Explaining the Single Factor Bias of Arbitrage Pricing Models in Finite Samples*

Matthew C. Harding,†

Final Version, *Economics Letters*, May 2007

Abstract

This paper shows that in finite samples it is not possible to distinguish all the latent factors from the idiosyncratic noise and that this leads to a bias towards the identification of a single factor. It provides an approximation to this bias and the corresponding sampling distribution.

JEL: C33, G11, C46
Keywords: factor analysis, arbitrage pricing theory, random matrix theory

1. Introduction

The Arbitrage Pricing Theory (APT) of Ross (1976) assumes the lack of arbitrage opportunities in capital markets and postulates a linear relationship between actual returns and a set of K common factors, with the implication that the expected returns will be linear functions of the common factor weights. This suggests the use of Factor Analysis (FA) developed by Spearman and Hotelling at the beginning of the last century as a potential tool for the extraction of the K common factors from a sample of returns.

Since factor analysis only identifies the factor loadings up to a non-trivial rotation, the task of extracting the K common factors and determining if they are priced by the market can be difficult. Determining the right rotation is a potentially very complicated task since the number of relevant options can be very large. In practice it is often preferred to fix the rotation on *a priori* grounds and this task is implicitly performed by the use of Principal Components Analysis (PCA) as a substitute for the more laborious Factor Analysis.

Asymptotic conditions for PCA to produce results which are close to FA are provided by Chamberlain and Rothschild (1983). They require for the first K eigenvalues of the covariance matrix of factor returns to grow without bound as the number of securities N in the portfolio increases,

*I am grateful to Xavier Gabaix, Jerry Hausman, Iain Johnstone and Raj Rao for comments.
†Department of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305; Phone: (650) 725-3266; Fax: (650) 725-5702; Email: `mch@stanford.edu`

while the remaining $N - K$ eigenvalues to remain bounded. This ensures that the returns are linearly related to the underlying K factors and even allows for weak forms of correlation between idiosyncratic shocks.

In practice one would consider the first K largest eigenvalues of the empirical covariance (or correlation) matrix of a panel of returns for a portfolio of N securities over T time periods, where both N and T are large. Since the number of underlying factors is unknown it is necessary to estimate a cut-off point which separates the K eigenvalues corresponding to the underlying factors from the remaining $N - K$ eigenvalues due to the idiosyncratic noise component.

Over the years numerous studies (Trzcinka, 1986; Connor and Korajczyk, 1993; Geweke and Zhou, 1996; Jones, 2001; Merville and Xu, 2001) have documented the dominance of one factor, labeled as the market factor, which explains most of the sample variation. More limited and inconclusive results have been obtained for identifying other factors such as industry specific factors. This has usually been attributed to the lack of formal criteria for choosing the number of factors from an empirical distribution of eigenvalues of the sample covariance matrix. In practice it is common to choose the number K by visual inspection of the scree plot or by the use of ad-hoc cut-off points of the distribution of eigenvalues.

More recently random matrix theory was employed to describe the distribution of the idiosyncratic noise component, which has a bounded support. Therefore, it is possible to choose the number of factors as the number of eigenvalues outside the finite support of the eigenvalues due to noise, as formalized by Onatski (2005). This approach is also found in the growing econophysics literature where numerous empirical investigations of different asset markets have been performed using this method (Plerou et. al., 2002; Bouchard and Potters, 2003). These studies also document the dominance of the market factor and remain inconclusive on the identification of further factors.

Brown (1989) provides Monte-Carlo evidence of an economy with K factors, each of which is priced and contributes equally to the returns. Moreover, the economy is by construction admissible under the framework of Chamberlain and Rothschild (1983) and calibrated to actual data from the NYSE. Nevertheless, he finds evidence that PCA is biased towards a single factor model. Thus, we cannot conclude that the empirical evidence presented in the numerous studies, some of which were mentioned above, is not also consistent with a setup where more than one factor is present in the economy, but where PCA fails to identify the entire set of relevant factors.

In this paper we use recent results from stochastic eigen-analysis to quantify the intuition of Brown (1989) and explain the single factor bias of arbitrage pricing models through the finite sample behavior of the factors estimated by PCA. We show that in finite samples of dimensions commonly found in these empirical investigations it is not possible to distinguish some of the factors from the idiosyncratic noise element either by heuristic methods or by a random matrix approach. This leads to the bias towards the identification of a single factor which is routinely reported in empirical investigations. Moreover, we find that the quantities estimated are severely biased, even when correctly identified, and provide an approximation to the finite sample bias and their sampling distribution. Overall, these results challenge the use of PCA, as it is commonly used in the empirical finance literature, as a suitable tool for the identification of the underlying factor structure of asset returns in situations where the number of assets is large relative to the number of time periods.

2. **Limiting Behavior of Sample Eigenvalues**

Let us now consider a version of the model introduced by Brown (1989). Assume that the relevant portfolio consists of N assets which are observed for T time periods. The asset returns are generated by an exact linear factor technology with K factors. Thus, the demeaned asset returns are given by:

$$R_N = \Lambda F' + \epsilon, \tag{2.1}$$

where R_N is a matrix of dimensions $N \times T$ of asset returns, F is an $T \times K$ matrix of factor scores, Λ is a $N \times K$ matrix of factor loadings and ϵ is a $N \times T$ matrix of idiosyncratic noise components. The covariance matrix of returns is given by the $N \times N$ matrix $\Sigma_N = (1/T)R_N R'_N$. We also assume that $E(\epsilon) = 0$,$E(\epsilon\epsilon') = \sigma_\epsilon^2 I_N$ and $E(|\epsilon_{\text{ii}}|^4) < \infty$. Notice that we do not assume normality for most of the results in this paper except for Proposition 2.

Furthermore, we assume that $N \to \infty$, $T \to \infty$, and $N/T \to c \in (0, \infty)$, where c is a constant. The asymptotic framework is similar to that used in other panel data studies (e.g. Hahn and Kuersteiner, 2002) and will facilitate the derivation of finite sample results. It corresponds to a setup where the number of cross-sectional units is large relative to the number of available time periods. We will first characterize the distribution of population eigenvalues and then employ recent results from random matrix theory to characterize their limiting distribution under the large N, large T asymptotic framework. This will allow us to explore the conditions under which the factors are identified and also derive the distribution of the correctly identified factors.

It is important to note the large N, large T asymptotic framework is the appropriate framework to model the samples encountered in practical finance. Traditionally PCA results were derived under the assumption that N is fixed while T goes to infinity. In practice however portfolio managers tend to have only a limited number of time periods available while looking to measure risk factors from a large number of securities.

Furthermore, assume that $E[(1/T)F'F] = \sigma_F^2 I_K$ and $E[(1/T)\epsilon F] = 0$. The then population covariance of factor returns is given by

$$\Sigma_N = \sigma_F^2 \Lambda\Lambda' + \sigma_\epsilon^2 I_{N.} \tag{2.2}$$

We now impose additional assumptions of the $N \times K$ matrix of factor loadings, Λ. Let $\Lambda = [\Lambda_1 : \Lambda_2 : ... : \Lambda_K]$, where each of the columns Λ_i of Λ is given by $\Lambda_i = \beta\iota + \sqrt{\sigma_\beta^2} e_i$. Let β be a constant, ι be the $N \times 1$ vector $(1, 1, 1, ..., 1)'$ and e_i be an i.i.d. vector random variable with mean 0 and variance 1 and finite fourth moments.

First let us consider the eigenvalue behavior of this model under the assumption that $T \to \infty$ much faster than N. The resulting eigenvalues are labeled as population eigenvalues, since they correspond to a setup where the number of time-periods for which the model is observed is very large.

Let us now compute the population eigenvalues of Σ_N from the decomposition $U_N \Sigma_N U_N^{-1} = \text{diag}\{\lambda_1, \lambda_2, ..., \lambda_N\}$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N$. Since $\Lambda\Lambda'$ is rank deficient it will have K non-zero eigenvalues and $N - K$ zero eigenvalues. But since the non-zero eigenvalues of the $N \times N$ matrix $\Lambda\Lambda'$ are the same as the eigenvalues of the $K \times K$ matrix $\Lambda'\Lambda$, it is sufficient to consider only the latter one. Furthermore, notice that in the large T limit, $\Lambda_i'\Lambda_j = N(\beta^2 + \sigma_b^2)$ if $i = j$ and $\Lambda_i'\Lambda_j = N\beta^2$ is $i \neq j$. Hence, $\Lambda'\Lambda = N\sigma_b^2 I_K + N\beta^2 J_K$, where J_K is the $K \times K$ matrix of ones. It follows by the Sherman-Morrison theorem (Graybill, 1983) that the eigenvalues $l_i$of $\Lambda'\Lambda$ are given by $l_1 = N(\sigma_b^2 + K\beta^2)$ and $l_j = N\sigma_b^2$, for $j = 2...K$. Furthermore, from this we obtain the population eigenvalues of Σ_N as:

$$\lambda_1 = N\sigma_F^2 \left(\sigma_\beta^2 + K\beta^2\right) + \sigma_\epsilon^2 \tag{2.3}$$

$$\lambda_i = N\sigma_F^2 \sigma_\beta^2 + \sigma_\epsilon^2, \text{ for i = 2...K} \tag{2.4}$$

$$\lambda_j = \sigma_\epsilon^2, \text{ for j = K + 1...N.} \tag{2.5}$$

Let us now consider the sample covariance matrix derived from a panel of returns of a portfolio of size N observed over T time periods where both N and T are large and $N/T - c = o(N^{-1/2})$. Recent advances in random matrix theory have made it possible to relate the distribution of sample eigenvalues to that of the population eigenvalues described above. Below we will use a set of results for "spiked covariance matrices" as derived by Baik and Silverstein (2005), Onatski (2005) and Paul (2005). Identical results were derived independently by these authors based on very similar assumptions. The results of Baik and Silverstein (2005) were derived under slightly more general conditions.

For the purposes of this note we will assume that at least the first factor can be identified. This is guaranteed if the mean factor loading is high enough in relation to the idiosyncratic component. Our main focus is on the identification of the remaining $K-1$ factors.

Applying the random matrix theory results quoted above we can obtain the expected value of the sample eigenvalues for large N, large T, which are summarized in the following proposition.

Proposition 1. *Let S_N be a sample covariance matrix for a set of observations with the population covariance Σ_N described above. Furthermore, let $V_N S_N V_N^{-1} = \mathrm{diag}\{\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_N\}$, be its eigenvalue decomposition with $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_N$. Then if $N/T - c = o(N^{-1/2})$ we have the following almost sure limits:*

$$(2.6) \qquad a) \quad \hat{\lambda}_1 \overset{a.s.}{\to} \left\{N\sigma_F^2\left(\sigma_\beta^2 + K\beta^2\right) + \sigma_\epsilon^2\right\}\left\{1 + \frac{1}{T}\frac{\sigma_\epsilon^2}{\sigma_F^2\left(\sigma_\beta^2 + K\beta^2\right)}\right\},$$

$$(2.7) \qquad b) \quad \hat{\lambda}_i \overset{a.s.}{\to} \left\{N\sigma_F^2 \quad \sigma_\beta^2 + \sigma_\epsilon^2\right\}\left\{1 + \frac{1}{T}\frac{\sigma_\epsilon^2}{\sigma_F^2\sigma_\beta^2}\right\}, \text{for} \quad i = 2...K \quad \text{and } N \geq \frac{1}{T}\left(\frac{\sigma_\epsilon^2}{\sigma_F^2\sigma_\beta^2}\right)^2,$$

$$(2.8) \qquad c) \quad \hat{\lambda}_i \overset{a.s.}{\to} \sigma_\epsilon^2(1 + \sqrt{N/T})^2, \text{for} \quad i = 2...K \quad \text{and } N < \frac{1}{T}\left(\frac{\sigma_\epsilon^2}{\sigma_F^2\sigma_\beta^2}\right)^2,$$

$$(2.9) \qquad d) \quad \hat{\lambda}_j \overset{a.s.}{\to} \sigma_\epsilon^2(1 + \sqrt{N/T})^2, \ \text{for} \quad j = K + 1...N.$$

Notice that the sample eigenvalues are biased estimates of the corresponding population eigenvalues. Moreover, the bias is always positive and does not disappear as we add more securities to the portfolio. The bias only disappears as $T \to \infty$, that is as we add more time periods to the sample. The quantity $\sigma_\epsilon^2(1 + \sqrt{N/T})$ corresponds to the upper support bound of the Marcenko-Pastur distribution, which characterizes the sample eigenvalues of the sample covariance matrix with mean zero and variance $\sigma_\epsilon^2 I_N$. We shall label this as the Marcenko-Pastur bound.

To exemplify the results of Proposition 1 let us consider the calibration of Brown (1989) based on the NYSE. Thus, let $\beta = 1$,σ_β^2=0.01, σ_F^2=0.000158 and σ_ϵ^2=0.0045. Furthermore, we let $T = 80$ and simulate the factor economy for portfolio sizes between $N = 50$ and $N = 200$. For each value of N we simulate the portfolio 300 times and compute the corresponding sample covariance matrix. This procedure is then used to extract the sample eigenvalues. The results for the first 10 eigenvalues are plotted in Figure 1, where we report the mean eigenvalue over the simulations for each portfolio

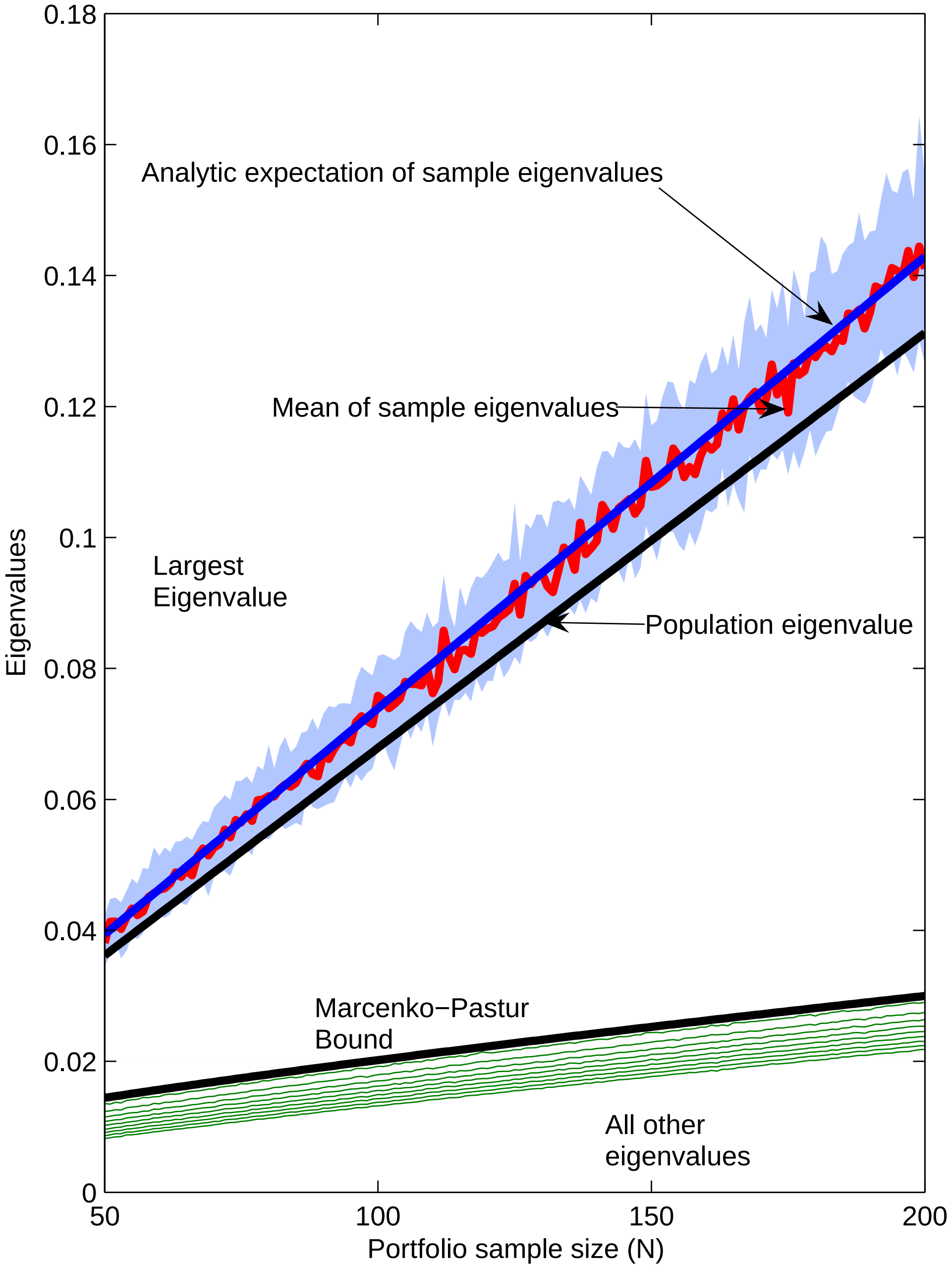

FIGURE 1. Limiting behavior of 10 largest eigenvalues as a function of portfolio size.

size. Furthermore, for the largest eigenvalue we compute the interquartile range of the distribution which is reported in the figure as the shaded area around the mean value of the first eigenvalue for each N. We also report the Marcenko-Pastur bound introduced above which characterizes the largest eigenvalue compatible with a pure idiosyncratic noise model. Additionally, we plot the population eigenvalues for the calibration at each N and the corresponding almost sure limit derived in Proposition 1.

Notice that the sample eigenvalues are biased away from the population eigenvalues at each portfolio size and that the extent of the bias is correctly estimated by the limits derived in Proposition 1. The bias is substantial and the population eigenvalues correspond the lower 25th percentile of the distribution of sample eigenvalues. Additionally, notice that all the remaining $N-1$ eigenvalues are bounded from above by the Marcenko-Pastur bound. This implies that in this setting it is not possible to identify any of the $K-1$ factors even though they contribute equally to the asset returns. No test based on the sample eigenvalues will be able to distinguish the second through $K-$th eigenvalues from the remaining $N-K$ eigenvalues due to the idiosyncratic noise component over the given range of portfolio sizes. Factor estimation based on PCA will only uncover a market factor corresponding to the first eigenvalue which will have high explanatory power, yet it will be a biased estimate of the true first factor.

Although the parameter values used in this simulation are similar to those used in many empirical application we can enquire further as to the minimum portfolio size $N_{\min}$ required to correctly identify all factors of the economy for the given value of T. We obtain $T_{\min} \geq \frac{1}{N}\left(\frac{\sigma_\epsilon^2}{\sigma_F^2\sigma_\beta^2}\right)^2 \cong 40,000$. This is an extremely large value (over 120 years of daily data) and shows the difficulties involved in identifying the non-diversifiable sources of risk in actual portfolio analysis using PCA. In particular notice that we face the a trade-off between the need for large samples in order to identify all latent factors and model stability over extended periods of time.

Using the results in Onatski (2005) and Paul (2005) we can also give the distribution of the first eigenvalue under the large N, large T asymptotic framework.

Proposition 2. *If* $\epsilon \sim^d N(0, \sigma_\epsilon^2 I_N)$ *and if* $N/T - c = o(N^{-1/2})$ *we have* $\hat{\lambda}_1 \sim^d N(m, q)$, *where*

$$m = \left\{N\sigma_F^2\left(\sigma_\beta^2 + K\beta^2\right) + \sigma_\epsilon^2\right\}\left\{1 + \frac{1}{T}\frac{\sigma_\epsilon^2}{\sigma_F^2\left(\sigma_\beta^2 + K\beta^2\right)}\right\}, \tag{2.10}$$

$$q = 2\left\{N\sigma_F^2\left(\sigma_\beta^2 + K\beta^2\right) + \sigma_\epsilon^2\right\}^2\left\{1 - \frac{1}{NT}\frac{\sigma_\epsilon^2}{\left[\sigma_F^2\left(\sigma_\beta^2 + K\beta^2\right)\right]^2}\right\}. \tag{2.11}$$

3. Conclusion

In this note we have explained the reasons for the bias of APT models estimated by PCA towards a single factor model. We have shown that unless the period of time over which the portfolio is observed is extremely large, it is not possible to identify all the factors of the economy. This is due to the finite sample bias of the estimated eigenvalues of the sample covariance matrix when the number of periods over which the portfolio is observed is of similar orders of magnitude to that of the number of securities in the portfolio. Using recent results in random matrix theory we have characterized the limiting behavior of the sample eigenvalues and the distribution of the largest eigenvalue.

This note challenges the use of PCA in its standard form as a tool for factor analysis in finite samples. The correct estimation of factors requires the use of bias corrections in finite samples. Moreover, it seems that the need arises to explore the finite sample properties of other factor estimation procedures other than PCA if we are to identify the full set of factors which determine the asset returns in a portfolio observed only for a finite period of time.

References

Baik, J., and J. W. Silverstein (2005): "Eigenvalues of Large Sample Covariance Matrices of Spiked Population Models," University of Michigan, Ann Arbor.

Bouchard, J.-P., and M. Potters (2003): *Theory of Financial Risk and Derivative Pricing*. Cambridge.

Brown, S. J. (1989): "The Number of Factors in Security Returns," *Journal of Finance*, 54, 1247–1262.

Chamberlain, G., and M. Rothschild (1983): "Arbitrage, Factor Structure, and Mean-Variance Analysis of Large Asset Markets," *Econometrica*, 51, 1281–1304.

Connor, G., and R. A. Korakjczyk (1993): "A Test for the Number of Factors in an Approximate Factor Model," *Journal of Finance*, 48, 1263–1291.

Geweke, J., and G. Zhou (1996): "Measuring the Pricing Error of the Arbitrage Pricing Theory," *Review of Financial Studies*, 9, 557–587.

Graybill, F. A. (1983): *Matrices with Applications in Statistics*. Wadsworth.

Hahn, J., and G. Kuersteiner (2002): "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects when Both N and T are Large," *Econometrica*, 70, 1639–1657.

Jones, C. S. (2001): "Extracting Factors from Heteroskedastic Asset Returns," *Journal of Financial Economics*, 62, 293–325.

Merville, L. J., and Y. Xu (2002): "The Changing Factor Structure of Equity Returns," University of Texas.

Onatski, A. (2005): "Asymptotic Distribution of Principal Components Estimators of Large Spherical Factor Models," Columbia University.

Paul, D. (2005): "Asymptotics of the Leading Sample Eigenvalues for a Spiked Covariance Model," Stanford University.

Plerou, V., V. Gopikrishnan, B. Rosenau, L. Amaral, T. Guhr, and E. Stanley (2002): "Random Matrix Approach to Cross-Correlations in Financial Data," *Physical Review E*, 65, 1–17.

Ross, S. (1976): "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, 3, 343–362.

Trzcinka, C. (1986): "On the Number of Factors in the Arbitrage Pricing Model," *Journal of Finance*, 41, 347–368.