A Replication Based on Principal Component Analysis of High Frequency Data

Wanjun Li

May 2017

1 Introduction

The purpose of this replication is to testify the method of Principal Component Analysis in the context of high frequency data. There are three main motivations for Aït-Sahalia and Xiu to do such a study. Theoretically, higher sampling frequency generates better variance and covariance estimation. It makes high frequency asymptotic analysis plausible with fixed cross-sectional dimension, and large time series dimension. Secondly, there is no need to assume the existence of population moments. It means that general PCA is applicable for Itô semimartingales. Thirdly, using local (or instantaneous) windows, this PCA captures general nonlinear relationships.

2 Methodology

2.1 Realized Spectral Functions

$$\hat{c}_{ik_n\Delta_n} = \frac{1}{k_n\Delta_n}\sum_{j=1}^{k_n}\left(\Delta^n_{ik_n+j}X\right)\left(\Delta^n_{ik_n+j}X\right)^T\mathbb{1}\{\|\Delta^n_{ik_n+j}X\| \leq u_n\}$$

where $u_n = \alpha\Delta_n^{\bar{\omega}}$, and $\Delta_l^n X = X_{l\Delta_n} - X_{(l-1)\Delta_n}$. In consistent with many literature, I choose $u_{i,n} = 3(\int_0^t c_{ii,s}ds/t)^{0.5}\Delta_n^{0.47}$, for $1 \leq i \leq d$, and $\int_0^t c_{ii,s}ds$ is estimated by bipower variations.

I then estimate the eigenvalues of $\hat{c}_{ik_n\Delta_n}$. The estimator of the integrated spectral function is given by

$$V(\Delta_n, X; F) = k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} f(\hat{\lambda}_{ik_n\Delta_n}).$$

2.2 Realized Eigenvalues

To deal with the problem of non-smoothness of repeated eigenvalues, the authors use the averages of repeated eigenvalues. A spectral function is constructed as follows:

$$F^{\lambda}(\cdot) = (\frac{1}{g_1}\sum_{j=1}^{g_1}\lambda_j(\cdot), \frac{1}{g_2 - g_1}\sum_{j=g_1+1}^{g_2}\lambda_j(\cdot), ..., \frac{1}{g_r - g_{r-1}}\sum_{j=g_{r-1}+1}^{g_r}\lambda_j(\cdot))^T$$

where

$$\lambda_{g_{l-1}}(\cdot) > \lambda_{g_{l-1}+1}(\cdot) = ... = \lambda_{g_l}(\cdot) > \lambda_{g_{l+1}}(\cdot)$$

Corollary 1. Suppose $k_n \asymp \Delta_n^{-\varsigma}$ and $u_n \asymp \Delta_n^{\bar{\omega}}$ for some $\varsigma \in \left(\frac{\gamma}{2} \vee \frac{1}{3}, \frac{1}{2}\right)$ and $\bar{\omega} \in \left[\frac{1-\varsigma}{2-\gamma}, \frac{1}{2}\right)$.

(i) Under Assumption 1, the estimator of integrated eigenvalue vector given by

$$V(\Delta_n, X; \lambda) = k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \lambda(\hat{c}_{ik_n\Delta_n}).$$

is consistent

(iii) Under Assumption 1 and 4, our estimator with respect to the gth simple eigenvlue $\lambda_{g.}$

is given by

$$\tilde{V}(\Delta_n, X; \lambda_g) = k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \left\{ \hat{\lambda}_{g,ik_n\Delta_n} - \frac{1}{k_n} Tr\Big((\hat{\lambda}_{g,ik_n\Delta_n} - \hat{c}_{ik_n\Delta_n})^+ \hat{c}_{ik_n\Delta_n} \Big) \hat{\lambda}_{g,ik_n\Delta_n} \right\},$$

which satisfies:

$$\frac{1}{\sqrt{\Delta_n}} \Big(\tilde{V}(\Delta_n, X; \lambda_g) - \int_0^t F^{\lambda}(c_s) ds \Big) \xrightarrow{\mathcal{L}-s} \mathcal{W}_t^{\lambda_g},$$

where $\mathcal{W}_t^{\lambda_g}$ is a continuous process defined on an extension of the original probability space, which conditionally on $\mathcal{F}$, is a continuous centered Gaussian martingale with its variance $\int_0^t \lambda_{g,s}^2 ds$.

2.3 Realized Eigenvectors

Corollary 2. *Suppose Assumptions 1 and 4 hold. In addition, $\gamma_{g,s}$ is a a vector-valued funcin that corresponds to the eigenvector of c_s with respect to a simple root $\lambda_{g,s}$ for each $s \in [0, t]$. Then we have as $\Delta_n \to 0$*

$$\frac{1}{\sqrt{\Delta_n}} \left(k_n\Delta_n \sum_{i=0}^{[t/(k_n\Delta_n)]} \left(\hat{\gamma}_{g,ik_n\Delta_n} + \frac{1}{2k_n} \sum_{p \neq g} \frac{\hat{\lambda}_{g,ik_n\Delta_n} \hat{\lambda}_{p,ik_n\Delta_n}}{(\hat{\lambda}_{g,ik_n\Delta_n} - \hat{\lambda}_{p,ik_n\Delta_n})^2} \hat{\gamma}_{g,ik_n\Delta_n} \right) - \int_0^t \gamma_{g,s} ds \right)$$

where $\mathcal{W}_t^{\lambda_g}$ is a continuous process defined on an extension of the original probability space, which conditionally on $\mathcal{F}$, is a continuous centered Gaussian martingale

3 Simulation Evidence

As stated before, high frequency asymptotic framework enables nonparametric analysis. Specifically the setups for the simulations are as follows

$$dX_{i,t} = \sum_{j=1}^{r} \beta_{ij,t} dF_{j,t} + dZ_{i,t}, \quad dF_{j,t} = \mu_j dt + \sigma_{j,t} dW_{j,t} + dJ^F_{j,t}, \quad dZ_{i,t} = \gamma_t dBi, t + dJ^Z i, t,$$

where i = 1, 2, ..., d, and j = 1, 2, ..., r. The correlation matrix of dW is denoted as ρ^F. $\sigma_{j,t}, \gamma_t dB$ and $\beta_{ij,t}$ are allowed to be time-varying and evolve as:

$$d\sigma^2_{j,t} = \kappa_j(\theta_j - \sigma^2_{j,t})dt + \eta_j \sigma_{j,t} d\widetilde{W}_{j,t} + dJ^{\sigma^2}_{j,t}, \quad d\gamma^2_t = \kappa(\theta - \gamma^2_t)dt + \eta\gamma_t d\bar{B}_t$$

$$d\beta_{ij,t} = \begin{cases} \tilde{\kappa}_j \left(\tilde{\theta}_{ij} - \beta_{ij,t}\right) dt + \tilde{\xi}_j \sqrt{\beta_{ij,t}} d\widetilde{B}_{ij,t} & if\ the\ jth\ factor\ is\ the\ market \\ \tilde{\kappa}_j \left(\tilde{\theta}_{ij} - \beta_{ij,t}\right) dt + \tilde{\xi}_j d\widetilde{B}_{ij,t} & otherwise \end{cases}$$

the correlation between dW_j, and $d\widetilde{W}_j$, is $\rho_j, \{J^F_j\}_{1 \le j \le r}$ and $\{J^Z_i\}_{1 \le i \le d}$ follows two compound Poisson Processes with arrival rates λ^F, λ^Z, respectively. And the jump sizes follow double exponential distributions with means $\mu^F_+, \mu_- F, \mu^Z_+, \mu^Z_-$, respectively. $\{J^{\sigma^2}_j\}_{1 \le j \le d}$ co-jumps with J^F, and their jump sizes follow exponential distibutions with the mean equal to μ^{σ^2}. The model parameters are given in Table 1.

I report the mean, bias and root-mean-square errors of the standardized estimates with d=5, 10, 20, 30, 50 stocks, using the returns sampled every $/delta_n = 5$ seconds, 1 minute, and 5 minutes over one week horizons. The eigenvalues are recovered with very small bias and standard errors.

Then to verify the accuracy of the asymptotic distribution, Iincluded the histograms of the standard estimates of integrated eigenvalues using 10 stocks with 5-second returns.

4 Conclusions

PCA at high frequency enables the application to semimartingales. The simulation results perform well. The estimates have very small bias and standard errors, for different dimensions and frequencies. The curse of dimensionality doesn't really hurt these estimates. On the other hand, the replication provides evidence for the asymptotic theories.

One thing I noticed that the true eigenvalues changes with frequencies. I infer there are three reasons. First is that I change the number of simulated observations every time in accordance with the frequency. The author use the average of higher frequency observations to get lower frequency observations. The constant matrix $\tilde{\theta}_{i,j}$ is generated randomly from the described distribution and is fixed at each frequency and stock dimension. But the author fixed them through out all replications. And thirdly, there are many mean reversion processes. To avoid start up effects, I should have deleted first 300 observations for example. I did notice that different starting values for some parameters like σ can change the true eigenvalues.

The replication provide evidence for this paper, with very small estimation bias, and Gaussian distributions. But to get similar numbers as the paper, more clarifications are needed from the author on the simulation process, initial values, and which factor used as the "market".

Appendix Figures and Tables

	κ_j	θ_j	η_j	ρ_j	μ_j	$\tilde{\kappa}_j$	$\tilde{\theta}_{i,j}$	$\tilde{\xi}_j$
j = 1	3	0.05	0.3	-0.6	0.05	1	U[0.25,1.75]	0.5
j = 2	4	0.04	0.4	-0.4	0.03	2	$N(0,0.5^2)$	0.6
j = 3	5	0.03	0.3	-0.25	0.02	3	$N(0,0.5^2)$	0.7
	λ^F	$\mu^F_{+/-}$	λ^Z	$\mu^Z_{+/-}$	μ^{σ^2}	ρ^F_{12}	ρ^F_{13}	ρ^F_{23}
	1/t	$4\sqrt{\Delta}$	2/t	$6\sqrt{\Delta}$	$\sqrt{\Delta}$	0.05	0.1	0.15
						κ	θ	η
						4	0.3	0.06

Table 1: Parameters in Monte Carlo Simulations

The constant matrix $\tilde{\theta}_{i,j}$ is generated randomly from the described distribution and is fixed at each frequency and stock dimension. The author fixed it throughout ALL replications.

1 Week, 5 Seconds				1 Week, 1 Minute			
# Stocks	TRUE	Bias	Stdev	# Stocks	TRUE	Bias	Stdev
5	0.425774	-0.000028	0.000018	5	0.232890	-0.000836	0.000025
10	0.807342	0.000003	0.000035	10	0.357888	-0.001247	0.000029
20	1.772680	0.000137	0.000081	20	0.579031	-0.001912	0.000085
30	2.736691	0.000484	0.000157	30	0.749635	-0.002379	0.000118
50	5.069696	0.002041	0.000301	50	1.248376	-0.003556	0.000228
1 Week, 5 Minutes				1 Month, 5 Minutes			
# Stocks	TRUE	Bias	Stdev	# Stocks	TRUE	Bias	Stdev
5	0.084517	0.001128	0.000307	5	0.181979	0.000851	0.002342
10	0.235776	0.002338	0.000678	10	0.278917	0.000803	0.002984
20	0.495761	0.007656	0.002110	20	0.578620	0.001273	0.007132
30	0.379170	0.007750	0.002183	30	0.859230	0.002130	0.010380
50	0.549394	0.011595	0.002954	50	1.503761	0.016736	0.015205

Table 2: Simulation Results: First Eigenvalue Estimation

Note: In this table, I report the summary statistics of 1,000 Monte Carlo simulations for estimating the first integrated eigenvalue. Column “TRUE” corresponds to the average of the true integrated eigenvalue. Column “Bias” corresponds to the mean of the estimation error; Column “Stdev” is the standard deviation of the estimation error.

1 Week, 5 Seconds				1 Week, 1 Minute			
# Stocks	TRUE	Bias	Stdev	# Stocks	TRUE	Bias	Stdev
5	0.073638	0.000006	0.000004	5	0.232890	-0.000128	0.000025
10	0.230410	0.000020	0.000011	10	0.357888	-0.000188	0.000029
20	0.348740	0.000080	0.000021	20	0.579031	-0.000440	0.000085
30	0.579345	0.000149	0.000028	30	0.749635	-0.000623	0.000118
50	1.255471	0.000628	0.000088	50	1.248376	-0.000943	0.000228
1 Week, 5 Minutes				1 Month, 5 Minutes			
# Stocks	TRUE	Bias	Stdev	# Stocks	TRUE	Bias	Stdev
5	0.014004	0.000042	0.000038	5	0.077270	0.001040	0.000545
10	0.046742	0.000288	0.000127	10	0.066685	0.001333	0.000606
20	0.115671	0.000909	0.000340	20	0.119953	0.003413	0.001045
30	0.073866	0.000993	0.000296	30	0.168516	0.005052	0.001344
50	0.127695	0.001647	0.000511	50	0.284999	0.009062	0.000511

Table 3: Simulation Results: Second Eigenvalue Estimation

Note: In this table, I report the summary statistics of 1,000 Monte Carlo simulations for estimating the Second integrated eigenvalue. Column “TRUE” corresponds to the average of the true integrated eigenvalue. Column “Bias” corresponds to the mean of the estimation error; Column “Stdev” is the standard deviation of the estimation error.

1 Week, 5 Seconds				1 Week, 1 Minute			
# Stocks	TRUE	Bias	Stdev	# Stocks	TRUE	Bias	Stdev
5	0.014638	0.000003	0.000002	5	0.006760	0.000023	0.000008
10	0.131355	0.000002	0.000006	10	0.028550	-0.000100	0.000021
20	0.286077	0.000023	0.000012	20	0.090280	-0.000299	0.000055
30	0.283042	0.000067	0.000016	30	0.131761	-0.000418	0.000084
50	0.966831	0.000354	0.000059	50	0.263394	-0.000760	0.000157
1 Week, 5 Minutes				1 Month, 5 Minutes			
# Stocks	TRUE	Bias	Stdev	# Stocks	TRUE	Bias	Stdev
5	0.000634	0.000004	0.000005	5	0.047708	-0.000968	0.000289
10	0.025950	-0.000111	0.000036	10	0.028412	-0.000098	0.000210
20	0.028819	-0.000534	0.000097	20	0.073304	-0.00046	0.000408
30	0.053172	-0.000480	0.000106	30	0.092287	-0.000039	0.000694
50	0.094516	-0.000874	0.000189	50	0.179446	-0.000708	0.000189

Table 4: Simulation Results: Third Eigenvalue Estimation

Note: In this table, I report the summary statistics of 1,000 Monte Carlo simulations for estimating the third integrated eigenvalue. Column “TRUE” corresponds to the average of the true integrated eigenvalue. Column “Bias” corresponds to the mean of the estimation error; Column “Stdev” is the standard deviation of the estimation error.

1 Week, 5 Seconds				1 Week, 1 Minute			
# Stocks	TRUE	Bias	Stdev	# Stocks	TRUE	Bias	Stdev
5	0.000378	0.000007	0.000002	5	0.000009	0.000002	0.000000
10	0.000593	0.000009	0.000002	10	0.000250	0.000001	0.000000
20	0.000407	0.000012	0.000002	20	0.000189	0.000003	0.000000
30	0.000564	0.000014	0.000003	30	0.000174	0.000004	0.000001
50	0.000674	0.000018	0.000003	50	0.000199	0.000006	0.000001
1 Week, 5 Minutes				1 Month, 5 Minutes			
# Stocks	TRUE	Bias	Stdev	# Stocks	TRUE	Bias	Stdev
5	0.000021	-0.000001	0.000001	5	0.001020	-0.000033	0.000009
10	0.000608	-0.000001	0.000000	10	0.000619	-0.000020	0.000005
20	0.000854	-0.000001	0.000000	20	0.000423	-0.000013	0.000002
30	0.000074	-0.000001	0.000000	30	0.000494	-0.000016	0.000003
50	0.000072	-0.000001	0.000000	50	0.000456	-0.000015	0.000000

Table 5: Simulation Results: Fourth and beyond Eigenvalue Estimation

Note: In this table, I report the summary statistics of 1,000 Monte Carlo simulations for estimating the repeated integrated eigenvalue. Column “TRUE” corresponds to the average of the true integrated eigenvalue. Column “Bias” corresponds to the mean of the estimation error; Column “Stdev” is the standard deviation of the estimation error.

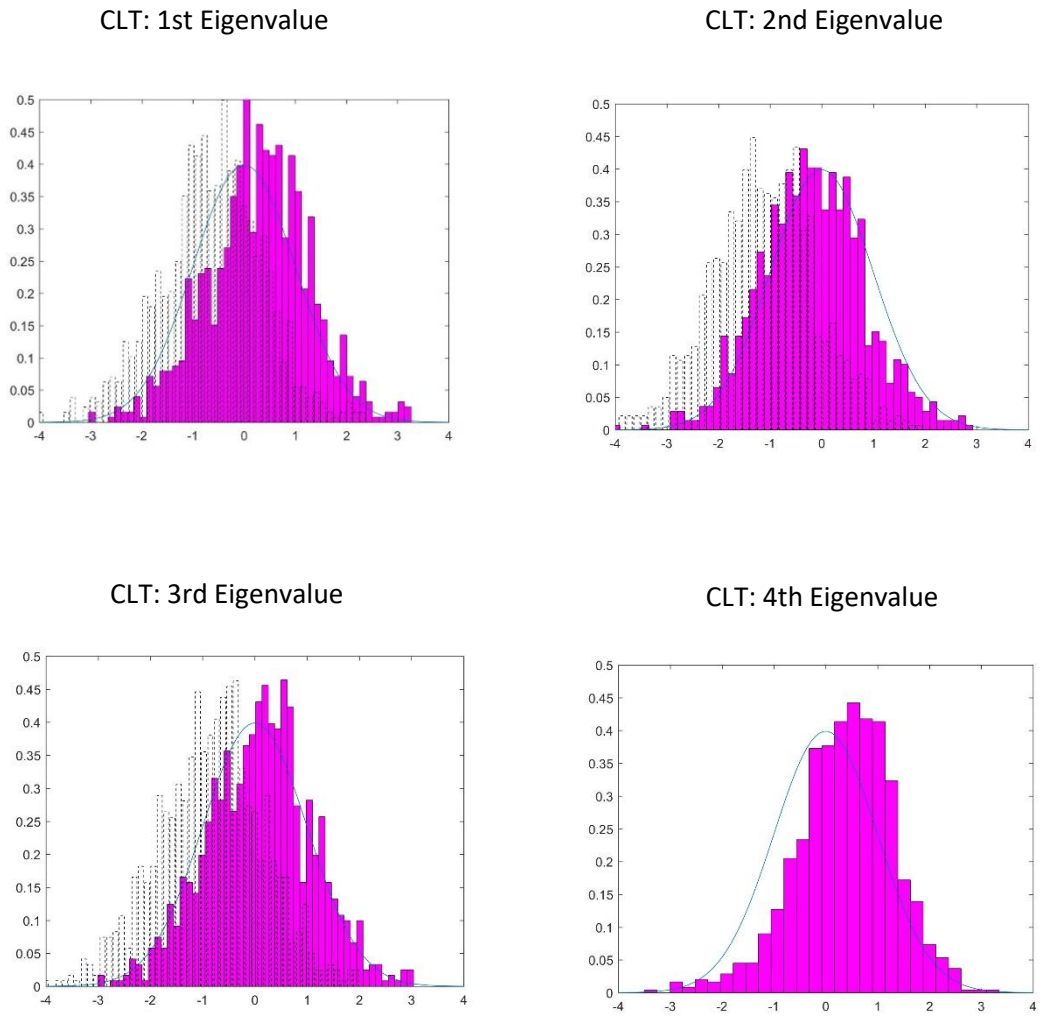

Figure 1: Finite Sample Distribution of the Standardized Statistics.

Note: In this figure, I report the histograms of the 1000 simulation results for estimating the first four integrated eigenvalues using 5-second returns for 5 stocks over one week. The Solid blue lines plot the standard normal density; the dashed histograms report the distribution of the estimates before bias corrections; the solid histograms report the distribution of the estimates after bias correction is applied and is to be compared to the asymptotic standard normal. Because

the fourth eigenvalue is small, the dashed histogram on the fourth subplot is out of the x-axis range, to the right.

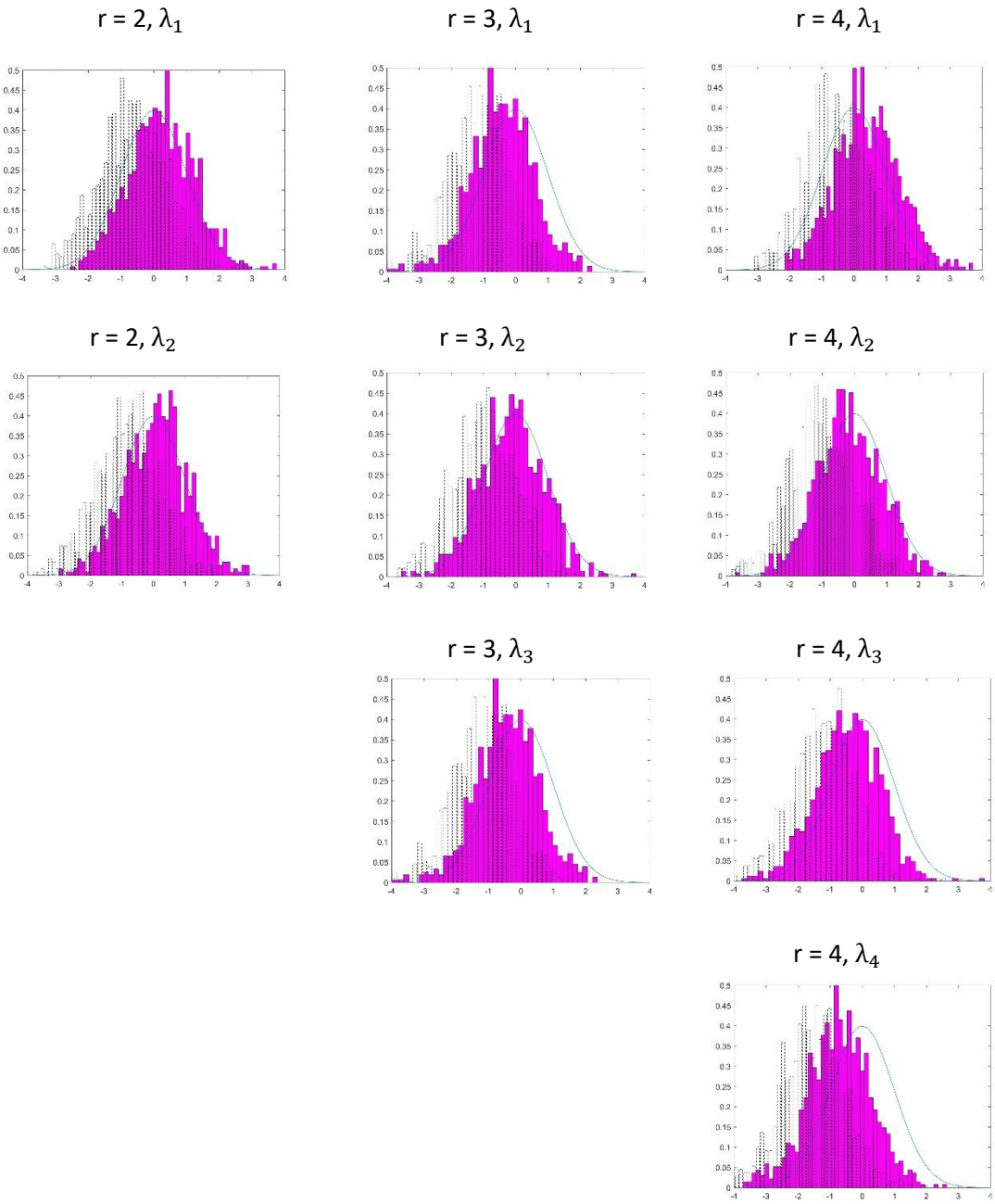

Figure 2: Finite Sample Distribution of the Standardized Statistics.

Note: In this figure, I report the histograms results with r=2, 3, and 4 common factors. Using weekly one-minute returns of 30 stocks.