A Non-Normal Principal Components Model for Security Returns

Sander Gerber Babak Javid Harry Markowitz Paul Sargen
David Starer

February 21, 2019

Abstract

We introduce a principal components model for securities' returns. The components are non-normal, exhibiting significant skewness and kurtosis. The model can explain a large proportion of the variance of the securities' returns with only one or two components. Third and higher-order components individually contribute so little that they can be considered to be noise terms.

1 Introduction

In this paper, we propose a non-normal principal component model of the stock market. We create the model from a statistical study of a broad cross-section of approximately 5,000 US equities daily for 20 years.

In our analysis, we find that only a small number of components can explain a significant amount of the variance of the securities' return. Generally, third and higher-order components (essentially, the idiosyncratic terms) individually contribute so little to the variance that they can be considered to be noise terms.

Importantly, we find that neither the significant components nor the idiosyncratic terms are normally distributed. Both sets exhibit significant skewness and kurtosis.

Therefore, traditional models based on normal distributions are not fully descriptive of security returns. They can neither represent the extreme movements and comovements that security returns often exhibit, nor the low-level economically insignificant noise that security returns also often exhibit. However, these characteristics can be represented by the model presented in this paper.

The finding of non-normality implies that meaningful security analysis requires statistical measures (1) that are insensitive to extreme moves, (2) that are also not influenced by small movements that may be noise, but (3) that still capture information in movements when such movements are meaningful. In Gerber et al. [2019], we introduced the Gerber statistic (GS), which is a robust measure of correlation that satisfies all three of these requirements.

Precursors

Portfolio construction [Markowitz, 1952, 1959] relies heavily on the availability of the matrix of covariances between securities' returns. Often the sample covariance matrix is used as an estimate for the actual covariance matrix. But as early as Sharpe [1963], it has been known that a single factor approximation of this matrix leads to portfolios that outperform those constructed from the sample covariance matrix.

Factor models originated with Spearman [1904], who showed that one can reduce the dimension of a model by expressing the model variables as linear combinations of underlying common factors plus random idiosyncratic terms. Quantitative analysts have embraced this methodology, and factor models for security returns now abound. These models range from the early single factor model of Sharpe [1963], through the three and more factor models of Fama and French [1992, 1993], past the multifactor model of Rosenberg and Maranthe [1976], to modern models using tens or even hundreds of quantitative and categorical explanatory variables.

This diversity in models is possible because of an important property of factor analysis: The factors and the factor loadings are not unique. That is, the factors can be rotated using any orthonormal matrix and the model remains identical. Therefore, in the abstract, any model can be expressed as a rotated version of any other model.

Nevertheless, despite the numerous studies of these factors, we are unaware of research focused on the statistical distributions of the factors themselves. As described by Cont [2001], a study by Laloux et al. [2000] showed that principal components, apart from the ones corresponding to the largest few eigenvalues, "do not seem to contain any information: in fact, their marginal distribution closely resembles the spectral distribution of a positive symmetric matrix with random entries whose distribution is the 'most random possible'—i.e., entropy maximizing. These results strongly question the validity of the use of the sample covariance matrix as an input for portfolio optimization ... and support the rationale behind factor models ... where the correlations between a large number of assets are represented through a small number of factors." We will use such a low-order principal component model, but will examine in more detail the statistics of the principal components.

2 Theory

Our objective is to examine the characteristics of securities' returns through the lens of a factor model. The structure of the model follows the general form

$$r_{tj} = \sum_{k=1}^{K} f_{tk} x_{jk} + \varepsilon_{tj} \tag{1}$$

where x_{jk} is the exposure of security j to a component k, f_{tk} is the return of component k for time period t, and ε_{tj} is an idiosyncratic or noise term. The f_{tk} terms can be considered to be the "drivers" of the securities' returns. With obvious notation, Equation (1) in matrix form is

$$\boldsymbol{R} = \boldsymbol{F}\boldsymbol{X}^{\top} + \mathcal{E} \tag{2}$$

The model in Equations (1) and (2) is extremely versatile and includes factor models, smart-beta models, econometric models, time series models, statistical models, and many others as special cases. For example, in the case of a factor model, the components could be financial statement data such as earnings yield, dividend yield, and so on. Here, x_{jk} would represent the financial data itself (centered and scaled to a standard deviation of one across the investment universe) and f_{tk} would represent the return obtained in period t from a one-standard deviation exposure to factor k. The model places no restriction on whether any of the returns should be raw, excess, or active.

The model can operate in several modes. For example, with the model in an identification mode, the exposures and security returns are assumed to be known, and the returns to the factors are found by linear or generalized regression. With the model in a prediction or data generating mode, the f_{tk} are assumed to be known, the ε_{tj} are replaced by their expected values of zero, and the security returns r_{tj} are computed as linear combinations of the component returns f_{tk}.

We can use the model in the data identification mode to gain a better understanding of market characteristics. For this purpose, we perform principal component analysis on a broad range of stocks to find the statistical distributions of important components.

Principal Component Analysis (PCA) [Pearson, 1901, Hotelling, 1933] produces a parsimonious summary of data in terms of orthogonal sets of standardized linear combinations of the original data.

Consider again the return matrix $\boldsymbol{R}$ whose columns represent different securities and whose rows represent different time intervals. We remove the mean of each column of $\boldsymbol{R}$ to obtain the "centered" return matrix $\boldsymbol{R}_c$. The sample covariance matrix of the returns is then

$$\boldsymbol{C} = \frac{1}{M-1}\boldsymbol{R}_c^\top \boldsymbol{R}_c,$$

where M is the number of time samples.

The singular value decomposition (SVD) [Golub and Van Loan, 2013] of $\boldsymbol{R}_c$ is

$$\boldsymbol{R}_c = \boldsymbol{W}_{\text{TOT}}\boldsymbol{S}_{\text{TOT}}\boldsymbol{X}_{\text{TOT}}^\top$$

where $\boldsymbol{W}_{\text{TOT}}$ and $\boldsymbol{X}_{\text{TOT}}$ are unitary matrices (i.e., matrices whose inverses equal their conjugate transposes) and $\boldsymbol{S}_{\text{TOT}}$ is a diagonal matrix whose elements $s_k;\ k = 1, \ldots, K$ are the singular values of $\boldsymbol{R}_c$. The singular values are non-negative real numbers.

In terms of the singular value decomposition, the sample covariance matrix is

$$\begin{aligned}\boldsymbol{C} &= \frac{1}{M-1}\boldsymbol{X}_{\text{TOT}}\boldsymbol{S}_{\text{TOT}}\boldsymbol{W}_{\text{TOT}}^\top\boldsymbol{W}_{\text{TOT}}\boldsymbol{S}_{\text{TOT}}\boldsymbol{X}_{\text{TOT}}^\top \\ &= \frac{1}{M-1}\boldsymbol{X}_{\text{TOT}}\boldsymbol{S}_{\text{TOT}}^2\boldsymbol{X}_{\text{TOT}}^\top,\end{aligned}$$

which can be rearranged to give

$$\boldsymbol{C}\boldsymbol{X}_{\text{TOT}} = \boldsymbol{X}_{\text{TOT}}\left(\frac{1}{M-1}\boldsymbol{S}_{\text{TOT}}^2\right).$$

The latter expression is the eigendecomposition of the covariance matrix. Therefore, the singular values are related to the eigenvalues $\lambda_k;\ k = 1, \ldots, K$, of the sample covariance $\boldsymbol{C}$ by the identity

$$\lambda_k = \frac{1}{M-1} s_k^2.$$

Each eigenvalue is equal to the variance of its respective principal component.

Letting $\boldsymbol{W}_{\text{TOT}} \boldsymbol{S}_{\text{TOT}} = \boldsymbol{F}_{\text{TOT}}$, the centered return matrix $\boldsymbol{R}_c$ can be written in the principal component form

$$\boldsymbol{R}_c = \boldsymbol{F}_{\text{TOT}} \boldsymbol{X}_{\text{TOT}}^\top \tag{3}$$

where the matrix $\boldsymbol{F}_{\text{TOT}}$ (called the score) has columns that are mutually orthogonal. The matrix $\boldsymbol{X}_{\text{TOT}}$ is a rotation matrix called the coefficient or loading matrix. Importantly, each column of $\boldsymbol{F}_{\text{TOT}}$ is called a principal component and can be considered to be a time series. The centered return matrix $\boldsymbol{R}_c$, therefore, is a linear combination of mutually orthogonal time series. These principal components are entirely analogous to factor return time series.

Equation (3) is an identity; that is, the return matrix on the left hand side is exactly equal to the decomposition on the right hand side. If, however, we consider a small number of principal components (say m components) to describe the data with sufficient accuracy, we can categorize the remaining $K - m$ components as noise. Accordingly, we can partition $\boldsymbol{F}_{\text{TOT}}$ and $\boldsymbol{X}_{\text{TOT}}$ into signal and noise parts as follows:

$$\boldsymbol{R}_c = \begin{bmatrix} \boldsymbol{F}_{\text{SIG}} & \boldsymbol{F}_{\text{NOISE}} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_{\text{SIG}}^\top \\ \boldsymbol{X}_{\text{NOISE}}^\top \end{bmatrix} = \boldsymbol{F}_{\text{SIG}} \boldsymbol{X}_{\text{SIG}}^\top + \mathcal{E} \tag{4}$$

where $\mathcal{E} = \boldsymbol{F}_{\text{NOISE}} \boldsymbol{X}_{\text{NOISE}}^\top$ is a noise matrix, and can be considered to be the "idiosyncratic" part of the decomposition. That is, the entry in the t-th row and the j-th column of $\mathcal{E}$ is the idiosyncratic return of the j-th security over interval t. Note the direct correspondence between the representations in Equation (4) and (2).

Making use of the decomposition of R_c into its signal and noise components, and the properties of the unitary matrices, we find that the covariance matrix is

$$\begin{aligned} \boldsymbol{C} &= \boldsymbol{X}_{\text{SIG}} \boldsymbol{\Lambda}_{\text{SIG}} \boldsymbol{X}_{\text{SIG}}^\top + \boldsymbol{X}_{\text{NOISE}} \boldsymbol{\Lambda}_{\text{NOISE}} \boldsymbol{X}_{\text{NOISE}}^\top, \\ &= \boldsymbol{C}_{\text{SIG}} + \boldsymbol{C}_{\text{NOISE}}, \end{aligned}$$

where $\boldsymbol{\Lambda}_{\text{SIG}}$ and $\boldsymbol{\Lambda}_{\text{NOISE}}$ are diagonal matrices containing the eigenvalues of the signal and noise parts, respectively. Note that $\boldsymbol{\Lambda}_{\text{SIG}}$ is an $m \times m$ matrix. In particular, if $m = 1$, it is a scalar.

Notice that the noise covariance matrix $\boldsymbol{X}_{\text{NOISE}} \boldsymbol{\Lambda}_{\text{NOISE}} \boldsymbol{X}_{\text{NOISE}}^\top$ is not diagonal. Therefore, the idiosyncratic terms are not orthogonal, but are mutually correlated.

3 Empirical Results

In our empirical tests, we used twenty years of daily returns from approximately 5,000 US stocks. We truncated the absolute value of returns to 30% to prevent our results being unduly influenced by outliers. We separated the period from the beginning of 1998 to the

end of 2017 into ten non-overlapping two-year intervals. For each interval we performed 1,000 repetitions of the following test.

In each test, we chose 100 stocks randomly with replacement from the available universe. For each two-year sample of 100 stocks, we formed a centered return matrix $\boldsymbol{R}_c$ as described above, and computed the principal component score matrix $\boldsymbol{F}$ and loading matrix $\boldsymbol{X}$. Recall that the columns of $\boldsymbol{F}$ represent orthogonal time series. The first column is the vector that best explains all columns of $\boldsymbol{R}_c$. The second column of $\boldsymbol{F}$ is the vector that is orthogonal to the first column, and best explains the remainder of the variance in $\boldsymbol{R}_c$. Similarly, for $n > 1$, the nth column of $\boldsymbol{F}$ is the one that is orthogonal to all preceding $n - 1$ columns and best explains the remaining variance in $\boldsymbol{R}_c$.

Table 1 lists the summary statistics (pooled over all two-year periods and all experiments) of the variance explained by the first ten principal components. This shows that the median variance explained by the first two principal components (PC01 and PC02) are 12.8% and 7.4%, respectively. Beyond the third principal component, the variance explained falls below 5%.

	min	Q1	med	Q3	max
PC01	5.56	11.09	12.82	15.75	37.61
PC02	3.20	6.07	7.44	9.06	19.54
PC03	2.51	4.73	5.53	6.55	12.74
PC04	2.21	4.02	4.58	5.24	10.08
PC05	2.11	3.53	3.97	4.44	7.91
PC06	1.95	3.17	3.52	3.89	6.24
PC07	1.61	2.86	3.18	3.49	5.66
PC08	1.53	2.62	2.92	3.18	4.56
PC09	1.35	2.41	2.69	2.91	4.25
PC10	1.32	2.23	2.50	2.70	3.87

Table 1: Pooled Summary Statistics of Variance Explained by the First 10 Principal Components

Figure 1 gives a graphical representation of the data summarized in Table 1. From it, we see again that the first component explains the most variance in the returns. In many cases, the first component explains more than 30% of the variance in the returns. Using a 10% cutoff for significance, we believe that the returns can be explained by a single-factor model; i.e., a model using only the first principal component.

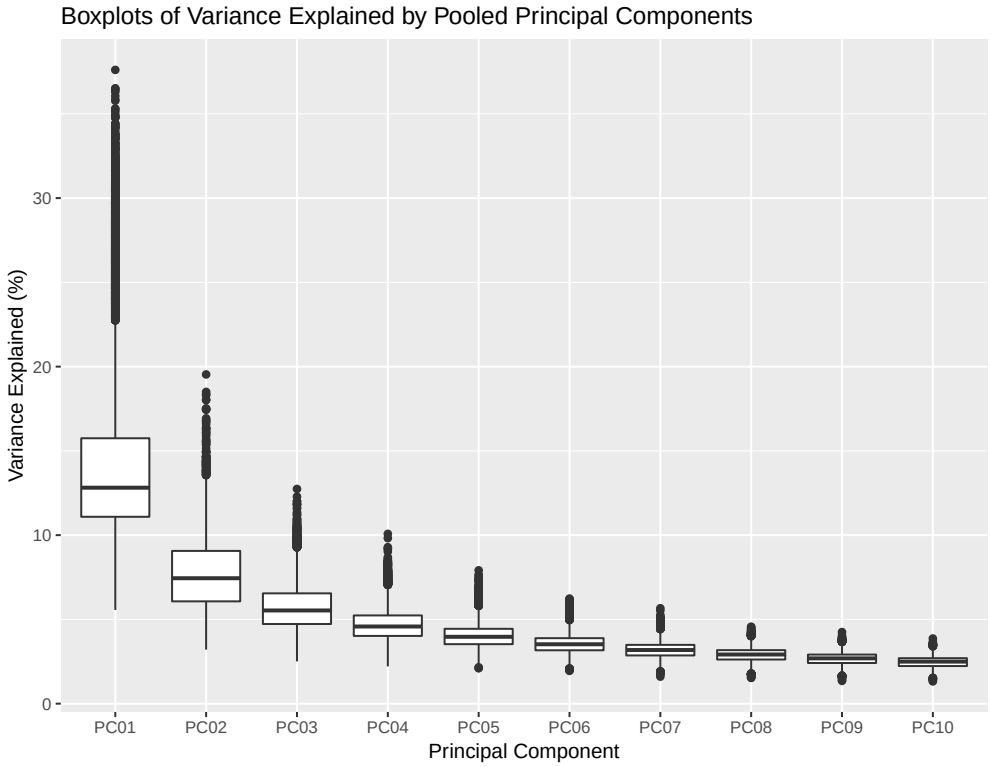

Figure 1: Box Plots of Variance Explained by the First 10 Principal Components

The variance explained by the principal components varies by time period. Table 2 shows the median variance explained by the first five principal components for the ten two-year periods from the beginning of 1998 to the end of 2017. Naturally, the first principal component always explains more variance than the other components. However, it is clear that in the 2008–2009 period, PC01 explains more variance than other times, and much more than is explained by other components at the same time. This, of course, corresponds to a period in which stock returns were highly correlated. Thus, a single component explains stock returns, and that single component is the principal component analog of the market.

	PC01	PC02	PC03	PC04	PC05
1998–1999	8.76	6.90	5.65	4.80	4.17
2000–2001	12.15	6.54	4.90	4.18	3.75
2002–2003	12.22	8.28	5.91	4.82	4.11
2004–2005	11.21	8.49	6.36	5.02	4.28
2006–2007	13.91	8.49	5.97	4.63	3.84
2008–2009	26.91	5.03	4.18	3.60	3.17
2010–2011	24.70	9.00	5.43	4.07	3.36
2012–2013	12.35	9.13	6.78	5.36	4.42
2014–2015	13.17	6.77	5.42	4.71	4.17
2016–2017	11.96	6.75	5.38	4.65	4.12

Table 2: Median Variance Explained by First Five Principal Components over Time.

Figure 2 shows histograms of the statistics of the first principal component for each of the 1,000 experiments conducted for 100 companies for every two-year period from 1998 to 2000.

1. The top panel of the figure shows the variance explained in each experiment. The variance explained is skewed left with a median of about 8.5%.

2. The second panel shows the standard deviation of the first principal component. The distributional shape of the standard deviations is not clear, but lie in the range of 9% to 14%.

3. The third panel shows the skewness of the first component. This is clearly bimodal. The reason for the bimodality is that principal components are unique only up to a change of sign. Therefore, the sign of all odd moments is indeterminate.

4. The bottom panel shows the kurtosis of the first principal component. A normal distribution has a kurtosis of 3. Here, we see that in the vast majority of cases, the kurtosis is greater than 3, and the distributions are therefore leptokurtic.

Figures 2 through 11 show the histograms for all two-year periods studied. In each figure, the layout is the same as that described above.

The ambiguity in the signs of the principal components is an important issue when one tries to compute statistics of these components. We have tried to resolve the ambiguity for the first principal component at least. We believe that the first principal component represents an estimate of the market. Therefore, this component should be positively correlated with a broad market index. Accordingly, we computed the correlation between the Russell 3000 Index returns and the first principal component in every experiment. We multiplied the first principal component by the sign of this correlation, in this way attempting to ensure that the first principal component and the market were positively correlated. This should have removed the ambiguity in the signs of the odd order moments and resulted in unimodal odd-order moments.

The results of the transformation were largely successful, although periods 1998–2000, 2004–2005, 2006–2007, and 2012–2013 still show some bimodal behavior. Nevertheless, visual inspection of the graphs shows the following:

- The standard deviation of the first principal component is approximately 10%. Recall that the component itself is constrained to have a norm of one.

- Looking at only the most prominent mode in each case, the skewness is approximately positive 25% or negative 25%.

- The kurtosis is significantly greater than the normal kurtosis of 3.

Although not shown in the figures, the mean of the principal component was indistinguishable from zero in each case.

In addition to the results described above, for each test, we also computed the first four moments (M1 through M4) of the first principal components. The first moment was zero, and the second through fourth are listed in Table 3. Note that the third moment is contaminated because of the sign ambiguity discussed above.

	M2	M3	M4
1998–1999	1.32E-02	-1.00E-04	8.00E-04
2000–2001	2.34E-02	7.00E-04	2.40E-03
2002–2003	1.60E-02	2.00E-04	9.00E-04
2004–2005	9.10E-03	-1.00E-04	4.00E-04
2006–2007	1.03E-02	-2.00E-04	5.00E-04
2008–2009	6.24E-02	-9.00E-04	2.06E-02
2010–2011	2.52E-02	-1.00E-03	3.50E-03
2012–2013	9.90E-03	-1.00E-04	4.00E-04
2014–2015	1.05E-02	-2.00E-04	4.00E-04
2016–2017	1.09E-02	-2.00E-04	5.00E-04

Table 3: Moments of the First Principal Component.

Conclusion

In this paper, we propose a model that accurately mimics the statistical properties of security returns. We find that realistic security returns can be generated by a low-order principal component model. We examined the statistics of a large cross section of US equities for the ten two-year periods from 1998 to 2017. In all periods, the principal components were highly skewed and leptokurtic.

Previously [Gerber et al., 2019], we introduced the Gerber statistic, which is a robust measure of correlation between two time series. The statistics reported in the current paper, and the return model proposed, show that characteristics of the market may make the Gerber statistic a better comovement measure for portfolio construction than traditional correlation.

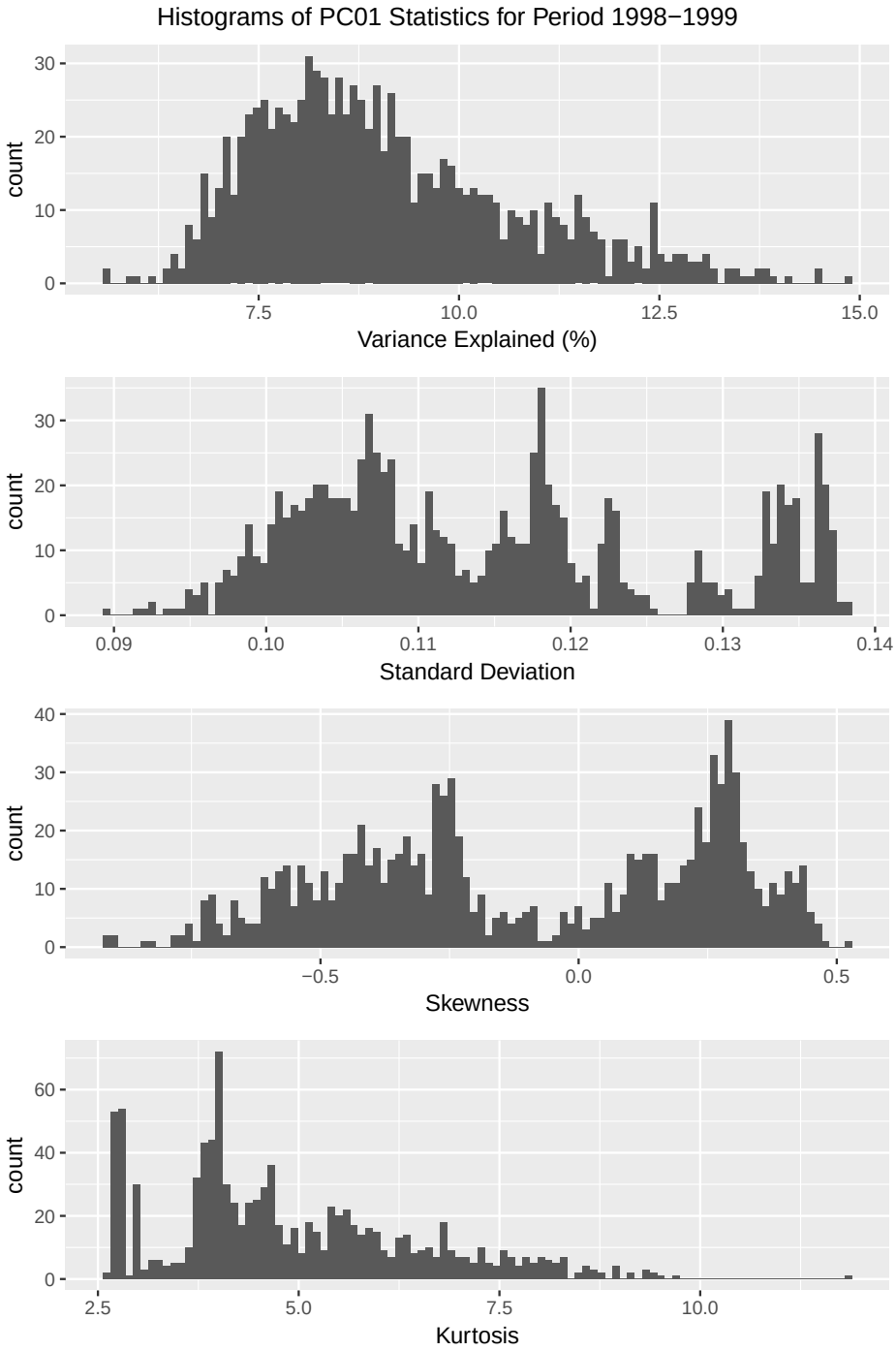

Figure 2: Histograms for the Period 1998–1999

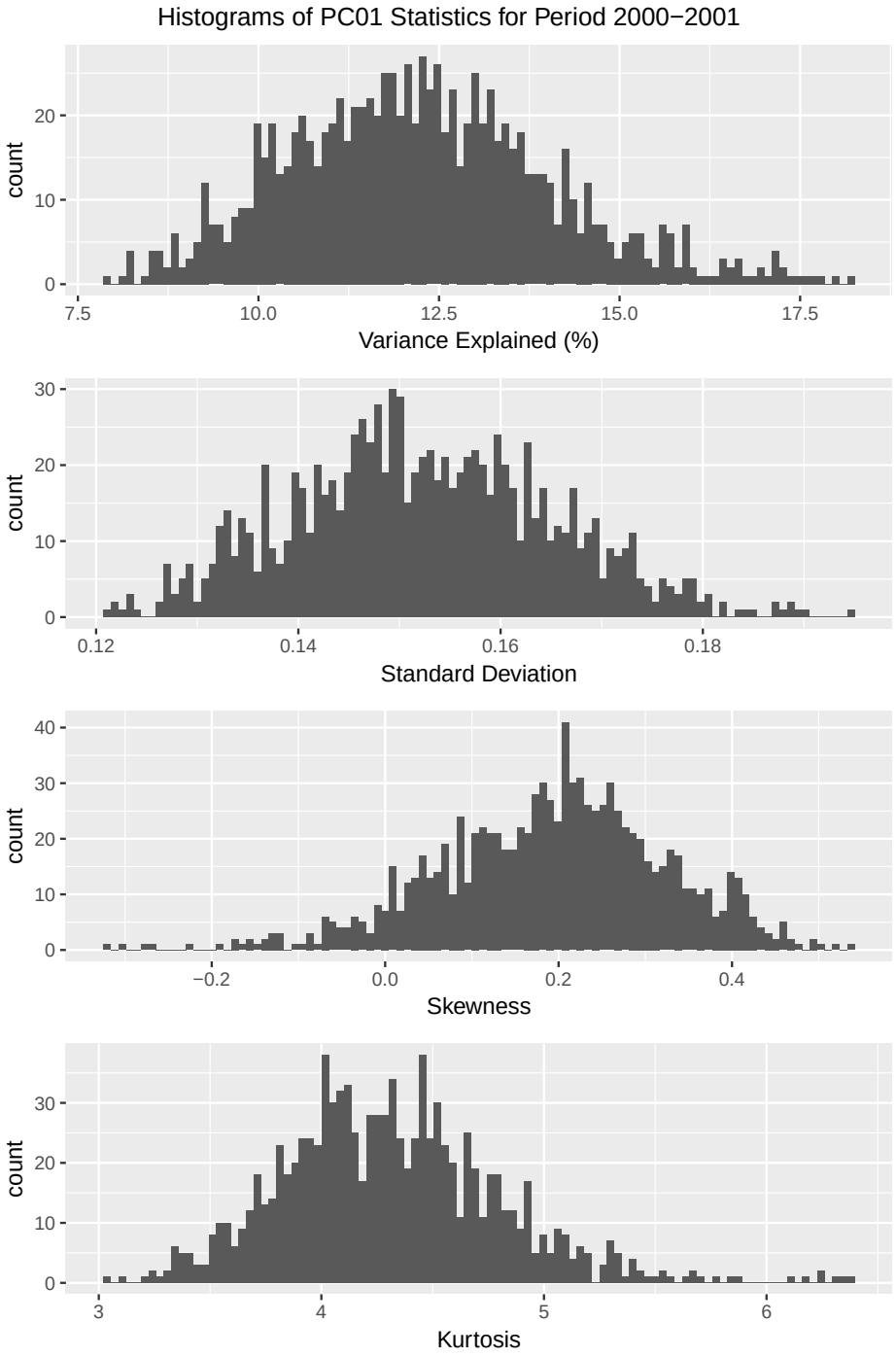

Figure 3: Histograms for the Period 2000–2001

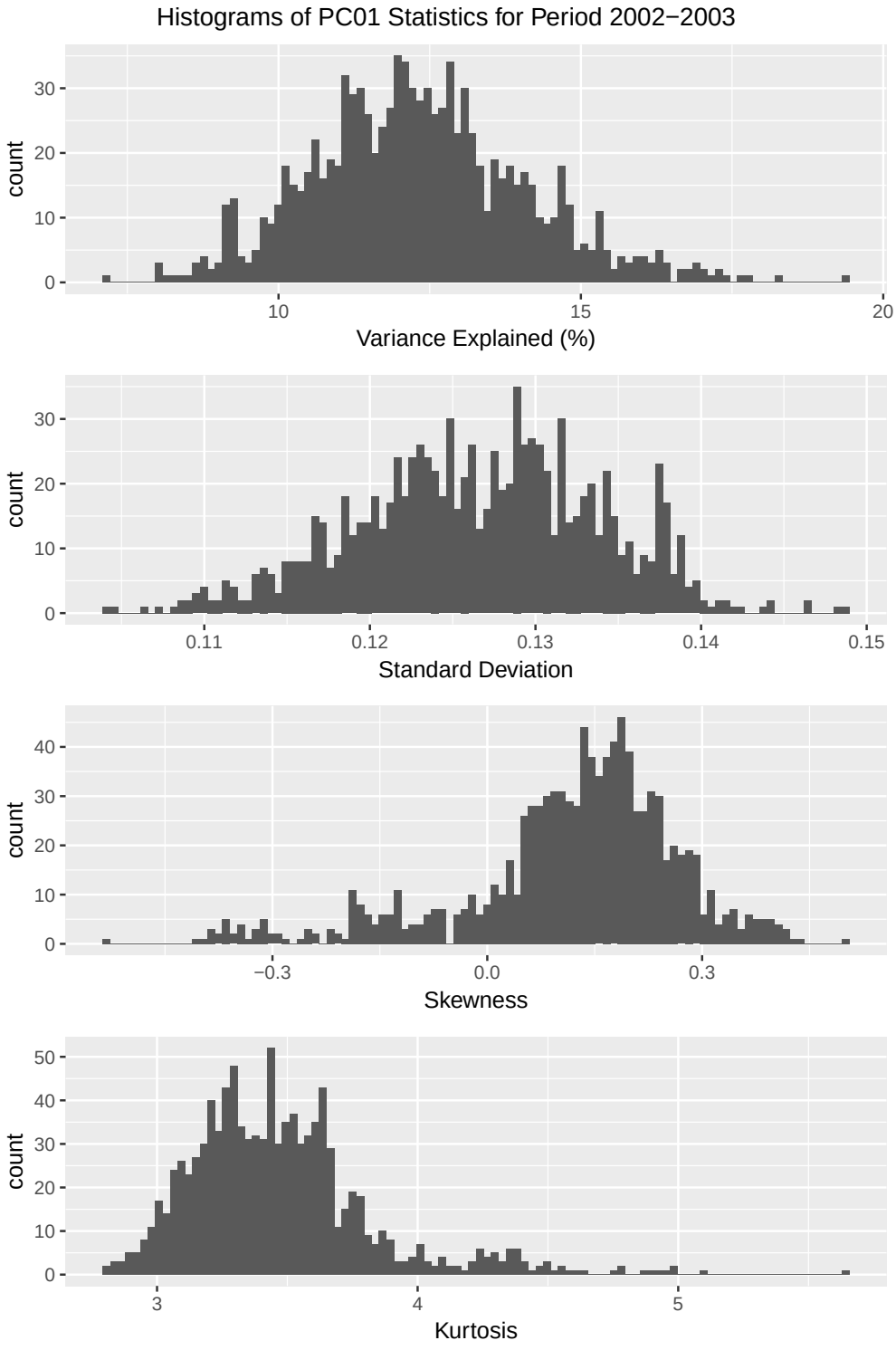

Figure 4: Histograms for the Period 2002–2003

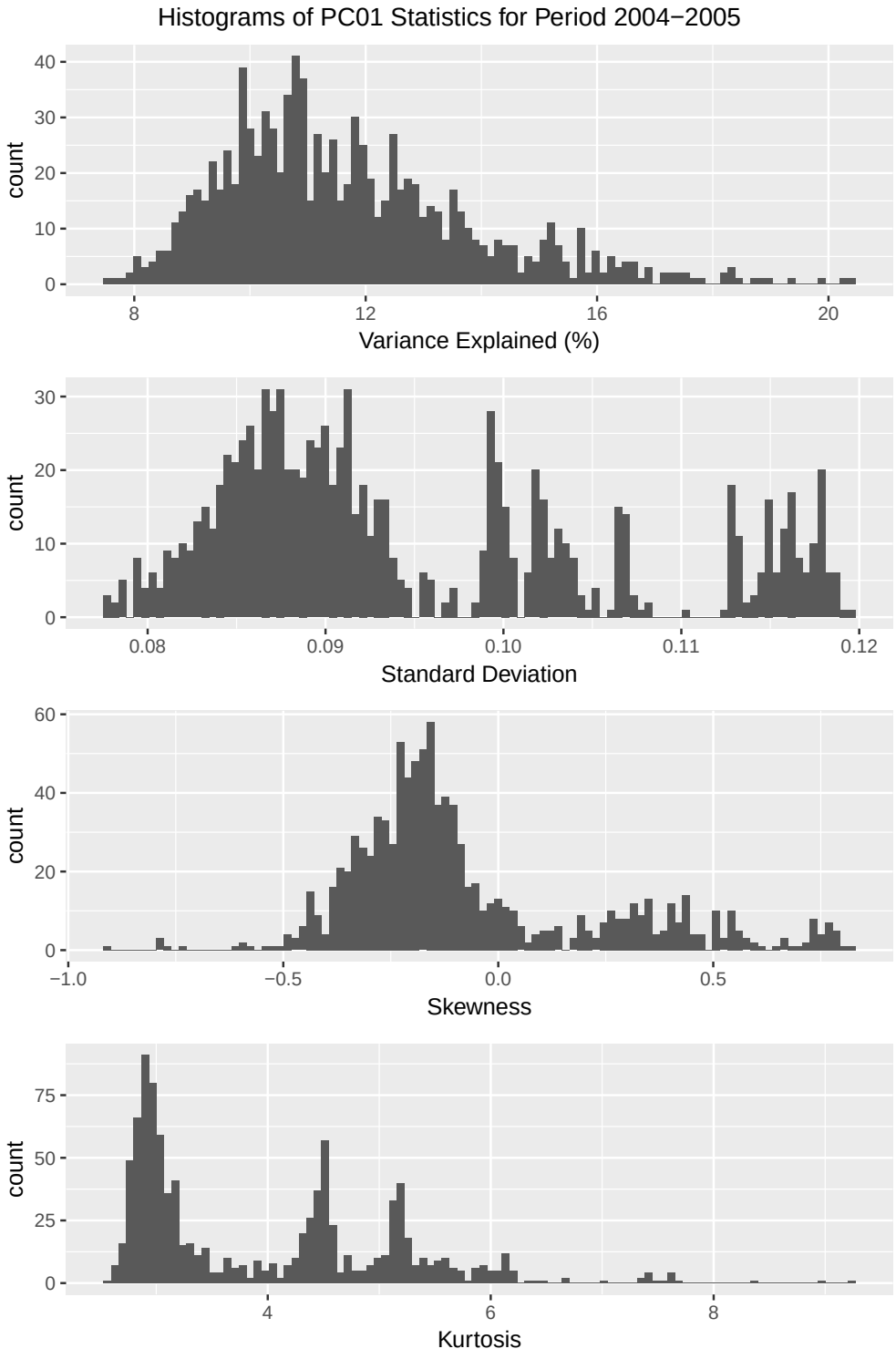

Figure 5: Histograms for the Period 2004–2005

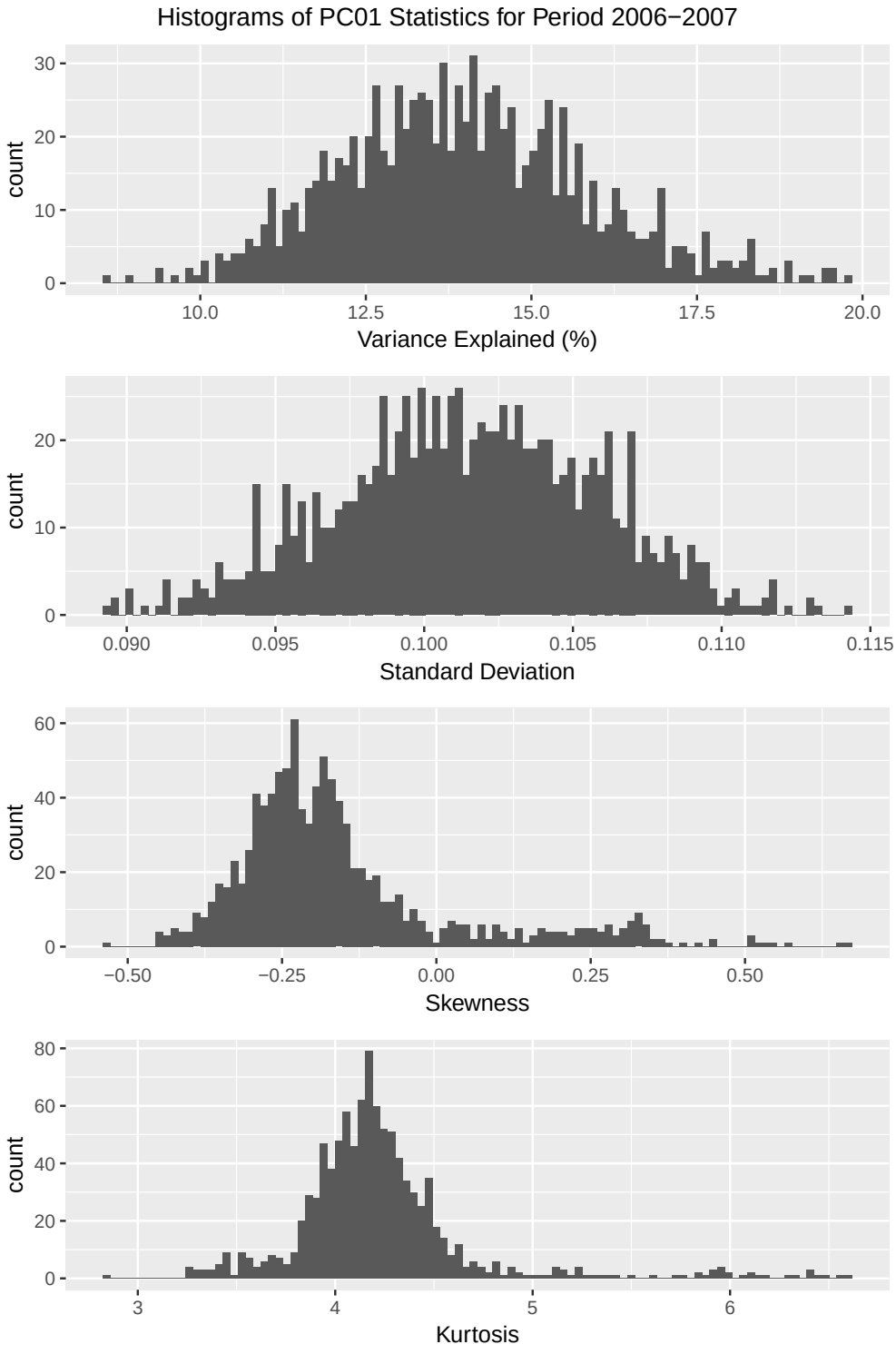

Figure 6: Histograms for the Period 2006–2007

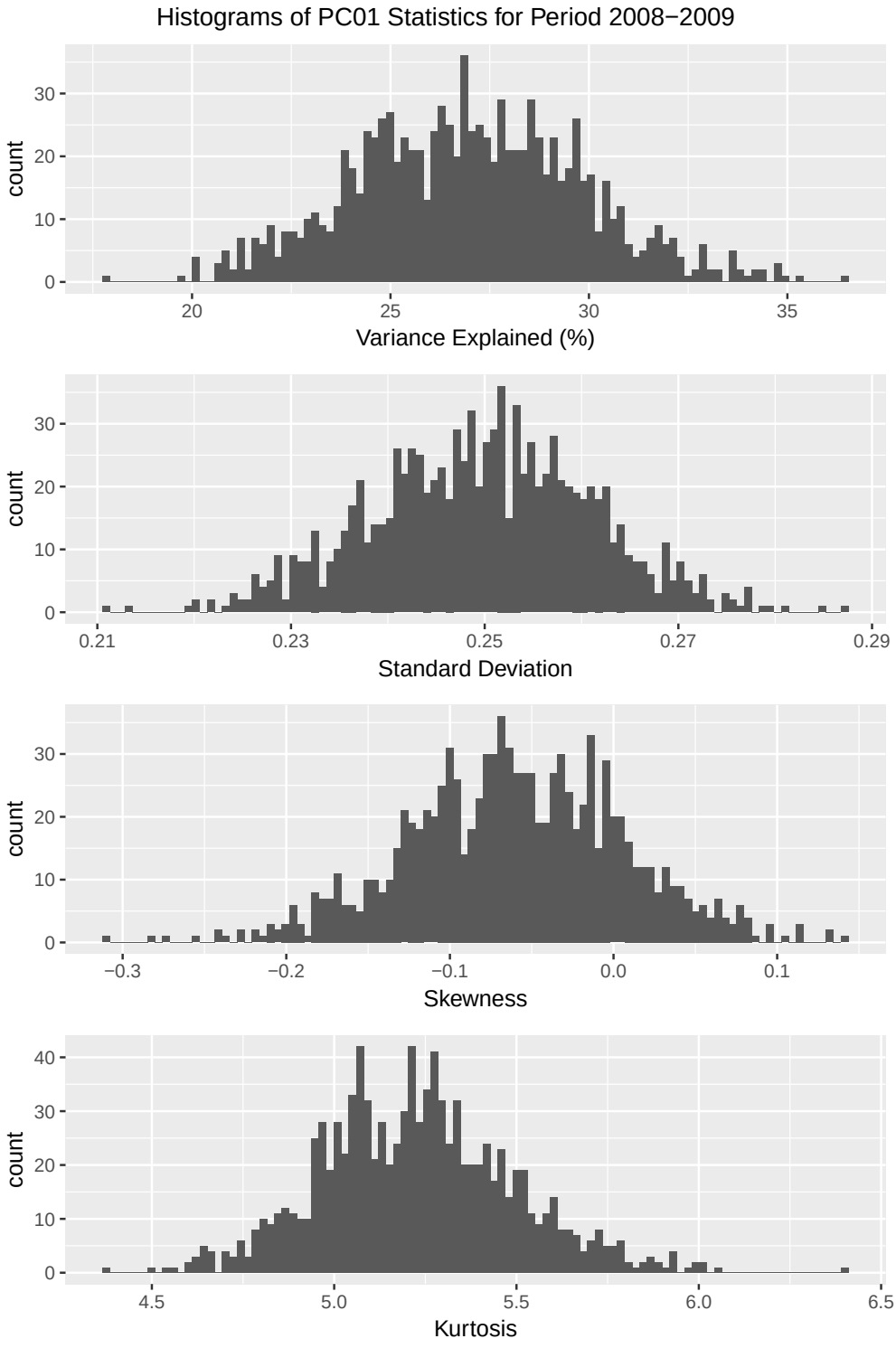

Figure 7: Histograms for the Period 2008–2009

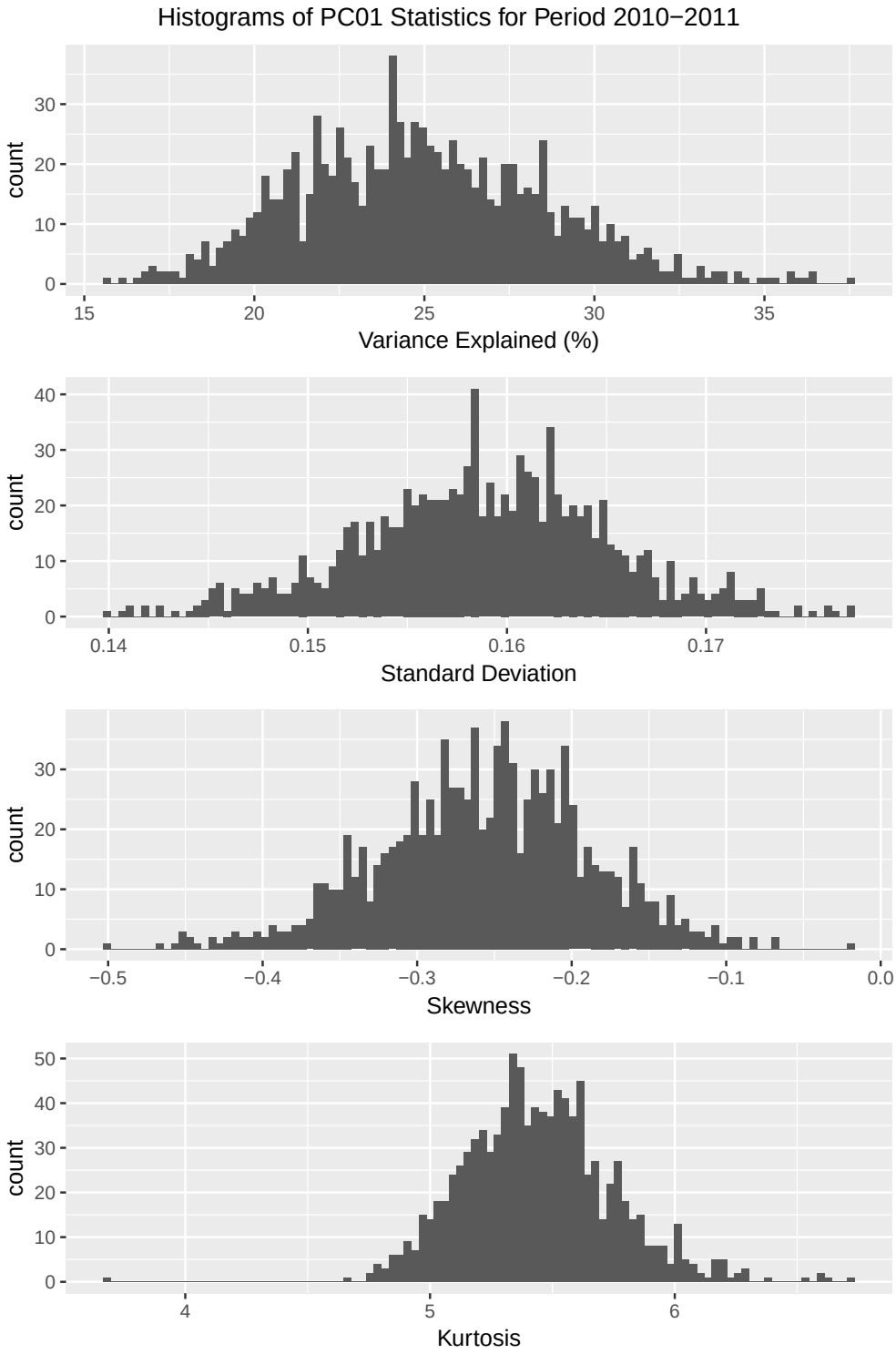

Figure 8: Histograms for the Period 2010–2011

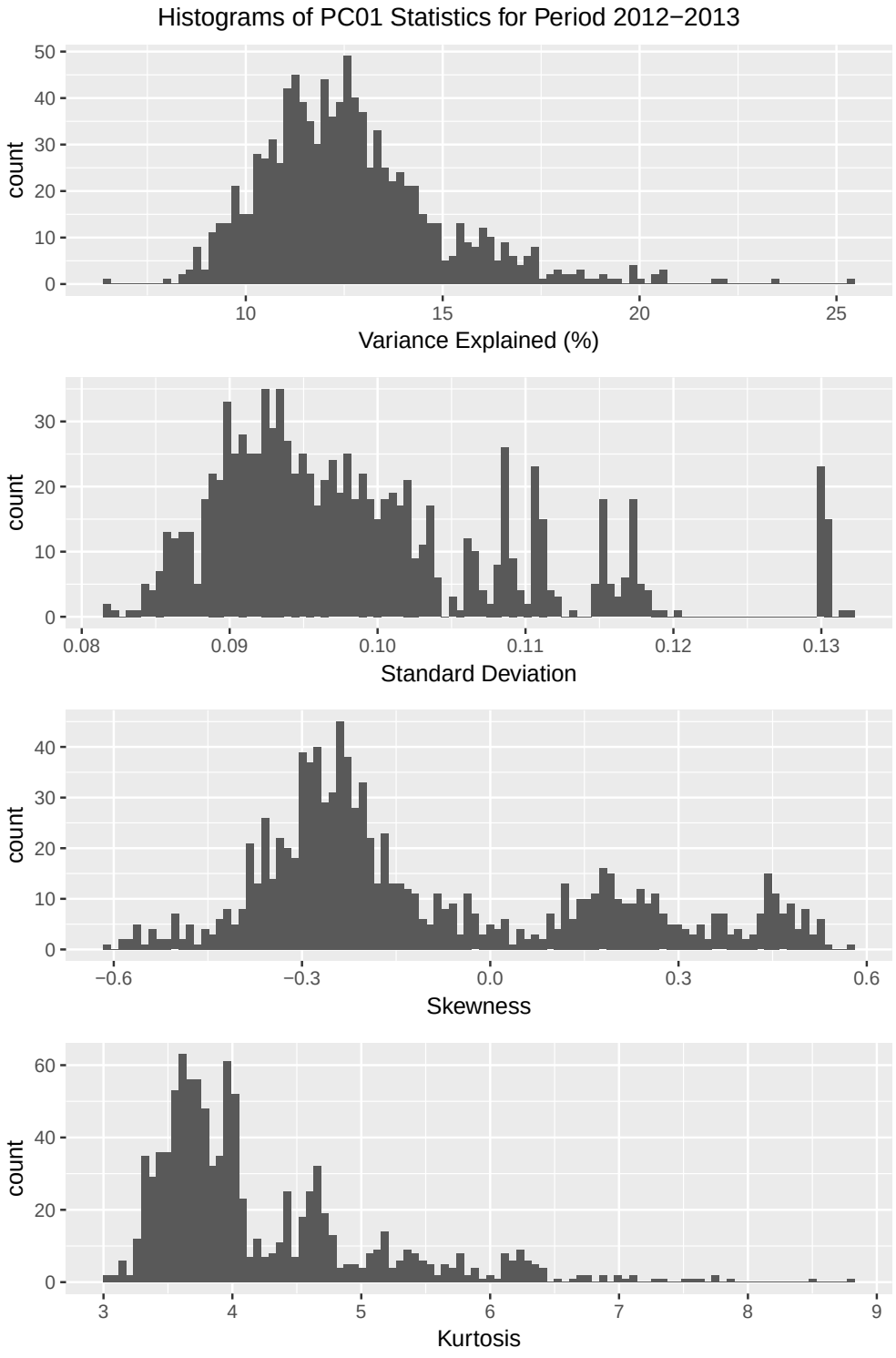

Figure 9: Histograms for the Period 2012–2013

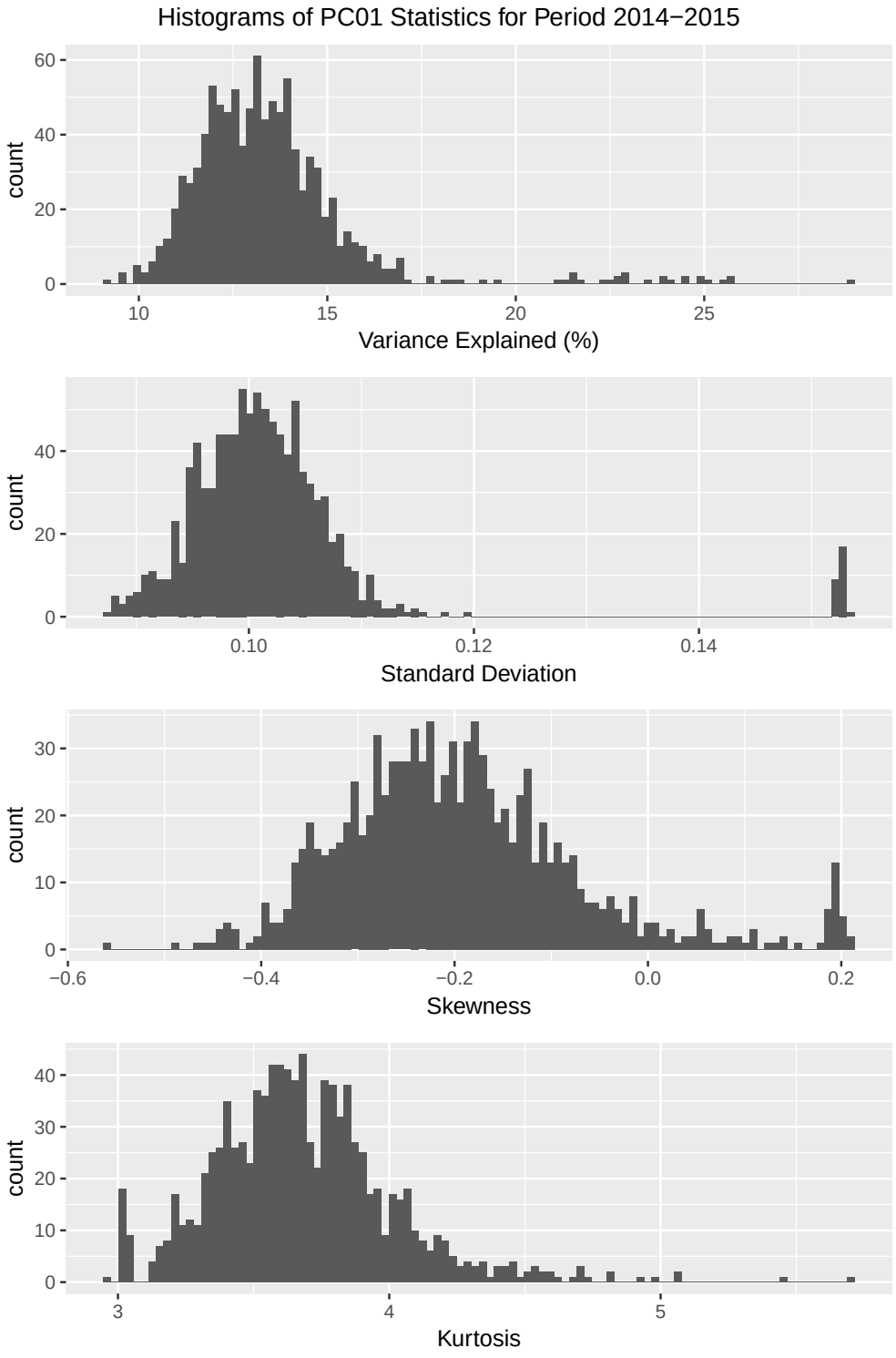

Figure 10: Histograms for the Period 2014–2015

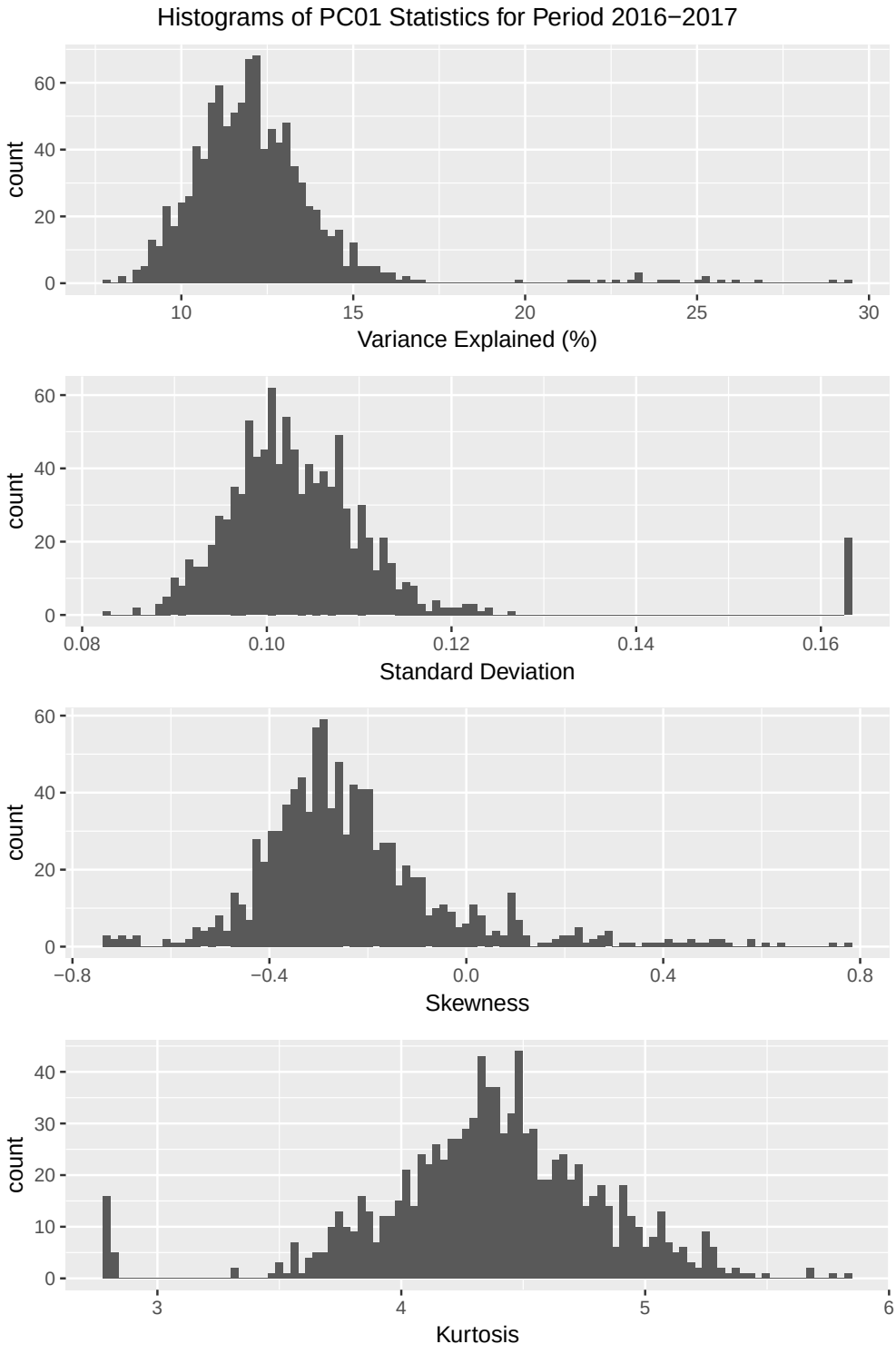

Figure 11: Histograms for the Period 2016–2017

References

Rama Cont. Empirical Properties of Asset Returns: Stylized Facts and Statistical Issues. *Quantitative Finance*, 1:223–236, 2001.

Eugene F Fama and Kenneth R French. The Cross-Section of Expected Stock Returns. *Journal of Finance*, 47(2):427–465, June 1992.

Eugene F. Fama and Kenneth R. French. Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33(1):3 –56, February 1993.

Sander Gerber, Babak Javid, Harry Markowitz, Paul Sargen, and David Starer. The Gerber Statistic: A Robust Measure of Correlation. Technical report, Hudson Bay Capital Management, 2019.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Fourth edition, 2013.

Harold Hotelling. Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

Laurent Laloux, Pierre Cizeau, Marc Potters, and Jean-Philippe Bouchaud. Random Matrix Theory and Financial Correlations. *International Journal of Theoretical & Applied Finance*, 3(3):391–397, 2000. ISSN 02190249.

Harry M. Markowitz. Portfolio Selection. *Journal of Finance*, 7(1):77–91, 1952.

Harry M. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. Basil Blackwell, Cambridge, MA, 1959.

Karl Pearson. LIII. On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559–572, 1901.

Barr Rosenberg and Vinay Maranthe. Common Factors in Security Returns: Microeconomic Determinants and Macroeconomic Correlates. In *Proceedings of the Seminar on the Analysis of Security Prices*, pages 61–115. University of Chicago, 1976.

William F. Sharpe. A Simplified Model for Portfolio Analysis. *Management Science*, 9(2): 277–293, January 1963.

Charles E. Spearman. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72–101, January 1904.