t-Statistic Based Correlation and Heterogeneity Robust Inference

Rustam Ibragimov
Harvard University

Ulrich K. Müller
Princeton University

May 8, 2008

Introduction

- Most economic data is observational

  $\Rightarrow$ data is plausibly not i.i.d.: think of countries, firms, or time series data

- Correlations and heterogeneity do not necessarily invalidate standard estimators, but standard errors and inference become a challenge

Consistent Variance Estimators

- White (1980) for heterogeneous and independent disturbances
- Newey and West (1987), Andrews (1991), etc. for time series disturbances
- Rogers (1993) and Arellano (1987) for clustered and panel data
- Conley (1999) for spatially correlated data

$\Rightarrow$ based on a Law of Large Numbers, and thus require "an infinite amount of independence"

$\Rightarrow$ poor small sample properties in many instances of interest

Inconsistent Variance Estimators

- Part of the problem is that sample variability of 'consistent' variance estimators is neglected

- Approaches that account for sample variability of variance estimator:
  - time series: Kiefer, Vogelsang and Bunzel (2000) and Kiefer and Vogelsang (2002, 2005), Müller (2007b)
  - panel data: Donald and Lang (2004), Hansen (2005)

$\Rightarrow$ We add to this literature and develop a general approach to robust inference when little is known about correlations and heterogeneity

Small Sample Result

- Bakirov and Székely (2005): If one applies the usual small sample t-test to independent Gaussian observations of possible heterogenous variance, then (two-sided) tests of level 5% or lower are conservative (i.e. rejection probability under the null hypothesis becomes smaller for unequal variances)

Our Approach

- Assume data can be classified in a finite number q of groups that allow asymptotically independent normal inference about the (scalar) parameter of interest β, so that $\hat{\beta}_j \overset{a}{\sim} id\mathcal{N}(\beta, v_j^2)$ for $j = 1, \cdots, q$. Time series example: Divide data into $q = 8$ consecutive blocks, and estimate the model 8 times.

- Treat $\hat{\beta}_j$ as observations for the usual t-statistic, and reject a 5% level test if t-statistic larger than usual critical value for $q - 1$ degrees of freedom. Results in valid inference by small sample result.

- Exploits information $\hat{\beta}_j \overset{a}{\sim} id\mathcal{N}(\beta, v_j^2)$ in an efficient way.

- Does not rely on single asymptotic model of sampling variability for estimated standard deviation.

- Important precursor: Fama–MacBeth (1973) method in finance.

Spatially Correlated Data

- In absence of more specific knowledge, exploit the default assumption that correlations become weaker as the distance between observations increases: Divide the sample of size n into q (approximately) equal sized blocks of neighboring observations.

- Monte Carlo design:
  - Inference about the mean of $n = 128$ observations located on a rectangular array of unit squares with 8 rows and 16 columns
  - Disturbances are $\chi^2_1 - 1$, with correlation of two observations given by $\exp(-\phi d)$ for some $\phi > 0$, where d is their Euclidian distance.
  - For t-statistic approach, $q = 2, 4, 8$ and 16 groups of spatial dimensions 8×8, 8×4, 4×4 and 2×4.

Monte Carlo Results

	t-statistic (q)				$\hat{\omega}^2_{UA}(b)$			
	2	4	8	16	0	2	4	8
				Size				
$\phi = \infty$	5.1	5.4	5.8	5.7	6.6	7.1	8.6	13.6
$\phi = 2$	5.3	6.0	7.0	8.9	16.2	12.2	11.8	16.3
$\phi = 1$	5.6	8.4	11.2	17.4	40.5	27.2	20.1	23.0
				Size Adjusted Power				
$\phi = \infty$	14.0	35.9	53.1	64.4	69.9	69.5	65.5	59.9
$\phi = 2$	15.5	38.3	55.9	63.4	71.3	70.0	65.2	59.3
$\phi = 1$	14.8	35.4	53.7	63.6	70.7	68.5	61.7	53.3

$\hat{\omega}^2_{UA}(b)$: Conley's (1999) unweighted average of local covariances within distance b

Plan of Talk

1. Introduction
2. The Small Sample t-Test
   (a) Conservativeness for Heterogenous Variances
   (b) Optimality
3. t-Statistic Based Large Sample Robust Inference
   (a) Basic Idea
   (b) Optimality
   (c) Comparison with Standard Inference and Known Asymptotic Variance
4. Applications
5. Conclusions

The Small Sample t-Test

- Let $X_j, j = 1, \cdots, q$ with $q \geq 2$, be distributed independent $\mathcal{N}(\mu, \sigma_j^2)$
- We are interested in testing $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$
- Use t-statistic

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

$\bar{X} = q^{-1}\sum_{j=1}^{q} X_j$ and $s_X^2 = (q-1)^{-1}\sum_{j=1}^{q}(X_j - \bar{X})^2$ and reject for large values of $|t|$

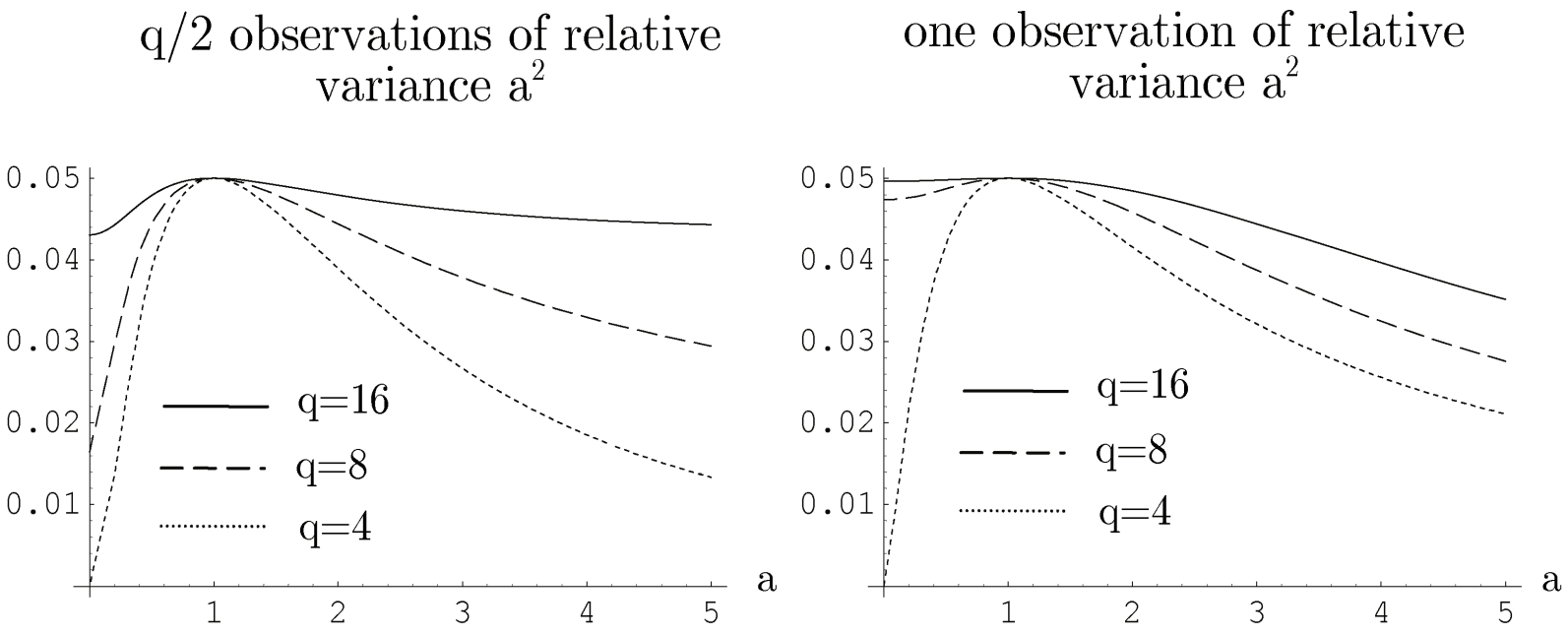
Conservativeness

Theorem (Bakirov and Székely 2005): Let $\mathsf{cv}_q(\alpha)$ be the critical value of the usual two-sided t-test based on $|t|$ of level α, i.e. $P(|T_{q-1}| > \mathsf{cv}_q(\alpha)) = \alpha$, where T_k is student-t distributed with k degrees of freedom, and let Φ denote the cumulative density function of a standard normal random variable.

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

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

Effective Rejection Probabilities



Small Sample Optimality of t-Statistic

- Let $X_j, j = 1, \cdots, q$ with $q \geq 2$, be distributed independent $\mathcal{N}(\mu, \sigma_j^2)$, and consider the hypothesis test

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

- Theorem: Let cv be such that $P(|T_{q-1}| > \text{cv}) = \alpha \leq 0.05$. For any $q \geq 2$, a test that rejects the null hypothesis for $|t| >$ cv is the uniformly most powerful scale invariant level α test.

- Proof:
  - For equal variance case, t-test is UMP scale invariant.
  - Conservativeness of t-test implies that equal variance case under the null hypothesis is least favorable.

Asymptotic t-Statistic Based Inference

- Partition the data 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 (where $\max_{1 \leq j \leq q} \sigma_j^2 > 0$). Satisfied for many models as long as $\min_j n_j \to \infty$, linear or nonlinear.
  - $\sqrt{n}(\hat{\beta}_i - \beta)$ and $\sqrt{n}(\hat{\beta}_j - \beta)$ are asymptotically independent for $i \neq j$.

  $\Rightarrow$ this amounts to

$$\sqrt{n}(\hat{\beta}_1 - \beta, \cdots, \hat{\beta}_q - \beta)' \Rightarrow \mathcal{N}(0, \mathrm{diag}(\sigma_1^2, \cdots, \sigma_q^2))$$

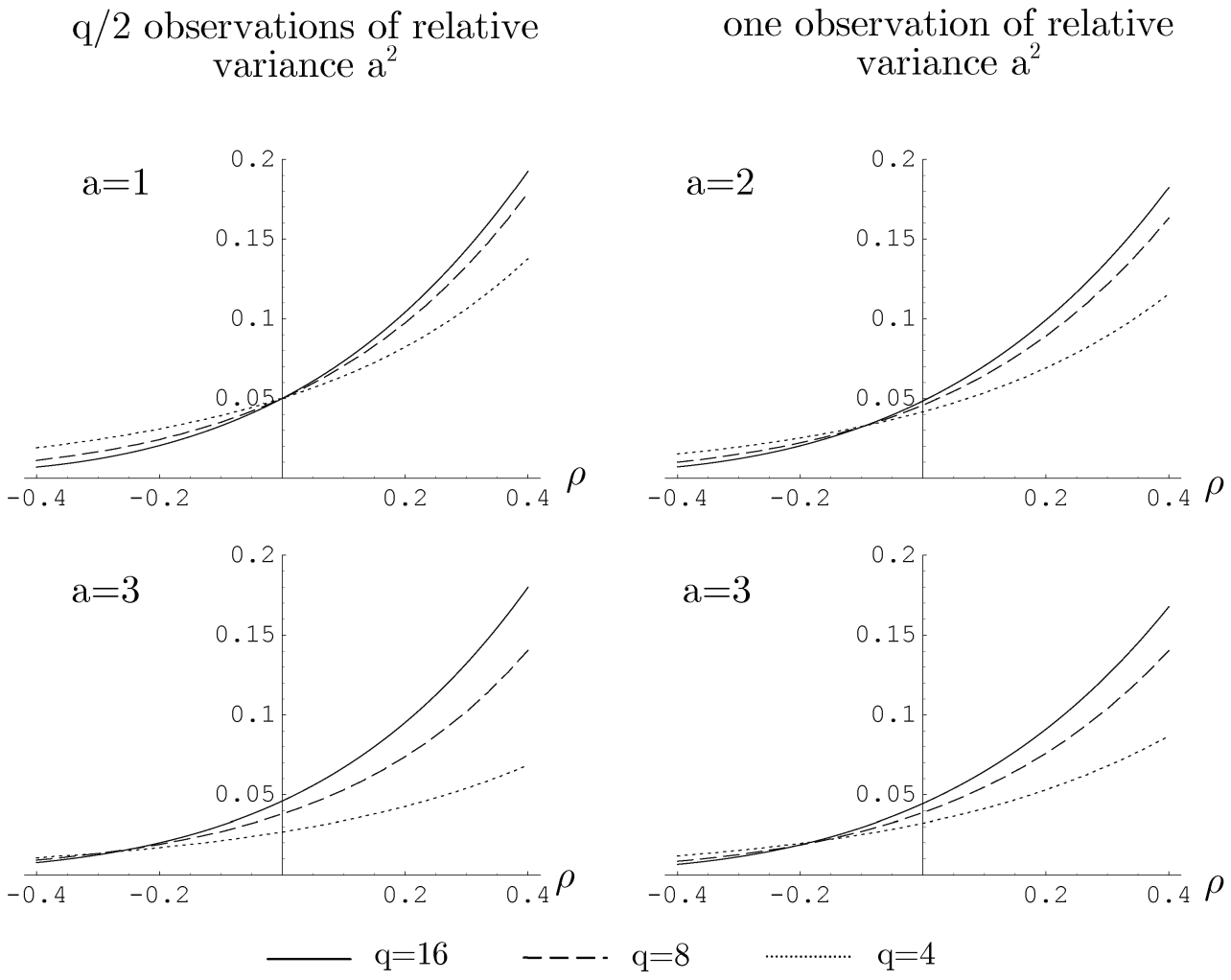
Asymptotic t-Statistic Based Inference

- Rejection of $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$ if $|t_\beta|$ 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

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

with $\overline{\hat{\beta}} = q^{-1}\sum_{j=1}^{q}\hat{\beta}_j$ and $s^2_{\hat{\beta}} = (q-1)^{-1}\sum_{j=1}^{q}(\hat{\beta}_j - \overline{\hat{\beta}})^2$, is asymptotically valid inference by small sample t-test result and the Continuous Mapping Theorem.

Size Control under AR(1) Correlation



Asymptotic Optimality I

- Key assumption about underlying data $\mathbf{Y}_T$: Under local alternatives of $H_0 : \beta = \beta_0$ of the form $\beta = \beta_n = \beta_0 + \mu/\sqrt{n}$

$$\{\sqrt{n}(\hat{\beta}_j - \beta_0)\}_{j=1}^q \Rightarrow \{X_j\}_{j=1}^q \text{ where } X_j \sim \mathcal{N}(\mu, \sigma_j^2) \qquad (1)$$

- Basic motivation: valid inference when little is known about correlation structure in data $\mathbf{Y}_T$

- Formalize this motivation with notion of robustness: Call a 5% level test of $H_0 : \beta = \beta_0$ *asymptotically robust* if its asymptotic rejection probability is at most 5% for all data generating processes for $\mathbf{Y}_T$ that satisfy (1) with $\beta = \beta_0$, i.e. whenever

$$\{\sqrt{n}(\hat{\beta}_j - \beta_0)\}_{j=1}^q \Rightarrow id\mathcal{N}(0, \sigma_j^2)$$

Asymptotic Optimality II

- Müller (2007a) considers asymptotically robust tests in this sense and finds that best robust test is simply given by best test in "limiting problem" with limiting random variables X_j assumed observed, evaluated at sample analogues
  - among all scale invariant asymptotically robust tests, t-statistic approach maximizes asymptotic local power uniformly against all local alternatives where $\beta = \beta_n = \beta_0 + \mu/\sqrt{n}$ for some $\mu \neq 0$ and $\sigma_i^2 = \sigma_j^2$ for all i, j
  - for any 5% level test φ_T that has asymptotically higher power under (1) than the t-statistic approach, there exists a data generating process with $\{\sqrt{n}(\hat{\beta}_j - \beta_0)\}_{j=1}^q \Rightarrow id\mathcal{N}(0, \sigma_j^2)$ for which φ_T has asymptotic rejection probability greater than 5%
  - it is impossible to use data-driven methods to select the groups while maintaining robustness

Comparison with Inference with Known Asymptotic Variance

- Typically,

$$\sqrt{n}(\hat{\beta}_1 - \beta, \cdots, \hat{\beta}_q - \beta)' \Rightarrow \mathcal{N}(0, \text{diag}(\sigma_1^2, \cdots, \sigma_q^2))$$

  is weaker than what is required to consistently estimate the asymptotic variance. Consistent estimation not only requires more assumptions on correlation structure, but typically also higher moments.

- What are the efficiency cost of this additional robustness, i.e. how does the t-statistic approach compare to an approach based on consistent variance estimation (if the variance can indeed be consistently estimated)?

Comparison with Inference with Known Asymptotic Variance

- Consider question in linear regression

$$y_i = x_i'\theta + u_i, \quad i = 1, \cdots, n$$

where $E[x_i u_i] = 0$, θ is $k \times 1$ and we are interested in conducting inference about the first element of θ, denoted by β.

- Paper considers general exactly identified GMM problem, with (almost) identical results.

Properties of Group Estimators

- Let $\mathcal{G}_j$ the set of indices of group j observations. Suppose the OLS estimator $\hat{\theta}_j$ based on group $j = 1, \cdots, q$ data satisfies

$$\sqrt{n}(\hat{\theta}_j - \theta) = \Gamma_j^{-1} Q_j + o_p(1) \Rightarrow \mathcal{N}(0, \Gamma_j^{-1} \Omega_j \Gamma_j^{-1})$$

where $n^{-1} \sum_{i \in \mathcal{G}_j} x_i x_i' \xrightarrow{p} \Gamma_j$ and $Q_j = n^{-1/2} \sum_{i \in \mathcal{G}_j} x_i u_i \Rightarrow \mathcal{N}(0, \Omega_j)$, and

$$(Q_1', \cdots, Q_q')' \Rightarrow \mathcal{N}(0, \text{diag}(\Omega_1, \cdots, \Omega_q)).$$

- The simple average of the group estimators $\bar{\hat{\theta}}$ then satisfies

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

where $\bar{\Sigma}_q = q^{-2} \sum_{j=1}^{q} \Gamma_j^{-1} \Omega_j \Gamma_j^{-1}$.

Full Sample Estimator

- In contrast, full sample OLS estimator $\hat{\theta} = (\sum_{i=1}^{n} x_i x_i')^{-1} \sum_{i=1}^{n} x_i y_i$ satisfies

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

  where $\Sigma_q = \left(\sum_{j=1}^{q} \Gamma_j\right)^{-1} \left(\sum_{j=1}^{q} \Omega_j\right) \left(\sum_{j=1}^{q} \Gamma_j\right)^{-1}$.

- Estimator $\hat{\theta}$ not efficient under group heterogeneity. Efficient estimator would be Generalized Least Squares.

- Feasible GLS estimator depends on $\{\Omega_j\}_{j=1}^{q}$, which is assumed difficult to estimate. We thus focus on comparison of t-statistic approach with $\overline{\hat{\beta}}$ in the numerator with inference based on $\hat{\beta}$ and known σ (the (1,1) element of Σ_q).

General Comparison

- In general, $\bar{\Sigma}_q$ and Σ_q are not identical. t-statistic approach and inference based on $\hat{\beta}$ with σ^2 known differ not only in the denominator, but also in the numerator.

- Both tests are consistent against fixed alternatives and have power against the same local alternatives $\beta = \beta_0 + \mu/\sqrt{n}$.

- Theorem: (Almost) nothing else can be said in general without knowing $\{\Omega_j\}_{j=1}^q$

Special Case

- Consider the special case $\Gamma_j = \Gamma$ for $j = 1, \cdots, q$. This naturally arises when the groups have an equal number of observations n/q, and the average of $x_i x_i'$ is homogenous across groups (leading example: i.i.d. data). Then

$$\sqrt{n}(\bar{\hat{\theta}} - \theta) = \sqrt{n}(\hat{\theta} - \theta) + o_p(1)$$

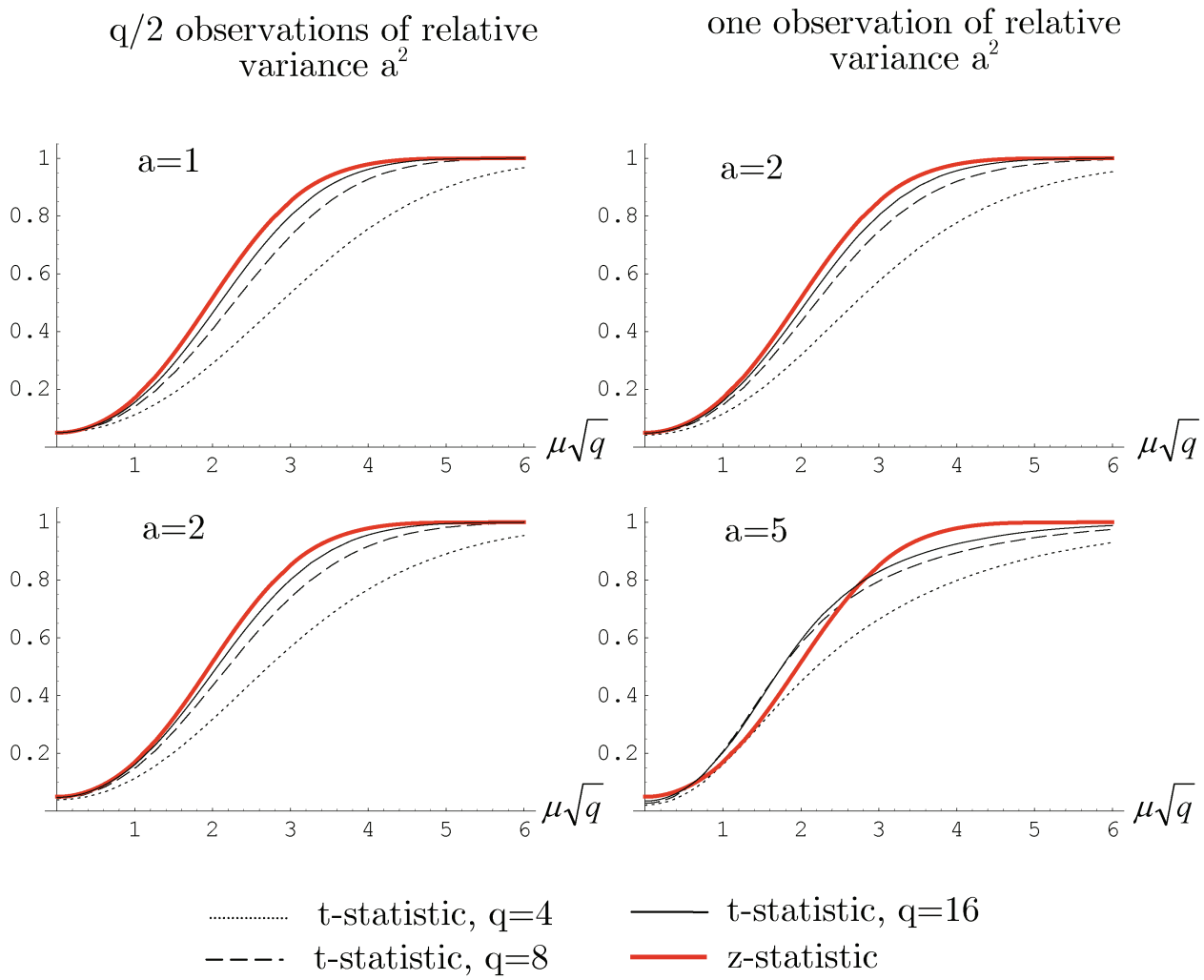
  and $\hat{\beta}$ and $\bar{\hat{\beta}}$ are asymptotically equivalent to order $\sqrt{n}$.

- The asymptotic local power of tests based on t_β and $\hat{\beta}$ with σ^2 known simply reduces to the small sample power of the small sample t-statistic and z-statistic of $H_0 : \mu = 0$ when $X_i \sim \mathcal{N}(\mu, \sigma_i^2)$, that is

$$z = \frac{\sum_{i=1}^q X_i}{\sqrt{\sum_{i=1}^q \sigma_i^2}}$$

  where σ_i^2 is the (1,1) element of $\Gamma^{-1}\Omega_i\Gamma^{-1}$.

Numerical Comparison



Applications

- t-statistic approach requires

$$\sqrt{n}(\hat{\beta}_1 - \beta, \cdots, \hat{\beta}_q - \beta)' \Rightarrow \mathcal{N}(0, \mathsf{diag}(\sigma_1^2, \cdots, \sigma_q^2)).$$

- Discussion of plausible group choices for

  1. Panel Data
  2. Time Series Data
  3. Spatially Correlated Data

  and some small sample results, with focus on linear regression examples considered in other studies.

Panel Data

- Panel with potential time series correlation and few individuals N

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

where $\{x_{i,t}, u_{i,t}\}_{t=1}^{T}$ are independent across i and $E[x_{i,t}u_{i,t}] = 0$ for all i, t.

- t-statistic approach asymptotically valid if $T^{-1}\sum_{t=1}^{T} x_{i,t}x'_{i,t} \xrightarrow{p} \Gamma_i$ and $T^{-1/2}\sum_{t=1}^{T} x_{i,t}u_{i,t} \Rightarrow \mathcal{N}(0, \Omega_i)$ for all i as $T \to \infty$ and N fixed for some full rank matrices Γ_i and Ω_i.

- Hansen (2007) shows that usual t-statistic with Rogers (1993) standard errors converges under the null to a scaled t-statistic with $q - 1$ degrees of freedom under 'asymptotic homogeneity', i.e. when $\Gamma_i = \Gamma$ and $\Omega_i = \Omega$ for all i.

Monte Carlo Results

Same design as in Kézdi (2004): Linear Regression, one nonconstant regressor that follows Gaussian AR(1) with coefficient ρ_x, disturbances are AR(1) with coefficient ρ_u, $N = 10$, $T = 25$. 5% level test about coefficient of one nonconstant regressor.

	homoskedastic			heteroskedastic		
ρ_x	0	0.9	1	0	0.9	1
ρ_u	0	0.9	0.5	0	0.9	0.5
	Size					
t-statistic	5.0	5.0	4.4	4.6	4.0	3.8
clustered	5.2	7.2	8.7	4.9	7.9	14.7
clustered, FE	5.1	6.2	6.8	4.9	6.2	8.6
	Size Adjusted Power					
t-statistic	58.7	39.4	42.5	54.6	62.9	73.7
clustered	60.8	31.9	83.6	51.9	33.6	52.1
clustered, FE	59.7	38.3	50.8	52.0	46.2	45.1

Panel Data II

- For applications in Finance, concern about cross section correlation. Our results justify Fama–MacBeth method where regression is run cross sectionally for each time period, and inference is based on t-statistic of resulting estimators $\hat{\beta}_j$, $j = 1, \cdots, T$, even for small T and potential heterogeneity of variances as long as no correlation across t.

- For corporate Finance applications, uncorrelatedness in time is often implausible. Rather than to try to consistently estimate the long-run variance with few observations, the approach taken here suggests forming groups of more than one unit in time to achieve approximate independence.

- Alternatively, assume independence in cross section dimension, say, across industries, as in Froot (1989). Combinations are possible.

- Same possibilities for long-run event studies, country panel data, city panel data and so forth.

Monte Carlo Results

Same design as in Thompson (2006): Linear Regression, one nonconstant regressor, $N = 50$, $T = 25$.

"individual persistence": $u_{i,t} = \xi_t + \eta_{i,t}$, $\eta_{i,t} = \rho\eta_{i,t-1} + \varepsilon_{i,t}$

"common persistence": factor structure $u_{i,t} = h_i f_t + \varepsilon_{i,t}$, $f_t = \rho f_{t-1} + \xi_t$, $h_i \sim N(1, 0.25)$

	Individual Persistence			Common Persistence		
ρ	0	0.7	0.9	0	0.7	0.9
			Size			
t-statistic $q = 2$	4.9	5.0	6.0	4.9	5.3	6.3
t-statistic $q = 4$	4.9	5.4	9.8	4.1	5.3	10.4
t-statistic $q = 8$	4.6	6.4	17.1	3.9	7.1	16.8
FM with Newey-West	12.6	19.8	34.8	11.4	14.2	23.4
cluster by i and t	9.3	8.8	7.0	10.2	29.9	49.5
cluster by i and t + cp	16.3	14.9	12.1	17.0	26.4	38.3
			Size Adjusted Power			
t-statistic $q = 2$	12.9	16.2	14.9	20.3	18.1	20.8
t-statistic $q = 4$	30.5	45.5	45.3	58.4	58.4	60.6
t-statistic $q = 8$	50.9	67.6	61.3	59.5	67.3	68.2
FM with Newey-West	100	91.6	58.9	57.3	47.4	47.1
cluster by i and t	46.8	67.6	74.8	86.3	66.2	70.2
cluster by i and t + cp	31.7	52.7	69.8	69.6	53.6	60.1

Time Series Data

- In absence of more specific knowledge, exploit the default assumption that correlations between observations become weaker the further apart in time they are: Divide the sample of size T into q (approximately) equal sized groups of consecutive observations.

- Under a wide range of assumptions on the underlying model and observations, OLS model satisfies

$$\sup_{0\leq r\leq 1} ||T^{-1}\sum_{t=1}^{\lfloor rT\rfloor} x_t x_t' - \int_0^r \Gamma(\lambda)d\lambda|| \xrightarrow{p} 0 \tag{2}$$

$$T^{-1/2}\sum_{t=1}^{\lfloor \cdot T\rfloor} x_t u_t \Rightarrow \int_0^{\cdot} h(\lambda)dW(\lambda) \tag{3}$$

where $\Gamma(\cdot)$ is a positive definite $k\times k$ matrix function and $h(\cdot)$ is nonzero.

Time Series Data

- With that convergence, t-statistic approach is asymptotically valid, since

$$\sqrt{T}\begin{pmatrix} \hat{\theta}_1 - \theta \\ \hat{\theta}_2 - \theta \\ \vdots \\ \hat{\theta}_q - \theta \end{pmatrix} \Rightarrow \begin{pmatrix} \left(\int_0^{1/q} \Gamma(\lambda)d\lambda\right)^{-1} \int_0^{1/q} h(\lambda)dW(\lambda) \\ \left(\int_{1/q}^{2/q} \Gamma(\lambda)d\lambda\right)^{-1} \int_{1/q}^{2/q} h(\lambda)dW(\lambda) \\ \vdots \\ (\int_{(q-1)/q}^{1} \Gamma(\lambda)d\lambda)^{-1} \int_{(q-1)/q}^{1} h(\lambda)dW(\lambda) \end{pmatrix}$$

- In contrast, no other known way of conducting asymptotically valid inference under (2) and (3):
  - Kiefer and Vogelsang (2002, 2005) approach requires $\Gamma(\cdot)$ and $h(\cdot)$ to be constant
  - Müller (2007b) shows that no long-run variance estimator can be consistent for $\mathsf{Var}[\int_0^1 h(\lambda)dW(\lambda)]$ for all processes that satisfy (3)

Monte Carlo Results

Same design as in Andrews (1991): Linear Regression, 5 regressors, 4 non-constant regressors are independent draws from stationary Gaussian AR(1), as are the disturbances, + heteroskedasticity. $T = 128$, 5% level test about coefficient of one nonconstant regressor.

	t-statistic (q)			$\hat{\omega}^2_{QA}$	$\hat{\omega}^2_{PW}$	$\hat{\omega}^2_{BT}(b)$			
	2	4	8			0.05	0.1	0.3	1
ρ	Size								
0	4.9	4.7	4.6	7.1	8.1	6.7	6.6	6.0	6.2
0.5	4.8	4.6	4.6	10.4	9.9	9.4	8.4	7.5	7.0
0.8	4.8	4.9	5.4	19.1	17.3	18.6	15.6	12.8	11.9
0.9	4.9	5.1	6.1	28.9	25.4	29.9	24.9	20.5	18.8
ρ	Size Adjusted Power								
0	15.1	38.4	53.7	62.7	60.6	60.7	58.6	51.9	47.2
0.5	14.5	38.2	55.9	57.0	56.2	56.0	53.5	48.4	44.2
0.8	15.4	45.1	66.0	52.9	51.7	54.0	52.6	46.9	42.4
0.9	17.2	56.7	77.6	57.5	54.6	58.7	57.5	51.4	46.6

Conclusion

1. General and simple approach to correlation robust inference in large samples without consistent variance estimation, not restricted to time series.

2. Valid inference even under pronounced heterogeneity, unlike other time series tests based on inconsistent variance estimators.

3. Method imposes only a 'finite amount of independence' through the assumption that estimators from different groups are independent and Gaussian, and approach exploits this assumption in an efficient way.

4. Challenge to choose groups in practice. But inference requires some assumption on correlation structure, and other methods make more implicit and even less interpretable assumptions.