Modelling Correlations in Portfolio Credit Risk

Bernd Rosenow*[1], Rafael Weißbach*[2], and Frank Altrock[3]

[1] *Institut für Theoretische Physik, Universität zu Köln, 50923 Köln*

[2] *Institut für Wirtschafts- und Sozialstatistik, Universität Dortmund, 44221 Dortmund*

[3] *Risk Management Support & Control, WestLB AG, 40217 Düsseldorf*

The risk of a credit portfolio depends crucially on correlations between the probability of default (PD) in different economic sectors. Often, PD correlations have to be estimated from relatively short time series of default rates, and the resulting estimation error hinders the detection of a signal. We present statistical evidence that PD correlations are well described by a (one-)factorial model. We suggest a method of parameter estimation which avoids in a controlled way the underestimation of correlation risk. Empirical evidence is presented that, in the framework of the CreditRisk+ model with integrated correlations, this method leads to an increased reliability of the economic capital estimate.

Managing portfolio credit risk in a bank requires a sound and stable estimation of the loss distribution with a special emphasis on the high quantiles denoted as Credit Value-at-Risk (CreditVaR). The difference between the CreditVaR and the expected loss has to be covered by the economic capital, a scarce resource of each bank. From a risk management perspective, the definition of industry sectors allows to diversify credit risk. The degree to which this diversification is successful depends on the strength of correlations between the sectors. Moreover, the correlations between sector PDs crucially influence the CreditVaR and hence the economic capital.

In large banks, the concentration risk in industry sectors is a key risk driver. In CreditRisk+ [1], concentration risk is modelled as a multiplicative random effect on the PD per counterpart in a given sector. In the original version of CreditRisk+, the loss distribution is calculated for independent sector variables. Correlations between PD fluctuations in different sectors can be integrated into CreditRisk+ with the method of Bürgisser et al.

* The first two authors have contributed equally.

[2]. For the calculation of the CreditVaR it is important whether input parameters like the correlation coefficients between sector PDs are known or must be estimated. In the latter case, this estimation leads to an additional variability of the target estimate, in our case the portfolio loss. In this way, uncertainty in the estimation of PD correlations translates itself into uncertainty of the economic capital of a bank.

The estimation of cross–correlations is difficult due to the "curse of dimensionality": if the length T of the available time series is comparable to the number K of industry sectors, the number of estimated correlation coefficients is of the same order as the number of input parameters with the result of large estimation errors. A way out of this dilemma is the use of a factor model with a reduced dimensionality of the parameter space. We present evidence that the PD correlations for $K = 20$ industry sectors are well captured by a one–factor model. Surprisingly, even the parameter estimation for the one–factor model is subject to large statistical fluctuations and gives rise to a considerable uncertainty in the CreditVaR. We discuss these fluctuations in detail and suggest a bootstrap method which allows to find an upper limit for the parameters. We assess the impact of different conservative estimates with respect to the CreditVaR of a realistic portfolio.

Description of data set

As the economic activity and the probability of default in a given industry sector is not directly observable, we approximate it by the insolvency rate in that sector over the last T years. The probability of insolvency PD_{kt} of sector k in year t is calculated as the ratio of the number of insolvencies in that sector to the total number of companies in the sector

$$\hat{PD}_{kt} = \frac{\sum_{A \in \text{ sector k in year t}} I_{\{A \text{ fails}\}}}{\sum_{A \in \text{ sector k in year t}}} \quad . \tag{1}$$

With the help of insolvency rates, the default probability for a given company A can be factorized into an individual expected PD p_A and the sector specific relative PD movement X_k with expectation $\langle X_k \rangle = 1$ according to

$$P(A \text{ fails}) = \mathrm{p_A X_k} \quad \text{with} \quad \mathrm{X_{kt}} = \frac{\hat{\mathrm{PD}}_{\mathrm{kt}}}{\frac{1}{\mathrm{T}} \sum_{\mathrm{t}} \hat{\mathrm{PD}}_{\mathrm{kt}}} \quad . \tag{2}$$

For this study, we use sector specific default histories as supplied by the federal statistical office of Germany. We analyze default rates for a segmentation of the economy into 20 sectors and estimate the sample covariance matrix Σ^{emp} and sample correlation matrix $\mathsf{C}^{\mathrm{emp}}$

as

$$\Sigma_{ij}^{\rm emp} = \frac{1}{T-1}\sum_{t=1}^{T}(X_{it}-1)(X_{jt}-1),\ C_{ij}^{\rm emp} = \Sigma_{ij}^{\rm emp}/\sigma_{X_i}\sigma_{X_j} \tag{3}$$

with $\sigma_{X_i}^2 = \Sigma_{ii}^{\rm emp}$.

Test for independent sectors

We first ask whether the sample correlation matrix of the PD time series is compatible with the hypothesis of zero correlations. Ideas for testing this hypothesis for covariance matrices date back to the seventies [3], and were recently generalized to situations where the number of time series is larger than the sample size [4]. Here, we use an adaption of the tests [3, 4] to test for the equivalence of correlation matrix to the unit matrix. The test statistics

$$\tilde{R} = \frac{1}{K}{\rm tr}\left[\mathsf{C}^2\right] - 1\ , \tag{4}$$

for a correlation matrix C is both K– and T–consistent with the T–limiting distribution $R := (T-1)K\tilde{R}/2 \stackrel{D}{\rightarrow} \chi^2_{K(K-1)/2}$ [5]. The factor $T-1$ rather than T is chosen to improve the finite T properties of the test. For our example with $T = 7$ and $K = 20$, we find $R = 348.31$, whereas the critical value for $\alpha = 0.05$ is $R_{\rm crit} = 223.16$. Hence, the independence of sector PDs must not be assumed and a model describing sector correlations is needed.

Description of one–factor model

We diagonalize the empirical cross correlation matrix $\mathsf{C}^{\rm emp}$ and rank order its eigenvalues $\lambda_{i,{\rm emp}} < \lambda_{i+1,{\rm emp}}$. As we are interested in modelling correlations rather than covariances, we normalize the X_{it} such that they have the same, namely the average variance $\sigma_X^2 = (1/K)\sum_{i=1}^{K}\sigma_{X_i}^2$ and subtract the mean

$$\tilde{X}_{it} = (X_{it}-1)\frac{\sigma_X}{\sigma_{X_i}}\ . \tag{5}$$

We use the components of the eigenvector $\mathbf{u}_{\rm emp}^{(K)}$ corresponding to the largest eigenvalue $\lambda_{K,{\rm emp}} = 10.38$ to define a factor time series

$$Y_t = \sum_{i=1}^{K} u_{i,{\rm emp}}^{(K)}\tilde{X}_{it}\ \ . \tag{6}$$

As compared to averaging the sector variables without prior information, the definition of the factor time series from the eigenvector with largest eigenvalue makes sure that the factor

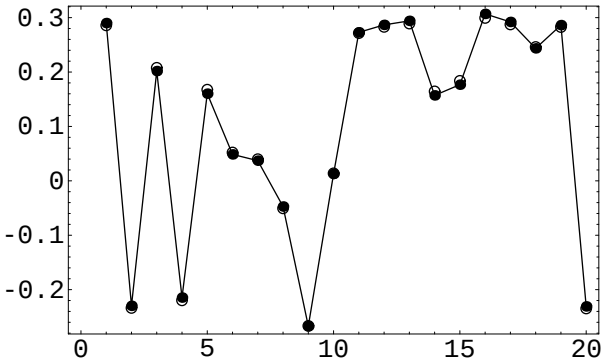

FIG. 1: The components of the eigenvector $\mathbf{u}^{(K)}_{\rm emp}$ of the empirical correlation matrix (connected full circles) are almost identical to the components of the eigenvector $\mathbf{u}^{(K)}_{\rm point}$ of the point estimator $\mathsf{C}^{\rm point}$ (open circles) .

explains a maximum amount of correlations. The idea arises from factor analysis, see e.g. [6]. In the context of stock returns, a time series defined according to the prescription of Eq. 6 was found to agree well with a value weighted stock index [7]. We expect that the factor time series Eq. 6 describes economy wide changes of relative PD, possibly weighted by the economic relevance of the individual sectors.

We model the correlations between relative PD movements by a one–factor model

$$\tilde{X}_{it} = b_i Y_t + \epsilon_{it} \quad . \tag{7}$$

The coefficients $\{b_i\}$ are found by performing a linear regression. To see whether a one–factor model fully describes the correlations between the $\{\tilde{X}_{it}\}$, we apply the test Eq. 4 to the correlation matrix of the residuals $\{\epsilon_{it}\}$. Taking into account that the regression reduces the effective length of the residual time series from T to $T-1$, we find $R = 220.46$ slightly below the threshold $R_{\rm crit} = 223.16$. As the assumption of uncorrelated residuals is not rejected, no further factors are needed for the description of correlations.

The point estimator can now be calculated under the assumption that the residua $\{\epsilon_{i,t}\}$ are iid observations from uncorrelated random variables ϵ_i $i = 1, \ldots, K$, i.e. $\langle \epsilon_i \epsilon_j \rangle = 0$ for $i \neq j$. Defining the factor variance $\sigma_Y^2 = \frac{1}{T-1} \sum_{t=1}^{T} Y_t^2$, one finds the point estimator for the cross correlation matrix as

$$C_{ij}^{\rm point} = \delta_{ij} + (1 - \delta_{ij}) b_i b_j \sigma_Y^2 / \sigma_X^2 \quad . \tag{8}$$

The largest eigenvalue of $\mathsf{C}^{\rm point}$ is found to be $\lambda_{K,{\rm point}} = 10.66$ in good agreement with the

original largest eigenvalue. In addition, the corresponding eigenvector $\mathbf{u}^{(K)}_{\rm point}$ is found to be very close to the original eigenvector (Fig.1).

Fluctuations in empirical correlation matrices – a toy model

In this section, we use the results of Monte Carlo simulations to study the relation between the cross correlation matrix $\mathsf{C}^{\rm model}$ resulting from infinitely long model time series and matrices $\mathsf{C}^{\rm sim}$ numerically calculated from finite time series of length T. We find that the $\{\mathsf{C}^{\rm sim}\}$ differ from $\mathsf{C}^{\rm model}$ both in a systematic way, for example a shift of the largest eigenvalue towards larger values, and a random way, i.e. an individual member of the simulated ensemble deviates significantly from the average [8, 9].

To simplify simulations, we rewrite Eq. 7 as

$$\tilde{X}_{it} = \alpha\beta_i F_t + c_i\eta_{it} \quad . \tag{9}$$

The random variables are rescaled according to $c_i\eta_{it} = \epsilon_{it}$ and $\alpha\beta_i F_t = b_i Y_t$ such that their variances are $\mathrm{var}(F) = \sigma_X^2$ and $\mathrm{cov}(\eta_i\eta_j) = \sigma_X^2\delta_{ij}$. In addition, the $\{\beta_i\}$ obey the normalization condition $\sum_{i=1}^K \beta_i^2 = 1$, which makes them comparable to the components of the eigenvector $\mathbf{u}^{(K)}$. The model parameters are subject to the constraint $c_i^2 = 1 - \alpha^2\beta_i^2$ in order to enforce $\mathrm{var}(\tilde{X}_{it}) = \sigma_X^2$.

The model is completely defined by i) the parameter α determining the largest eigenvalue, ii) the parameters $\{\beta_i\}$ and iii) by the probability distribution of the random variables F and $\{\eta_i\}$. As the empirical PD movements are commonly assumed to follow a gamma distribution, we model the random variables F and η_i by gamma distributions as well [10]. In addition, we use normal distributions for the random variables and find that the deviation from a simulation with gamma distributed variables is smaller than 3%. As the simulation of gaussian random variables is computationally much more efficient than the simulation of gamma distributed variables, we use normally distributed variables for the computationally demanding selfconsistent calculations described in the next section.

The infinite time series correlation matrix of the model is given by

$$C_{ij}^{\rm model} = \delta_{ij} + (1-\delta_{ij})\alpha^2\beta_i\beta_j \ . \tag{10}$$

In this section, we study the outcome of model simulations for the particularly simple hypothetical case $u^{(K)}_{i,\rm model} = \beta_i \equiv 1/\sqrt{K}$ in order to gain qualitative insight into the occurring

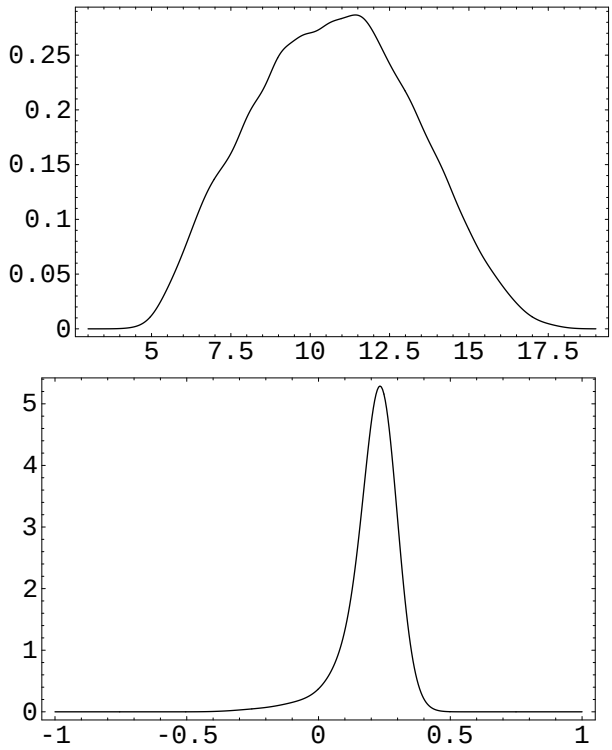

FIG. 2: Distribution of the largest eigenvalue and of all components of the corresponding eigenvector from simulations of the one–factor model with $\lambda_{K,\mathrm{model}} = 10.38$.

fluctuations. For a given value of α corresponding to $\lambda_{K,\mathrm{model}} = 10.38$, we perform a Monte Carlo simulation of Eq. 9. We compute the pdf for the largest eigenvalue $\lambda_{K,\mathrm{sim}}$ of $\mathsf{C}^{\mathrm{sim}}$ and the corresponding eigenvector $\mathbf{u}^{(K)}_{\mathrm{sim}}$. We find that both quantities have broad distributions (see Fig. 2). The distribution of eigenvalues has an average $\langle \lambda_{K,\mathrm{sim}} \rangle = 10.72$, which is significantly larger than the true eigenvalue $\lambda_{K,\mathrm{model}} = 10.38$. In addition, one finds simulated eigenvalues as low as $\lambda_{K,\mathrm{sim}} = 5$. We quantify the systematic shift of eigenvalues by the average $\Delta\lambda = \langle \lambda_{K,\mathrm{sim}} \rangle - \lambda_{K,\mathrm{model}}$. The magnitude of eigenvalue fluctuations is described by the standard deviation

$$\sigma_\lambda = \sqrt{\langle \lambda^2_{K,\mathrm{sim}} \rangle - \langle \lambda_{K,\mathrm{sim}} \rangle^2} \quad . \tag{11}$$

For the distribution shown in Fig. 2 we find $\sigma_\lambda = 2.42$.

There are significant fluctuations of eigenvector components as well. For theoretical eigenvector components $u^{(K)}_{i\mathrm{model}} = 0.224 \ \forall \ i$ one even finds negative empirical components indicating spurious anticorrelations, which would lead to dangerous hedges in credit portfo-

lios. Specifically, we calculate the standard deviation

$$\sigma_{u_i} = \sqrt{\langle (u_{i,\mathrm{sim}}^{(K)})^2 \rangle - \langle u_{i,\mathrm{sim}}^{(K)} \rangle^2} \tag{12}$$

and find $\sigma_{u_i} = 0.083$. Since the $u_{i,\mathrm{model}}^{(K)}$ do not vary across i we only need to estimate one σ_{u_i}.

As a conclusion, even if the generating process for relative PD movements is a simple one–factor model, the empirically found parameters can deviate significantly from the theoretical ones. We advocate the point of view that the empirical $\mathsf{C}^{\mathrm{emp}}$ has to be viewed as a member of such a fluctuating ensemble in that its eigenvalues and eigenvectors can deviate significantly from the unknown "true" correlation matrix of PD movements [8, 9]. Then, the statistical properties of the ensemble $\{\mathsf{C}^{\mathrm{sim}}\}$ can be used to derive error bars for both the largest eigenvalue and the components of the corresponding eigenvector.

Conservative estimates

How can we use these results to make a reliable estimate for the correlation matrix of relative PD movements? A bank needs to act in a conservative manner to prevent insolvency. Using the empirical correlation matrix, the bank risks that the correlations are "accidently" low. The most conservative approach would be to assume all correlations to be 1, i.e. $u_i^{(K)} = \frac{1}{\sqrt{K}}\ \forall\ i$. But now the model would effectively one-sector model. Any possibility to measure concentration risk in certain industry sectors would be prevented. The model would not enforce diversifying the business across sectors.

As a controlled mediation we introduce "cases" of add-ons of $x = 1, 2, 3$ standard deviations to the fluctuating quantities such that the predicted risk for a portfolio is increased. This means correcting the eigenvalue towards larger values and the eigenvector components towards the value $u_i^{(K)} \equiv 1/\sqrt{K}$ indicating the same correlation strength for all sectors and the absence of hedge possibilities.

Specifically, we let $u_{i,\mathrm{case}}^{(K)} = 1/\sqrt{K}$ if

$$|u_{i,\mathrm{emp}}^{(K)} - 1/\sqrt{K}| < x \cdot \sigma_{u_i} \tag{13}$$

and $u_{i,\mathrm{case}}^{(K)} = u_{i,\mathrm{emp}}^{(K)} \pm x \cdot \sigma_{u_i}$ otherwise. The sign is chosen such that the overall risk increases, i.e. such that $u_{i,\mathrm{case}}^{(K)}$ falls between the empirical value and $1/\sqrt{K}$. After applying these corrections, the eigenvector is normalized. For this calculation, we fix the parameter α such

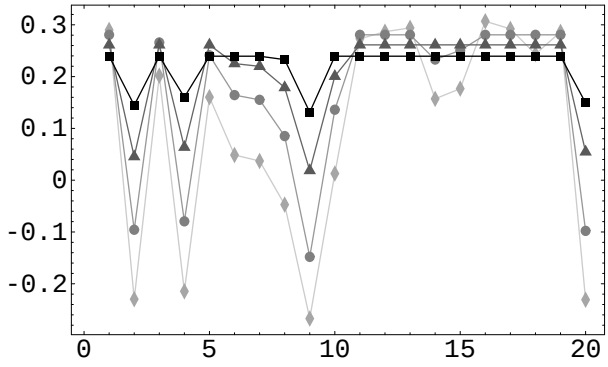

FIG. 3: Comparison between the empirical eigenvector $\mathbf{u}^{(K)}_{\rm emp}$ (diamonds) and the conservative estimates $\mathbf{u}^{(K)}_{1\sigma}$ (circles), $\mathbf{u}^{(K)}_{2\sigma}$ (triangles), and $\mathbf{u}^{(K)}_{3\sigma}$ (squares).

that the simulated largest eigenvalue $\lambda_{K,\rm sim}$ is equal to the empirically observed one. We calculate the $\{\sigma_{u_i}\}$ selfconsistently, i.e. we calculate σ_{u_i} for the $u^{(K)}_{i,\rm case}$ which solves Eq.13. The results are shown in Fig.3. We see that for increasing $x = 1, 2, 3$, the model eigenvector comes closer to the null hypothesis of an eigenvector with identical components. While the empirical eigenvector has significant negative components indicating anticorrelations between some of the sectors, the negative components in $\mathbf{u}^{(K)}_{1\sigma}$ are already strongly reduced and completely gone in $\mathbf{u}^{(K)}_{2\sigma}$.

Similarly, we add a fluctuation margin to the model eigenvalue [11] such that

$$\lambda_{K,\rm case} = \lambda_{K,\rm emp} + x \cdot \sigma_\lambda \quad . \tag{14}$$

Here, x specifies the width of the confidence interval for the estimation of $\lambda_{K,\rm model}$. We perform this calculation selfconsistently, i.e. we calculate $\Delta\lambda$ and σ_λ for the $\lambda_{K,\rm case}$ which solves Eq.14. We find $\lambda_{K,1\sigma} = 11.17$, $\lambda_{K,2\sigma} = 12.93$, and $\lambda_{K,3\sigma} = 15.42$.

Economic implications of the different correlation matrices

So far, we have described five different estimates for the cross correlation matrix, i.e. $\mathsf{C}^{\rm emp}$, $\mathsf{C}^{\rm point}$, $\mathsf{C}^{\rm model}_{1\sigma}$, $\mathsf{C}^{\rm model}_{2\sigma}$, and $\mathsf{C}^{\rm model}_{3\sigma}$. To judge the economic implications of these estimates, we study the differences in the loss distribution resulting from these correlation estimations. The CreditVaR is a key quantity in banking when it comes to risk management. Reduced by the expected loss, it quantifies the capital needed to prevent insolvency for a given level of security. As capital is a resource it must be considered in the pricing of credit and trading products. Therefor, we quantify the impact of the different correlation estimates by

correlation matrix	CreditVaR (in billion Euro)
C	2.872
$\mathsf{C}^{\text{point}}$	2.825
$\mathsf{C}^{\text{model}}_{1\sigma}$	3.172
$\mathsf{C}^{\text{model}}_{2\sigma}$	3.465
$\mathsf{C}^{\text{model}}_{3\sigma}$	3.665

TABLE I: Analysis of CreditVaR for different correlation matrices

calculating their influence on CreditVaR.

The portfolio we study is realistic – although fictitious – for an international bank. It consists of 4934 risk units distributed asymmetrically over 20 sectors with 20 to 500 counterparts per sector. The total exposure is 70 bn Euro with a largest exposure of 1.5 bn Euro and a smallest exposure of 0.25 mn Euro. The counterpart specific default probability varies between 0.03% and 7%, the expected loss for the total portfolio is 373.3 mn Euro. Table I shows the CreditVaR calculated by using CreditRisk+ and the method of Bürgisser et al. [2] for integrating correlations.

We note that the use of a one–factor model changes the CreditVaR only by two percent as compared to the sample cross correlation matrix. Thus, the assumption of a one–factor description and the increase of estimation confidence achieved with this assumption yields portfolio risk estimation compatible with the parameter free estimation.

Our aim is to estimate a quantile of a probability distribution – namely the CreditVaR of the portfolio loss distribution. In the presence of an unknown parameter, it is a well established statistical result (see [12]) that the use of the point estimate for the parameter – derived by a model or not – leads to an underestimation of the quantile estimate. To account for this additional estimation insecurity, we add a volatility σ to the parameter estimate, i.e. the correlation matrix. When applying a one–σ estimate, the CreditVaR increases by 400 mn Euro, for the two–σ estimate there is another increase by 300 mn Euro, and using the three–σ estimate the CreditVaR increases by yet another 200 mn Euro. To put these numbers in perspective, we note that the CreditVaR without including correlations is found to be 2.27 bn Euro, and that the assumption of full correlations among all sectors leads to a CreditVaR of 3.952 bn Euro. Because negative PD correlations are not plausible from

an economic point of view, the use of the two–σ estimate guarantees a sufficient forecast reliability on the one hand and allows for some guidance for economical decision on the other hand.

In summary, we have shown that correlations between empirical default rates for economic sectors are statistically significant and must be taken into account. We have described these correlations with a one–factor model and found that this description reproduces well the empirical correlations. However, when using the model to generate short time series and calculating their correlation matrix, one typically observes large statistical fluctuations in the correlation structure. Due to these fluctuations, the parameter estimation for a one–factor model is plagued by large uncertainties. When estimating the model parameters in such a way that the empirically observed ones appear as a worst case scenario, the reliability of the estimate is increased in a systematic way, leading to a moderately increased CreditVaR.

Acknowledgement: We would like to thank A. Müller-Groeling for initiating this project, and J.-H. Schmidt, A. Wilch, and C. von Lieres und Wilkau for useful discussions. The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

---

[1] Credit Suisse First Boston (CSFB) : Credit Risk +: A Credit Risk Management Framework, *Technical document*, 1997.

[2] Bürgisser, P., A. Kurth , A. Wagner, and M. Wolf, *Integrating Correlations*, Risk, 07/1999.

[3] John, C.,*Some optimal multivariate tests*, Biometrika **58**, 123-127 (1971).

[4] Ledoit, O. and M. Wolf, *Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size*, Annals of Statistics **30**, 1081-1102 (2002).

[5] B. Rosenow, to be published. For correlation matrices, the test statistics R is equivalent to the statistics W studied in [4]. R has a limiting distribution $\chi^2_{K(K-1)}$ as compared to the limiting distribution $\chi^2_{K(K+1)}$ for W as the diagonal elements of a correlation matrix are fixed and do not fluctuate.

[6] Mardia, K.V., Kent, J.T., Bibby, J.M., *Multivariate Analysis*, Academic Press, 2000.

[7] P. Gopikrishnan, B. Rosenow, V. Plerou, and H.E. Stanley, *Quantifying and interpreting collective behavior in financial markets*, Phys. Rev. E **64**, 035106(R) (2001).

[8] Laloux, L., P. Cizeau, J.-P. Bouchaud, and M. Potters, *Random Matrix Theory*, Risk Magazine **12**, 69 (1999); see also L. Laloux et al., Phys. Rev. Lett. **83**, 1467 (1999).

[9] Plerou, V., P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, and H.E. Stanley, *Universal and non–universal properties of cross–correlations in financial time series*, Phys. Rev. Lett. **83**, 1471 (1999).

[10] As gamma distributed variables have nonzero mean, the model must be rewritten as $\tilde{X}_{it}+1 = \alpha\beta_i F_t + c_i\eta_{it}$, i.e. variables $\tilde{X}_{it}+1$ with the original expectation value one are modelled. The expectation values of the random variables are $\langle F\rangle = 1/(\alpha\frac{1}{K}\sum_{i=1}^{K}\beta_i + \frac{1}{K}\sum_{i=1}^{K}c_i)$ and $\langle\eta_i\rangle = (1-\alpha\beta_i\langle F\rangle)/c_i$ such that $\langle\tilde{X}_i\rangle = 0$ for all i.

[11] For the numerical calculation, we also account for the systematic eigenvalue shift $\Delta\lambda$.

[12] Lehmann, E.L., *Testing statistical hypotheses*, Chapmann & Hall (1993).