Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: http://www.tandfonline.com/loi/ubes20

t-Statistic Based Correlation and Heterogeneity Robust Inference

Rustam Ibragimov & Ulrich K. Müller

To cite this article: Rustam Ibragimov & Ulrich K. Müller (2010) t-Statistic Based Correlation and Heterogeneity Robust Inference, Journal of Business & Economic Statistics, 28:4, 453-468, DOI: 10.1198/jbes.2009.08046

To link to this article: https://doi.org/10.1198/jbes.2009.08046

t-Statistic Based Correlation and Heterogeneity Robust Inference

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

Ulrich K. MÜLLER
Economics Department, Princeton University, Fisher Hall, Princeton, NJ 08544 (*umueller@princeton.edu*)

We develop a general approach to robust inference about a scalar parameter of interest 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 of interest. 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.

KEY WORDS: Dependence; Fama–MacBeth method; Least favorable distribution; t-test; Variance estimation.

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 heteroscedasticity 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 nonparametric standard errors for datasets that exhibit spatial correlations.

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: heteroscedasticity 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. Also, block bootstrap and subsampling techniques derive their asymptotic validity from averaging over an infinite number of essentially uncorrelated blocks. When correlations are pervasive and pronounced enough, these methods are inapplicable or yield poor results.

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 significance levels of 8.3% or lower, the usual t-test remains conservative 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 there exists estimators $\hat{\beta}_j$, estimated from groups $j = 1, \ldots, q$, so that approximately $\hat{\beta}_j \sim \mathcal{N}(\beta, v_j^2)$, and $\hat{\beta}_j$ is approximately independent of $\hat{\beta}_i$ for $j \neq i$. Typically, the estimator $\hat{\beta}_j$ will simply be the element of interest of the vector $\hat{\boldsymbol{\theta}}_j$, where $\hat{\boldsymbol{\theta}}_j$ is the estimator of the model's parameter vector using group j data only. The observations $\hat{\beta}_1, \ldots, \hat{\beta}_q$ can then approximately be treated as independent normal observations with common mean β (but not necessarily equal variance), and the usual t-test concerning β constructed from $\hat{\beta}_1, \ldots, \hat{\beta}_q$ (with $q - 1$ degrees of freedom) is conservative. If the number 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.

Knowledge about the correlation structure in the data is embodied in the assumption that $\hat{\beta}_1, \ldots, \hat{\beta}_q$ are (approximately) independent. In contrast to consistent variance estimators, only this finite amount of uncorrelatedness is directly required for

Journal of Business & Economic Statistics
October 2010, Vol. 28, No. 4
DOI: 10.1198/jbes.2009.08046

the validity of the t-statistic approach. What is more, by invoking the results of Müller (2008), we show that the t-statistic approach in some sense efficiently exploits the information contained in the assumption of asymptotically independent and Gaussian estimators $\hat{\beta}_1, \ldots, \hat{\beta}_q$. Of course, a stronger (correct) assumption, that is, 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 iid observations, a 5% level t-statistic based test with $q = 16$ equal sized groups loses at most 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 versus power trade-off is thus especially acute only for datasets where an even coarser partition is required to yield independent information about the parameter of interest.

In applications of the t-statistic approach, the key question will nevertheless often be the adequate number and composition of groups. Some examples and Monte Carlo evidence for group choices are discussed in Section 3 below. The efficiency result mentioned above formally shows that one cannot delegate the decision about the adequate group choice 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 challenging 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 its validity is based on such a fairly explicit regularity condition. At the end of the day, inference requires some assumption about potential correlations, and empirical researchers should agonize about the appropriate amount of regularity that is imposed on the data.

Our paper is related to previous work on inference procedures 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 (2002, 2005) and Jansson (2004). Müller (2007) 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 (2007) 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 iid 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." Recent work by Bester, Conley, and Hansen (2009) builds on our paper and proposes inference about both scalar and vector valued parameters based on the usual full sample estimator, using clustered standard errors with groups chosen as suggested here and critical values derived in Hansen (2007). This approach requires the design matrix to be (asymptotically) identical across groups. Under this homogeneity, their procedure for inference about a scalar parameter is asymptotically numerically identical to the t-statistic approach under both the null and local alternatives. At the same time, the t-statistic approach remains valid even without this homogeneity, so from the usual first-order asymptotic point of view, the t-statistic approach is strictly preferable.

The t-statistic approach 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 our suggested method, 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 t-statistic approach generalizes the previous literature on large sample inference without consistent variance estimation to a generic strategy that can be employed in different settings, such as in time series data, panel data, or spatially correlated data. Due to the small sample conservativeness result, the approach allows for unknown and unmodeled heterogeneity. In a time series context, for instance, this means that unlike Kiefer and Vogelsang (2005), we can allow for low frequency variability in second moments, and in a panel context, we do not require the asymptotic homogeneity as in Hansen (2007). Also, the t-statistic approach is very easy to implement, and does not require any new tables of critical values. The crucial regularity condition—the assumption that $\hat{\beta}_1, \ldots, \hat{\beta}_q$ are approximately independent and distributed $\mathcal{N}(\beta, v_j^2)$—is more explicit and may be 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). 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 rest of the paper is organized as follows: Section 2 reviews the small sample result by Bakirov and Székely (2005), and discusses the large sample validity and consistency of the t-statistic approach. Section 3 gives examples of group choices and provides Monte Carlo evidence for time series, panel, clustered, and spatially correlated data. We discuss the efficiency properties in Section 4, followed by concluding remarks in Section 5.

2. VALIDITY OF t–STATISTIC BASED INFERENCE

2.1 Small Sample Result

Let X_j, $j = 1, \ldots, q$, 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

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

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 $|t|$. [To be precise, we define t in (2) to be equal to zero if $s_X = 0$.] 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 following theorem:

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, that is, $P(|T_{q-1}| > \mathrm{cv}_q(\alpha)) = \alpha$, and let Φ denote the cumulative density function of a standard normal random variable.

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

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

(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 $q \geq k_i + 1$, $\sup_{\{\sigma_1^2,\ldots,\sigma_q^2\}} P(|t| > \widetilde{\mathrm{cv}}_q(\alpha_i)|H_0) = \alpha_i$.

The usual 5% level two-sided test of (1) based on the usual t-test thus remains valid for all values of $\{\sigma_1^2, \ldots, \sigma_q^2\}$, 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.

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 iid Gaussian observations has effective coverage level of at least C for all values of $\{\sigma_1^2, \ldots, \sigma_q^2\}$. As long as the realized value of $|t|$ is larger than the smallest $\mathrm{cv}_q(\alpha)$ 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. As stressed by Bakirov and Székely (2005), a further implication of Theorem 1 is the conservativeness of the usual t-test against iid observations that are scale mixtures of Gaussian variates: Let $Y_j = \mu + Z_j V_j$ where $Z_j \sim \text{iid}\mathcal{N}(0, 1)$ and V_j is iid 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 iid 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.

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 $\{V_j\}_{j=1}^q$.

Figure 1 depicts the null rejection probability of the 5% level two-sided t-test for $q = 4, 8$, and 16 when (i) there are two equal sized groups of iid 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 in Sections 2.3 and 4.3, respectively) 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 null rejection probability can be much lower than the nominal level, but for $q = 16$, it does not drop much below 4% in either scenario.

2.2 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

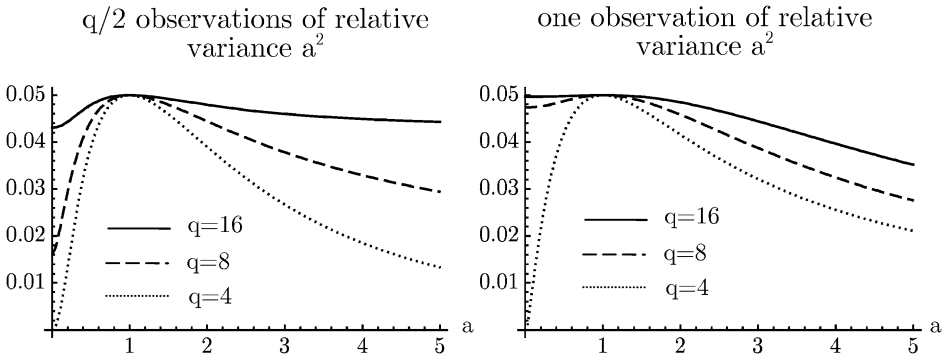

Figure 1. Null rejection probabilities of 5% level t-tests with q independent observations.

a large dataset 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 of the original dataset 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

$$\sqrt{n}(\hat{\beta}_1-\beta,\ldots,\hat{\beta}_q-\beta)' \Rightarrow \mathcal{N}(\mathbf{0},\operatorname{diag}(\sigma_1^2,\ldots,\sigma_q^2)), \qquad \max_{1\leq j\leq q}\sigma_j^2 > 0. \quad (4)$$

The asymptotic Gaussianity of $\sqrt{n}(\hat{\beta}_j-\beta), j=1,\ldots,q$, 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.

Under (4), for large n, the q estimators $\hat{\beta}_j$, $j=1,\ldots,q$, are approximately independent Gaussian random variables with common mean β and variances σ_j^2/n. Thus, by Theorem 1 above, one can perform an asymptotically valid test of level α, $\alpha \leq 0.083$ 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

$$t_\beta = \sqrt{q}\frac{\overline{\hat{\beta}}-\beta_0}{s_{\hat{\beta}}} \quad (5)$$

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 (4) holds, irrespective of the values of $\sigma_j^2, j=1,\ldots,q$. Also, by implication, the confidence interval $\overline{\hat{\beta}} \pm \mathrm{cv}\, s_{\hat{\beta}}$ 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 $C \geq 0.917$.

An important class of models that typically induce (4) with an appropriate choice of groups are Hansen's (1982) Generalized Method of Moments (GMM) models. Suppose the moment condition is $E[g(\boldsymbol{\theta},\mathbf{y}_i)] = \mathbf{0}$, where g is a known $k \times 1$ vector valued function, $\boldsymbol{\theta}$ is a $l \times 1$ vector of parameters ($l \leq k$) and $\mathbf{y}_i$, $i=1,\ldots,n$, are possibly vector-valued observations. Without loss of generality, assume that the first element of $\boldsymbol{\theta}$ is the parameter of interest β, so that the last $l-1$ elements of $\boldsymbol{\theta}$ are nuisance parameters. Denote by $\mathcal{G}_j$ the set of indices of group j observations, such that $\mathbf{y}_i$ is in group j if and only if $i \in \mathcal{G}_j$. Assume that the GMM estimator $\hat{\boldsymbol{\theta}}_j$ based on group j observations $\mathcal{G}_j$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}) = (\boldsymbol{\Gamma}_j'\boldsymbol{\Psi}_j\boldsymbol{\Gamma}_j)^{-1}\boldsymbol{\Gamma}_j'\boldsymbol{\Psi}_j\mathbf{Q}_j + o_p(1), \quad (6)$$

where $n^{-1}\sum_{i\in\mathcal{G}_j}\frac{\partial g(\mathbf{a},\mathbf{y}_i)}{\partial \mathbf{a}}|_{\mathbf{a}=\hat{\boldsymbol{\theta}}_j} \xrightarrow{p} \boldsymbol{\Gamma}_j$, with $\boldsymbol{\Gamma}_j$ of full rank and nonstochastic for all j, $\boldsymbol{\Psi}_j$ is the nonstochastic full rank limit of the weighting matrix for the GMM estimator $\hat{\boldsymbol{\theta}}_j$, and $\mathbf{Q}_j = n^{-1/2}\sum_{i\in\mathcal{G}_j} g(\boldsymbol{\theta},\mathbf{y}_i) \Rightarrow \mathcal{N}(\mathbf{0},\boldsymbol{\Omega}_j)$. In addition, suppose that the GMM estimators are asymptotically independent, which requires $(\mathbf{Q}_1',\ldots,\mathbf{Q}_q') \Rightarrow \mathcal{N}(\mathbf{0},\operatorname{diag}(\boldsymbol{\Omega}_1,\ldots,\boldsymbol{\Omega}_q))$. These assumptions follow from the usual linearization arguments under appropriate conditions. As a consequence, (4) holds, so that the t-statistic approach yields valid inference about β.

For some applications, a slightly more general regularity condition than (4) is useful: Suppose

$$\{m_n(\hat{\beta}_j-\beta)\}_{j=1}^{q} \Rightarrow \{Z_jV_j\}_{j=1}^{q} \quad (7)$$

for some positive sequence $m_n \to \infty$, where $Z_j \sim \text{iid}\mathcal{N}(0,1)$, the random variables $\{V_j\}_{j=1}^{q}$ are independent of $\{Z_j\}_{j=1}^{q}$ and $\max_j|V_j| > 0$ almost surely. As discussed in Section 2.1, (7) accommodates convergences (at an arbitrarily slow rate) to independent but potentially heterogeneous scale mixtures of standard normal distributions, such as the family of strictly 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 unconditionally.

Under the fixed alternative $\beta \neq \beta_0$, (4) or (7) imply that $s_{\hat{\beta}} = o_p(1)$ and $\bar{\hat{\beta}} - \beta = o_p(1)$, so that $P(|t_\beta| > \text{cv}) \to 1$ for all $\text{cv} > 0$ and a test based on $|t_\beta|$ is consistent at any level of significance. Under fixed heterogeneous alternatives of the null hypothesis $\beta = \beta_0$, 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 $j = 1, \ldots, q$, and a test based on $|t_\beta|$ with critical value cv is consistent if

$$\frac{q(\bar{\beta} - \beta_0)^2}{(q-1)^{-1} \sum_{j=1}^{q} (\beta_j - \bar{\beta})^2} > \text{cv}^2, \tag{8}$$

where $\bar{\beta} = q^{-1} \sum_{j=1}^{q} \beta_j$. Especially for small q and large cv, (8) might not be satisfied when $\{\beta_j - \beta_0\}_{j=1}^{q}$ are very heterogenous, even when all $\beta_j - \beta_0$ 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 $\{\beta_j - \beta_0\}_{j=1}^{q}$ of equal sign that are no more heterogeneous (in the majorization sense, see Marshall and Olkin 1979) than $\beta_1 - \beta_0 = \cdots = \beta_{\lfloor q/2 \rfloor} - \beta_0 = 0$ and $\beta_{\lfloor q/2 \rfloor + 1} - \beta_0 = \cdots = \beta_q - \beta_0 \neq 0$, where $\lfloor \cdot \rfloor$ denotes the greatest lesser integer function.

2.3 Size Control Under Dependence

Tests of level 5% or lower based on t_β are asymptotically valid whenever (4) 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 application, 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)], that is, 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, that is, 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

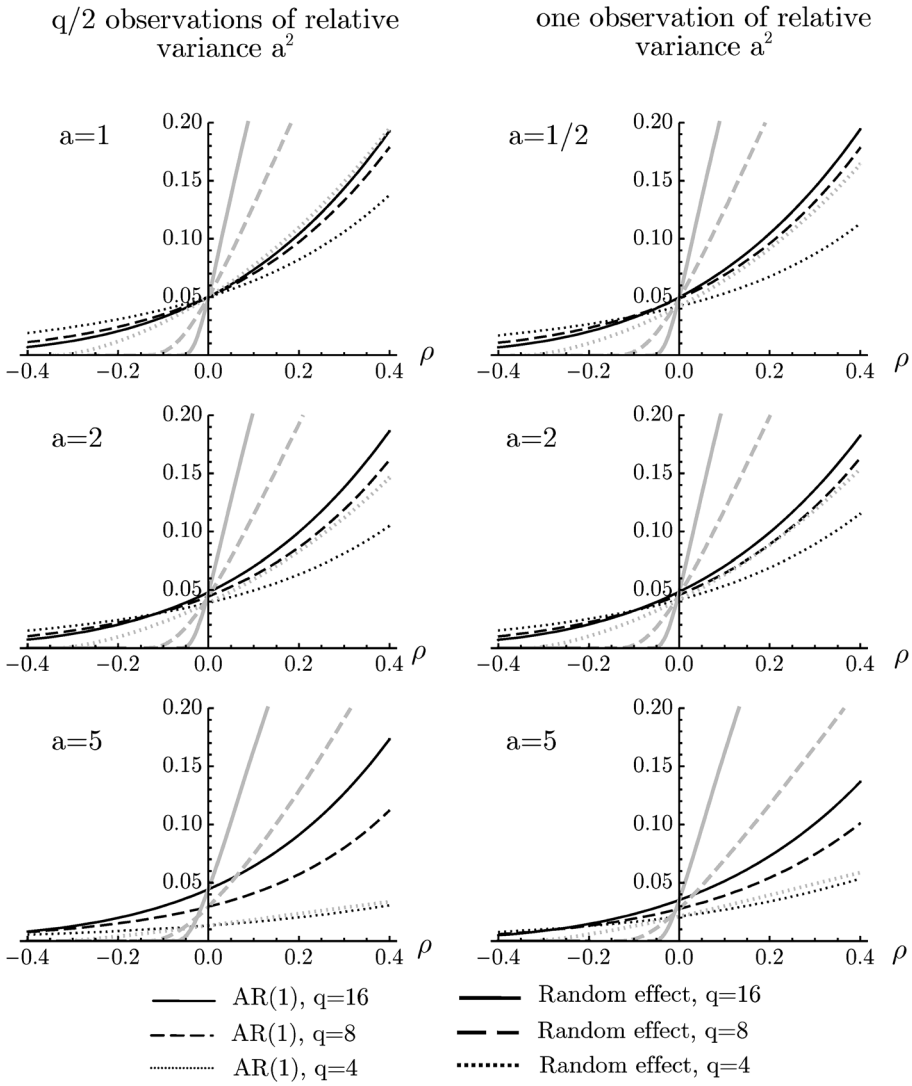

Figure 2. Null rejection probabilities of 5% level t-tests with q correlated observations.

depicts the effective size of a 5% level two-sided t-tests under these four scenarios. As might be expected, a negative ρ leads 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 the null rejection probability 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 (4) is the essential assumption of the approach suggested here, inference based on t_β continues to have quite reasonable properties as long as the dependence in $\{\hat\beta_j\}_{j=1}^q$ is weak, especially when q is small.

3. APPLICATIONS

We now discuss applications of the t-statistic approach, and provide some Monte Carlo evidence on its performance relative 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 focuses 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, and it also represents the case where the theoretical results above predict size control to be most difficult for the t-statistic approach.

3.1 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, $t = 1, \dots, T$, is element of group j if $t \in \mathcal{G}_j = \{s : (j-1)T/q < s \le jT/q\}$ for $j = 1, \dots, q$. The smaller q, the less approximate independence in time is imposed by this group choice.

Consider a GMM setup, as discussed in Section 2.2 above, with a $k \times 1$ moment condition, a $l \times 1$ parameter vector $\boldsymbol\theta$, and the scalar parameter of interest β is the first element of $\boldsymbol\theta$. Under a wide range of assumptions on the underlying model and observations, the partial-sample GMM estimator $\hat{\boldsymbol\theta}(r, s)$ computed from the observations $t = \lfloor rT \rfloor, \dots, \lfloor sT \rfloor$ satisfies (see, e.g., Andrews 1993 and Hansen 2000)

$$\sqrt{T}(\hat{\boldsymbol\theta}(r,s) - \boldsymbol\theta) \Rightarrow \left(\int_r^s \boldsymbol\Upsilon(\lambda)\, d\lambda\right)^{-1} \int_r^s \mathbf{h}(\lambda)\, d\mathbf{W}(\lambda) \qquad (9)$$

for all $0 \le r < s \le 1$, where $\boldsymbol\Upsilon(\cdot)$ is a positive definite $l \times l$ nonstochastic function, $\mathbf{h}(\cdot)$ is a $l \times k$ nonstochastic function, and $\mathbf{W}$ is a $k \times 1$ standard Wiener process [cf. Equation (6) above]. For the groups chosen as above, by the Continuous Mapping Theorem, we obtain

$$\sqrt{T}\begin{pmatrix} \hat{\boldsymbol\theta}_1 - \boldsymbol\theta \\ \hat{\boldsymbol\theta}_2 - \boldsymbol\theta \\ \vdots \\ \hat{\boldsymbol\theta}_q - \boldsymbol\theta \end{pmatrix} \Rightarrow \begin{pmatrix} (\int_0^{1/q} \boldsymbol\Upsilon(\lambda)\, d\lambda)^{-1} \int_0^{1/q} \mathbf{h}(\lambda)\, d\mathbf{W}(\lambda) \\ (\int_{1/q}^{2/q} \boldsymbol\Upsilon(\lambda)\, d\lambda)^{-1} \int_{1/q}^{2/q} \mathbf{h}(\lambda)\, d\mathbf{W}(\lambda) \\ \vdots \\ (\int_{(q-1)/q}^{1} \boldsymbol\Upsilon(\lambda)\, d\lambda)^{-1} \int_{(q-1)/q}^{1} \mathbf{h}(\lambda)\, d\mathbf{W}(\lambda) \end{pmatrix}$$

so that also $\{\sqrt{T}(\hat\beta_j - \beta)\}_{j=1}^q$ are asymptotically independent and Gaussian. Therefore, whenever (9) holds, t-statistic based inference is asymptotically valid for any $q \ge 2$. The t-statistic approach can hence allow for asymptotically time varying information [nonconstant $\boldsymbol\Upsilon(\cdot)$] and pronounced variability in second moments [nonconstant $\mathbf{h}(\cdot)$]. In fact, using (7), the t-statistic approach remains valid even if $\boldsymbol\Upsilon(\cdot)$ and $\mathbf{h}(\cdot)$ are stochastic, as long as they are independent of $\mathbf{W}$. In contrast, the approach of Kiefer and Vogelsang (2002, 2005) requires $\boldsymbol\Upsilon(\cdot)$ and $\mathbf{h}(\cdot)$ 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), and Müller and Watson (2008). 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 (9) holds, as least under double-array asymptotics: Müller (2007) 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 (9). The general validity of the t-statistic approach under (9) is thus an analytical reflection of its robustness.

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 $\mathcal{G}_j = \{s : (j-1)T/q < s \le jT/q\}$, we include in our study the approach developed by Kiefer and Vogelsang (2005) and usual inference 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 $b \le 1$ of

Table 1. Small sample results in a time series regression with $T = 128$

	t-statistic (q)						$\hat{\omega}^2_{QS}(b)$			$\hat{\omega}^2_{BT}(b)$		
	2	4	8	16	$\hat{\omega}^2_{QA}$	$\hat{\omega}^2_{PW}$	0.05	0.3	1	0.05	0.3	1
Size, AR(1)												
ρ												
−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
Size, MA(1)												
ϕ												
−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
Size adjusted power, AR(1)												
ρ												
−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
Size adjusted power, MA(1)												
ϕ												
−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

NOTE: The entries are rejection probabilities of nominal 5% level two-sided t-tests about the coefficient β of the first element of $\mathbf{X}_t$ in the linear regression $y_t = \mathbf{X}_t'\boldsymbol{\theta} + u_t$, $t = 1, \ldots, T$, where $\mathbf{X}_t = (x_t', 1)'$, $x_t = (T^{-1}\sum_{s=1}^T \tilde{x}_s\tilde{x}_s')^{-1/2}\tilde{x}_t$, $\tilde{x}_t = \tilde{x}_t - T^{-1}\sum_{s=1}^T \tilde{x}_s$, 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 ϕ, respectively. The disturbances u_t are an independent draw from the same model as the (pretransformed) regressors, multiplied by the first element of $\mathbf{X}_t$. Unreported simulations for the other forms of heteroscedasticity considered by Andrews (1991) yield qualitatively similar results. 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.

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 in highly dependent scenarios.

The group estimators $\hat{\beta}_j$, $j = 1, \ldots, q$, may fail to be approximately normal because the underlying random variables are too heavy tailed, so that the limiting law is instead a α-stable distribution with $\alpha < 2$. McElroy and Politis (2002) stress that few options are available for inference in time series models with such innovations, even for inference about the location parameter. But as long as, for some sequence m_T, $m_T(\hat{\beta}_j - \beta)$, $j = 1, \ldots, q$, are asymptotically independent with strictly α-stable *symmetric* limiting distributions, the discussion around Equation (7) implies that the t-statistic approach remains applicable. Note that this approach does not require knowledge of m_T, which typically depends on the tail index α. We provide some Monte Carlo evidence on the favorable properties of the t-statistic approach relative to subsampling studied by McElroy and Politis (2002) in the supplementary materials.

Table 2. Small sample results in a panel with $N = 10$, $T = 50$, and time series correlation

		Homoskedastic				Heteroskedastic			
ρ_x	ρ_u	$\frac{\beta-\beta_0}{\sqrt{nT}}$	t-stat	clust	cl.FE	$\frac{\beta-\beta_0}{\sqrt{nT}}$	t-stat	clust	cl.FE
				Size					
0	0	0	5.0	5.2	5.1	0	4.6	4.9	4.9
0.5	0.5	0	4.9	5.3	5.1	0	4.6	5.5	5.3
0.9	0.5	0	5.3	6.5	6.2	0	4.4	7.7	6.4
0.9	0.9	0	5.0	7.2	6.2	0	4.0	7.9	6.2
1	0.5	0	4.4	8.7	6.8	0	3.8	14.7	8.6
				Size adjusted power					
0	0	2.5	58.7	60.8	59.7	5	54.6	51.9	52.0
0.5	0.5	3	53.9	55.0	55.1	8	64.7	56.6	58.2
0.9	0.5	8	56.7	71.5	60.8	8	62.0	49.1	48.9
0.9	0.9	0.8	39.4	31.9	38.3	25	62.9	33.6	46.2
1	0.5	12	42.5	83.6	50.8	180	73.7	52.1	45.1

NOTE: 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 $y_{i,t} = \mathbf{X}_{i,t}'\boldsymbol{\theta} + u_{i,t}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$, where $\mathbf{X}_{i,t} = (x_{i,t}, 1)'$, $x_{i,t} = \rho_x x_{i,t-1} + \varepsilon_{i,t}$, $x_{i,0} = 0$, $\varepsilon_{i,t} \sim \text{iid}\mathcal{N}(0, 1)$, $u_{i,t} = \rho_u u_{i,t-1} + \eta_{i,t}$, $u_{i,0} = 0$, where under homoscedasticity, $\eta_{i,t} \sim \text{iid}\mathcal{N}(0, 1)$ independent of $\{\varepsilon_{i,t}\}$, and under heteroscedasticity, $\eta_{i,t} = (0.5 + 0.5x_{i,t}^2)\tilde{\eta}_{i,t}$ and $\tilde{\eta}_{i,t} \sim \text{iid}\mathcal{N}(0, 1)$ independent of $\{\varepsilon_{i,t}\}$. The considered tests are the t-statistic approach with groups defined by individuals ("t-stat"); OLS coefficient based tests with Rogers (1993) standard errors ("clust"); and OLS coefficient based test which includes individual Fixed Effects and Arellano (1987) standard errors ("cl.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.

3.2 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. 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

$$y_{i,t} = \mathbf{X}_{i,t}'\boldsymbol{\theta} + u_{i,t}, \qquad i = 1, \ldots, N, t = 1, \ldots, T, \tag{10}$$

where $\{\mathbf{X}_{i,t}, u_{i,t}\}_{t=1}^T$ are independent across i and $E[\mathbf{X}_{i,t}u_{i,t}] = \mathbf{0}$ for all i, t. Suppose that under $T \to \infty$ asymptotics with N fixed, $T^{-1}\sum_{t=1}^T \mathbf{X}_{i,t}\mathbf{X}_{i,t}' \xrightarrow{p} \boldsymbol{\Gamma}_i$ and $T^{-1/2}\sum_{t=1}^T \mathbf{X}_{i,t}u_{i,t} \Rightarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_i)$ for all i for some full rank matrices $\boldsymbol{\Gamma}_i$ and $\boldsymbol{\Omega}_i$. These assumptions are enough to guarantee that the OLS coefficient estimators $\hat{\beta}_i$ using data from individual i 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 $\boldsymbol{\Gamma}_i = \boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}_i = \boldsymbol{\Omega}$, 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 (5) 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 2 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 $\{\mathbf{X}_{i,t}\}$, 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, $T^{-2}\sum_{t=1}^T \mathbf{X}_{i,t}\mathbf{X}_{i,t}'$ does not converge to the same limit across i, so that despite the iid sampling across i, asymptotic homogeneity fails. These asymptotic considerations successfully explain the size results in Table 2. The t-statistic approach has higher size adjusted power for heteroscedastic disturbances, but this is not true under homoscedasticity.

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 (2009) 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 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 nonparametrically 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 conducted a Monte Carlo study of the performance of this approach relative to clustering and Fama–MacBeth standard errors in a panel with both cross sectional and time series dependence. We found substantially better size control of the t-statistic approach, but also somewhat smaller size-adjusted power. See the supplementary materials for details.

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 interindustry 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 4.3 below 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, e.g., Jegadeesh and Karceski 2009) for a discussion of inference based on consistent variance estimation), 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 (10) with fixed effects in both dimensions. In a typical application, the individuals $i = 1, \ldots, N$ 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 (2007) show that if $u_{i,t}$ has an iid Gaussian random effect structure for each (potential) preintervention and postintervention 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 $\hat{\beta}_j$, $j = 1, \ldots, q$, 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 preintervention and postintervention specific random effects in $u_{i,t}$ are independent, but not necessarily identically distributed scale mixture of standard normals. This is a considerable weakening of the homogeneous Gaussian assumption required for the approach of Donald and Lang (2007). 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 standard normals.

3.3 Clustered Data

A further potential application of our approach is to draw inferences about a population based on a two-stage (or multistage) sampling design with a small number of independently sampled primary sampling units (PSUs). PSUs could be villages in a development study (see, e.g., 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 $j = 1, \ldots, q$, 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, $j = 1, \ldots, q$. Of course, this will only be possible if the parameter of interest is identified in each PSU; in a regression context, a coefficient about a regressor that only vary across PSUs cannot be estimated from one PSU only, at least 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.

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 setup 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)

$$Y_{i,j} = \mathbf{X}_{i,j}'\boldsymbol{\theta}_j + u_{i,j} = \mathbf{X}_{i,j}'\boldsymbol{\theta} + \mathbf{X}_{i,j}'(\boldsymbol{\theta}_j - \boldsymbol{\theta}) + u_{i,j}$$

for individual $i = 1, \ldots, n_j$ in PSU $j = 1, \ldots, q$, $E[\mathbf{X}_{i,j}u_j] = \mathbf{0}$ and $\boldsymbol{\theta}_j$ are iid draws from some population with mean $\boldsymbol{\theta}$. Thought of as part of the disturbance term, $\mathbf{X}_{i,j}'(\boldsymbol{\theta}_j - \boldsymbol{\theta})$ induces intra-PSU correlations. Now under sufficient regularity conditions, $\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j \xrightarrow{p} \mathbf{0}$ as $n_j \to \infty$, and the t-statistic approach for inference about β (the first element of $\boldsymbol{\theta}$) remains valid as long as the distribution of $\beta_j - \beta$ can be written as a scale mixture of standard normals. This is a wide class of distributions, as noted in Section 2.1 above. If n_j is not large enough to make $\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j \xrightarrow{p} \mathbf{0}$ a good approximation, but instead $\hat{\beta}_j|\beta_j \sim \text{iid}\mathcal{N}(\beta_j, v_j^2)$, then the t-statistic approach remains valid as long as $\beta_j - \beta$ is a scale mixture of standard normals, as the induced unconditional distribution of $\hat{\beta}_j - \beta$ then is a scale mixture of standard normals, too.

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 iid 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 iid 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 iid sample from U.S. households is no longer iid 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.

3.4 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 decay of correlations between distant observations, 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 Dupor (2003), for instance, use metrics based on input-output relations 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. Bester, Conley, and Hansen (2009) provide such sufficient primitive conditions for linear regressions.

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 3, 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.

4. EFFICIENCY OF t–STATISTIC BASED INFERENCE

We now turn to a discussion of the efficiency properties of the t-statistic based approach to large sample inference. We start by establishing a small sample optimality result for the t-statistic in Section 4.1. This in turn yields a corresponding large sample "efficiency under robustness" result by an application of the recent results in Müller (2008), as discussed in Section 4.2. In particular, these results imply the impossibility of using data dependent methods to automatically select the number or composition of groups while maintaining robustness. Finally, in Section 4.3, we compare the efficiency of the t-statistic based approach to inference to the benchmark case of known (or, equivalently, correctly consistently estimated) asymptotic variance.

4.1 Small Sample Result

Theorem 1 provides conditions under which the usual small sample t-test remains a valid test. We now turn to a discussion of the small sample optimality of the t-statistic (2) when the underlying Gaussian variates $X_j \sim \mathcal{N}(\mu, \sigma_j^2)$ are not necessarily

Table 3. 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

NOTE: The entries are rejection probabilities of nominal 5% level two-sided t-tests about β in the model $y_{i,j} = \beta + u_{i,j}$, $i = 1, \ldots, 8$, $j = 1, \ldots, 16$. 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 $(y_{i,j} - \bar{y})(y_{k,l} - \bar{y})$, $i, k = 1, \ldots, 8$, $j, l = 1, \ldots, 16$, of Euclidian distance $d \le b$, 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.

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 a significance level of 5% or lower, Theorem 1 shows that the null rejection probability for the t-test never exceeds the nominal level for heterogeneous variances. So if we consider the hypothesis test

$$\begin{aligned} &H_0 : \mu = 0 \text{ and } \{\sigma_j^2\}_{j=1}^q \text{ arbitrary} \quad \text{against} \\ &H_1 : \mu \neq 0 \text{ and } \sigma_j^2 = \sigma^2 \text{ for all } j \end{aligned} \tag{11}$$

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 (11) for any given alternative $\mu \neq 0$ when the level constraint is most difficult to satisfy. By theorem 7 of Lehmann (1986, pp. 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 (11).

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.

4.2 Asymptotic Efficiency and the Choice of Groups

Suppose the n observations in the potentially correlated large data set are of dimension $\ell \times 1$, so that the overall data $\mathbf{Y}_n$ is an element of $\mathbb{R}^{\ell n}$. In general, tests φ_n of $H_0 : \beta = \beta_0$ are sequences of (measurable) functions from $\mathbb{R}^{\ell n}$ to the unit interval, where $\varphi_n(\mathbf{y}_n) \in [0, 1]$ 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 hypothesis becomes $H_0 : \mu = 0$. Under such local alternatives, (4) implies

$$\{\sqrt{n}(\hat{\beta}_j - \beta_0)\}_{j=1}^q \Rightarrow \{X_j\}_{j=1}^q \tag{12}$$

where X_j, $j = 1, \ldots, q$, are as in the small sample Section 4.1 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

$$\left\{\frac{\hat{\beta}_j-\beta_0}{\bar{\hat{\beta}}-\beta_0}\right\}_{j=1}^{q} \Rightarrow \{R_j\}_{j=1}^{q}=\left\{\frac{X_j}{\bar{X}}\right\}_{j=1}^{q}, \tag{13}$$

where $\bar{X}=q^{-1}\sum_{j=1}^{q}X_j$. The interest of (13) over (12) 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 $j=1,\ldots,q$ as long as $\alpha\leq 0.083$.

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 (12) and (13) 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 (12) 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 plausibility of this assumption in any given application depends on the functional relationship between $\{\hat{\beta}_j\}_{j=1}^{q}$ and the data $\mathbf{Y}_n$, and the properties of $\mathbf{Y}_n$. The convergence (13) 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 (12) and (13) 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 (12) and (13) hold pointwise for every $\mu\neq 0$. 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,

$$\limsup_{n\to\infty}\int\varphi_n\,dF_n(m,0,\{\sigma_j^2\}_{j=1}^{q})\leq\alpha$$
$$\text{for all } m\in\mathcal{M}_0 \text{ and } \{\sigma_j^2\}_{j=1}^{q} \text{ with } \max_j\sigma_j^2>0 \tag{14}$$

for $\mathcal{M}_0=\mathcal{M}_0^X$ or $\mathcal{M}_0=\mathcal{M}_0^R$. The robustness constraint (14) is strong, as the asymptotic size requirement is imposed for *all* $m\in\mathcal{M}_0$, which is attractive only if (12) or (13) summarize all knowledge about properties of $\mathbf{Y}_n$ that are relevant for inference about β.

Denote by $\varphi_n^*(\alpha)=\mathbf{1}[|t_\beta|>\mathrm{cv}_q(\alpha)]$ the $\mathbb{R}^{\ell n}\mapsto\{0,1\}$ test of asymptotic size α that rejects for large values of $|t_\beta|$ as defined in (5), and note that, by scale invariance, t_β can also be computed from the observations $\{(\hat{\beta}_j-\beta_0)/(\bar{\hat{\beta}}-\beta_0)\}_{j=1}^{q}$. As discussed in Section 2.2, the test $\varphi_n^*(\alpha)$ satisfies (14) for $\mathcal{M}_0=\mathcal{M}_0^X$ as long as $\alpha\leq 0.083$, and scale invariance implies that (14) also holds for $\mathcal{M}_0=\mathcal{M}_0^R$. What is more, for any data generating process satisfying (13) under the alternative with $\mu\neq 0$, that is, for any $m\in\mathcal{M}_1^R$ or $m\in\mathcal{M}_1^X$, $\varphi_n^*(\alpha)$ has local asymptotic power $\lim_{n\to\infty}\int\varphi_n^*(\alpha)\,dF_n(m,\mu,\{\sigma_j^2\}_{j=1}^{q})$ for $\mu\neq 0$ equal to the power of the small sample t-test $\mathbf{1}[|t|>\mathrm{cv}(\alpha)]$ in the "limiting problem" (1) 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 (14) has higher local asymptotic power. The following theorem, which follows straightforwardly from the general results in Müller (2008) 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 (14) 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 $\mu\neq 0$, $\sigma^2>0$ 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 (14) 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 $\mu\neq 0$, $\sigma^2>0$, and $m\in\mathcal{M}_1^X$.

Part (i) of Theorem 3 shows the t-statistic approach to be asymptotically most powerful against the benchmark alternative of equal asymptotic variances among all tests that provide asymptotically valid inference under the regularity condition (13). Part (ii) contains the same claim under the slightly more natural condition (12) 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 (12) and (13) holds with $\mu\neq 0$. In other words, for any test that has higher asymptotic power for some data-generating process for which (12) and (13) holds with $\mu\neq 0$ and equal asymptotic variances, there exists a data generating process satisfying (12) and (13) 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 (13) to hold for some small $q=q_0$, but the actual data is much more regular in the sense that (13) 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 (13) holds with $q=q_0$, and this method overrejects asymptotically. Thus, with (13) viewed as a regularity condition on the underlying large dataset, the t-statistic approach efficiently exploits the available information, with highest possible power in the benchmark case of equal asymptotic variances.

4.3 Comparison With Inference Under Known Variance

We now turn to a discussion of the relative performance of the t-statistic approach as outlined in Section 2.2 and inference based on the full sample estimator $\hat{\beta}$ with $\sqrt{n}(\hat{\beta}-\beta) \Rightarrow \mathcal{N}(0,\sigma^2)$ and known $\sigma^2 > 0$. When (12) or (13) 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 testing the null hypothesis $\beta = \beta_0$ is to reject when $|z_\beta|$ exceeds the critical value for a standard normal, where z_β is given by

$$z_\beta = \sqrt{n}\frac{\hat{\beta}-\beta_0}{\hat{\sigma}} = \sqrt{n}\frac{\hat{\beta}-\beta_0}{\sigma} + o_p(1) \tag{15}$$

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 consider the class of GMM models as discussed in Section 2.2 under exact identification ($k = l$). Under the assumptions made there, the simple average of the q group estimators $\bar{\hat{\boldsymbol{\theta}}} = q^{-1}\sum_{j=1}^q \hat{\boldsymbol{\theta}}_j$ satisfies

$$\sqrt{n}(\bar{\hat{\boldsymbol{\theta}}} - \boldsymbol{\theta}) = q^{-1}\sum_{j=1}^q \boldsymbol{\Gamma}_j^{-1}\mathbf{Q}_j + o_p(1) \Rightarrow \mathcal{N}(\mathbf{0}, \bar{\boldsymbol{\Sigma}}_q), \tag{16}$$

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

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \left(\sum_{j=1}^q \boldsymbol{\Gamma}_j\right)^{-1}\sum_{j=1}^q \mathbf{Q}_j + o_p(1) \Rightarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_q), \tag{17}$$

where $\boldsymbol{\Sigma}_q = (\sum_{j=1}^q \boldsymbol{\Gamma}_j)^{-1}(\sum_{j=1}^q \boldsymbol{\Omega}_j)(\sum_{j=1}^q \boldsymbol{\Gamma}_j')^{-1}$. 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 $E[g(\boldsymbol{\theta}, \mathbf{y}_i)] = \mathbf{0}$ for $i \in \mathcal{G}_j$, $j = 1, \ldots, q$. But this efficient full-sample estimator requires the consistent estimation of the optimal weighting matrix, which involves $\boldsymbol{\Omega}_j$, $j = 1, \ldots, q$. This is unlikely to be feasible or appropriate in applications with pronounced correlations and heterogeneity, so that the relevant comparison for $\bar{\hat{\boldsymbol{\theta}}}$ is with $\hat{\boldsymbol{\theta}}$ as characterized in (17).

Comparing $\bar{\boldsymbol{\Sigma}}_q$ with $\boldsymbol{\Sigma}_q$, we find that while $\sqrt{n}$-consistent and asymptotically Gaussian, the estimators $\hat{\boldsymbol{\theta}}$ and $\bar{\hat{\boldsymbol{\theta}}}$ (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 $\bar{\boldsymbol{\Sigma}}_q$ and $\boldsymbol{\Sigma}_q$ 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 $\boldsymbol{\iota}$ be the $k \times 1$ vector with 1 in the first row and zeros elsewhere. Then $\inf_{\{\boldsymbol{\Gamma}_j\}_{j=1}^q \in \mathcal{Q}_1^q, \{\boldsymbol{\Omega}_j\}_{j=1}^q \in \mathcal{P}_1^q} \bar{\boldsymbol{\Sigma}}_q / \boldsymbol{\Sigma}_q = 0$,

$$\inf_{\{\boldsymbol{\Gamma}_j\}_{j=1}^q \in \mathcal{P}_k^q, \{\boldsymbol{\Omega}_j\}_{j=1}^q \in \mathcal{P}_k^q} \frac{\boldsymbol{\iota}'\boldsymbol{\Sigma}_q\boldsymbol{\iota}}{\boldsymbol{\iota}'\bar{\boldsymbol{\Sigma}}_q\boldsymbol{\iota}} = 0 \qquad \text{and}$$

$$\inf_{\{\boldsymbol{\Gamma}_j\}_{j=1}^q \in \mathcal{P}_k^q, \{\boldsymbol{\Omega}_j\}_{j=1}^q \in \mathcal{P}_k^q} \frac{\boldsymbol{\iota}'\bar{\boldsymbol{\Sigma}}_q\boldsymbol{\iota}}{\boldsymbol{\iota}'\boldsymbol{\Sigma}_q\boldsymbol{\iota}} = \begin{cases} 1/q^2 & \text{if } k = 1 \\ 0 & \text{if } k \geq 2. \end{cases}$$

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

(iii) If $\boldsymbol{\Gamma}_j = \boldsymbol{\Gamma}$ for $j = 1, \ldots, q$, then $\bar{\boldsymbol{\Sigma}}_q = \boldsymbol{\Sigma}_q$ for all $\{\boldsymbol{\Omega}_j\}_{j=1}^q$.

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 $\boldsymbol{\Gamma}_j$ 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 $\hat{\boldsymbol{\theta}}$ is more efficient than $\bar{\hat{\boldsymbol{\theta}}}$ without knowledge of $\{\boldsymbol{\Omega}_j\}_{j=1}^q$, and vice versa in the important special case where $\boldsymbol{\Gamma}_j$ are symmetric and positive definite. (There exist $\{\boldsymbol{\Gamma}_j\}_{j=1}^q \notin \mathcal{P}_k^q$ that make $\bar{\hat{\boldsymbol{\theta}}}$ the more efficient estimator for all possible values of $\{\boldsymbol{\Omega}_j\}_{j=1}^q$; for instance, for $k = 1$ and $q = 2$, let $\boldsymbol{\Gamma}_1 = 1$ and $\boldsymbol{\Gamma}_2 = -1/2$.)

When $\boldsymbol{\Gamma}_j = \boldsymbol{\Gamma}$ 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 setup with this feature is the case of underlying iid observations. With $\boldsymbol{\Gamma}_j = \boldsymbol{\Gamma}$, $\sqrt{n}(\bar{\hat{\boldsymbol{\theta}}} - \boldsymbol{\theta}) = \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + o_p(1)$ 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 reestimation of the last $k-1$ elements of $\boldsymbol{\theta}$ 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) 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 $\boldsymbol{\Gamma}^{-1}\boldsymbol{\Omega}_j\boldsymbol{\Gamma}^{-1}$ 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

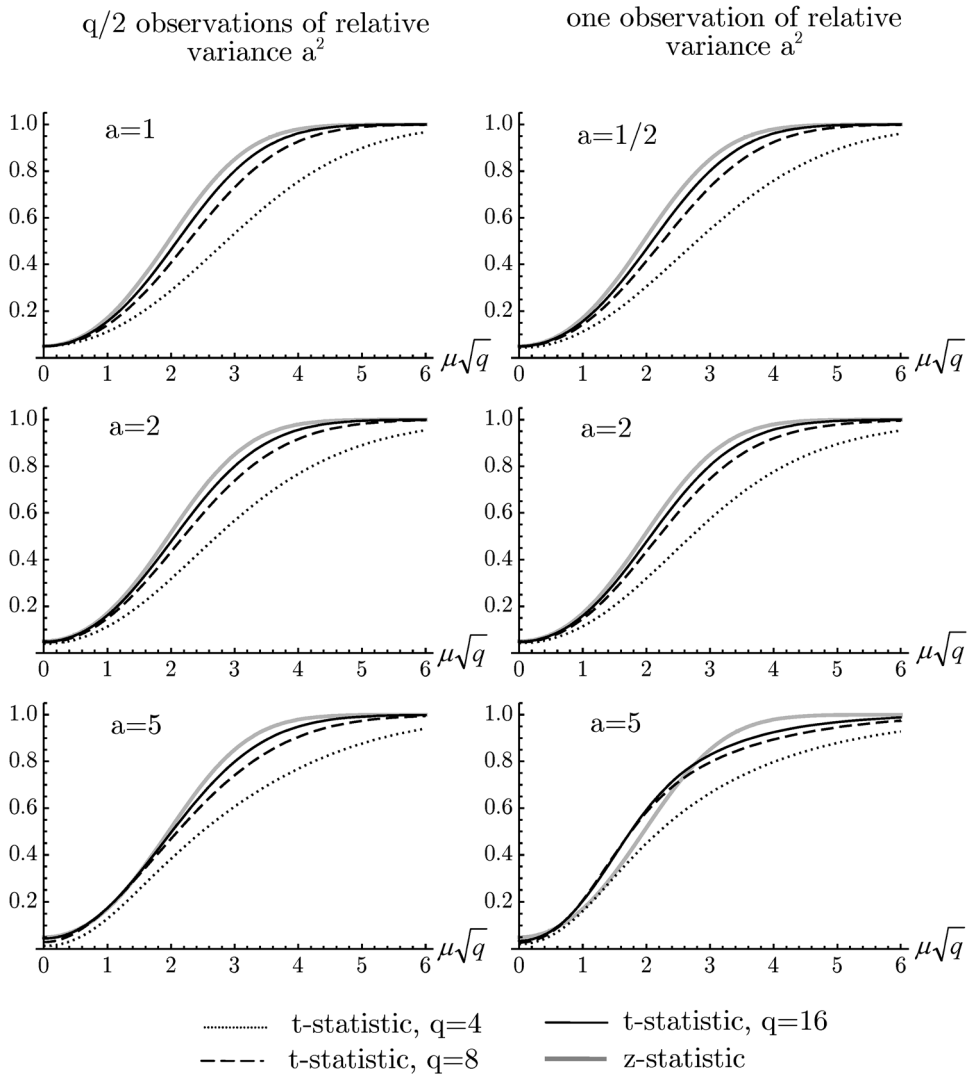

Figure 3. Power of 5% level t-tests and z-tests with q independent observations.

$q = 8$, 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 heterogeneous. When $\mathbf{\Gamma}_j = \mathbf{\Gamma}$ 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 iid Gaussian setup. 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 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 $\mathbf{\Gamma}_j$—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 $\mathbf{\Omega}_j$, the group j asymptotic covariance. In the important special case where $\mathbf{\Gamma}_j = \mathbf{\Gamma}$ 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 iid 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.

5. CONCLUSION

This 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, $j = 1, \ldots, q$, are approximately independent, unbiased and Gaussian, but not necessarily of equal variance. The t-statistic approach to inference provides in some sense efficient inference under this regularity condition. What

is more, this inference remains valid also when the group estimators have a joint distribution that can be written as a mixture of independent Gaussian distributions with means equal to the parameter of interest. The t-statistic approach may therefore be used also for group estimators that are approximately symmetric stable or of statistically dependent stochastic variances. Despite its simplicity, the proposed method is thus applicable in a wide range of models and settings.

APPENDIX

Proof of Theorem 4

(i) For the first claim, let $\mathbf{\Gamma}_1 = \xi - (q-1)$, $\mathbf{\Omega}_1 = 1$ and $\mathbf{\Gamma}_j = \mathbf{\Omega}_j = 1$ for $j = 2, \ldots, q$ for some $\xi > 0$. Then $\bar{\mathbf{\Sigma}}_q / \mathbf{\Sigma}_q = \xi^2 (q - 1 + (1 - q + \xi)^{-2})/q^2$, so that $\bar{\mathbf{\Sigma}}_q / \mathbf{\Sigma}_q \to 0$ as $\xi \to 0$.

For the second claim, let $\mathbf{\Gamma}_1 = \mathbf{\Omega}_1 = \mathbf{I}_k$, and $\mathbf{\Omega}_j = \xi \mathbf{I}_k$, $\mathbf{\Gamma}_j = \varsigma \mathbf{I}_k$ for $j = 2, \ldots, q$, for some $\varsigma > 0$, $\xi > 0$, so that $\sum_{j=1}^q \mathbf{\Gamma}_j = ((q-1)\varsigma + 1)\mathbf{I}_k$. Then

$$\mathbf{\Sigma}_q = \frac{\xi(q-1)+1}{((q-1)\varsigma+1)^2}\mathbf{I}_k \qquad \text{and}$$

$$\bar{\mathbf{\Sigma}}_q = \frac{1 + (q-1)\xi/\varsigma^2}{q^2}\mathbf{I}_k.$$

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

Also, for $k \geq 2$, let $\mathbf{\Gamma}_1 = \operatorname{diag}(\mathbf{A}, \mathbf{I}_{k-2}) \in \mathcal{P}_k$ with $\mathbf{A} = ((1, \frac{1}{2})', (\frac{1}{2}, 1)')$, $\mathbf{\Omega}_1 = \mathbf{\Gamma}_1 \operatorname{diag}(1, \xi, \mathbf{I}_{k-2})\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_j = \mathbf{\Omega}_j = \mathbf{I}_k$ for $j = 2, \ldots, q$. Then

$$\boldsymbol{\iota}' \bar{\mathbf{\Sigma}}_q \boldsymbol{\iota} = 1/q \qquad \text{and}$$

$$\boldsymbol{\iota}' \mathbf{\Sigma}_q \boldsymbol{\iota} = \frac{-3 - 4q + 16q^3 + 4\xi(q-1)^2}{(1-4q^2)^2}$$

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

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

$$\mathbf{\Sigma}_q = \left(\sum_{j=1}^q \mathbf{\Gamma}_j\right)^{-2} \sum_{j=1}^q \mathbf{\Omega}_j \leq \left(\sum_{j=1}^q \mathbf{\Gamma}_j^2\right)^{-1} \sum_{j=1}^q \mathbf{\Omega}_j$$

$$\leq \sum_{j=1}^q \mathbf{\Gamma}_j^{-2} \mathbf{\Omega}_j = q^2 \bar{\mathbf{\Sigma}}_q.$$

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

For the first claim, let $\bar{\mathbf{\Omega}}_j = \mathbf{\Gamma}_j \mathbf{\Gamma}_j'$. Then $\bar{\mathbf{\Sigma}}_q = q^{-1}\mathbf{I}_k$, and $\mathbf{\Sigma}_q = (\sum_{j=1}^q \mathbf{\Gamma}_j)^{-1} (\sum_{j=1}^q \mathbf{\Gamma}_j \mathbf{\Gamma}_j') (\sum_{j=1}^q \mathbf{\Gamma}_j')^{-1}$. It suffices to show that $\mathbf{\Sigma}_q^{-1} - \bar{\mathbf{\Sigma}}_q^{-1}$ is negative semidefinite, and this follows from the above result with $\mathbf{Y} = (\mathbf{I}_k, \ldots, \mathbf{I}_k)'$ and $\mathbf{X} = (\mathbf{\Gamma}_1, \ldots, \mathbf{\Gamma}_q)'$.

For the second claim, let $\underline{\mathbf{\Omega}}_j = \mathbf{\Gamma}_j$. Then $\bar{\mathbf{\Sigma}}_q = q^{-2} \sum_{j=1}^q \mathbf{\Gamma}_j^{-1}$ and $\mathbf{\Sigma}_q = (\sum_{j=1}^q \mathbf{\Gamma}_j)^{-1}$, and the result follows by setting $\mathbf{Y} = (\mathbf{\Gamma}_1^{-1/2}, \ldots, \mathbf{\Gamma}_q^{-1/2})'$ and $\mathbf{X} = (\mathbf{\Gamma}_1^{1/2}, \ldots, \mathbf{\Gamma}_q^{1/2})$.

(iii) Immediate from $\bar{\mathbf{\Sigma}}_q = q^{-2}\mathbf{\Gamma}(\sum_{j=1}^q \mathbf{\Omega}_j)\mathbf{\Gamma}'$ and $\sum_{j=1}^q \mathbf{\Gamma}_j = q\mathbf{\Gamma}$.

ACKNOWLEDGMENTS

The authors would like to thank the Editor, Arthur Lewbel, an anonymous Associate Editor, a referee, seminar, and conference participants at Berkeley, Brown, Cambridge, N. G. Chebotarev Research Institute of Mathematics and Mechanics, Chicago GSB, Columbia, Harvard/MIT, Institute of Mathematics of Uzbek Academy of Sciences, LSE, Ohio State, Penn State, Princeton, Rice, Risk Metrics Group London, UC Riverside, Rochester, Stanford, Tashkent State University, Tashkent State University of Economics, Queen Mary, Vanderbilt, the Greater New York Metropolitan Area Econometrics Colloquium 2006, the NBER Summer Institute 2006, the CIREQ Time Series Conference 2006, the 62nd European Meeting of the Econometric Society 2007 and the 2008 Far Eastern and South Asian Meeting of the Econometric Society for valuable comments and discussions. Ibragimov gratefully acknowledges partial support by the NSF grant SES-0820124, a Harvard Academy Junior Faculty Development grant and the Warburg Research Funds (Department of Economics, Harvard University), and Müller gratefully acknowledges support by the NSF via grant SES-0518036.

[Received February 2008. Revised April 2009.]

REFERENCES

Andrews, D. W. K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858. [453,454,458,459]

——— (1993), "Tests for Parameter Instability and Structural Change With Unknown Change Point," *Econometrica*, 61, 821–856. [458]

Andrews, D. W. K., and Monahan, J. C. (1992), "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," *Econometrica*, 60, 953–966. [458,459]

Arellano, M. (1987), "Computing Robust Standard Errors for Within-Groups Estimators," *Oxford Bulletin of Economics and Statistics*, 49, 431–434. [460]

Bakirov, N. K., and Székely, G. J. (2005), "Student's t-Test for Gaussian Scale Mixtures," *Zapiski Nauchnyh Seminarov POMI*, 328, 5–19. [453,455]

Bertrand, M., Duflo, E., and Mullainathan, S. (2004), "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics*, 119, 249–275. [461]

Bester, C. A., Conley, T. G., and Hansen, C. B. (2009), "Inference With Dependent Data Using Cluster Covariance Estimators," working paper, Chicago GSB. [454,462]

Bollerslev, T., Engle, R. F., and Nelson, D. B. (1994), "ARCH Models," in *Handbook of Econometrics*, Vol. IV, eds. R. F. Engle and D. McFadden, Amsterdam: Elsevier Science. [458]

Case, A. C. (1991), "Spatial Patterns in Household Demand," *Econometrica*, 59, 953–965. [462]

Conley, T. G. (1999), "GMM Estimation With Cross Sectional Dependence," *Journal of Econometrics*, 92, 1–45. [453,462,463]

Conley, T. G., and Dupor, B. (2003), "A Spatial Analysis of Sectoral Complementarity," *Journal of Political Economy*, 111, 311–352. [462]

Conley, T. G., and Taber, C. (2005), "Inference With 'Difference in Differences' With a Small Number of Policy Changes," Technical Working Paper 312, NBER. [461]

Deaton, A. (1997), *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*, Baltimore: John Hopkins University Press. [461]

Donald, S. G., and Lang, K. (2007), "Inference With Difference-in-Differences and Other Panel Data," *The Review of Economics and Statistics*, 89, 221–233. [454,461]

Fama, E. F., and MacBeth, J. (1973), "Risk, Return and Equilibrium: Empirical Tests," *Journal of Political Economy*, 81, 607–636. [454]

Ferguson, T. S. (1967), *Mathematical Statistics—A Decision Theoretic Approach*, New York and London: Academic Press. [463]

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. [461]

Hansen, B. E. (2000), "Testing for Structural Change in Conditional Models," *Journal of Econometrics*, 97, 93–115. [458]

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. [454,460]

Hansen, L. P. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054. [456]

Imhof, J. P. (1961), "Computing the Distribution of Quadratic Forms in Normal Variables," *Biometrika*, 48, 419–426. [455]

Jansson, M. (2004), "The Error in Rejection Probability of Simple Autocorrelation Robust Tests," *Econometrica*, 72, 937–946. [454]

Jegadeesh, N., and Karceski, J. (2009), "Long-Run Performance Evaluation: Correlation and Heteroskedasticity-Consistent Tests," *Journal of Empirical Finance*, 16, 101–111. [461]

Kézdi, G. (2004), "Robust Standard Error Estimation in Fixed-Effects Panel Models," *Hungarian Statistical Review*, 9, 95–116. [460]

Kiefer, N., and Vogelsang, T. J. (2002), "Heteroskedasticity-Autocorrelation Robust Testing Using Bandwidth Equal to Sample Size," *Econometric Theory*, 18, 1350–1366. [454,458]

——— (2005), "A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests," *Econometric Theory*, 21, 1130–1164. [454,458,459]

Kiefer, N. M., Vogelsang, T. J., and Bunzel, H. (2000), "Simple Robust Testing of Regression Hypotheses," *Econometrica*, 68, 695–714. [454,458]

Kim, C.-J., and Nelson, C. R. (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. [458]

Lehmann, E. L. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York: Wiley. [463]

Marshall, A. W., and Olkin, I. (1979), *Inequalities: Theory of Majorization and Its Applications*, New York: Academic Press. [457]

McConnell, M. M., and Perez-Quiros, G. (2000), "Output Fluctuations in the United States: What Has Changed Since the Early 1980's," *American Economic Review*, 90, 1464–1476. [458]

McElroy, T., and Politis, D. N. (2002), "Robust Inference for the Mean in the Presence of Serial Correlation and Heavy-Tailed Distributions," *Econometric Theory*, 18, 1019–1039. [459]

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. [462]

Müller, U. K. (2007), "A Theory of Robust Long-Run Variance Estimation," *Journal of Econometrics*, 141, 1331–1352. [454,458]

——— (2008), "Efficient Tests Under a Weak Convergence Assumption," working paper, Princeton University. [454,462,464]

Müller, U. K., and Watson, M. W. (2008), "Testing Models of Low-Frequency Variability," *Econometrica*, 76, 979–1016. [458]

Newey, W. K., and West, K. D. (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708. [453,461]

Petersen, M. A. (2009), "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches," *The Review of Financial Studies*, 22, 435–480. [460]

Rogers, W. H. (1993), "Regression Standard Errors in Clustered Samples," *Stata Technical Bulletin*, 13, 19–23. [453,460]

Swamy, P. A. V. B. (1970), "Efficient Inference in a Random Coefficient Regression Model," *Econometrica*, 38, 311–323. [462]

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–830. [453]

Wooldridge, J. M. (2003), "Cluster-Sample Methods in Applied Econometrics," *American Economic Review*, 93, 133–138. [461]