group design matrices (which induce underrejection) for the considered data generating process. In contrast, the tests based on the consistent estimators and the fixed-b asymptotic approach lead to much more severe overrejections.

For the computations of size adjusted power, the magnitude of the alternative was chosen to highlight differences. For moderate degrees of dependence, tests based on $\hat{\omega}^2_{QA}$ and $\hat{\omega}^2_{PW}$, as well as on $\hat{\omega}^2_{QS}(b)$ and $\hat{\omega}^2_{BT}(b)$ with b small have larger size corrected power than the $t-$statistic, with especially large differences for q small. On the other hand, the $t-$statistic approach can be substantially more powerful than any of the other tests

Table 1: Small Sample Results in a Time Series Regression with $T = 128$

	$t-$statistic (q)				$\hat{\omega}^2_{QA}$	$\hat{\omega}^2_{PW}$	$\hat{\omega}^2_{QS}(b)$			$\hat{\omega}^2_{BT}(b)$		
	2	4	8	16			0.05	0.3	1	0.05	0.3	1
ρ						Size, AR(1)						
-0.5	4.7	4.7	5.0	5.1	10.1	9.4	8.5	6.9	6.1	9.0	7.7	7.5
0	4.9	4.7	4.6	4.8	7.1	8.1	7.3	5.5	5.2	6.7	6.0	6.2
0.5	4.8	4.6	4.6	4.9	10.4	9.9	9.0	6.0	6.1	9.4	7.5	7.0
0.9	4.9	5.1	6.1	7.8	28.9	25.4	26.4	15.2	11.5	29.9	20.5	18.8
0.95	5.1	5.3	7.0	10.2	37.8	32.4	36.3	21.2	14.7	40.3	28.2	25.5
θ						Size, MA(1)						
-0.5	4.5	5.0	4.8	4.9	8.4	8.3	7.7	6.0	5.7	7.6	6.9	6.8
0.5	5.0	5.1	5.2	5.4	8.9	8.6	7.9	6.2	6.1	8.1	6.9	6.6
0.9	5.0	4.8	5.0	5.1	9.1	8.3	8.1	6.4	6.1	8.3	6.9	6.8
0.95	4.9	4.8	5.0	5.1	9.1	8.3	8.1	6.4	6.0	8.4	7.0	6.8
ρ						Size Adjusted Power, AR(1)						
-0.5	14.6	37.3	54.0	56.1	56.5	55.2	55.6	37.3	24.3	56.1	46.4	42.3
0	15.1	38.4	53.7	50.0	62.7	60.6	59.0	41.3	27.2	60.7	51.9	47.2
0.5	14.5	38.2	55.9	54.3	57.0	56.2	54.4	40.3	24.9	56.0	48.4	44.2
0.9	17.2	56.7	77.6	78.3	57.5	54.6	58.3	42.7	27.7	58.7	51.4	46.6
0.95	22.1	72.0	88.5	90.6	70.0	65.5	71.5	56.0	35.1	72.0	63.3	57.5
θ						Size Adjusted Power, MA(1)						
-0.5	17.6	45.2	64.9	62.4	71.4	70.2	68.9	49.0	31.6	70.5	59.7	53.6
0.5	16.5	44.5	63.3	64.8	69.3	67.2	67.1	47.1	28.0	68.5	57.7	53.2
0.9	15.5	42.8	60.4	63.3	65.0	63.6	63.1	43.3	26.8	64.1	53.6	49.4
0.95	15.7	42.8	60.4	63.3	64.8	63.6	63.0	43.3	27.0	64.0	53.4	49.1

Notes: The entries are rejection probabilities of nominal 5% level two-sided $t-$tests about the coefficient β of the first element of X_t in the linear regression $y_t = X_t'\theta + u_t$, $t = 1, \cdots, T$, where $X_t = (x_t', 1)'$, $x_t = (T^{-1}\sum_{s=1}^T \bar{x}_s\bar{x}_s')^{-1/2}\bar{x}_t$, $\bar{x}_t = \tilde{x}_t - T^{-1}\sum_{s=1}^T \tilde{x}_s$ and the elements of $\tilde{x}_t$ are four independent draws from a mean-zero, Gaussian, stationary AR(1) and MA(1) process of unit variance and common coefficients ρ and θ, respectively. The disturbances u_t are an independent draw from the same model as the (pretransformed) regressors, multiplied by the first element of X_t. Under the alternative, the difference between the true and hypothesized coefficient of interest was chosen as $4/\sqrt{T(1-\rho^2)}$ in the AR(1) model and as $5/\sqrt{T}$ in the MA(1) model. See text for description of test statistics. Based on 10,000 replications.

in highly dependent scenarios. Simulations for other forms of heteroskedasticity yield qualitatively similar results and are not reported for brevity.

The argument for the asymptotic normality of the group estimators $\hat{\beta}_j$, $j = 1, \cdots, q$, can fail if the underlying random variables are too heavy-tailed. For $\alpha \in (0, 2]$, denote by $\mathfrak{D}(\alpha)$ the domain of attraction of a strictly α-stable distribution. Then suitably scaled averages of i.i.d. random variables in $\mathfrak{D}(\alpha)$ converge in distribution to a strictly α-stable law, which coincides with the Gaussian law only for $\alpha = 2$. Heavy-tailed distributions have been used to model financial data (see, for instance, Loretan and Phillips (1994) and Rachev, Menn, and Fabozzi (2005) for a review and additional references), catastrophe risks and economic losses from natural disasters (see Ibragimov, Jaffee, and Walden (2006) and references therein).

Inference with heavy-tailed random variables is complicated by the fact that the rate of convergence and the limiting distribution depends on the unknown tail index α. These difficulties are compounded in the presence of additional time series dependence. To fix ideas, consider the problem of conducting inference on β in the scalar linear process

$$y_t = \beta + \sum_{j=-\infty}^{\infty} \psi_j S_{t-j}, \quad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty, \quad S_t \sim \text{i.i.d. in } \mathfrak{D}(\alpha) \tag{17}$$

so that $\beta = E[y_t]$ when $\alpha > 1$.

McElroy and Politis (2002) stress that few options are available for conducting inference on β in this model. They show that with $\alpha \in (1, 2]$, the usual $t-$statistic in y_t converges in distribution under the null hypothesis, and prove that under the additional assumption of y_t being strong mixing, subsampling can be employed to approximate the limiting distribution. Also see Kokoszka and Wolf (2004) for extensions of this approach to GARCH-type time series models.

The $t-$statistic approach outlined in Section 3.1 above provides an alternative, at least when S_t in (17) are in the domain of attraction of a *symmetric* stable law, which is weaker than the assumption of a symmetric distribution of S_t: For $q \geq 2$, let $\mathcal{G}_j = \{t : (j-1)T/q < t \leq jT/q\}$ as before, and define $\hat{\beta}_j = \sum_{t \in \mathcal{G}_j} y_t / \sum_{t \in \mathcal{G}_j} 1$. Lemma 1 of Avram and Taqqu (1992) about the joint convergence of the finite dimensional distributions of the partial sum process $m_T T^{-1} \sum_{t=1}^{\lfloor \cdot T \rfloor} (y_t - \beta)$ under (17) and $\alpha \in (0, 2]$ for some sequence m_T imply the convergence in distribution of $\{m_T(\hat{\beta}_j - \beta)\}_{j=1}^{q}$ to independent $\alpha-$stable symmetric random variables. Since symmetric stable distributions can be written as scale mixtures of normals, (7) holds, and the $t-$statistic approach is valid.

Table 2 reports the small sample performance of the subsampling and $t-$statistic approach in the same Monte Carlo design as considered in McElroy and Politis (2002).[11] As McElroy and Politis (2002), we find that the size control of the subsampling approach strongly depends on the choice of the subsample length b, on which there is little theoretical guidance. In contrast, the $t-$statistic based tests are mostly moderately undersized, with pronounced distortions only in the case of the MA process and a sample size of $T = 100$. We report non-size adjusted power because the $t-$statistic approach has asymptotic rejection probability under the null hypothesis below the nominal level in this data generating process, with potentially adverse effects on power. For the considered sample sizes, however, the $t-$statistic approach is typically more powerful than subsampling when both control size, at least for $q \geq 8$.

4.2 Panel Data

Many empirical studies in economics are based on observing N individuals repeatedly over T time periods, and correlations are possible in either (or both) dimensions. In applications, it is typically assumed that, possibly after the inclusion of fixed effects, one of the dimension is uncorrelated, and inference is based on consistent standard errors that allows for arbitrary correlation in the other dimension (Arellano (1987) and Rogers (1993)). The asymptotic validity of these procedures stems from an application of a law of large numbers across the uncorrelated dimension.[12] So if the uncorrelated dimension is small, one would expect these procedures to have poor finite sample properties, and our approach to inference is potentially attractive.

To fix ideas, consider a linear regression for the case where N is small and T is large

$$y_{i,t} = X'_{i,t}\theta + u_{i,t},\ i = 1, \cdots, N,\ \ t = 1, \cdots, T \tag{18}$$

[11] We were unable to replicate some of the Monte Carlo results reported in McElroy and Politis (2002). The differences do not alter any of the qualitative conclusions of McElroy and Politis (2002), however.

[12] Interestingly, Hansen (2007) shows that under weak regularity conditions, this remains true even if the dependent dimension is allowed to increase with the independent dimension in the underlying asymptotics.

Table 2: Small Sample Results in a Time Series Location Model with Symmetric α-Stable Disturbances

				$t-$statistic (q)				subsampled $t-$statistic (b)			
				2	4	8	16	2	4	8	16
DGP	α	T	β	Size							
MA	1.2	100	0	3.8	3.4	4.7	9.8	0.0	0.5	3.4	18.7
MA	1.2	1000	0	3.6	3.1	2.9	3.1	2.0	3.3	7.1	16.5
MA	1.8	100	0	4.9	5.0	6.7	12.1	0.0	0.1	1.6	9.4
MA	1.8	1000	0	4.9	4.7	5.0	5.0	0.1	0.1	0.2	0.9
AR	1.2	100	0	3.9	3.2	3.7	5.4	0.2	7.0	18.6	32.6
AR	1.2	1000	0	4.1	3.0	2.5	3.0	2.9	10.5	20.5	26.3
AR	1.8	100	0	4.7	4.4	5.3	6.9	0.0	1.1	6.5	14.9
AR	1.8	1000	0	4.7	4.6	4.8	4.8	0.1	0.3	1.9	4.8
DGP	α	T	β	Non-Size Adjusted Power							
MA	1.2	100	1.0	10.4	30.1	45.3	56.4	0.6	10.5	28.6	49.4
MA	1.2	1000	1.0	14.2	45.7	58.9	63.8	2.5	19.0	44.1	59.1
MA	1.8	100	0.4	14.2	38.4	58.9	73.5	0.0	2.9	20.6	49.2
MA	1.8	1000	0.2	19.0	56.8	76.0	82.0	0.1	0.1	15.8	51.1
AR	1.2	100	2.0	10.8	29.2	42.5	50.5	3.5	36.2	51.3	60.2
AR	1.2	1000	2.0	14.9	45.0	58.3	63.1	5.2	52.3	65.7	69.5
AR	1.8	100	0.8	14.2	37.7	55.4	65.6	0.1	24.1	51.2	66.4
AR	1.8	1000	0.4	18.7	57.2	75.6	81.3	0.1	34.6	70.0	78.6

Notes: Rejection probabilities of nominal 5% level two-sided tests about β in the model $y_t = \beta + u_t$, $t = 1, \cdots, T$, where $u_t = 0.5u_{t-1} + S_t$ and $u_0 = 0$ (AR) and $u_t = \sum_{j=0}^{10} \psi_j S_{t-j}$ with $\{\psi_j\}_{j=0}^{10} = \{.03, .05, .07, .1, .15, .2, .15, .1, .07, .05, .03\}$ (MA), and S_t are i.i.d. mean zero α-symmetric stable distributed. The subsampled $t-$statistic rejects if the full sample OLS $t-$statistic falls outside the 2.5% and 97.5% quantiles of the empirical distribution function of OLS $t-$statistics computed on all $T - b + 1$ consecutive subsamples of length b, as described in detail in McElroy and Politis (2002). Based on 10,000 replications.

where $\{X_{i,t}, u_{i,t}\}_{t=1}^{T}$ are independent across i and $E[X_{i,t}u_{i,t}] = 0$ for all i, t. Suppose that under $T \to \infty$ asymptotics with N fixed, $T^{-1}\sum_{t=1}^{T} X_{i,t}X_{i,t}' \xrightarrow{p} \Gamma_i$ and $T^{-1/2}\sum_{t=1}^{T} X_{i,t}u_{i,t} \Rightarrow \mathcal{N}(0, \Omega_i)$ for all i for some full rank matrices Γ_i and Ω_i. These assumptions are enough to guarantee that the OLS coefficient estimators $\hat{\beta}_i$ using data from individual $i = 1, \cdots, N$ only are asymptotically independent and Gaussian, so the $t-$statistic approach with $q = N$ groups is valid. Hansen (2007) derives a closely related result under "asymptotic homogeneity across i", that is if $\Gamma_i = \Gamma$ and $\Omega_i = \Omega$ for all i: in that case, the standard $t-$statistic for $\hat{\beta}$ based on the usual Rogers (1993) standard errors converges in distribution to a $t-$distributed random variable with $N - 1$ degrees of freedom, scaled by $\sqrt{N/(N-1)}$, under the null hypothesis. In fact, it is not hard to see that under asymptotic homogeneity across i, $\bar{\hat{\beta}}$ and $s_{\hat{\beta}}$ in (6) of our approach are first order asymptotically equivalent to $\hat{\beta}$ and the appropriately scaled Rogers (1993) standard error under the null and local alternatives, so both approaches have the same asymptotic local power. The advantage of our approach is that it does not require asymptotic homogeneity to yield valid inference.

Table 3 provides some small sample evidence for the performance of these two approaches, with the same data generating process as considered by Kézdi (2004), with an AR(1) in both the regressor and the disturbances. Since $\hat{\beta}_i$, conditionally on $\{X_{i,t}\}$, is Gaussian with mean β, the $t-$statistic approach is exactly small sample conservative for this DGP. Hansen's (2007) asymptotic result is formally applicable for $|\rho_x| < 1$ and $|\rho_u| < 1$, as this DGP then is asymptotically homogeneous in the sense defined above. With a unit root in the regressors, however, $T^{-2}\sum_{t=1}^{T} X_{i,t}X_{i,t}'$ does not converge to the same limit across i, so that despite the i.i.d. sampling across i, asymptotic homogeneity fails. These asymptotic considerations successfully explain the size results in Table 3. The $t-$statistic approach has higher size adjusted power for heteroskedastic disturbances, but this is not true under homoskedasticity.

For panel applications in finance with individuals that are firms, it is often the cross-section dimension for which uncorrelatedness is an unattractive assumption (see Petersen (2005) for an overview of popular standard error corrections in finance). As noted in the introduction, if one is willing to assume that there is no time series correlation, which is empirically plausible at least for stock returns, then our approach with time periods as

Table 3: Small Sample Results in a Panel with $N = 10$, $T = 50$ and Time Series Correlation

	homoskedastic					heteroskedastic				
ρ_x	0	0.5	0.9	0.9	1	0	0.5	0.9	0.9	1
ρ_u	0	0.5	0.5	0.9	0.5	0	0.5	0.5	0.9	0.5
					Size					
$t-$statistic	5.0	4.9	5.3	5.0	4.4	4.6	4.6	4.4	4.0	3.8
clustered	5.2	5.3	6.5	7.2	8.7	4.9	5.5	7.7	7.9	14.7
clustered, FE	5.1	5.1	6.2	6.2	6.8	4.9	5.3	6.4	6.2	8.6
					Size Adjusted Power					
$(\beta - \beta_0)/\sqrt{nT}$	2.5	3	8	0.8	12	5	8	8	25	180
$t-$statistic	58.7	53.9	56.7	39.4	42.5	54.6	64.7	62.0	62.9	73.7
clustered	60.8	55.0	71.5	31.9	83.6	51.9	56.6	49.1	33.6	52.1
clustered, FE	59.7	55.1	60.8	38.3	50.8	52.0	58.2	48.9	46.2	45.1

Notes: The entries are rejection probabilities of nominal 5% level two-sided $t-$tests about the coefficient β of $x_{i,t}$ in the linear regression $y_{i,t} = X_{i,t}'\theta + u_{i,t}$, $i = 1, \cdots, N$, $t = 1, \cdots, T$, where $X_{i,t} = (x_{i,t}, 1)'$, $x_{i,t} = \rho_x x_{i,t-1} + \varepsilon_{i,t}$, $x_{i,0} = 0$, $\varepsilon_{i,t} \sim i.i.d.\ \mathcal{N}(0,1)$, $u_{i,t} = \rho_u u_{i,t-1} + \eta_{i,t}$, $u_{i,0} = 0$, where under homoskedasticity, $\eta_{i,t} \sim i.i.d.\ \mathcal{N}(0,1)$ independent of $\{\varepsilon_{i,t}\}$, and under heteroskedasticity, $\eta_{i,t} = (0.5 + 0.5x_{i,t}^2)\tilde{\eta}_{i,t}$ and $\tilde{\eta}_{i,t} \sim i.i.d.\ \mathcal{N}(0,1)$ independent of $\{\varepsilon_{i,t}\}$. The considered tests are the $t-$statistic approach with groups defined by individuals ("$t-$statistic"); OLS coefficient based tests with Rogers (1993) standard errors ("clustered"); and OLS coefficient based test which includes individual Fixed Effects and Arellano (1987) standard errors ("clustered, FE"). The critical value for the clustered test statistic was chosen from the appropriate quantile of a Student$-t$ distribution with $N - 1$ degrees of freedom, scaled by $\sqrt{N/(N-1)}$. Based on 10,000 replications.

groups becomes the so-called Fama-MacBeth approach: Estimate the model of interest for each time period j cross sectionally to obtain $\hat{\beta}_j$, and compute the usual $t-$statistic for the resulting $q = T$ coefficient estimates. Our results formally justify this approach for T small and possible heterogeneity in the variances of $\hat{\beta}_j$. Note that the variances may be stochastic and dependent even in the limit, as in (7), which would typically arise when regression errors follow a stochastic volatility model with some common volatility component.

In corporate finance applications, or with overlapping long-term returns as dependent variable, one would typically not want to rule out additional dependence in the time dimension. Under the assumption that the correlation dies out over time, one could try to non-parametrically estimate the long-run variance of the sequence $\{\hat{\beta}_j\}_{j=1}^T$ using, say, the Newey and West (1987) estimator. However, this will require a long panel (T large) to yield reasonable inference. Our results suggest an alternative approach: Divide the data in fewer groups that span several consecutive time periods. For instance, with $T = 24$ yearly sampling frequency, one might form 8 groups of 3 year blocks, or, more conservatively, 4 groups of 6 year blocks. If the time series correlation is not too pronounced, then parameter estimators from different groups will have little correlation, and the $t-$statistic approach yields approximately valid inference.

We investigate the empirical performance of this approach for two data generating processes considered by Thompson (2006) for $N = 50$ and $T = 25$, both of which generate some dependence in both dimensions. As noted by Thompson (2006), an approach that clusters in both dimensions (also see Cameron, Gelbach, and Miller (2006)) has poor small sample properties for these values of N and T, even in absence of any time series correlation ($\rho = 0$). In contrast, the $t-$statistic approach has reasonable size control as long as the time series dependence is not extreme ($\rho = 0.9$), and has favorable size control properties compared to parametric or non-parametric corrections to the Fama-MacBeth approach. Unreported results show that this remains true also under the inclusion of additional fixed effects in either or both dimensions. As can be seen from Table 4, these advantages in size control of the $t-$statistic approach come at a certain cost in size adjusted power, though, especially for q small. The higher power of the Fama-MacBeth approaches when ρ is small stems from the inherent time fixed effect in that estimator; the other approaches have

Table 4: Small Sample Results in a Panel with $N = 50$, $T = 25$ and Correlation in Both Dimensions

	Individual Persistence				Common Persistence			
ρ	0	0.5	0.7	0.9	0	0.5	0.7	0.9
				Size				
$t-$statistic $q = 2$	4.9	5.3	5.0	6.0	4.9	5.1	5.3	6.3
$t-$statistic $q = 4$	4.9	5.2	5.4	9.8	4.1	5.0	5.3	10.4
$t-$statistic $q = 8$	4.6	5.3	6.4	17.1	3.9	4.9	7.1	16.8
Fama-MacBeth with Newey-West	12.6	13.6	19.8	34.8	11.4	12.3	14.2	23.4
Fama-MacBeth with AR(1) corr.	9.6	9.9	13.5	22.5	8.8	9.1	10.4	18.0
cluster by i and t	9.3	8.9	8.8	7.0	10.2	19.0	29.9	49.5
cluster by i and t + common pers.	16.3	16.2	14.9	12.1	17.0	21.3	26.4	38.3
				Size Adjusted Power				
$(\beta - \beta_0)/\sqrt{nT}$	7	7	7	7	25	30	30	45
$t-$statistic $q = 2$	12.9	13.0	16.2	14.9	20.3	20.9	18.1	20.8
$t-$statistic $q = 4$	30.5	35.2	45.5	45.3	58.4	63.5	58.4	60.6
$t-$statistic $q = 8$	50.9	57.7	67.6	61.3	59.5	73.2	67.3	68.2
Fama-MacBeth with Newey-West	100	99.5	91.6	58.9	57.3	58.4	47.4	47.1
Fama-MacBeth with AR(1) corr.	100	99.2	88.8	51.4	57.6	60.2	46.7	44.8
cluster by i and t	46.8	55.4	67.6	74.8	86.3	83.4	66.2	70.2
cluster by i and t + common pers.	31.7	39.4	52.7	69.8	69.6	70.0	53.6	60.1

Notes: The entries are rejection probabilities of nominal 5% level two-sided $t-$tests about the coefficient β of $x_{i,t}$ in the linear regression $y_{i,t} = X_{i,t}'\theta + u_{i,t}$, $i = 1, \cdots, N$, $t = 1, \cdots, T$, where $X_{i,t} = (x_{i,t}, 1)'$. The DGPs correspond to Panels B and C of Thompson (2006), where under "Individual Persistence", $u_{i,t} = \xi_t + \eta_{i,t}$, $\eta_{i,t} = \rho\eta_{i,t-1} + \varepsilon_{i,t}$, $\eta_{i,0} = 0$, ξ_t and $\varepsilon_{i,t}$ are mutually independent and distributed i.i.d. $\mathcal{N}(0,1)$, and under "Common Persistence" $u_{i,t} = h_i f_t + \varepsilon_{i,t}$, $f_t = \rho f_{t-1} + \xi_t$, $f_0 = 0$, and the disturbances are mutually independent and $\varepsilon_{i,t} \sim i.i.d.\ \mathcal{N}(0, 0.01)$, $h_i \sim i.i.d.\ \mathcal{N}(1, 0.25)$, $\xi_t \sim i.i.d.\ \mathcal{N}(0,1)$. In both cases, $x_{i,t}$ is an independent draw of the same distribution as $u_{i,t}$ (with the same h_i under common persistence). The considered tests are: the $t-$statistic approach with groups $\mathcal{G}_j = \{(i,t) : (j-1)T/q < t \leq jT/q\}$; Fama-MacBeth standard errors with a Newey West correction with 5 lags; Fama-MacBeth standard errors multiplied by $\sqrt{(1+\hat{\rho})/(1-\hat{\rho})}$, where $\hat{\rho}$ is the first order autocorrelation coefficient of $\hat{\beta}_j$, $j = 1, \cdots, T$ (see Fama and French (2002)); and inference based on clustering in both dimensions as suggested in Thompson (2006), where in the "+ common pers."-row, the clustering allows for a persistence common shock with lag length 2. For all approaches other than the $t-$statistic, critical values from a standard normal were employed. Based on 10,000 replications.

similar size adjusted power when time fixed effects are included (see Petersen (2005) for similar results on efficiency of alternative estimators).

If a panel is very short and potential autocorrelations are large, then it might be more appealing to assume some independence in the cross section. For instance, in finance applications, one might be willing to assume that there is little correlation between firms of different industries, as in Froot (1989). Under this assumption, one could collect all firms of the same industry in the same group to obtain as many groups as there are different industries. If the parameter of interest is a regression coefficient of a regressor that varies within industry, then one could add time fixed effects in each group to guard against inter-industry correlation from a yearly common shock that is independent of the other regressors. Alternatively, one can also combine independence assumptions in both dimensions by, say, forming twice as many groups as there are industries by splitting each industry group into two depending on whether $t < T/2$ or not. The theoretical results in Section 3.4 suggest that there are substantial gains in power (more than 10% for 5% level tests) of such an additional independence assumption as long as $q \leq 8$. Similar possibilities of group formation might be attractive for long-run performance evaluations in finance (see, for instance, Jegadeesh and Karceski (2004) for a discussion of inference based on consistent variance estimation that could be easily adapted for the $t-$statistic approach), and panel analyses with individuals as countries and trade-blocks or continents as one group dimension.

Recently, Bertrand, Duflo, and Mullainathan (2004) have also stressed the importance of allowing for time series correlation in panel difference-in-difference applications. This technique is popular to estimate causal effects, and it is usually implemented by a linear regression (18) with fixed effects in both dimensions. In a typical application, the individuals $i = 1, \cdots, N$ are U.S. states, and the coefficient of interest β multiplies a binary regressor that describes some area specific intervention, such as the passage of a law. Donald and Lang (2004) show that if $u_{i,t}$ has an i.i.d. Gaussian random effect structure for each (potential) pre- and post-intervention area group, then correct inference is obtained for fixed N by a two stage inference procedure using a Student$-t$ critical value with an appropriate degrees of freedom correction. See Wooldridge (2003) for further discussion, and Conley and Taber (2005) for a possible approach when only few states were subject to the intervention, but many others were not. With the time fixed effects, it is obviously not possible to apply the $t-$statistic approach with groups defined as states. However, by collecting states into groups defined as larger geographical areas so that at least one of the states in each group was subject to the intervention, it again becomes possible to obtain estimators $\hat{\beta}_j$, $j = 1, \cdots, q$, from each group and to apply the $t-$statistic approach. This leads to a loss of degrees of freedom, but it has the advantage of yielding correct inference when the pre- and post-intervention specific random effects in $u_{i,t}$ are independent, but not necessarily identically distributed scale mixture of normals. This is a considerable weakening of the homogeneous Gaussian assumption required for the approach of Donald and Lang (2004). Also, if larger geographical areas are formed by collecting neighboring states, the $t-$statistic approach becomes at least partially robust to moderate spatial correlations. When the number of states in each of these larger areas is not too small (which for the U.S. then implies a relatively small q), one might appeal to the central limit theorem to justify the $t-$statistic approach to inference when the underlying random effects cannot be written as scale mixtures of normals.

4.3 Clustered Data

A further potential application of our approach is to draw inferences about a population based on a two-stage (or multi-stage) sampling design with a small number of independently sampled primary sampling units (PSUs). PSUs could be villages in a development study (see, for instance, Deaton (1997), Chapters 1.4 and 2.2), or a small number of, say, city blocks in a large metropolitan area. One would typically expect that observations from the same PSU are more similar than those from different PSUs, which necessitates a correction of the standard errors. Note that PSUs are independent by sample design, so with PSUs as groups $j = 1, \cdots, q$, the only additional requirement of our approach is that the parameter of interest can be estimated by an approximately Gaussian and unbiased estimator $\hat{\beta}_j$ from each PSU, $j = 1, \cdots, q$. Of course, this will only be possible if the parameter of interest is identified in each PSU; in a regression context, a coefficient about regressors that only vary across PSUs cannot be estimated from one PSU only, as long as the regression contains a constant. In such cases, our approach is still applicable by collecting more than one PSU in each group.

As a stylized example, imagine a world where the only spacial correlation between household characteristics in the population arises through the fact that households in the same neighborhood are very similar to each other, and villages consist of, say, 30-80 neighborhoods. Consider a two stage sample design with a simple random sample of 400 households within 12 villages as PSUs. Sample means $\hat{\beta}_j$ of household characteristics of a single PSU are then approximately Gaussian with a mean that is equal to the national average β, and a variance that is a function of the number of neighborhoods. This variance is larger than that of a national simple random sample of the same size, so ignoring the clustering leads to incorrect inference, while our approach is approximately correct. What is more, our derivations in Section 3.4 above show the $t-$statistic approach results in a small loss in power only compared to inference based on the overall sample average with known variance, regardless whether or not there is indeed this neighborhood-type spatial correlation in the population.

In some instances, it will be more appropriate to assume that all individuals from the same PSU are similar—think of the extreme case where all households in the same village are identical. In this case, there is no equivalent to the averaging over the neighborhoods, and one cannot appeal to the central limit theorem to argue for the approximation $\hat{\beta}_j \sim \mathcal{N}(\beta, v_j^2)$. This set-up would naturally lead to a random parameter model, where the household characteristic β_j in PSU j is a random draw from the national distribution. In a slightly more general regression context, this leads to the random coefficient regression model (cf., for instance, Swamy (1970))

$$Y_{i,j} = X_{i,j}'\theta_j + u_{i,j} = X_{i,j}'\theta + X_{i,j}'(\theta_j - \theta) + u_{i,j}$$

for individual $i = 1, \cdots, n_j$ in PSU $j = 1, \cdots, q$, $E[X_{i,j}u_j] = 0$ and θ_j are i.i.d. draws from some population with mean θ. Thought of as part of the disturbance term, $X_{i,j}'(\theta_j - \theta)$ induces intra-PSU correlations. Now under sufficient regularity conditions, $\hat{\theta}_j - \theta_j \xrightarrow{p} 0$ as $n_j \to \infty$, and our approach for inference about β (the first element of θ), remains valid as long as the distribution of β_j can be written as a scale mixture of normals. This is a wide class of distributions, as noted in Section 2 above. If n_j is not large enough to make $\hat{\theta}_j - \theta_j \xrightarrow{p} 0$ a good approximation, but instead $\hat{\beta}_j|\beta_j \sim i.i.d.\ \mathcal{N}(\beta_j, v_j^2)$, then the unconditional distribution of $\hat{\beta}_j$ is given by a convolution of the distribution of β_j and mean zero normal, which can be written a scale mixture of normals if the distribution of β_j is one, so our approach remains applicable.

The need for clustering might arise in a more subtle way depending on the relationship between the sampling scheme and the population of interest. For example, suppose we want to study labor supply based on a large i.i.d. sample from U.S. households, which are located in, say, 12 different regions. Similar to the example above, assume that each region consists of, say, 30-80 different metropolitan and rural areas, and that the characteristics of these areas induce similar behavior of households, so that there are effectively about 500 different types of households. Of course, in a large sample, we will have many observations from the same area, which are quite similar to each other. Nevertheless, the usual (small) standard errors, based on the total number of observations, are applicable by definition of an i.i.d. sample for statements about labor supply in the current U.S. population. But if the study's results are to be understood as generic statements about labor supply, then the relevant population becomes households in all kinds of circumstances, and the i.i.d. sample from U.S. households is no longer i.i.d. in this larger population. Instead, it makes sense to think of the 12 regions as independently sampled PSUs of this superpopulation, and apply our approach with the regions as groups. As pointed out by Moulton (1990), ignoring this clustering often leads to very different results.

4.4 Spatially Correlated Data

Inference with spatially correlated data is usually justified by a similar reasoning as with time series observations: more distant observations are less correlated. With enough assumptions on the rate of decay of correlation as a function of their distance, consistent parametric and nonparametric variance estimators of spatially correlated data can be derived—see Case (1991) and Conley (1999). "Distance" here can mean physical distance between geographical units (country, county, city and so forth), but may also be thought of as distance in some economic sense. Conley and Topa (2002), for instance, considers spatial correlation as a function of socioeconomic distance, and Conley and Dupor (2003) uses metrics based on input-output relations as to measure the distance different sectors of the U.S. economy.

For the $t-$statistic approach suggested here, an assumption of correlations decaying as a function of distance suggests constructing the q groups out of blocks of neighboring observations. If the groups are carefully chosen, then under asymptotics where there are more and more observations in each of the q groups, most observations are sufficiently far away from the "borders". The variability of the group estimators is thus dominated by observations that are essentially uncorrelated with observations from other groups. Furthermore, the averaging within each group yields asymptotic Gaussianity for each $\hat{\beta}_j$, so that under sufficiently strong regularity conditions, the $t-$statistic based inference is valid.

We investigate the relative performance of the $t-$statistic approach and inference based on consistent variance estimators in a Monte Carlo exercise as follows: We are interested in conducting inference about the mean β of $n = 128$ observations which are located on a rectangular array of unit squares with 8 rows and 16 columns (two checker boards side by side). The observations are generated such that in the Gaussian case, the correlation of two observations is given by $\exp(-\phi d)$ for some $\phi > 0$, where d is the Euclidian distance between the two observations. We also consider disturbances with a mean corrected chi-squared distribution with one degree of freedom. As can be seen from Table 5, the $t-$statistic approach is more successful at controlling size than inference based on the consistent variance estimators. The asymmetry in the error distribution has only a relatively minor impact on size control. Size corrected power of the $t-$statistics increases in q, but is always smaller than the size corrected power of tests based on nonparametric spatial consistent variance estimators suggested by Conley (1999) with a small bandwidth $b \leq 2$, which includes the OLS variance estimator as a special case.

5 Conclusions

The paper develops a general strategy to deal with inference about a scalar parameter in data with pronounced correlations of largely unknown form. The key assumption is that it is possible to partition the data into q groups, such that estimators based on data from group j, $j = 1, \cdots, q$, are approximately independent and normal, but not necessarily of equal variance. As long as there are no pronounced common shocks and the group sizes are not too small, the normality assumption seems rather weak, as the central limit

Table 5: Small Sample Results in a Location Problem with Spatial Correlation, $n = 128$

	$t-$statistic (q)				$\hat{\omega}^2_{UA}(b)$				$\hat{\omega}^2_{WA}(b)$		
	2	4	8	16	0	2	4	8	2	4	8
					Size, Gaussian Errors						
$\phi = \infty$	5.0	5.0	5.1	5.1	5.5	6.2	7.9	13.2	8.1	14.9	19.1
$\phi = 2$	5.1	5.4	5.9	7.5	15.1	11.0	10.4	15.8	8.0	14.9	21.0
$\phi = 1$	5.6	7.5	10.6	16.9	39.8	26.4	19.6	22.8	16.5	17.4	25.0
					Size, Mean Corrected Chi-Squared Errors						
$\phi = \infty$	5.0	5.4	5.7	6.3	6.5	7.1	8.4	13.4	8.8	14.7	19.1
$\phi = 2$	5.5	6.5	7.0	8.0	13.5	10.9	11.1	16.2	9.5	16.0	21.7
$\phi = 1$	5.7	9.5	12.8	17.9	35.3	25.8	20.6	23.8	17.9	19.5	26.9
					Size Adjusted Power, Gaussian Errors						
$\phi = \infty$	15.4	40.0	56.8	64.1	68.8	67.7	65.1	60.8	62.5	41.1	31.3
$\phi = 2$	15.6	43.0	59.0	67.5	71.3	70.8	67.8	62.4	68.1	46.2	31.8
$\phi = 1$	15.4	41.1	57.1	64.0	69.6	67.7	63.9	57.8	66.4	49.7	30.8
					Size Adjusted Power, Mean Corrected Chi-Squared Errors						
$\phi = \infty$	15.5	34.8	52.0	60.3	67.8	67.0	63.0	58.8	59.7	36.2	29.4
$\phi = 2$	14.1	36.9	59.8	69.5	76.9	75.3	70.8	63.3	70.3	43.0	31.3
$\phi = 1$	15.2	35.7	52.3	64.7	79.3	72.4	62.2	53.3	65.0	39.4	28.4

Notes: The entries are rejection probabilities of nominal 5% level two-sided $t-$tests about β in the model $y_{i,j} = \beta + u_{i,j}$, $i = 1, \cdots, 8$, $j = 1, \cdots, 16$. Under Gaussian errors, $u_{i,j}$ are multivariate mean zero unit variance Gaussian with correlation between $u_{i,j}$ and $u_{l,k}$ given by $\exp(-\phi\sqrt{(i-l)^2 + (j-k)^2})$, and the mean corrected chi-squared errors were generated by $u_{i,j} = \Phi^{-1}_{\chi^2-1}(\Phi(\tilde{u}_{i,j}))$, where $\tilde{u}_{i,j}$ are the Gaussian model disturbances, Φ is the cdf of a standard normal and $\Phi^{-1}_{\chi^2-1}$ is the inverse of the cdf of a mean corrected chi-squared random variable. The considered tests are the $t-$statistic approach with groups of spatial dimension 8×8, 8×4, 4×4 and 2×4, at the obvious locations; and inference based on $\bar{y} = n^{-1} \sum_{i=1}^{8} \sum_{j=1}^{16} y_{i,j}$ with two versions of Conley's (1999) nonparametric spatial consistent variance estimators of bandwidth b: a simple average $\hat{\omega}^2_{SA}(b)$ of all cross products of $(y_{i,j} - \bar{y})(y_{k,l} - \bar{y})$, $i, k = 1, \cdots, 8$, $j, l = 1, \cdots, 16$, of Euclidian distance $d \leq b$, and a weighted average $\hat{\omega}^2_{WA}(b)$ of these cross products, with weights $w(i,j,k,l) = \mathbf{1}[\tilde{w}(i,j,k,l) > 0]\tilde{w}(i,j,k,l)$ and $\tilde{w}(i,j,k,l) = (1 - |i-k|/b)(1 - |j-l|/b)$ (cf. equation (3.14) of Conley (1999)). Alternatives where chosen as $\beta - \beta_0 = c/\sqrt{n}$ with $c = 2.5$, 3.4, 5.7 under Gaussian disturbances and $c = 3.5$, 4.7, 8 under chi-squared errors for $\phi = \infty$, 2, 1 respectively. Based on 10,000 replications.

theorem provides good approximations even for small samples as long as the underlying observations are not very fat-tailed or skewed.

The crucial assumption is the approximate independence of the q estimators, and in applications, it will be challenging to decide on an adequate number and composition of groups. As formally discussed in Section 3.2, it is impossible to delegate that decision to the data. On a fundamental level, some *a priori* knowledge about the correlation structure is required in order to be able to learn from the data. This is also true of other approaches to inference, although the assumed regularity tends to be more implicit. For instance, consider the problem of conducting inference about the mean real exchange rate in 40 years of quarterly data. It seems quite difficult to have a substantive discussion about the appropriateness of, say, a confidence interval based on Andrews' (1991) consistent long-run variance estimator (whose formal validity is based on primitive conditions involving mixing conditions and the like), or, for that matter, on Kiefer and Vogelsang's (2005) approach with a bandwidth of, say, 30% of the sample size. At the same time, it seems at least conceivable to debate whether averages from, say, 8 year blocks provide approximately independent information; business cycle frequency fluctuations of the real exchange rate, for instance, would rule out the appropriateness of 4 year blocks.

In our view, it is a strength of the $t-$statistic approach that it requires such an explicit statement of what drives the validity of inference. At the end of the day, inference requires some assumption about potential correlations, and to agonize about the appropriate amount of regularity is precisely what researchers in the field should be doing. The $t-$statistic approach offers simple and in some sense efficient inference for one general type of regularity condition.

6 Appendix

Proof of Theorem 4:

(i) For the first claim, let $\Gamma_1 = \xi - (q-1)$, $\Omega_1 = 1$ and $\Gamma_j = \Omega_j = 1$ for $j = 2, \cdots, q$ for some $\xi > 0$. Then $\bar{\Sigma}_q/\Sigma_q = \xi^2(q-1+(1-q+\xi)^{-2})/q^2$, so that $\bar{\Sigma}_q/\Sigma_q \to 0$ as $\xi \to 0$.

For the second claim, let $\Gamma_1 = \Omega_1 = I_k$, and $\Omega_j = \xi I_k$, $\Gamma_j = \varsigma I_k$ for $j = 2, \cdots, q$, for some $\varsigma > 0$, $\xi > 0$, so that $\sum_{j=1}^q \Gamma_j = ((q-1)\varsigma + 1)I_k$. Then

$$\Sigma_q = \frac{\xi(q-1)+1}{((q-1)\varsigma+1)^2} I_k \quad \text{and} \quad \bar{\Sigma}_q = \frac{1+(q-1)\xi/\varsigma^2}{q^2} I_k.$$

Letting $\xi = 1$ and $\varsigma \to 0$ proves the second claim, and with $\xi = \varsigma^4$ and $\varsigma \to 0$ we find $\iota'\bar{\Sigma}_q\iota/\iota'\Sigma_q\iota \to 1/q^2$.

Also, for $k \geq 2$, let $\Gamma_1 = \operatorname{diag}(A, I_{k-2}) \in \mathcal{P}_k$ with $A = ((1, \frac{1}{2})', (\frac{1}{2}, 1)')$, $\Omega_1 = \Gamma_1 \operatorname{diag}(1, \xi, I_{k-2})\Gamma_1$ and $\Gamma_j = \Omega_j = I_k$ for $j = 2, \cdots, q$. Then

$$\iota'\bar{\Sigma}_q\iota = 1/q \quad \text{and} \quad \iota'\Sigma_q\iota = \frac{-3-4q+16q^3+4\xi(q-1)^2}{(1-4q^2)^2}$$

so that $\iota'\bar{\Sigma}_q\iota/\iota'\Sigma_q\iota \to 0$ as $\xi \to \infty$.

We are thus left to show that for $k = 1$, $\bar{\Sigma}_q/\Sigma_q \geq 1/q^2$ for all positive numbers $\{\Gamma_j\}_{j=1}^q$ and nonnegative numbers $\{\Omega_j\}_{j=1}^q$. But

$$\Sigma_q = \left(\sum_{j=1}^q \Gamma_j\right)^{-2} \sum_{j=1}^q \Omega_j \leq \left(\sum_{j=1}^q \Gamma_j^2\right)^{-1} \sum_{j=1}^q \Omega_j \leq \sum_{j=1}^q \Gamma_j^{-2}\Omega_j = q^2\bar{\Sigma}_q.$$

(ii) Note that for any real full column rank matrix X, $X(X'X)^{-1}X'$ is idempotent, so that $I - X(X'X)^{-1}X'$ is positive semidefinite. Therefore, for any real matrix Y of suitable dimension, $Y'Y - Y'X(X'X)^{-1}X'Y$ is positive semidefinite.

For the first claim, let $\bar{\Omega}_j = \Gamma_j\Gamma_j'$. Then $\bar{\Sigma}_q = q^{-1}I_k$, and $\Sigma_q = \left(\sum_{j=1}^q \Gamma_j\right)^{-1}\left(\sum_{j=1}^q \Gamma_j\Gamma_j'\right)\left(\sum_{j=1}^q \Gamma_j'\right)^{-1}$. It suffices to show that $\Sigma_q^{-1} - \bar{\Sigma}_q^{-1}$ is negative semidefinite, and this follows from the above result with $Y = (I_k, \cdots, I_k)'$ and $X = (\Gamma_1, \cdots, \Gamma_q)'$.

For the second claim, let $\underline{\Omega}_j = \Gamma_j$. Then $\bar{\Sigma}_q = q^{-2}\sum_{j=1}^q \Gamma_j^{-1}$ and $\Sigma_q = \left(\sum_{j=1}^q \Gamma_j\right)^{-1}$, and the result follows by setting $Y = (\Gamma_1^{-1/2}, \cdots, \Gamma_q^{-1/2})'$ and $X = (\Gamma_1^{1/2}, \cdots, \Gamma_q^{1/2})$.

(iii) Immediate from $\bar{\Sigma}_q = q^{-2}\Gamma\left(\sum_{j=1}^q \Omega_j\right)\Gamma'$ and $\sum_{j=1}^q \Gamma_j = q\Gamma$.

References

Andersen, T., T. Bollerslev, P. Christoffersen, and F. X. Diebold (2007): *Volatility: Practical Methods for Financial Applications*. Princeton University Press, forthcoming, Princeton.

Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858.

Andrews, D. W. K., and J. C. Monahan (1992): "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 60, 953–966.

Arellano, M. (1987): "Computing Robust Standard Errors for Within-Groups Estimators," *Oxford Bulletin of Economics and Statistics*, 49, 431–434.

Avram, F., and M. S. Taqqu (1992): "Weak Convergence of Sums of Moving Averages in the α-stable Domain of Attraction," *Annals of Probability*, 20, 483–503.

Bakirov, N. K. (1989): "The Extrema of the Distribution Function of Student's Ratio for Observations of Unequal Accuracy are Found," *Journal of Soviet Mathematics*, 44, 433–440.

Bakirov, N. K., and G. J. Székely (2005): "Student's T-Test for Gaussian Scale Mixtures," *Zapiski Nauchnyh Seminarov POMI*, 328, 5–19.

Benjamini, Y. (1983): "Is the T Test Really Conservative When the Parent Distribution is Long-Tailed?," *Journal of the American Statistical Association*, 78, 645–654.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How Much Should We Trust Differences-in-Differences Estimates?," *The Quarterly Journal of Economics*, 119, 249–275.

Bollerslev, T., R. F. Engle, and D. B. Nelson (1994): "ARCH Models," in *Handbook of Econometrics Vol. IV*, ed. by R. F. Engle, and D. McFadden. Elsevier Science, Amsterdam.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2006): “Robust Inference with Multi-Way Clustering,” *NBER Technical Working Paper 327.*

Case, A. C. (1991): “Spatial Patterns in Household Demand,” *Econometrica*, 59, 953–965.

Conley, T. G. (1999): “GMM Estimation with Cross Sectional Dependence,” *Journal of Econometrics*, 92, 1–45.

Conley, T. G., and B. Dupor (2003): “A Spatial Analysis of Sectoral Complementarity,” *Journal of Political Economy*, 111, 311–352.

Conley, T. G., and C. Taber (2005): “Inference with "Difference in Differences" with a Small Number of Policy Changes,” *NBER Technical Working paper No. 312.*

Conley, T. G., and G. Topa (2002): “Socio-Economic Distance and Spatial Patterns in Unemployment,” *Journal of Applied Econometrics*, 17, 303–327.

Deaton, A. (1997): *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy.* John Hopkins University Press, Baltimore.

Donald, S., and K. Lang (2004): “Inference with Difference in Differences and Other Panel Data,” *Working paper, University of Texas.*

Fama, E. F., and K. R. French (2002): “Testing Trade-Off and Pecking Order Predictions About Dividends and Debt,” *The Review of Financial Studies*, 15, 1–33.

Fama, E. F., and J. MacBeth (1973): “Risk, Return and Equilibrium: Empirical Tests,” *Journal of Political Economy*, 81, 607–636.

Ferguson, T. S. (1967): *Mathematical Statistics — A Decision Theoretic Approach.* Academic Press, New York and London.

Froot, K. A. (1989): “Consistent Covariance Matrix Estimation with Cross-Sectional Dependence and Heteroskedasticity in Financial Data,” *The Journal of Financial and Quantitative Analysis*, 24, 333–355.

Hansen, C. B. (2007): "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When T is Large," *Journal of Econometrics*, 141, 597–620.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Ibragimov, R., D. Jaffee, and J. Walden (2006): "Non-diversification traps in catastrophe insurance markets," *Working paper, Harvard University and University of California at Berkeley.*

Imhof, J. P. (1961): "Computing the Distribution of Quadratic Forms in Normal Variables," *Biometrika*, 48, 419–426.

Jansson, M. (2004): "The Error in Rejection Probability of Simple Autocorrelation Robust Tests," *Econometrica*, 72, 937–946.

Jegadeesh, N., and J. Karceski (2004): "Long-Run Performance Evaluation: Correlation and Heteroskedasticity-Consistent Tests," *Working paper, University of Florida.*

Kézdi, G. (2004): "Robust Standard Error Estimation in Fixed-Effects Panel Models," *Hungarian Statistical Review*, 9, 95–116.

Kiefer, N., and T. J. Vogelsang (2002): "Heteroskedasticity-Autocorrelation Robust Testing Using Bandwidth Equal to Sample Size," *Econometric Theory*, 18, 1350–1366.

——— (2005): "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests," *Econometric Theory*, 21, 1130–1164.

Kiefer, N. M., T. J. Vogelsang, and H. Bunzel (2000): "Simple Robust Testing of Regression Hypotheses," *Econometrica*, 68, 695–714.

Kim, C.-J., and C. R. Nelson (1999): "Has the Economy Become More Stable? A Bayesian Approach Based on a Markov-Switching Model of the Business Cycle," *The Review of Economics and Statistics*, 81, 608–616.

Kokoszka, P., and M. Wolf (2004): "Subsampling the mean of heavy-tailed dependent observations," *Journal of Time Series Analysis*, 25, 217–234.

Lehmann, E. L. (1986): *Testing Statistical Hypotheses.* Wiley, New York, second edn.

Liang, K., and S. L. Zeger (1986): "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Loretan, M., and P. C. B. Phillips (1994): "Testing the covariance stationarity of heavy-tailed time series," *Journal of Empirical Finance*, 1, 211–248.

Marshall, A. W., and I. Olkin (1979): *Inequalities: theory of majorization and its applications.* Academic Press, New York.

McConnell, M. M., and G. Perez-Quiros (2000): "Output Fluctuations in the United States: What Has Changed Since the Early 1980's," *American Economic Review*, 90, 1464–1476.

McElroy, T., and D. N. Politis (2002): "Robust inference for the mean in the presence of serial correlation and heavy-tailed distributions," *Econometric Theory*, 18, 1019–1039.

Moulton, B. R. (1990): "An Illustration of a Pitfall in Estimating the Effects of Aggregrate Variables on Micro Units," *Review of Economics and Statistics*, 72, 334–338.

Müller, U. K. (2007a): "An Alternative Sense of Asymptotic Efficiency," *Working paper, Princeton University.*

——— (2007b): "A Theory of Robust Long-Run Variance Estimation," *Journal of Econometrics*, 141, 1331–1352.

Müller, U. K., and M. W. Watson (2007): "Testing Models of Low-Frequency Variability," *Working Paper, Princeton University.*

Newey, W. K., and K. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

Petersen, M. A. (2005): "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches," *NBER Working Paper 11280.*

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*, Springer Series in Statistics. Springer-Verlag, New York.

Rachev, S. T., C. Menn, and F. J. Fabozzi (2005): *Fat-tailed and skewed asset return distributions: implications for risk management, portfolio selection, and option pricing.* Wiley, Hoboken, NJ.

Rogers, W. H. (1993): "Regression Standard Errors in Clustered Samples," *Stata Technical Bulletin*, 13, 19–23.

Sun, Y., P. C. B. Phillips, and S. Jin (2006): "Optimal Bandwidth Selection in Heteroskedasticity-Autocorrelation Robust Testing," *Cowles Foundation Discussion Paper 1545.*

Swamy, P. A. V. B. (1970): "Efficient Inference in a Random Coefficient Regression Model," *Econometrica*, 38, 311–323.

Thompson, S. B. (2006): "Simple Formulas for Standard Errors That Cluster by Both Firm and Time," *Working paper, Harvard University.*

White, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–830.

Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data.* The MIT Press, Cambridge, Massachusetts.

——— (2003): "Cluster-Sample Methods in Applied Econometrics," *American Economic Review*, 93, 133–138.