A Performance Comparison of Large-n Factor Estimators

Zhuo Chen
Northwestern University

Gregory Connor
Maynooth University

Robert Korajczyk
Northwestern University

Motivation

- A number of methods for estimating large-scale factor models
  - The method accommodate alternative assumptions about factor structure
    - Cross-sectional heteroskedasticity
    - Time-series heteroskedasticity
    - Approximate versus strict factor models
    - Balanced vs. unbalanced panels
  - How do these asymptotic methods perform for various sample sizes? (How large is large?)
  - How sensitive are they to various data features (heteroskedasticity, unbalanced panels)?
  - We use calibrated panel data and a very large simulation study to compare their performance under various conditions.

Multi-factor Asset Pricing Model

- Return Generating Process; n assets, T time periods:

$$r_t = E(r_t) + Bf_t + \varepsilon_t \qquad \text{(RGP)}$$

- Asset Pricing Model:

$$E(r_t) = e^n r_{0,t} + B\mu_t \qquad \text{(APM)}$$

- RGP + APM:

$$R_t = r_t - e^n r_{0,t} = B(\mu_t + f_t) + \varepsilon_t$$

Calibration

- Number of assets: n = 250,500,750,1000, …, 10,000
- Number of time periods: T = 240 months
- Number of factors: k = 3
  - Calibrate the factors to the Fama/French Market, HML, and SMB factors over 1991-2010
    - $\mu_F \times 100 = [0.58; 0.31; 0.34]$
    - $\sigma_F \times 100 = [4.46; 3.51; 3.37]$
    - Draw from $N(\mu_F, \Sigma_F)$, $\Sigma_F = \text{Diag}(\sigma^2_F)$

Calibration

- Use data from the 1991-2010 period to calibrate the random draws of factor loadings, residual risk, and patterns of missing data
  - Sample 10,937 NYSE/NASDAQ/AMEX stocks with at least 36 observations
  - Estimate time series regressions of returns on Fama/French factors
- Calibrate factor loadings *B* from the panel of regression coefficents
  - μ_B = [1.001; 0.882; 0.210]
  - $\sigma_B \times 100$ = [0.824; 1.204; 1.315]
  - Draw from $N(\mu_B, \Sigma_B)$, $\Sigma_B = Diag(\sigma^2_B)$
- Idiosyncratic risk: estimate σ_i for each asset in CRSP sample

Simulation

- Simulate 24 different cases, 24=4x3x2, using different econometric assumptions
- Four alternative assumptions about idiosyncratic heteroskedasticity:
    1. Time series and cross-sectional homoskedasticity
    2. Cross-sectional heteroskedasticity
    3. Time series heteroskedasticity
    4. Cross-sectional and time series heteroskedasticity

Simulation

- Three alternative assumptions about idiosyncratic cross-correlations, ρ
  - No cross-correlation: $\rho_{i,i+1} = 0$ (strict factor model)
  - $\rho_{i,i+1} = 0.25$, and $\rho_{i,i+1} = 0.50$
- Two alternative assumptions on missing data
  - No missing data (balanced panel)
  - Missing data
    - For *j*=1,…, *n* draw from CRSP sample (with replacement) and use the pattern of missing data observed in the CRSP data.

Simulation

- 5000 samples each.
- 288 billion = 5,000x10,000x240x24 simulated returns are used in the study
- All of the estimation methodologies are reasonably computation-intensive (e.g., compute leading eigenvectors) and some also require iteration.
- None require numerical search methods
- Computer details here

Estimators – Balanced Panel

- Asymptotic Principal Components (APC)
  - Connor and Korajczyk (1986)
  - Approximate *k*-factor model
  - Allows general cross-sectional heteroskedasticity of idiosyncratic returns
  - Allows limited time-series heteroskedasticity of idiosyncratic returns
    - Can vary over time at the firm level with the cross-sectional average remaining constant.
  - Consider case with *T* fixed and *n* increasing

Estimator 1 – APC Balanced Panel

- Asymptotic Principal Components (APC)
  - *R: nxT* matrix of returns
  - *F: kxT* matrix of unobserved factors

$$\Omega = \frac{1}{n} R'R$$

$$\widehat{F} = eigvec_k(\Omega)$$

Estimator 2 – APC Balanced Panel with Efficiency Gain

- Asymptotic Principal Components (APC-X)
  - Connor and Korajczyk (1988)
  - Apply APC to weighted returns to improve estimation efficiency
  - Estimate factor loadings and diagonal idiosyncratic variance matrix, *D*

$$R^* = \widehat{D}^{-1/2} R$$

$$\Omega^* = \frac{1}{n} R^{*\prime} R^*$$

$$\widehat{F} = eigvec_k(\Omega^*)$$

Estimator 3 – EM-MLFA Balanced Panel

- EM-based Maximum Likelihood Factor Analysis (MLFA-S)
  - Stroyny (1992)
  - Apply Rubin-Thayer EM algorithm rather than the (unfeasible) Joreskog algorithm
    - No Heywood cases; no nxn-matrix inversion

Estimator 4 – HFA Balanced Panel

- Heterskedastic Factor Analysis (HFA)
  - Jones (2001)
  - *V: TxT* = $Diag(\widehat{\sigma}_t^2)$
  - Scaled Ω matrix: $V^{-1/2}\Omega V^{-1/2}$
  - Factors: $\widehat{F} = eigvec_k(V^{-1/2}\Omega V^{-1/2})$
  - Iterate between F and Ω estimation steps

Estimator 1 – APC Unbalanced Panel

- Asymptotic Principal Components over observed data (APC-M)
  - Connor and Korajczyk (1987)
  - Allows for an unbalanced panel
  - Estimate Ω^u by averaging component-by-component over available data for date pairs t,t^*
  - Factor estimates from eigenvectors of Ω^u

Estimator 2 – Unbalanced Panel

- Stock and Watson (1998) *EM*-based estimator (APC-EM)
  - Allows for unbalanced panels using EM for missing data
  - Estimate initial factor model (using ACP-M) on the balanced subset
  - At iteration *j* define the elements of the return matrix

$$\Omega^* = \frac{1}{n} R^{*\prime} R^*$$

$R^*_{i,t} = R_{i,t}$ *if not missing*

$R^*_{i,t} = B^{j-1} F^{j-1}{}_t$ *if missing*

$$\widehat{F}^j = eigvec_k(\Omega^*)$$

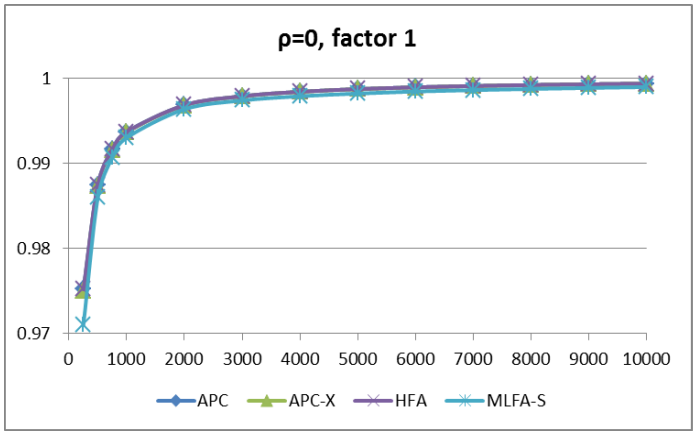
Estimator 3 – Unbalanced Panel

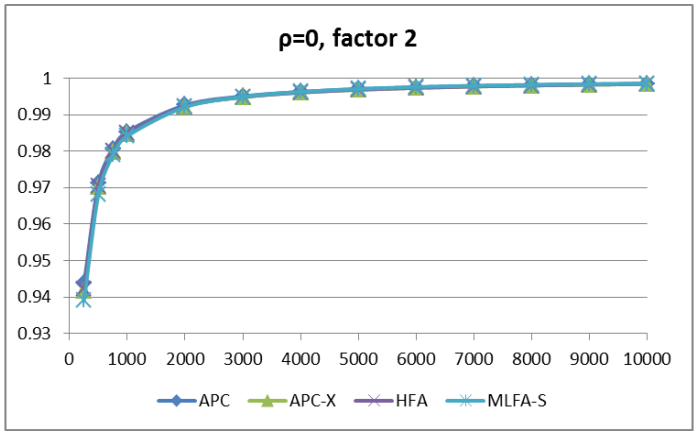
- Asymptotic Principal Components (APC-MX)
  - Connor and Korajczyk (1987, 1988)
  - Apply APC to obtain initial idiosyncratic covariance matrix D
  - Scale returns by $D^{-1/2}$
  - Calculate Ω^{u} over observed data using scaled returns
  - Does not weaken econometric assumptions, but may increase estimation efficiency

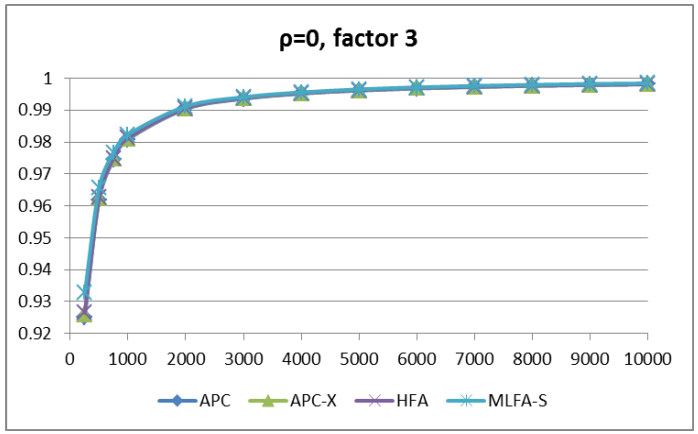
Estimator 4 – Unalanced Panel

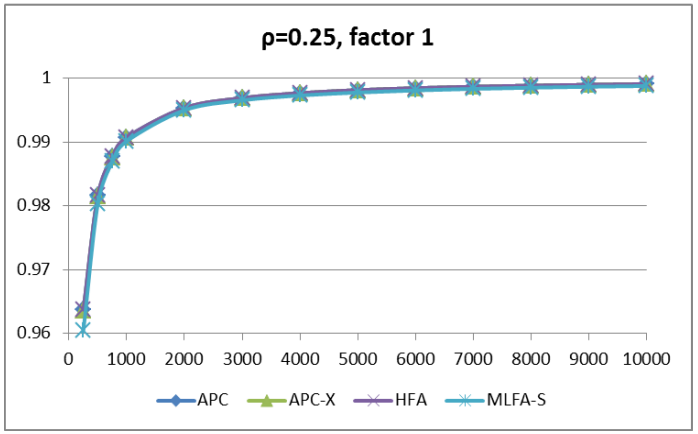
- Heterskedastic Factor Analysis (HFA-M)
  - Combines Jones (2001) HFA with Connor and Korajczyk (1987) missing data method
  - Estimate V and Ω over observed data

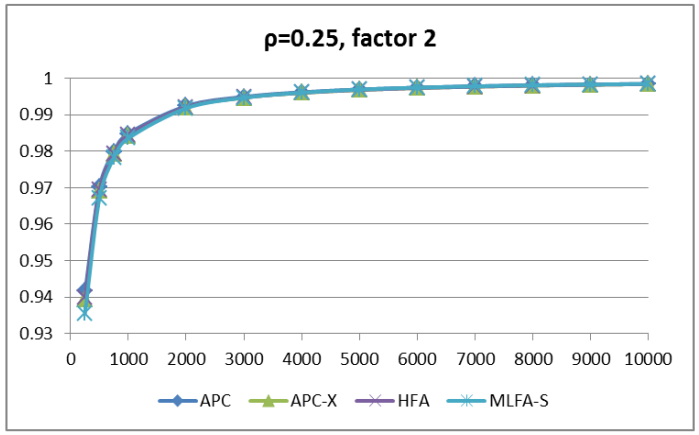
Cross-sectional homoskedastic and time series homoskedastic

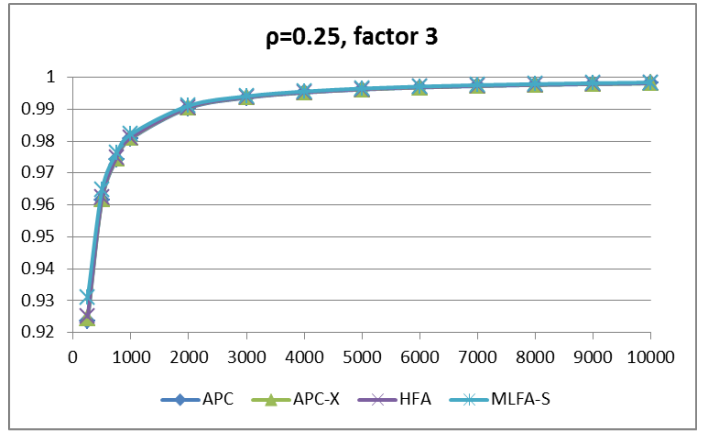


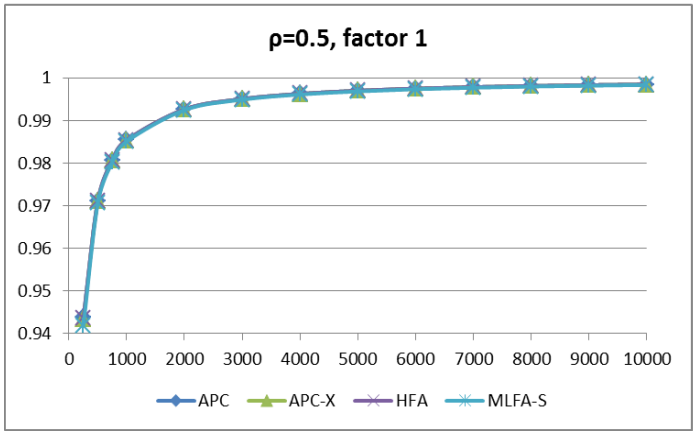


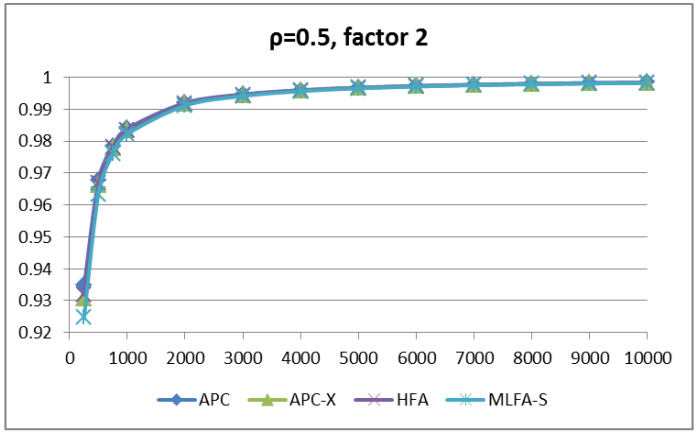


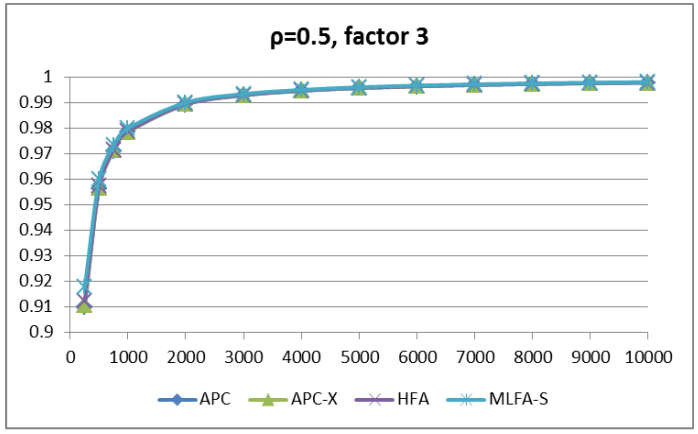




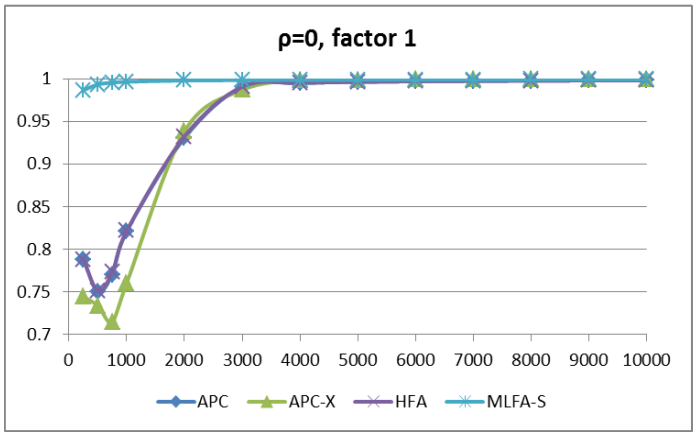


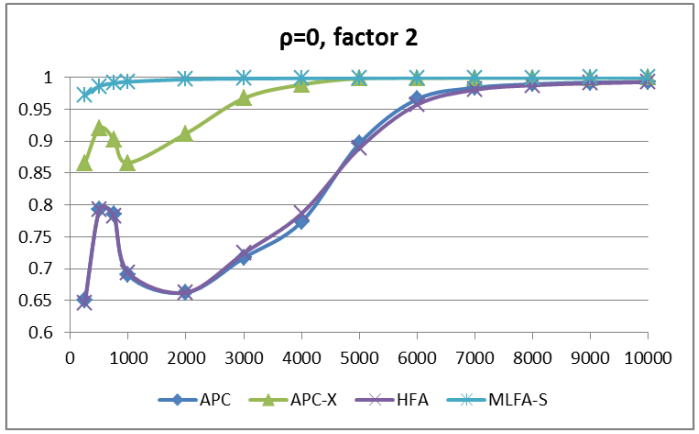


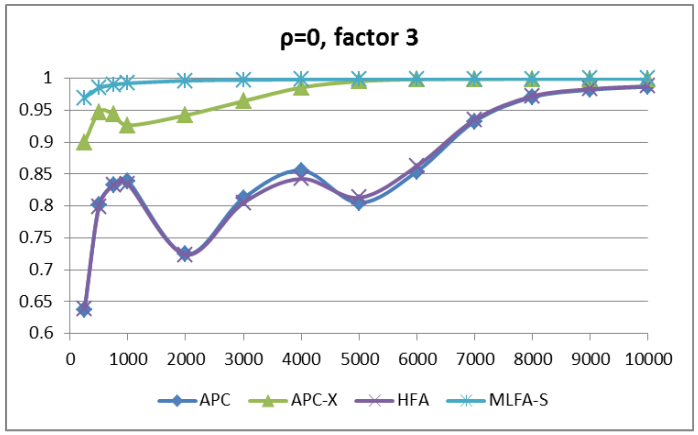


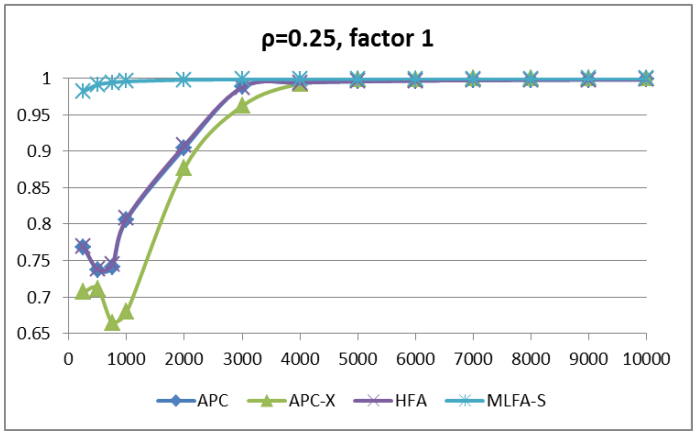


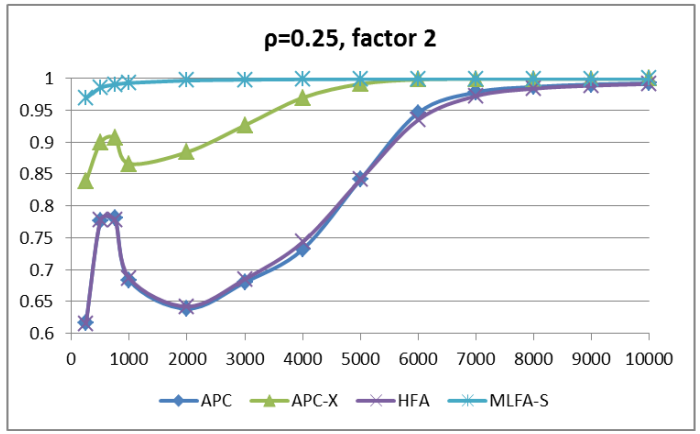
Cross-sectional heteroskedastic and time series homoskedastic

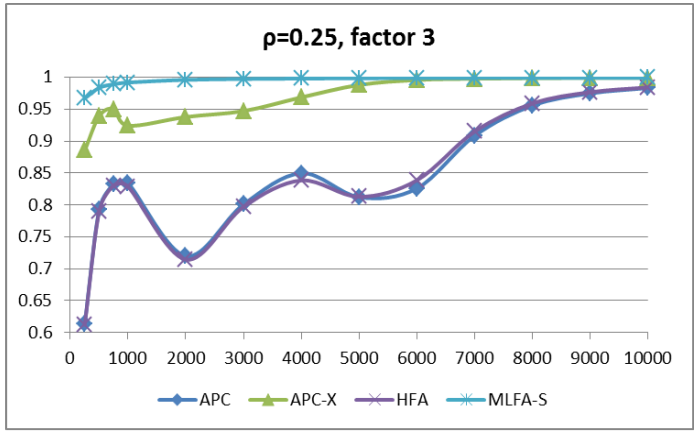


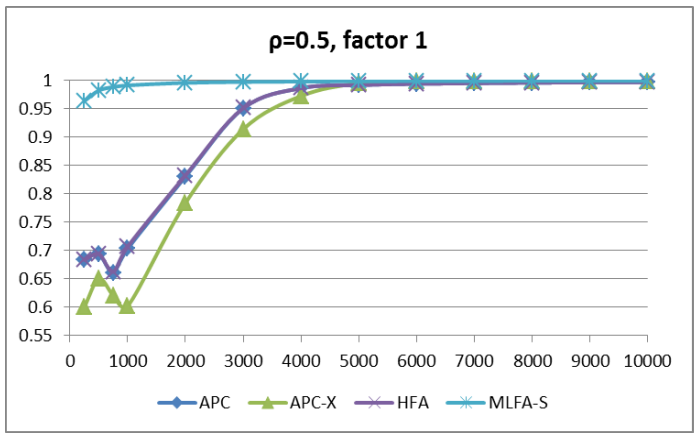


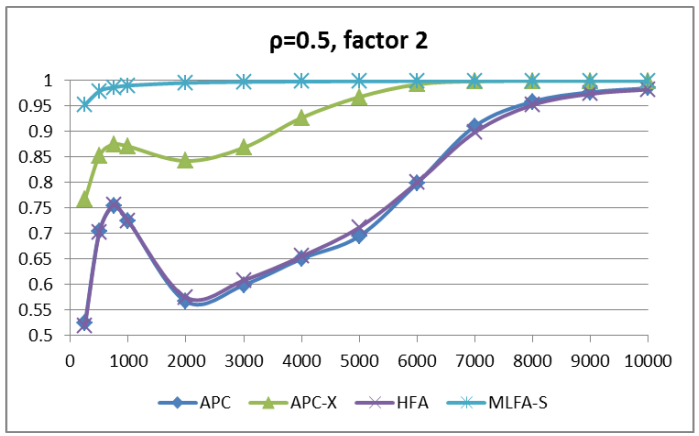


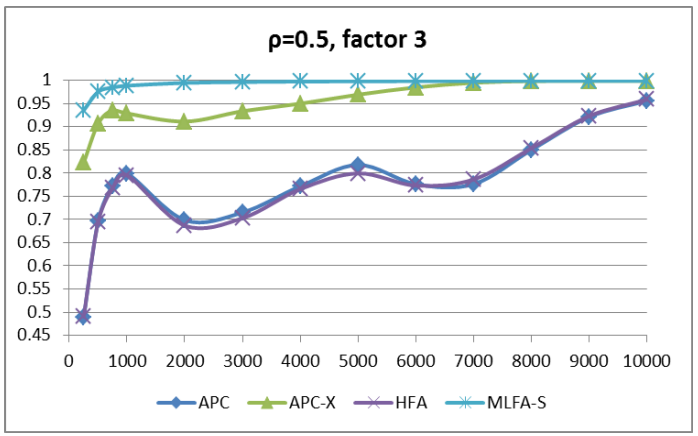




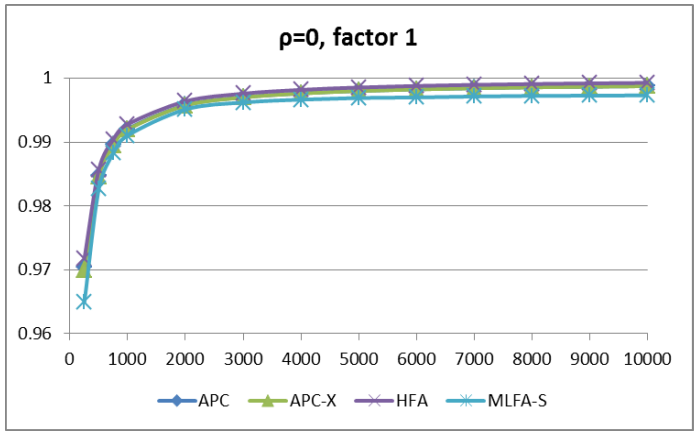


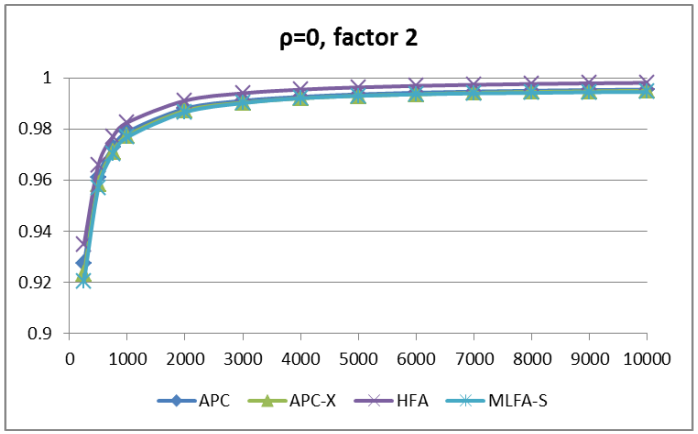


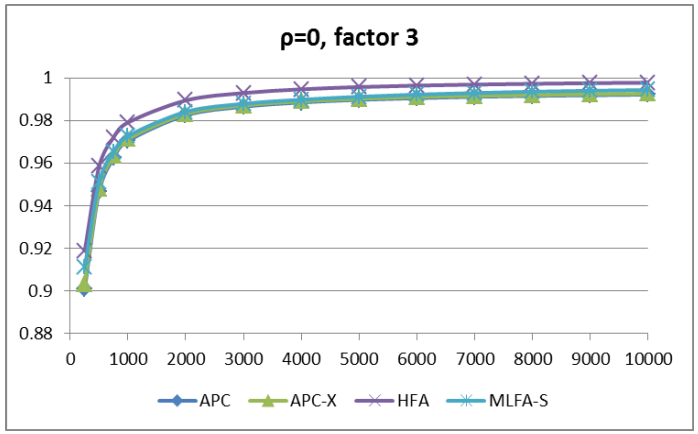


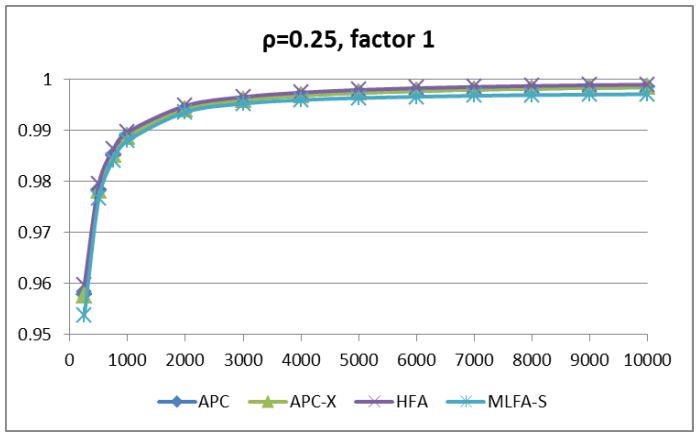


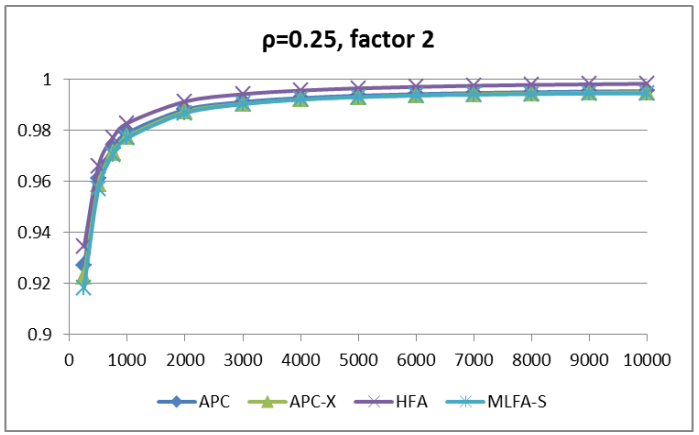
Cross-sectional homoskedastic and time series heteroskedastic

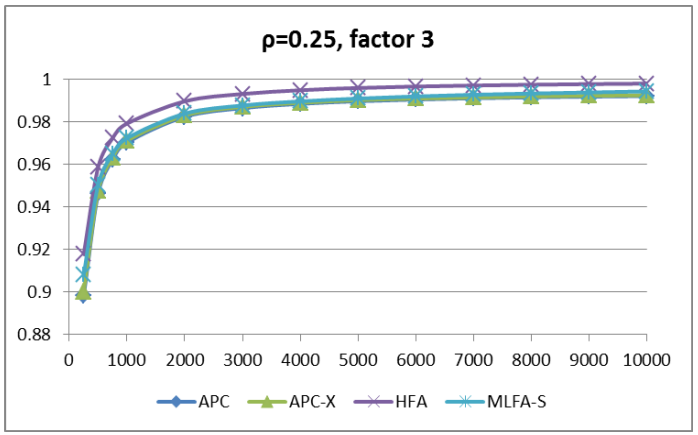


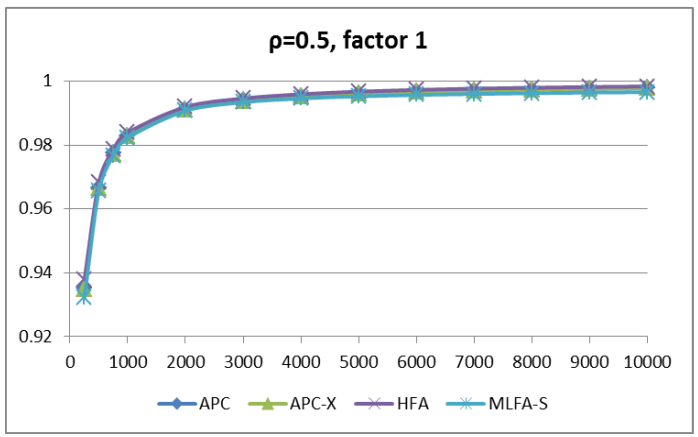


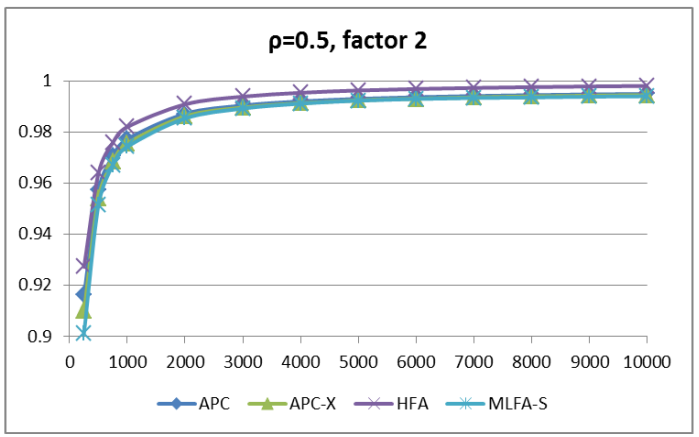


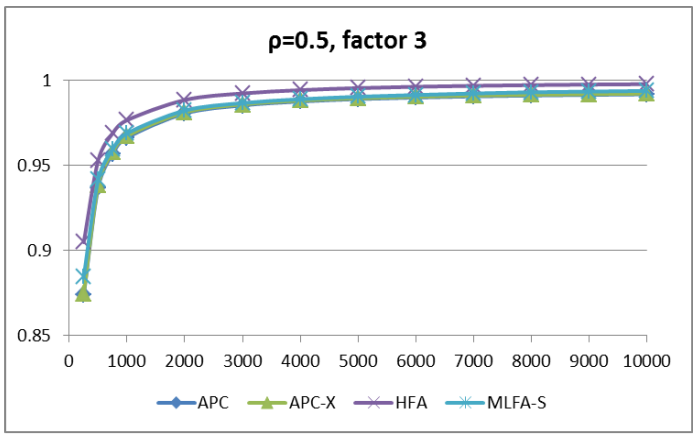




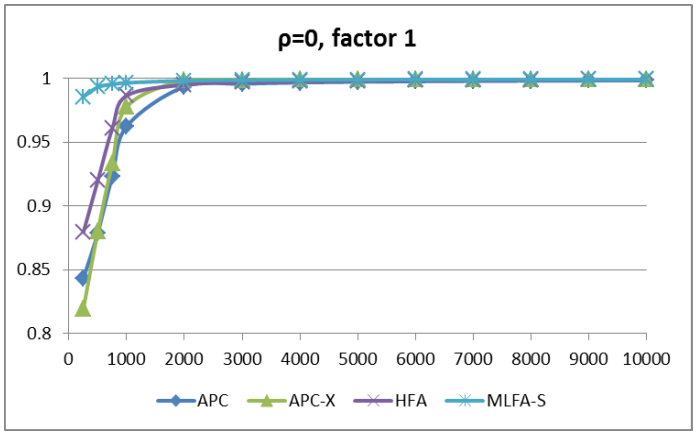


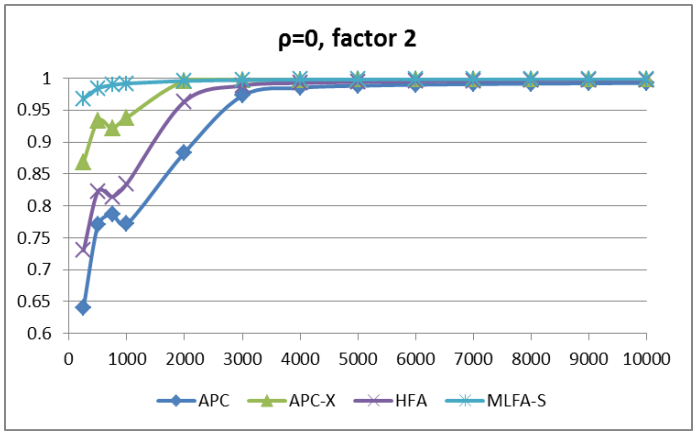


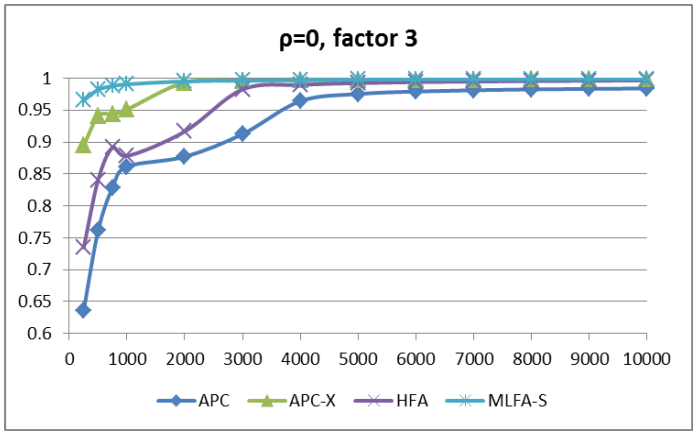


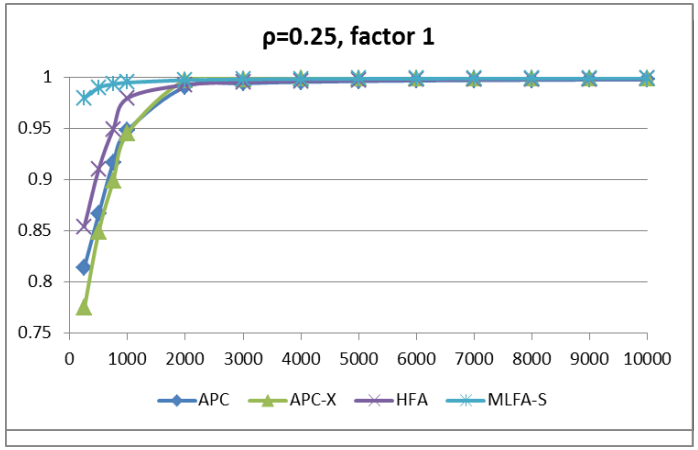


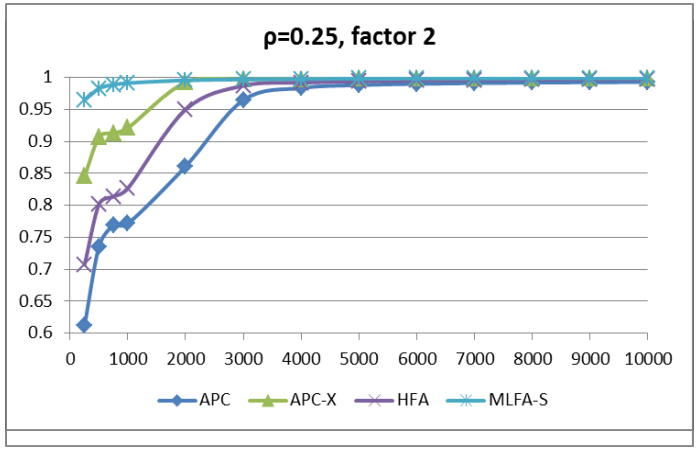
Cross-sectional heteroskedatic and time series heteroskedastic

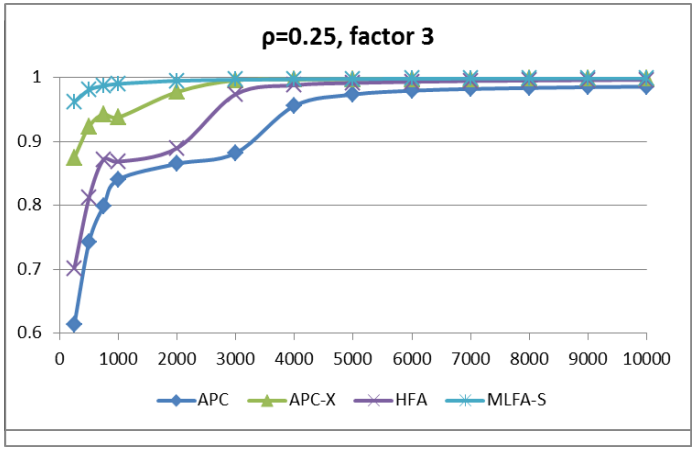


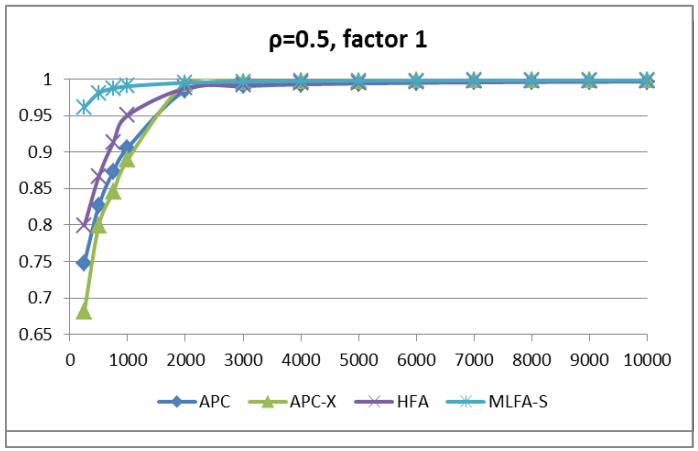


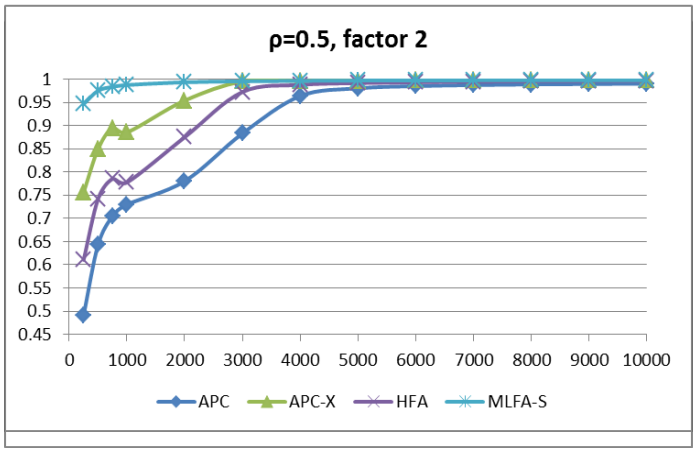


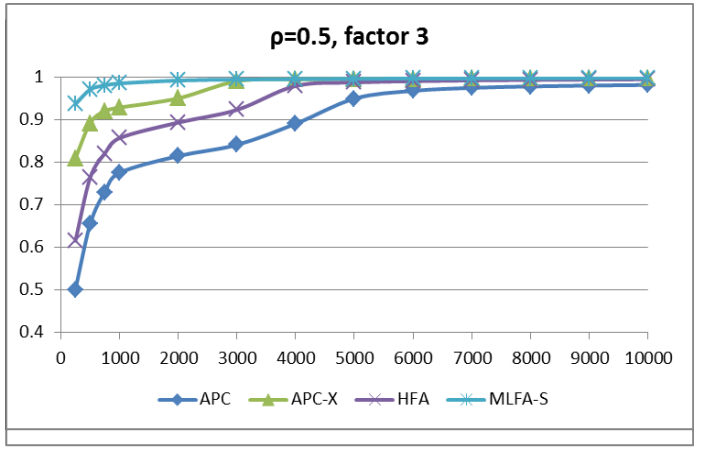




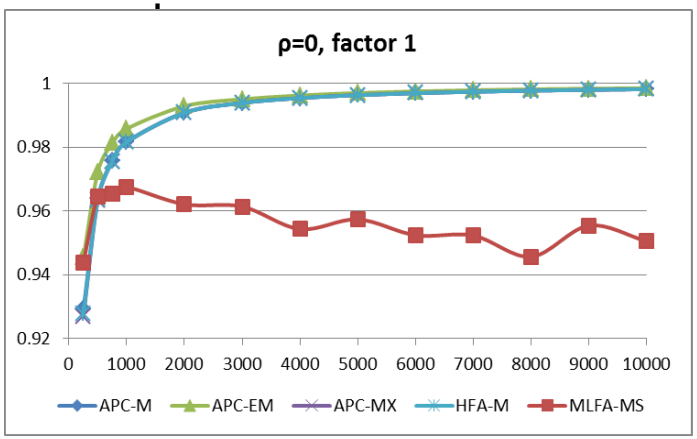


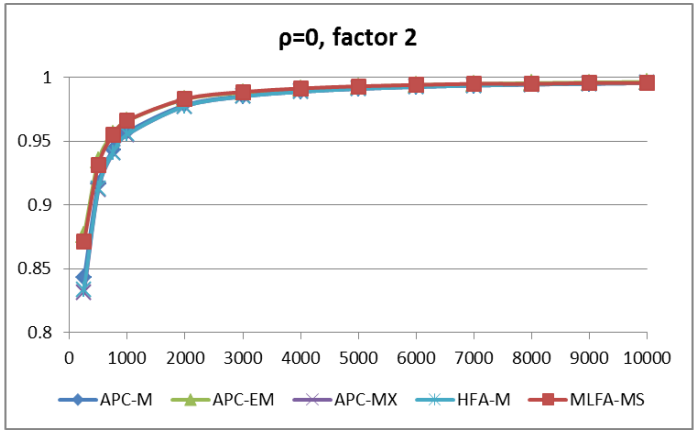


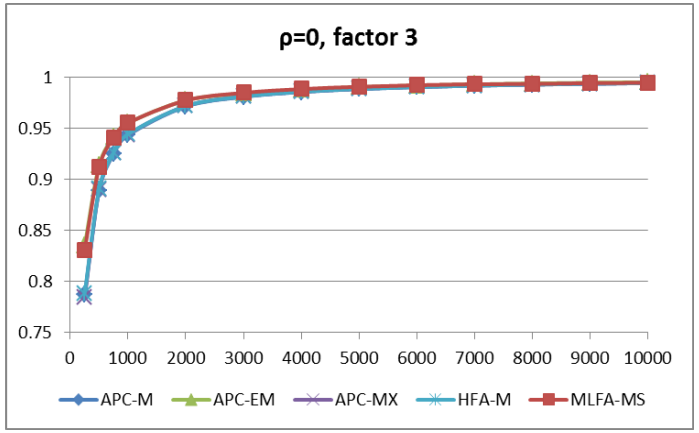


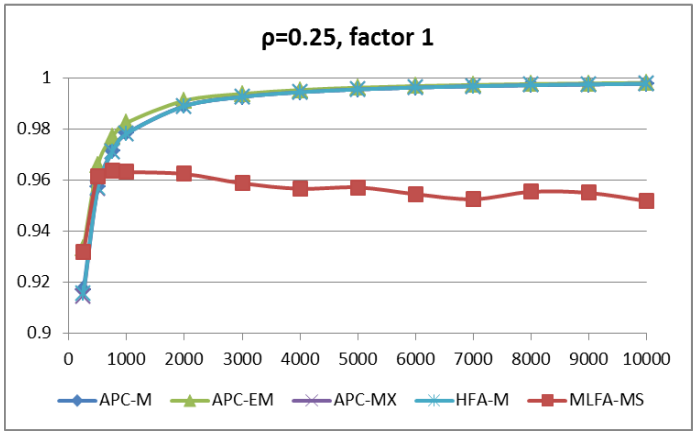


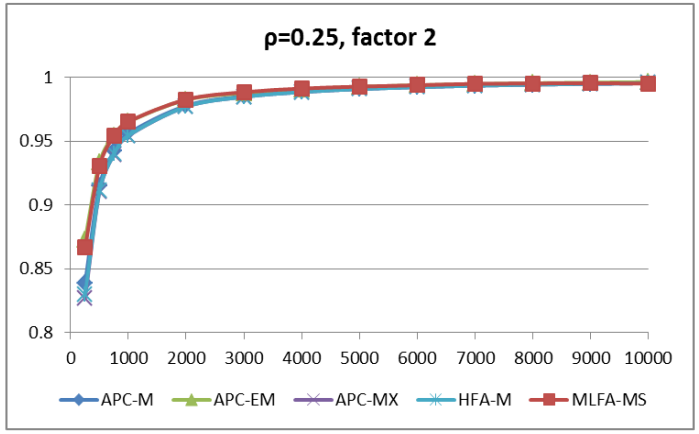
Cross-sectional homoskedastic and time series homoskedastic – unbalanced

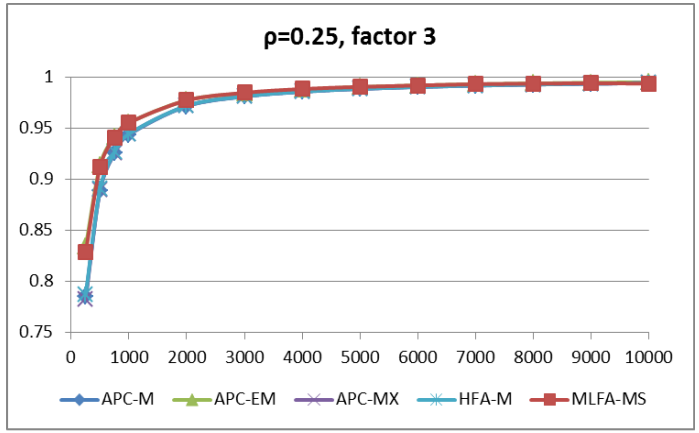


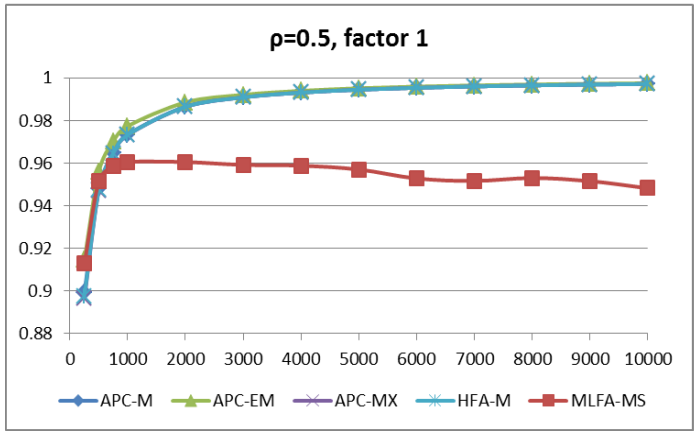


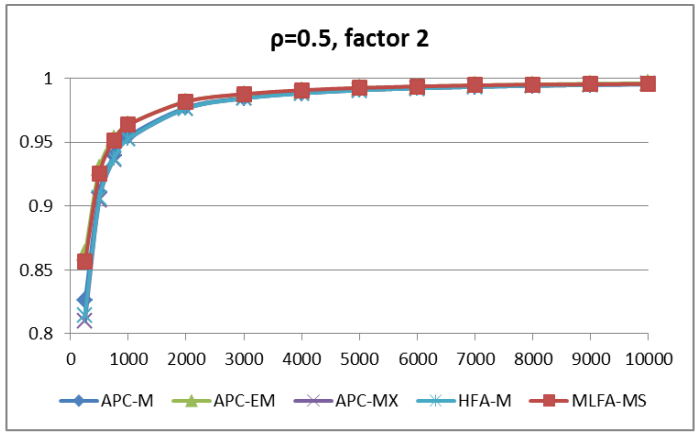


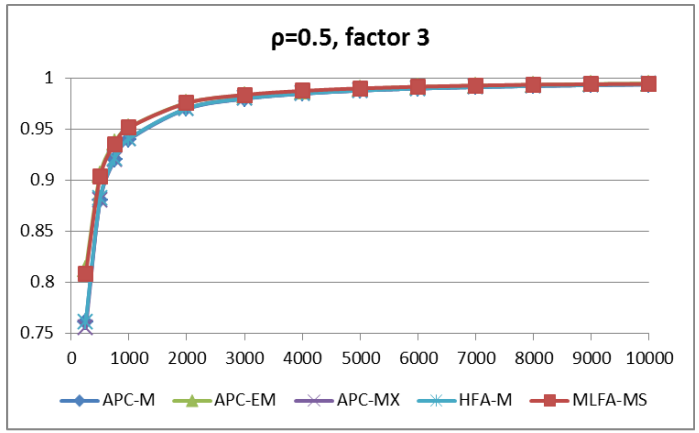




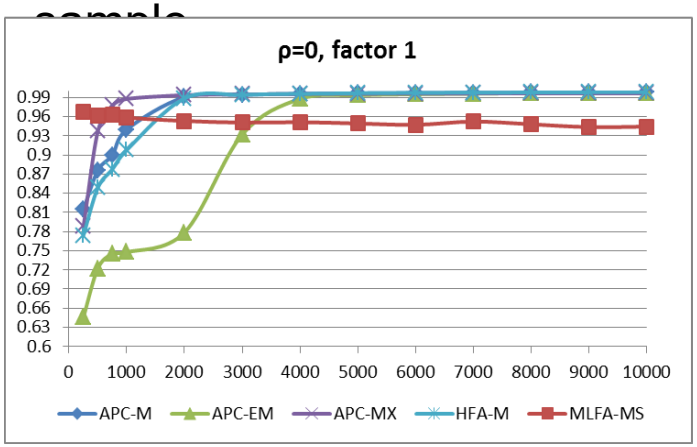


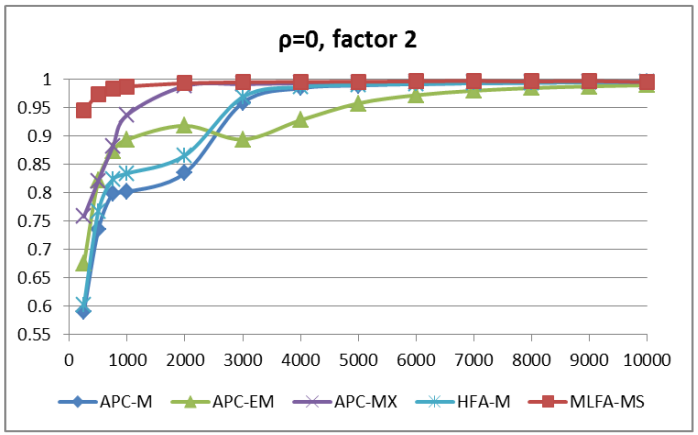


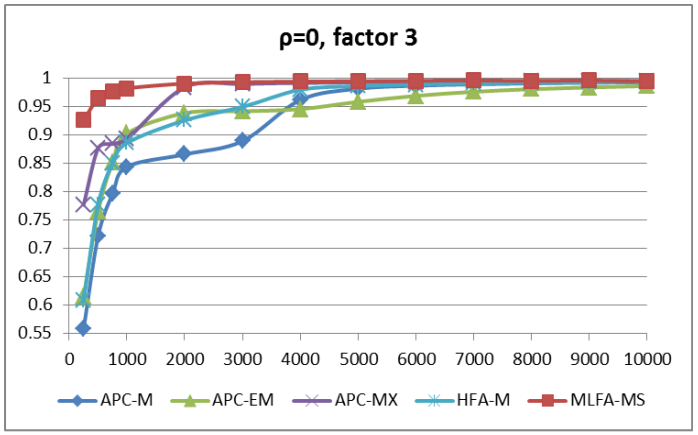


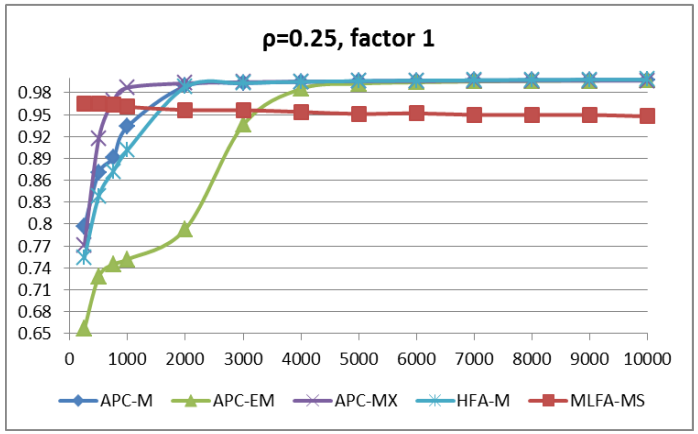


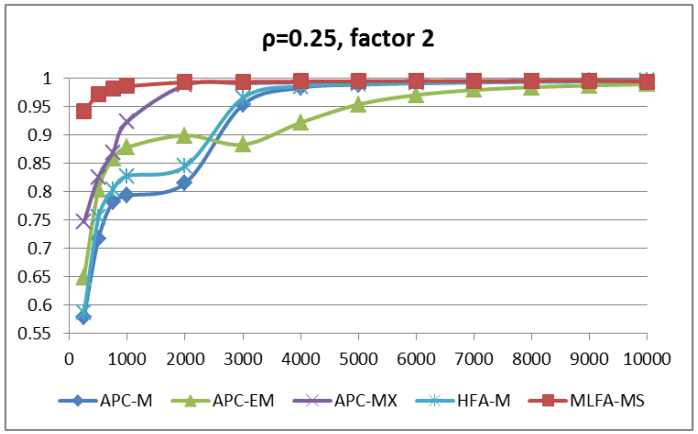
Cross-sectional heterskedastic and time series homoskedastic – unbalanced sample

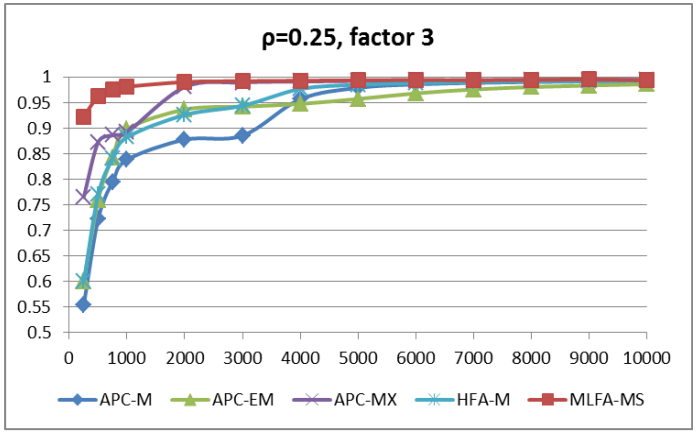


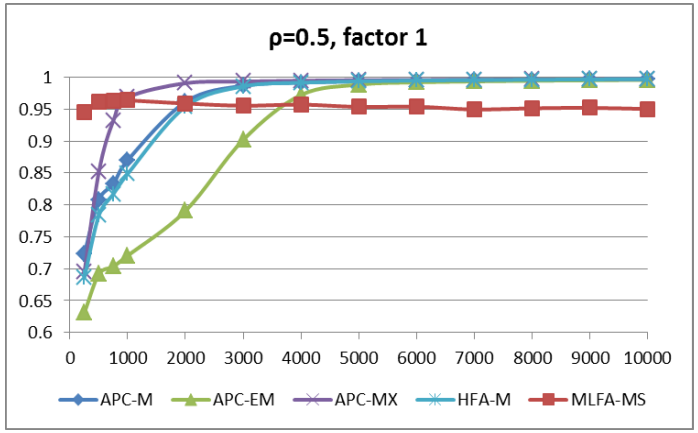


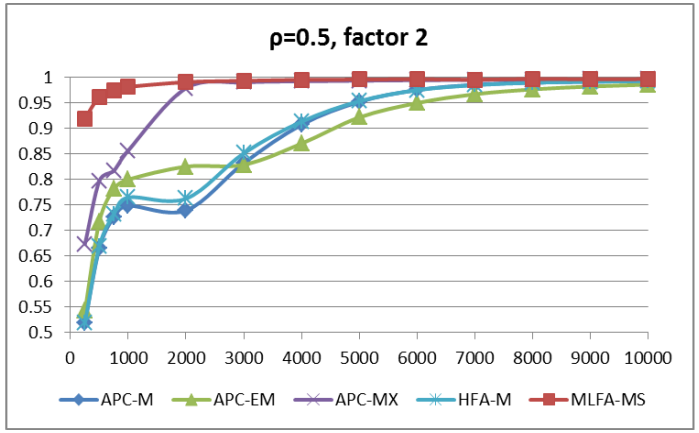


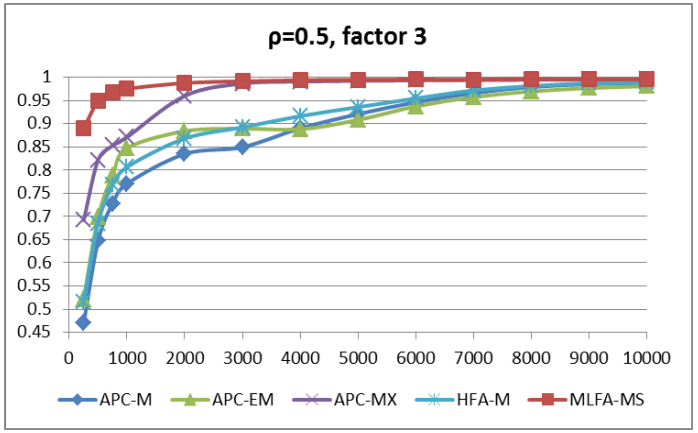




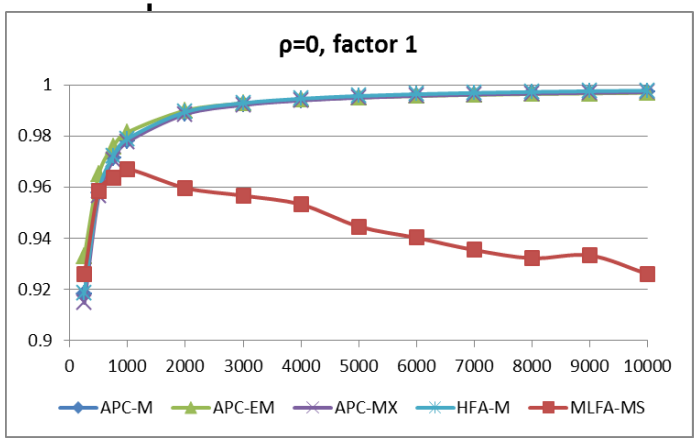


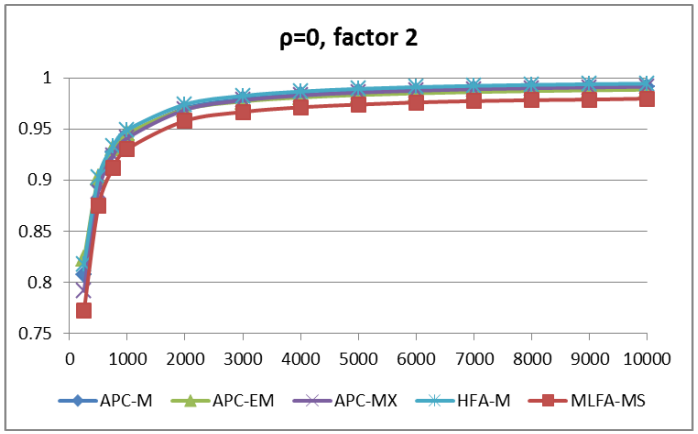


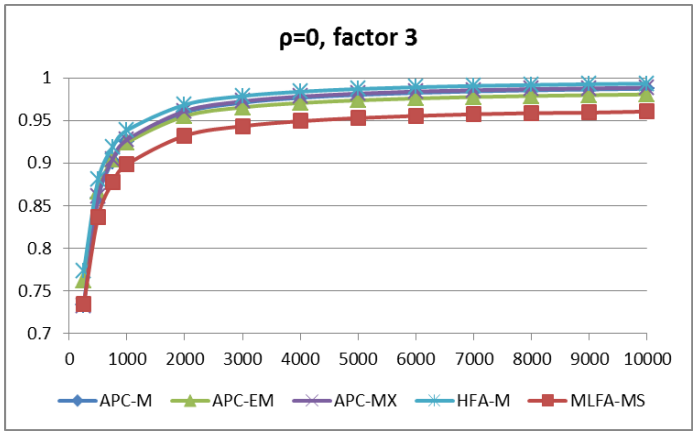


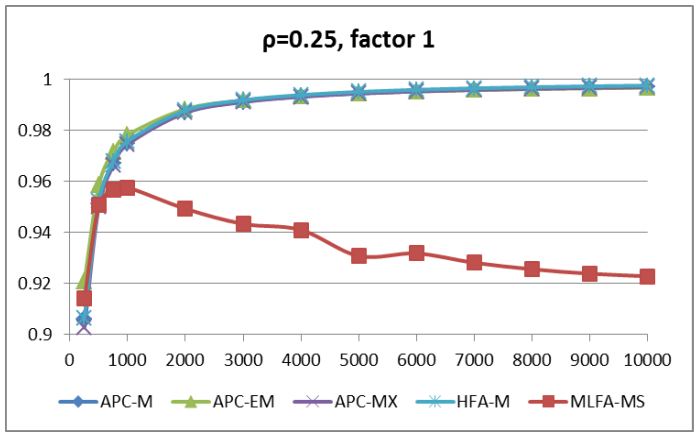


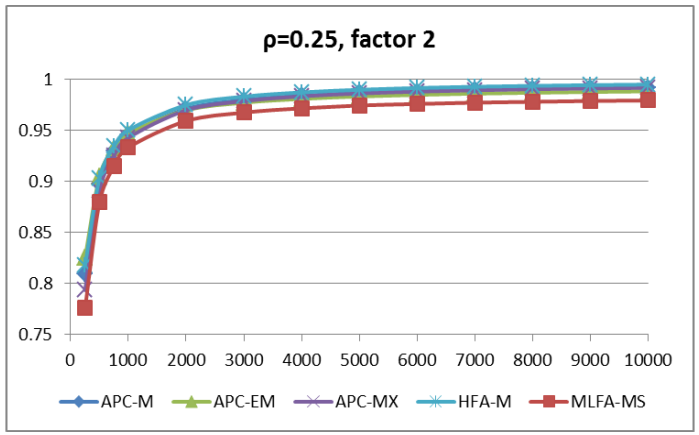
Cross-sectional homoskedastic and time series heteskedasticr – unbalanced

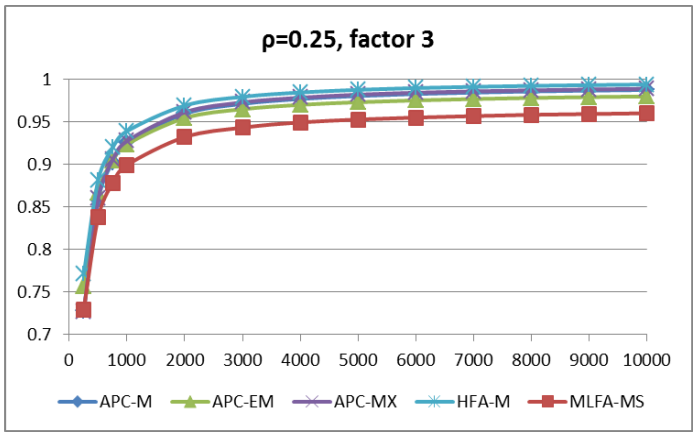


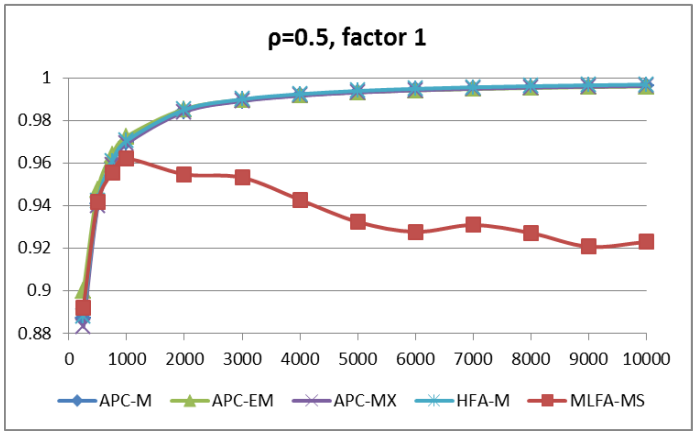


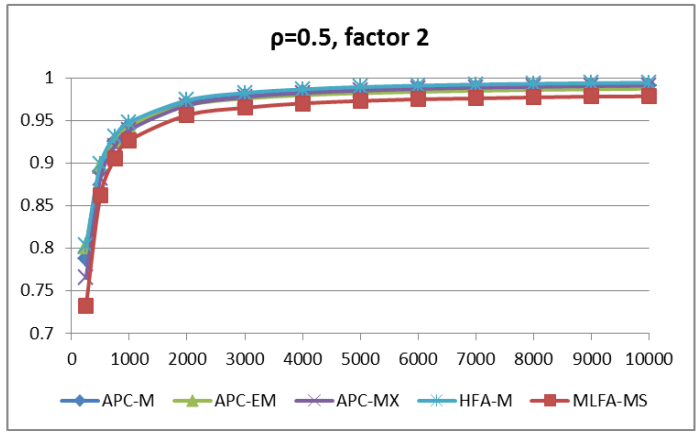


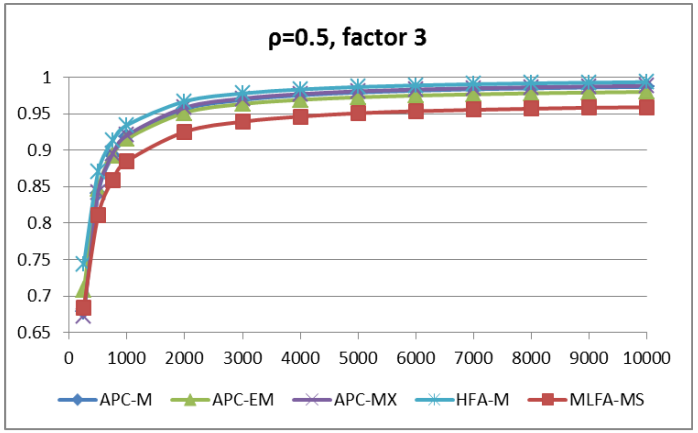




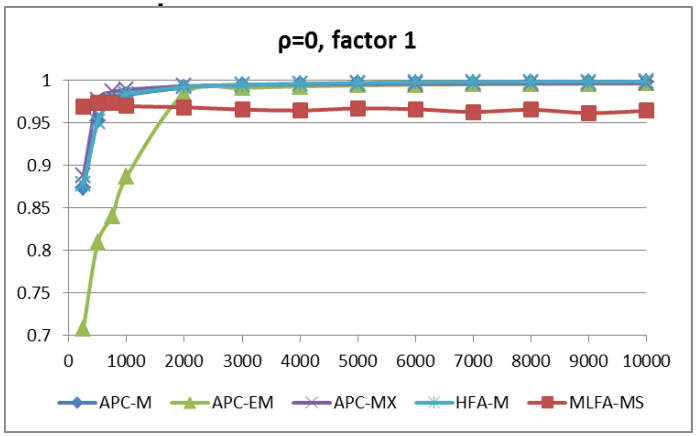


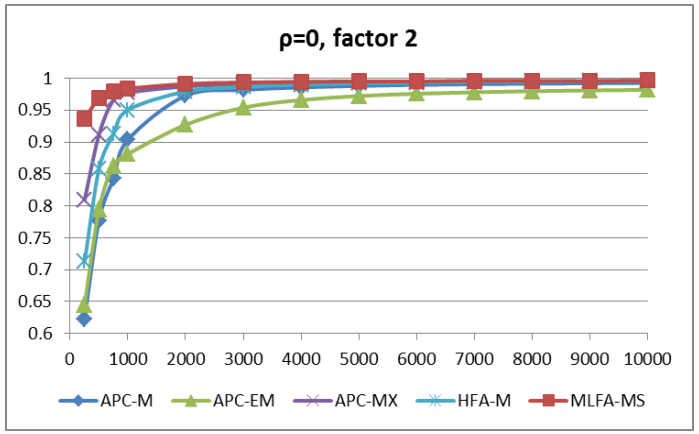


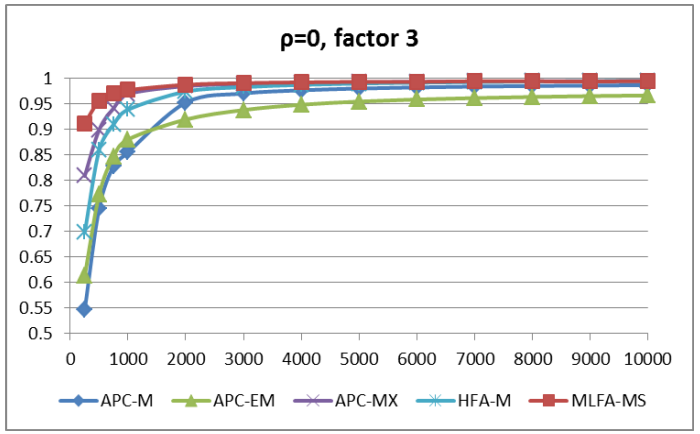


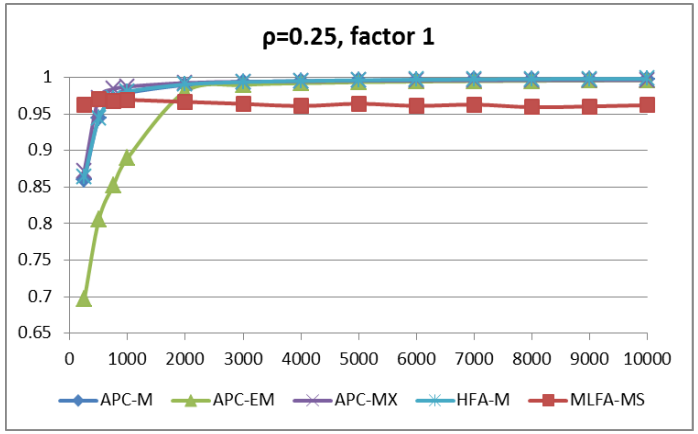


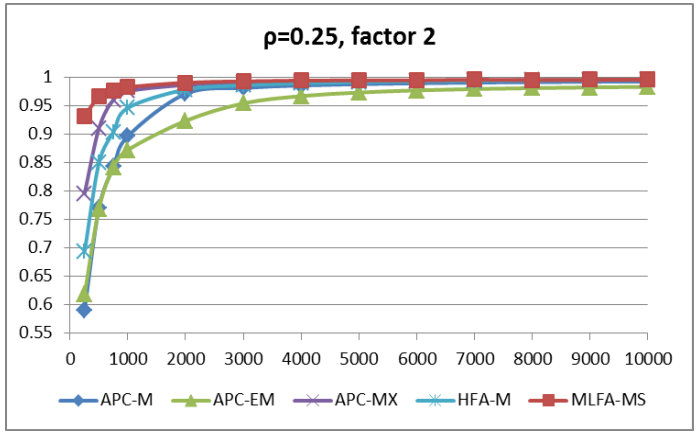
Cross-sectional heteroskedastic and time series heteroskedastic – unbalanced

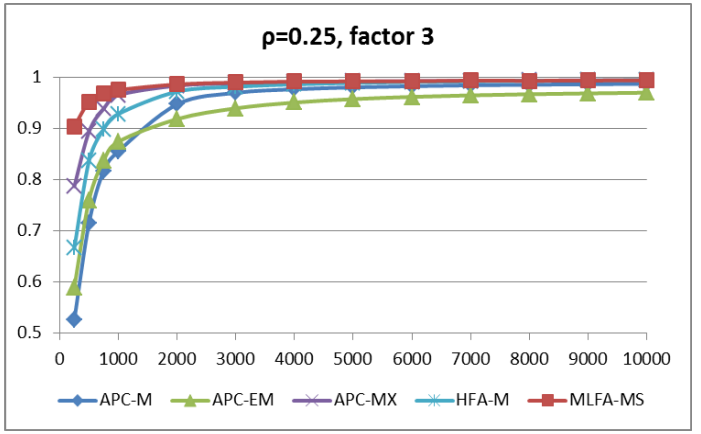


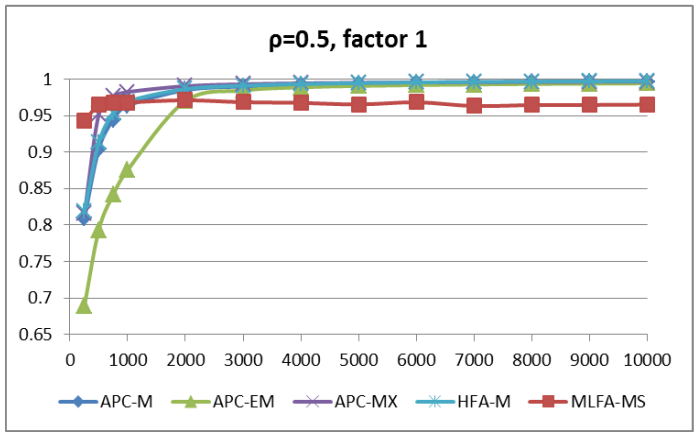


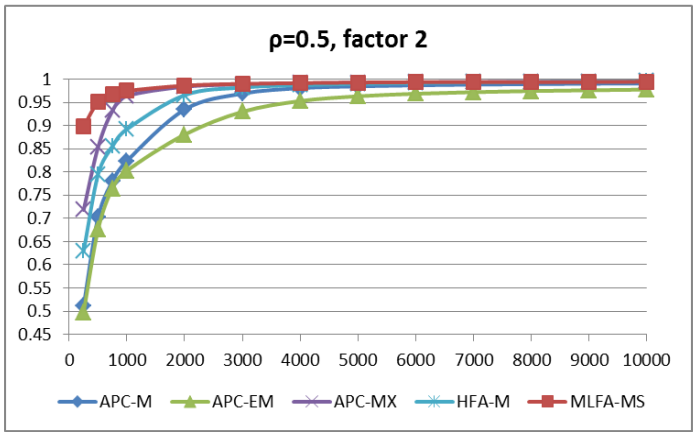


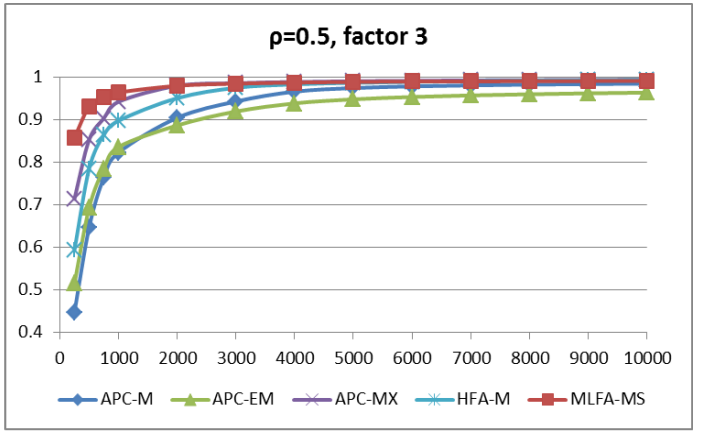












Summary and Conclusion

- In the simplest case of homoskedasticity and balanced panels, all estimators perform well.
  - Estimating extraneous parameters does not seem to degrade performance much.
- Estimators incorporating cross-sectional heteroskedasticity generally do better in that setting, particularly for smaller sample sizes.
- Estimators incorporating time series heteroskedasticity perform only slightly better in that setting.
- Cross-sectional correlation (approximate versus strict factor model) does not affect estimation performance much.
- In the unbalanced cases, APC-EM seems to require substantially larger samples than the other estimators.

Extensions

- Dynamics
  - Factors
  - Factor loadings
- Variation in *T*
- Performance of alternative tests for *k*
- *Deviations from Asset Pricing Model*

Details on Idiosyncratic Return Calibration

- Four alternative assumptions about idiosyncratic heteroskedasticity
  - Time series and cross-sectional homoskedasticity

$$\varepsilon_{i,t} \sim N(0, \overline{\sigma^2}), \overline{\sigma} = \frac{1}{N}\sum_{i=1}^{N} \widehat{\sigma}_i$$

    - With idiosyncratic cross-correlation

$$\varepsilon_{i,t} = \rho\varepsilon_{i-1,t} + u_{i,t}; u_{i,t} \sim N(0, (1-\rho^2)\overline{\sigma}^2)$$

  - Cross-sectional heteroskedasticity

$$\varepsilon_{i,t} \sim N(0, \widehat{\sigma}_i^2)$$

    - With idiosyncratic cross-correlation

$$\varepsilon_{i,t} = \rho\varepsilon_{i-1,t} + u_{i,t}; u_{i,t} \sim N(0, \widehat{\sigma}_i^2 - \rho^2\widehat{\sigma}_{i-1}^2)$$

Details on Idiosyncratic Return Calibration

- Time series heteroskedasticity

$$\varepsilon_{i,t} \sim N(0, \hat{\sigma}_t), \hat{\sigma}_t^2 = \frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}^2_{i,t}$$

  - With idiosyncratic cross-correlation

$$\varepsilon_{i,t} = \rho\varepsilon_{i-1,t} + u_{i,t}; u_{i,t} \sim N(0, (1-\rho^2)\bar{\sigma}_t^2)$$

- Cross-sectional and time-series heteroskedasticity

$$\ln(\hat{\varepsilon}^2_{i,t}) = a_i + b_i \ln(\hat{\sigma}_t^2) + e^2_{i,t}$$

$$\varepsilon_{i,t} \sim N(0, \exp(\hat{a}_i + \hat{b}_i \ln(\hat{\sigma}_t^2)))$$

  - With idiosyncratic cross-correlation

$$\varepsilon_{i,t} = \rho\varepsilon_{i,t-1} + u_{i,t};$$

$$u_{i,t} \sim N(0, \exp(\hat{a}_i + \hat{b}_i \ln(\hat{\sigma}_t^2)) - \rho^2 \exp(\hat{a}_{i-1} + \hat{b}_{i-1} \ln(\hat{\sigma}_t^2)))$$

Time Series of Average Idiosyncratic Variance