Modelling correlations in credit portfolio risk

Received (in revised form): 20th June, 2009

Bernd Rosenow*
received a PhD in physics from Würzburg University in 1998 and joined the Physics Department at Cologne University as a research associate in the same year. He obtained an assistant professorship at Cologne University in 2002. Since 2005, he has been a Heisenberg fellow of the German Research Council, conducting theoretical physics research at Harvard University and the Max-Planck Institute Stuttgart. His research focuses on the analysis of correlations in financial markets, multivariate volatility models, and market microstructure. He uses the theory of random matrices and other methods of mathematical physics to study these problems. His work has been published in international journals including *Quantitative Finance* and the *Journal of Economic Dynamics & Control.*

Rafael Weissbach
has a diploma in mathematics from Göttingen University and a PhD in statistics from the University of Dortmund. In 2004, he joined the University of Dortmund's Faculty of Statistics as a research assistant and was promoted to assistant professor for econometrics in 2007. Rafael has also acted as chair for econometrics at the University of Mannheim's Faculty of Economics. From 2001 to 2004, he worked as a risk analyst and portfolio manager in the credit risk management division of an international investment bank. His current interest is statistics in finance, especially estimation and the modelling of credit risk related parameters such as rating migration matrices and default correlations. His work has been published in international journals including the *Journal of the American Statistical Association.* He is currently associate professor for statistics at Rostock University.

*Physics Department, Harvard University, Cambridge, MA 02138, USA and Max-Planck Institute for Solid State Research, 70569 Stuttgart, Germany
Tel: +49 711 689 1514; Fax: +49 711 689 1702; E-mail: rosenow@fkf.mpg.de

Abstract A credit portfolio's risk level depends on correlations between latent covariates, such as the probability of default in different economic sectors. Correlations often have to be estimated from relatively short time series, and the resulting estimation error hinders the detection of a signal. This paper suggests a general method of parameter estimation which avoids, in a controlled way, the underestimation of correlation risk. The paper presents empirical evidence to show how, in the framework of the CreditRisk+ model with integrated correlations, this method leads to an increased economic capital estimate. In this way, the limits of detecting the portfolio's diversification potential are adequately reflected.

Keywords: *credit risk, portfolio risk, estimation risk, correlation matrix, random matrix, Bessel function*
JEL Classification: *C46, C15, C53, G32, G33*

INTRODUCTION

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 credit risk to be diversified. The degree to which this diversification is successful depends on the strength of correlations between the sectors. Moreover, the correlations between sector probabilities of default (PDs) crucially influence the CreditVaR and hence the economic capital.

In large banks, the concentration risk in industry sectors is a key risk driver. Recently, several approaches for describing and modelling concentration risk have been discussed.[1–4] In CreditRisk+,[5] 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+ according to the method of Bürgisser *et al.*[1] To calculate the CreditVaR, one must know or at least estimate input parameters such as the correlation coefficients between sector PDs. Estimation does however increase variability in the target estimate — in the present case, the portfolio loss. In this way, uncertainty in the estimation of PD correlations translates into uncertainty regarding the economic capital of a bank. The influence of estimation risk on portfolio performance is well known,[6,7] but usually studied in terms of optimal portfolio choice.

The estimation of cross-correlations is difficult due to the 'curse of dimensionality': if the length of the available time series (T) is comparable to the number of industry sectors (K), 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 minimal model with a reduced dimensionality of the parameter space. A reasonable choice is a parsimonious model with the global default rate as the latent factor.[8]

Despite the fact that the parameter space of a single-factor model has considerably lower dimensionality than that of the full correlation matrix, there are large statistical fluctuations in the parameter estimation resulting in a considerable uncertainty in the CreditVaR based on such a model. This paper discusses these fluctuations in detail and suggests a bootstrap method to find a level for the parameters that reflects the applicable risk aversion of the individual bank. The study exemplifies the impact of different conservative estimates on the CreditVaR and the expected shortfall of a realistic portfolio.

METHODOLOGY

As the economic activity and consequent probability of default in a given industry sector is not directly observable, it can be approximated by the insolvency rate in that sector over the last $T+1$ years. The probability of insolvency PD_{it} of sector i 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:

$$\hat{P}D_{it} = \frac{\sum_{A\in\text{sector i in year t}} I_{\{A\,\text{fails}\}}}{\sum_{A\in\text{sector i in year t}}}. \quad (1)$$

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

$$P(A\text{ fails}) = \mathrm{p_A X_i}. \quad (2)$$

When using CreditRisk+ with a time horizon of one year, what is of interest is the relative change of default probabilities.[9]

The individual PD, p_A in Equation 2, is usually taken to depend on the current economic activity at time $t-1$, ie it describes a point-in-time rating. For comments on estimating the p_A see Weissbach and Dette[10] and Weissbach *et al.*[11] and the references therein. The PD for the forthcoming period t is hence the product of the current individual p_A (depending on information available at time $t-1$) and the ratio of the future PD at time-step t and the current PD. The latter ratio is the relative change in economic activity,

$$X_{it} = \frac{\hat{P}D_{it+1}}{\hat{P}D_{it}} + 1 - \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{P}D_{it+1}}{\hat{P}D_{it}}, \quad (3)$$

which is normalised to $\langle X_i \rangle = 1$ in the above definition. As the correlations between relative PD movements in different sectors crucially influence the risk of a credit portfolio, it is important to estimate them in a reliable way.

CORRELATION ESTIMATE FROM EMPIRICAL DATA

To illustrate the given theoretical concept, the study uses sector-specific default histories for a segmentation of the German economy into $K = 20$ sectors selected by the authors. The study takes the viewpoint of a portfolio owner whose counterparts are to a large extent located in Germany. The environment for the remaining counterparts is assumed to be similar. The data — for a much finer segmentation[12] — are supplied by the Federal Statistical Office of Germany and date from 1994 to 2000.[13] In addition, the paper will analyse the insolvency rate for the whole German economy from 1962 to 2003 (that is, West Germany until 1994, and Germany as a whole subsequently).

To obtain credible correlation estimates from these data requires information concerning their stationarity. The use of relative changes according to Equation (3) eliminates the trend. For a visual assessment of stationarity, this paper displays the time series Y_t of the default rate growth factor for the national economy in Figure 1, and the default rate growth factors X_{it} of sectors in Figure 2. All time series appear to be stationary. With respect to inferential statistics, the study only analyses the 40-year time series Y_t as the length of the sector time series is insufficient for this analysis. The study tests for the hypothesis of a unit root indicating in-stationarity with the Dickey-Fuller test.[14] The p-value of 0.00079 (using SAS macro 'dftest') enables one to reject the AR(1) hypotheses and safely work under the stationarity assumption. The decision does not change (at level $\alpha = 5$ per cent) for larger models, ie lags $p = 2$ to 5. In addition to testing the

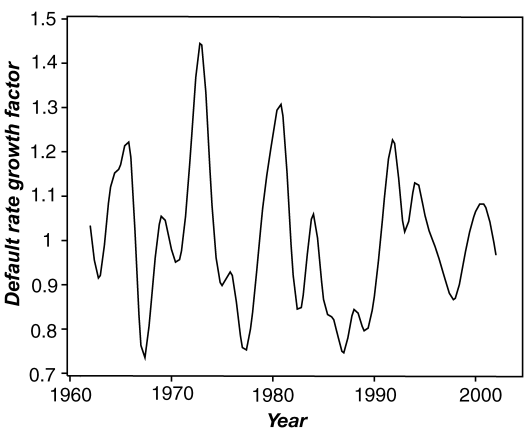

Figure 1: Default rate growth factor for the German economy from 1962 to 2003 (West Germany until 1994)

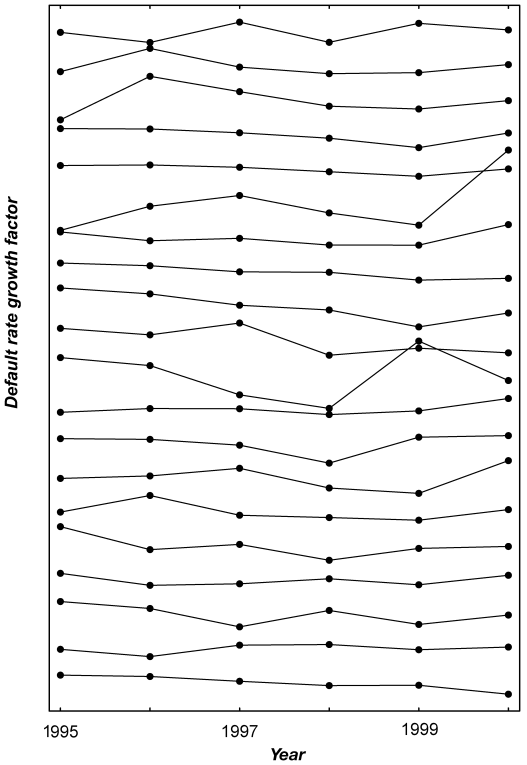

Figure 2: Default rate growth factors of $K = 20$ sectors from 1995 to 2000; for clarity, subsequent curves have an offset of 0.5 against each other

stationarity of the time series Y_t, one must assess the stationarity of its variance. The Lagrange-multiplier test for ARCH-effects of the volatility does not reject (for any order $0.5 < p\text{-}value < 0.8$). For this reason, the volatility may be considered as constant, in other words, the series is mean-variance stationary.

In view of the small sample size, the study uses a parsimonious single-factor model for the estimation of cross-correlations. As a factor, the study uses the relative changes Y_t of the national insolvency rate. The definition of Y_t is analogous to the definition of the X_{it} in Equation (3). The study decomposes the sector PDs according to

$$X_{it} = Y_t + \epsilon_{it}, \qquad (4)$$

where the residuals ϵ_{it} are defined by this equation. The economic reasoning behind this decomposition is that the study does not allow sectors' fortunes to be systematically linked other than via the single factor. Moreover, sectors are not differentiated according to the intensity of their relationship to the single factor. This has two major advantages: (i) one needs to estimate only $K + 1$ parameters as compared with the $2K + 1$ parameters for a standard single-factor model, and (ii) the factor variance can be reliably estimated over a long time interval spanning several economic cycles, as no sector-specific data are required.

As a consequence, the correlation between the systematic parts of the sectors' default rates now is uniformly equal to one. However, this systematic correlation is obscured by the residuals. As Equation (4) realises a variance decomposition, it creates a relation between the correlations and volatilities. For reasons of tractability, the fundamental assumption is made that the residuals are uncorrelated among each other and uncorrelated with the factor. One then obtains the correlation matrix $\mathbf{C}^{\text{var}}$ with the following elements:

$$C_{ij}^{\text{var}} = \delta_{ij} + (1 - \delta_{ij}) \times \frac{1}{\sqrt{1 + \sigma_{\epsilon_i}^2/\sigma_Y^2}\sqrt{1 + \sigma_{\epsilon_j}^2/\sigma_Y^2}}. \qquad (5)$$

$\mathbf{C}^{\text{var}}$ has an intuitive interpretation: according to Equation (4), the sector variance is decomposed into the factor variance and the residual variance. The smaller the influence of the factor on a given sector, the larger the residual variance of this sector, and according to Equation (5), the correlation coefficients

between this sector and other sectors become small. $\mathbf{C}^{\mathrm{var}}$ is a conservative and robust input for business applications. This is because the neglect of (negative) covariances between factor and residuals tends to result in an overestimation of correlations.

By applying the standard maximum likelihood estimator for the variances, the canonical correlation estimate $\mathbf{C}^{\mathrm{var}}_{\mathrm{canonical}}$ is generated. However, applying this estimation procedure for the variances leads to some non-desirable properties of the correlation estimate and produces a bias in further applications. This issue becomes relevant for small sample sizes and is investigated in a controlled environment in the next section. More precisely, the factor volatility σ_Y is estimated during the period 1962–2003, and the residual volatilities σ_{ϵ_i} are estimated over the time interval 1994–2000.

FLUCTUATIONS IN EMPIRICAL CORRELATION MATRICES: A SIMULATION STUDY

This section uses the results of Monte Carlo simulations to study the relation between the true cross-correlation matrix C and matrices $\mathbf{C}^{\mathrm{sim}}$ estimated from time series of length T. One finds that $\{\mathrm{C}^{\mathrm{sim}}\}$ differs from C in both a systematic way (eg a shift of the largest eigenvalue towards larger values), and a random way (ie an individual member of the simulated ensemble deviates significantly from the average).[14–16]

Assuming that the process Equation (5) with mutually independent time series Y_t, ϵ_{it} and ϵ_{jt} is valid, uncertainties in the determination of $\mathrm{C}^{\mathrm{var}}_{\mathrm{canonical}}$ arise from uncertainties in the estimation of σ_Y and σ_{ϵ_i}. For the simulations, normality of ϵ_{it} is assumed due to the increased computational efficiency as compared with other distributional assumptions. This gain in efficiency is especially important for the computationally demanding iterative calculations described in the next section. The different distributional assumptions have little influence on the result of simulations. For instance, the deviation between a simulation with normal distributed variables and a simulation with gamma distributed variables is smaller than 3 per cent for the standard deviations defined in Equations (8) and (9).

As σ_Y is calculated from a long time series including more than 40 years of data, its estimation error is negligible in comparison with that of the σ_{ϵ_i}. The ratio $\hat{\sigma}^2_{\epsilon_i}/\sigma^2_Y$ follows a chi-square distribution with $T-1$ degrees of freedom

$$f_i(z) = f_{\chi^2,T-1}\left(\frac{T-1}{\mu_i}z\right)\frac{T-1}{\mu_i}, \quad (6)$$

where $f_{\chi^2,n}$ is the density function of the chi-square distribution with n degrees of freedom, and unknown $\mu_i = \sigma^2_{\epsilon_i}/\sigma^2_Y$. As a consequence, one obtains $\mathrm{Var}(\hat{\sigma}^2_{\epsilon_i}/\sigma^2_Y) = 2\mu_i^2/(T-1)$. In the limit $T\to\infty$, statistical fluctuations disappear.

This section studies a situation in which the parameters $\{\mu_i\}$ determining the distributions of Equation (6) are known. The true correlation matrix for the Equation (4) is then:

$$C^{\mathrm{model}}_{ij} = \delta_{ij} + (1-\delta_{ij}) \times \frac{1}{\sqrt{1+\mu_i}\sqrt{1+\mu_j}}. \quad (7)$$

We now fix σ^2_Y and simulate time series $\{\epsilon_{it}\}$ with mean zero and variance $\mu_i\sigma^2_Y$,

calculate the variances $\sigma^2_{\epsilon_i}$ using the standard variance estimator, and use these estimated variances to calculate a correlation matrix $\mathbf{C}^{\text{sim}}$ for each simulation run according to Equation (5). The study now asks (i) whether the $\mathbf{C}^{\text{sim}}$ estimates are biased compared with the true correlation matrix $\mathbf{C}^{\text{model}}$, and (ii) how large are fluctuations from one simulation run to the next.

To answer these questions, it is useful to recall that most of the relevant information in the $\{C^{\text{sim}}\}$ is contained in the largest eigenvalue and the corresponding eigenvector. As the true correlation matrix C^{model}_{ij} is completely characterised by K parameters, it is clear that it can be fully described by one eigenvector and eigenvalue. In the spirit of a spectral decomposition approximation, the largest eigenvalue and corresponding eigenvector are most informative as they describe a larger part of the total correlation than any other eigenvalue/eigenvector. This point of view is corroborated by results from random matrix theory, which show that, due to the estimation uncertainty of the input parameters of the correlation matrix, the small eigenvalues contain no information at all.[15–17] Minimising information loss is formalised in principal component analysis by maximising the variance of a (normalised) linear combination of X. The variance is the largest eigenvalue and the linear weights are its eigenvector components.[18] This procedure also applies to the correlation matrix, because for given volatilities, the covariance matrix of the standardised observations is the correlation matrix.

As a specific choice for the model simulations, the paper considers the particularly simple hypothetical case where signal Y and noise ϵ_i have the same volatility, ie $\mu_i \equiv 1$, in order to gain qualitative insight into the occurring fluctuations. The corresponding infinite time series correlation matrix $C^{\text{model}}_{ij} = \delta_{ij} - (1 - \delta_{ij})/2$ has a largest eigenvalue $\lambda_{K,\text{model}} = 10.5$ and a corresponding eigenvector $u^{(K)}_{i,\text{model}} \equiv 1/\sqrt{K}$. The simulation is performed for $K = 20$ and $T = 7$.

Instead of simulating the time series $\{\epsilon_{it}\}$ and estimating their variance, it is recalled that, for normally distributed $\{\epsilon_{it}\}$, the variance estimator follows a chi-square distribution. If in addition σ^2_Y is known, then the ratios $\{\sigma^2_{\epsilon_i}/\sigma^2_Y\}$ are indeed distributed according to Equation (6). Hence, for each of the 500,000 simulation runs, the study (i) draws a set of $K = 20$ values for the ratios $\{\sigma^2_{\epsilon_i}/\sigma^2_Y\}$ from the chi-square distribution defined by Equation (6) with parameters $T = 7$ and $\mu_i = 1$; (ii) uses them to calculate a matrix $\mathbf{C}^{\text{sim}}$ according to Equation (5); and (iii) calculates the largest eigenvalue $\lambda_{K,\text{sim}}$ and the corresponding eigenvector $\mathbf{u}^{(K)}_{\text{sim}}$ from this matrix. Averaging over all simulation runs, one finally obtains the pdf for both quantities.

One finds that both eigenvalue and eigenvector components have broad distributions (see Figure 3). The distribution of eigenvalues has an average $\langle\lambda_{K,\text{sim}}\rangle = 11.314$ which is significantly larger than the true eigenvalue $\lambda_{K,\text{model}}$=10.5. Hence, the above described procedure for estimating the correlation matrix is indeed biased towards larger eigenvalues. The systematic shift of eigenvalues is quantified by the difference $\Delta\lambda = \langle\lambda_{K,\text{sim}}\rangle - \lambda_{K,\text{model}}$, which is 0.814 for the present simulation.

In addition, one sees from Figure 3 that there are significant fluctuations around the mean. The magnitude of

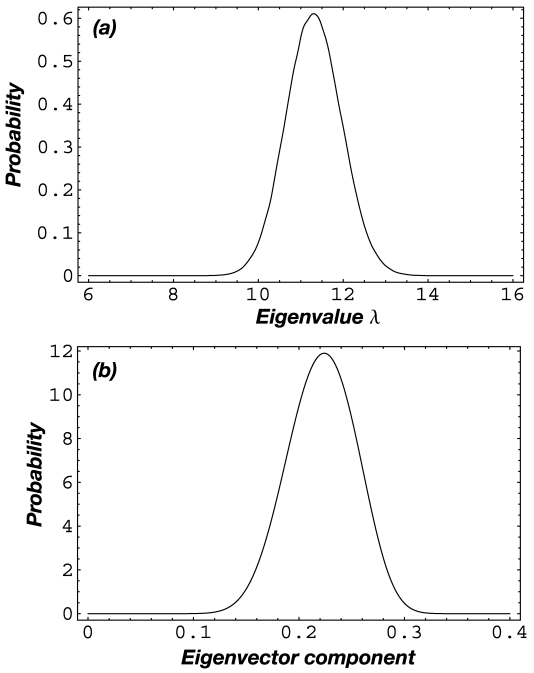

Figure 3: Distribution of (a) the largest eigenvalue and (b) all components of the corresponding eigenvector from simulations of the model with $\lambda_{K,\text{model}} = 10.5$

eigenvalue fluctuations is described by the standard deviation:

$$\sigma_\lambda = \sqrt{\langle \lambda_{K,\text{sim}}^2 \rangle - \langle \lambda_{K,\text{sim}} \rangle^2}. \quad (8)$$

For the distribution shown in Figure 3, one obtains $\sigma_\lambda = 0.65$.

There are also significant fluctuations of the eigenvector components. To quantify them, one calculates the standard deviation

$$\sigma_{u_i} = \sqrt{\langle (u_{i,\text{sim}}^{(K)})^2 \rangle - \langle u_{i,\text{sim}}^{(K)} \rangle^2} \quad (9)$$

and finds $\sigma_{u_i} = 0.0284$. As the $u_i^{(K)}$ values do not vary across I, one only needs to estimate one σ_{u_i}.

The aim is now to use this knowledge about the statistical fluctuations of eigenvalue and eigenvector components to construct a better estimator for $\mathbf{C}^{\text{var}}$. As the matrix $\mathbf{C}^{\text{var}}$ is calculated from a single-factor model, it is adequately described by its first principal component, the largest eigenvalue and its eigenvector. The model simulations show that the use of the maximum likelihood estimator for the variances $\{\sigma_{\epsilon_i}^2\}$ leads to a systematic overestimation of the largest eigenvalue as $\langle \lambda_{K,\text{sim}} \rangle = 11.314$ while the true model eigenvalue is $\lambda_{K,\text{model}} = 10.5$. In addition, estimates of the largest eigenvalue and eigenvector from a single simulation are subject to significant statistical fluctuations described by the variances σ_λ and σ_{u_i}.

As a first step, it is necessary to remove the bias from the estimate $\mathbf{C}_{\text{canonical}}^{\text{var}}$. To achieve this goal, one takes the point of view that its largest eigenvalue $\lambda_{K,\text{canonical}}$ is an estimator for the expectation value $\langle \lambda_{K,\text{sim}} \rangle$ over all Monte Carlo simulations. Based on a single observation, $\lambda_{K,\text{canonical}}$ estimates its mean. Its highly non-linear provenance from the maximum-likelihood estimators for the σ_{ϵ_i} does impede an analytical assessment, motivating a computational analysis. The hypothetical assumption $\mu_i \equiv 1 \; \forall i$ is relaxed, and knowledge of the bias generation is used to remove the bias via bootstrapping. Starting from the original volatility estimates, the study constructs a set of smaller model parameters $\{\mu_{i,\text{boot}}\}$ such that $\langle \lambda_{K,\text{sim}} \rangle = \lambda_{K,\text{canonical}}$. As the map $G : \{\mu_i\} \to \{\langle \lambda_{K,\text{sim}} \rangle, \langle u_{i,\text{sim}}^{(K)} \rangle\}$ is only defined via a Monte Carlo simulation, it cannot easily be inverted. The inversion of G is described in detail in Appendix B. The new parameters are defined by

$$\{\mu_{i,\text{boot}}\} = G^{-1}(\lambda_{K,\text{canonical}}, \{u_{i,\text{canonical}}^{(K)}\}) \quad (10)$$

The set $\{\mu_{i,\text{boot}}\}$ are used as optimal estimators (with respect to estimating the correlation matrix from finite-length time series) for the ratios $\{\sigma_{\epsilon_i}^2/\sigma_Y^2\}$ in

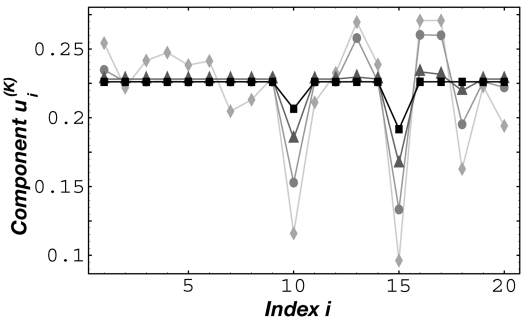

Figure 4: Comparison between the eigenvector $\mathbf{u}^{(K)}_{\text{boot}}$ (diamonds) and the conservative estimates $\mathbf{u}^{(K)}_{1\sigma}$ (circles), $\mathbf{u}^{(K)}_{2\sigma}$ (triangles) and $\mathbf{u}^{(K)}_{3\sigma}$ (squares)

Equation (5) to derive the 'true' infinite time series correlation matrix $\mathbf{C}^{\text{var}}_{\text{boot}}$ according to Equation (7).

The largest eigenvalue of $C^{\text{var}}_{\text{boot}}$ is $\lambda_{K,\text{boot}} = 11.838$, which is smaller than the previous estimate $\lambda_{K,\text{canonical}} = 12.348$. The difference between the two is due to the systematic eigenvalue shift explained above. The eigenvector $\mathbf{u}^{(K)}_{\text{boot}}$ corresponding to the largest eigenvalue of $\mathbf{C}^{\text{var}}_{\text{boot}}$ is displayed in Figure 4; it is almost identical to the eigenvector of $\mathbf{C}^{\text{var}}_{\text{canonical}}$.

As a conclusion, even if the generating process for relative PD movements is a simple single-factor model, the empirically found parameters can deviate significantly from the theoretical ones. The empirical $\mathbf{C}^{\text{var}}_{\text{canonical}}$ 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.[15–17] Then, the statistical properties of the ensemble $\{\mathbf{C}^{\text{sim}}\}$ can be used to derive error bars for both the largest eigenvalue and the components of the corresponding eigenvector.

CONSERVATIVE ESTIMATES

How can these results be used to make a reliable estimate for the correlation matrix of relative PD movements? A bank needs to act in a conservative manner to prevent its insolvency. Using the bias-corrected correlation estimate $\mathbf{C}^{\text{var}}_{\text{boot}}$ discussed in the last section, the bank risks that the correlations are 'accidentally' low. The most conservative approach would be to assume all correlations to be 1, ie $u_i^{(K)} = 1/\sqrt{K}\,\forall\, i$. But now the model would effectively be a single-sector model. Any possibility to measure concentration risk in certain industry sectors would be prevented. The model would not encourage diversifying the business across sectors.

As a controlled mediation, the paper introduced '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. To achieve this goal, the paper proceeds by determining parameters $\{\mu_{i,\text{case}}\}$ such that:

- the bias in the largest eigenvalue is removed;
- the expectation value $\langle\lambda_{K,\text{sim}}\rangle$ calculated from simulations with parameters $\{\mu_{i,\text{case}}\}$ is by x standard deviations σ_λ larger than the corresponding expectation value calculated based on the parameters $\{\mu_{i,\text{boot}}\}$;
- the eigenvector component expectation values $\langle u^{(K)}_{i,\text{sim}}\rangle$ calculated from simulations with parameters $\{\mu_{i,\text{case}}\}$ are x standard deviations σ_{u_i} closer to the most conservative value $1/\sqrt{K}$ than the corresponding expectation values from simulations with parameters $\{\mu_{i,\text{boot}}\}$.

Having found a set of parameters $\{\mu_{i,\text{case}}\}$ satisfying the above requirements, they are used to calculate conservative estimates $\mathbf{C}_{x\sigma}$ from Equation (7).

As the above requirements cannot be solved directly for the $\{\mu_{i,\text{case}}\}$, an

iterative routine is used to determine them. The details of this routine are described in Appendix A. The iterative routine is stopped when the total change resulting from ten successive iterations is smaller then 1 per cent.

Having derived the sets $\{\mu_{i,\mathrm{case}}\}$ which satisfy requirements (i) to (iii) to the desired accuracy, they are used in conjunction with Equation (7) to derive the 'true' infinite time series correlation matrices $\mathbf{C}_{1\sigma}$, $\mathbf{C}_{2\sigma}$ and $\mathbf{C}_{3\sigma}$ for each case. The results are shown in Figure 4. One can 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 components of the initial eigenvector fluctuate significantly, the components of $\{u_{i,1\sigma}^{(K)}\}$ fluctuate less, and $\{u_{i,3\sigma}^{(K)}\}$ is closest to the null hypothesis of equal components. For the largest eigenvalue, one obtains $\lambda_{K,1\sigma} = 12.443$, $\lambda_{K,2\sigma} = 13.050$ and $\lambda_{K,3\sigma} = 13.633$.

Here, the study takes into account the influence of estimation risk for a particularly simple single-factor model. Multiple loadings on the global risk factor Y (in 4), ie a single-factor model with multiple loadings $X_{it} = b_i Y_t + \epsilon_{it}$, could be established with a linear model, rather than univariate variance estimation. However, the data design would need to change. Further extending the approach to multi-factor models, commonly used in practice (PortfolioManager, CreditRisk+, CreditMetrics), one can estimate the coefficients to the factors via a linear model. However, given the orthogonality of the factors, there would now be as many eigenvalues as factors, and a conservative estimation might simultaneously shift them towards their maximum (relative to their individual standard deviations).

Out-of-sample testing, or back-testing, is an established procedure that prevents under-smoothing. Additionally, structural changes in the data-generating process can be detected by this method. Ideally, portfolio-dependent measures, like the value-at-risk (VaR) presented in Table 1, should be confronted with realised losses. For model parameters, like correlations, out-of-sample testing would in principle be possible, even without specifying a portfolio. However, back-testing is not feasible if either internally the industry sector definition changes in time, or as in the present case, if externally the data supplier changes its sector definition. The empirically observed insolvency rates for the years 1994–2000 were measured in the sector segmentation of 1993.[12] A novel segmentation was introduced in 2003,[19] such that there are insufficient independent observations for out-of-sample testing. Additionally, it must be mentioned that time series of insolvencies are not fractal, in the sense that the prediction of annual correlation matrices cannot be assessed by using monthly insolvency rates. This method of out-of-sample testing is only possible in the stock market where high-frequency data with a short

Table 1: Analysis of CreditVaR and expected shortfall at level 99.95 per cent (99.90 per cent) for different correlation matrices (€bn)

Correlation matrix	CreditVaR	Expected shortfall
Independence	1.078 (983)	1.209 (1.117)
$\mathbf{C}^{\mathrm{var}}_{\mathrm{canonical}}$	1.299 (1.186)	1.460 (1.348)
$\mathbf{C}^{\mathrm{var}}_{\mathrm{boot}}$	1.283 (1.172)	1.441 (1.331)
$\mathbf{C}^{\mathrm{var}}_{1\sigma}$	1.314 (1.200)	1.478 (1.365)
$\mathbf{C}^{\mathrm{var}}_{2\sigma}$	1.340 (1.223)	1.509 (1.392)
$\mathbf{C}^{\mathrm{var}}_{3\sigma}$	1.366 (1.246)	1.539 (1.419)
Single sector	1.561 (1.417)	1.769 (1.625)

auto-correlation time are readily available (eg see Rosenow[20]). However, for the model of equi-correlation considered in the preceding section, the influence of out-of-sample fluctuations was assessed via Monte-Carlo simulations. As the method for calculating conservative estimates replicates findings from Monte Carlo simulations in an approximate estimate, and as deviations due to these approximations are evaluated and found to be negligible for each step of the approximation, the conservative correlation estimates can rightly be expected to perform well under out-of-sample conditions.

ECONOMIC IMPLICATIONS OF THE DIFFERENT CORRELATION MATRICES

The last section described five different estimates for the cross-correlation matrix, ie $\mathsf{C}^{\text{var}}_{\text{canonical}}$, $\mathsf{C}^{\text{var}}_{\text{boot}}$, $\mathsf{C}_{1\sigma}$, $\mathsf{C}_{2\sigma}$ and $\mathsf{C}_{3\sigma}$. To judge the economic implications of these estimates, the paper studies the differences in the loss distribution resulting from these correlation estimations. To do this, the impact of the different correlation estimates is quantified by calculating their influence on CreditVaR and the conditional expectation over the CreditVaR, ie the expected shortfall.

The portfolio under study is realistic — although fictitious — for an international bank. It consists of 4,934 risk units distributed asymmetrically over 20 sectors with 20 to 500 counterparts per sector. The total exposure is in the double-digit billion euro range with a largest exposure of €750m and a smallest exposure of €0.13m. The counterpart-specific default probability varies between 0.03 per cent and 7 per cent, and the expected loss for the total portfolio is €187m. The primary aim here is to estimate a quantile and a lower partial moment of a probability distribution — namely the CreditVaR and the expected shortfall of the portfolio loss distribution.

The loss can be written as:

$$L = \sum_{\text{counterpart } A} \text{Exposure at default}_A \times \text{Loss given default}_A \times \text{Default indicator}_A \quad (11)$$

with the exposure at default and the loss given default known in advance. The defaults in this model are assumed to be, conditionally on a univariate X, independent. In a univariate simplification of the default probabilities to the default indicator (2), all depend on a univariate latent random variable X. The discrete loss distribution is analytically derived by mixing the probability generating function (pgf) of the loss distribution for independent defaults with the density of the latent variable X. For the model CreditRisk+, ie a gamma distributed X, this results in a pgf for the loss (11), from which the loss probabilities are calculated with the Panjer recursion.[21]

In the case of a multivariate latent variable, as in model (2), it is possible to calculate the variance of L dependent of the variances $\sigma_i^2 = Var(X_i)$ and the correlations $\rho_{ij} = Corr(X_i, X_j) = (\sqrt{1 + \sigma^2_{\epsilon_i}/\sigma^2_Y}\sqrt{1 + \sigma^2_{\epsilon_j}/\sigma^2_Y})^{-1}$ in model (4). Similarly, the variance of L can be calculated for a single-sector model, ie a univariate X with variance denoted by σ_X^2. Matching of the loss variances and solving for σ_X^2 results in:

$$\sigma_X^2 = \frac{1}{E(L)^2}\sum_{i,j=1}^{K}\rho_{ij}\sigma_i\sigma_j\varepsilon_i\varepsilon_j.$$

Here ϵ_i is the expected loss in sector i.[1] For the presented candidates of ρ_{ij}, Table 1 shows the CreditVaR and expected shortfall.

In the presence of an unknown parameter, it is a well established statistical result (see Lehmann[22]) that the use of the point estimate for the parameter — whether or not derived by a model — leads to an underestimation of the quantile estimate. To account for this additional estimation uncertainty, the bias-corrected point estimate $C^{\text{var}}_{\text{boot}}$ is used as a starting point, with volatilities 1σ, 2σ and 3σ added to the correlation estimate. (The bias correction accounts for a reduction of €16m capital as compared with $C^{\text{var}}_{\text{canonical}}$ on the 99.95 per cent level.) When applying a one-σ estimate, the CreditVaR increases by €31m, for the two-σ estimate there is another increase of €27m, and using the three-σ estimate, the CreditVaR increases by yet another €26m (all at the 99.95 per cent level). To put these numbers in perspective, note that without including correlations, the CreditVaR is found to be €1.078bn, and that the assumption of full correlations among all sectors leads to a CreditVaR of €1.561bn. The effects on the 99.90 per cent confidence level for the CreditVaR as well as for the expected shortfall are similar.

The use of the two-σ estimate guarantees a sufficient forecast reliability on the one hand and allows for some guidance for (11) economical decision on the other. Even more importantly, the conservative method of parameter estimation should provide smooth correlation estimates in the sense that new observations — occurring as time goes by — have only a small impact on the correlation estimate. In this way, one prevents the disruption of banking activities as a consequence of drastic changes in risk assessment, which are not proportional to the increase in information.

In summary, this paper has addressed the problem of estimating correlations between empirical default rates for economic sectors. Due to the short length of these time series, estimation errors are large and the use of a parsimonious model such as a single-factor model is necessary. However, when using such a model to calculate the corresponding correlation matrix, one still typically observes large statistical fluctuations in the correlation structure. Due to these fluctuations, the parameter estimation for an explanatory 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.

It is important to stress that the proposed methodology is neither specific for CreditRisk+ nor to model (4). It may be used in any credit portfolio model depending on a multivariate covariable following a specified model.

Acknowledgment

The authors would like to thank A. Müller-Groeling for initiating this project. We are indebted to F. Altrock for collaboration in an early stage of this research, and for important discussions later on. S. Lösch, C. von Lieres und Wilkau and A. Wilch are thanked for useful discussions.

References

1 Bürgisser, P., Kurth, A., Wagner, A. and Wolf, M. (1999) 'Integrating correlations', *Risk Magazine*, Vol. 12, pp. 57–60.

2 Nagpal, K. and Bahar, R. (2001) 'Measuring default correlation', *Risk Magazine*, Vol. 14, pp. 129–132.

3 Nagpal, K. and Bahar, R. (2001) 'Modelling default correlation', *Risk Magazine*, Vol. 14, pp. 85–89.

4 Frey, R., McNeil, A. and Nyfeler, M. (2001) 'Copulas and credit models', *Risk Magazine*, Vol. 14, pp. 111–114.

5 Credit Suisse First Boston (1997) 'Credit Risk+: A Credit Risk Management Framework', technical document.

6 Jorion, P. (1985) 'International portfolio diversification with estimation risk', *The Journal of Business*, Vol. 58, pp. 259–278.

7 Klein, R. and Bawa, V. (1976) 'The effect of estimation risk on optimal portfolio choice', *Journal of Financial Economics*, Vol. 3, pp. 215–231.

8 Gordy, M. B. (2001) 'A risk-factor foundation for ratings-based capital rules', *Journal of Financial Intermediation*, Vol. 12, pp. 199–232.

9 The specific choice of the variables X_i is adapted to the use of CreditRisk+, but all credit risk models require the specification of a correlation model for latent risk factors. For example, CreditMetrics assumes a linear factor model for the counterpart's asset return R_A — see Gupton, G. M., Finger, C. C. and Bhatia, M. (1997) 'CreditMetrics', JPMorgan, technical document. The systematic part of a company's return depends — with different weights $\omega_{A,i}$ — on economic indices X_i in the countries and industry sectors in which the company is active. An independent idiosyncratic part ε_A is added: $R_A = \sum_{i=1}^{K} \omega_{A,i} X_i + \varepsilon_A$. The risk factors X_i are directly observable in this model.

10 Weissbach, R. and Dette, H. (2007) 'Kolmogorov-Smirnov-type testing for the partial homogeneity of Markov processes — With application to credit risk', *Applied Stochastic Models in Business and Industry*, Vol. 22, pp. 223–234.

11 Weissbach, R., Tschiersch, P. and Lawrenz, C. (2009) 'Testing time-homogeneity of rating transitions after origination of debt', *Empirical Economics*, Vol. 36, pp. 575–596.

12 Statistisches Bundesamt, Wiesbaden, Klassifikation der Wirtschaftszweige mit Erläuterungen Ausgabe 1993, Metzler-Poeschel (2001).

13 Taking a look at the whole sample, the minimum number of entities under default risk per sector was 570, whereas the maximum number was 47,858. Averaging over time, the minimum was 648, and the maximum 47,084. The median was 2,569 and the mean was 10,296. For the insolvency frequencies, the time-average minimum for the sectors is 0.1 per cent, the maximum 2.1 per cent and the median 1.2 per cent.

14 Dickey, D. A. and Fuller, W. A. (1979) 'Distribution of estimators for autoregressive time series with unit root', *Journal of the American Statistical Association*, Vol. 74, pp. 427–431.

15 Laloux, L., Cizeau, P., Bouchaud, J.-P. and Potters, M. (1999) 'Random matrix theory', *Risk Magazine*, Vol. 12, pp. 69–73.

16 Laloux, L. *et al.* (1999) *Physical Review Letters*, Vol. 83, pp. 1467–1470.

17 Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L. A. N. and Stanley, H. E. (1999) 'Universal and non-universal properties of cross-correlations in financial time series', *Physical Review Letters*, Vol. 83, pp. 1471–1474.

18 Anderson, T. W. (2003) 'An Introduction to Multivariate Statistical Analysis' (3rd edn), John Wiley & Sons, Hoboken.

19 Statistisches Bundesamt, Wiesbaden, Klassifikation der Wirtschaftszweige mit Erläuterung 2003, Destatis (2003).

20 Rosenow, B. (2007) 'Determining the optimal dimensionality of multivariate volatility models with tools from random matrix theory', *Journal of Economic Dynamics and Control*, Vol. 32, p. 279.

21 Panjer, H. H. and Willmot, G. E. (1992) 'Insurance Risk Models', Society of Actuaries, Schaumberg, IL.
22 Lehmann, E. L. (1993) 'Testing Statistical Hypotheses', Chapman & Hall.
23 Abramowitz, M. and Stegun, I. A. (eds) (1970) 'Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables', Dover, New York.

APPENDIX A: ITERATIVE CALCULATION OF CONSERVATIVE ESTIMATES

First, choose $\{\mu_{i,\text{boot}}\}$ as initial values for the $\{\mu_{i,\text{case}}\}$. Next, decompose

$$\mu_i = \alpha\beta_i, \quad \text{with} \quad \sum_{i=1}^{K} \beta_i^2 = 1. \quad \text{(A1)}$$

Steps (a) to (c) of the following routine are iterated until convergence is reached. During these iterations, the $\{\beta_{i,\text{case}}\}$ is varied but $\alpha \equiv \alpha_{\text{boot}}$ is fixed. The iterative steps are as follows:

a) The parameters $\{\mu_{i,\text{case}}\}$ are used to calculate $\langle u_{i,\text{sim}}\rangle$, σ_λ and σ_{u_i} via a Monte Carlo simulation along the lines described in the previous section.

b) The ideal values for the expectation values of eigenvector components would be

$$\langle u_{i,\text{sim}}^{(K)}\rangle_{\text{ideal}} = \begin{cases} u_{i,\text{canonical}}^{(K)} + x\sigma_{u_i} & \text{if } \dfrac{1}{\sqrt{K}} - u_{i,\text{canonical}}^{(K)} > x\sigma_{u_i} \\ u_{i,\text{canonical}}^{(K)} - x\sigma_{u_i} & \text{if } u_{i,\text{canonical}}^{(K)} - \dfrac{1}{\sqrt{K}} > x\sigma_{u_i} \\ \sqrt{1/K} & \text{otherwise} \end{cases} \quad \text{(A2)}$$

As these 'ideal values' depend on the $\{\sigma_{ui}\}$ which in turn are functions of the $\{\mu_{i,\text{case}}\}$, it is not useful to impose the conditions of Equation (A2) directly. Instead, choose an iterative approach and define auxiliary quantities

$$\nu_i = \langle u_{i,\text{sim}}^{(K)}\rangle + \eta(\langle u_{i,\text{sim}}^{(K)}\rangle_{\text{ideal}} - \langle u_{i,\text{sim}}^{(K)}\rangle) \quad \text{(A3)}$$

which are normalised to unity before proceeding. For the calculations in this paper, the choice $\eta \approx 0.1$ turned out to be a good compromise between achieving a fast convergence (favours large values of η) and avoiding oscillatory limit cycles of the iterative algorithm (demands small values of η).

c) Next, calculate a new set of parameters $\{\mu_{i,\text{case}}\}$, which satisfy the equation $\langle u_{i,\text{sim}}^{(K)}\rangle = \nu_i$ when used as input parameters for a Monte Carlo simulation. The determination of these new parameter values is the most difficult part of the iterative routine, as the map $G : \{\mu_i\} \to \langle u_{i,\text{sim}}^{(K)}\rangle$ is only defined via a Monte Carlo simulation and hence cannot easily be inverted. The inversion of G is described in detail in Appendix B.
The new parameters are defined by

$$\{\mu_{i,\text{case}}\} = G^{-1}(\alpha_{Boot}, \{\nu_i\}). \quad \text{(A4)}$$

In the iterative loop, this new set of parameters is used as input for step A. To achieve both fast convergence and reliable results, increase the number N of Monte Carlo simulations in (a) and decrease η as the parameters $\{\mu_{i,\text{case}}\}$ converge to their final values. For each

value of $x = 1, 2, 3$, use this routine to obtain sets of parameters $\{\beta_{i,1\sigma}\}$, $\{\beta_{i,1\sigma}\}$, $\{\beta_{i,1\sigma}\}$, which are used in the routine for the calculation of the α_{case}. For the iterative calculation of the new eigenvalue, use again the decomposition of Equation (A1). Now, the goal is the calculation of the α_{case} and the $\{\beta_{i,\text{case}}\}$ is fixed. A reasonable initial value for α_{case} is α_{boot}. The iterative routine used for this study contains the steps below.

d) Parameters $\mu_i = \sqrt{\alpha_{\text{case}}}\beta_{i,\text{case}}$ are used to calculate $\langle\lambda_{K,\text{sim}}\rangle$ and σ_λ.

e) Incorporating the safety margin of $x\sigma_\lambda$, the ideal value of $\langle\lambda_{K,\text{sim}}\rangle$ would be

$$\langle\lambda_{K,\text{sim}}\rangle_{\text{ideal}} = \lambda_{K,\text{canonical}} + x\sigma_\lambda. \quad \text{(A5)}$$

Again, as σ_λ is a function of the $\{\mu_{i,\text{case}}\}$, it is not useful to enforce the relation of Equation (A5) directly. Instead, define an auxiliary 'largest eigenvalue' which contains a small correction

$$\kappa = \langle\lambda_{K,\text{sim}}\rangle + \eta(\langle\lambda_{K,\text{sim}}\rangle_{\text{ideal}} - \langle\lambda_{K,\text{sim}}\rangle). \quad \text{(A6)}$$

f) Using the inversion G^{-1} of the mapping $G : \{\mu_i\} \to \{\langle\lambda_{K,\text{sim}}\rangle, \langle u^{(K)}_{i,\text{sim}}\rangle\}$ defined above, one can now calculate a new set of parameters

$$\{\mu_{i,\text{case}}\} = G^{-1}(\kappa, \{\beta_{i,\text{case}}\}). \quad \text{(A7)}$$

These new parameters are used in step (e) again, until convergence is reached.

APPENDIX B: INVERSION OF THE FUNCTION G

For the calculation of both unbiased and conservative estimates of correlation matrices, it is important to find an efficient algorithm to invert the function $G : \mathbb{R}^K \to \mathbb{R}^{K+1}$, $\{\mu_i\} \to \{\langle\lambda_{K,\text{sim}}\rangle, \langle u^{(K)}_{i,\text{sim}}\rangle\}$ defined via Monte Carlo simulations in the section on 'Fluctuations in empirical correlation matrices'. Note that the range of G has dimension K because the eigenvector $u^{(K)}_{i,\text{sim}}$ is normalised to unity.

To find the inversion algorithm, an analytic approximation to G is needed. First, calculate the expectation value $\langle\mathbf{C}^{\text{var}}\rangle$ by taking the expectation value of Equation (5) with respect to the distribution Equation (6). This study has found that the largest eigenvalue and corresponding eigenvector of $\langle\mathbf{C}^{\text{var}}\rangle$ are good approximations (error of the order of 1 per cent) to $\langle\lambda_{K,\text{sim}}\rangle$ and $\{\langle u^{(K)}_{i,\text{sim}}\rangle\}$ and hence calculates the largest eigenvalue $\lambda_{K,\langle C^{\text{var}}\rangle}$ and the corresponding eigenvector $u^{(K)}_{i\langle C^{\text{var}}\rangle}$. To this end, the following parameterisation is introduced:

$$\langle\mathbf{C}^{\text{var}}_{ij}\rangle = \delta_{ij} + (1 - \delta_{ij})ab_ib_j, \text{ with}$$
$$\sum_{i=1}^{K} b_i^2 = 1. \quad \text{(B1)}$$

The parameters are given by

$$\sqrt{a}b_i = \left\langle \left(1 + \frac{\sigma^2_{\epsilon_i}}{\sigma^2_Y}\right)^{-1/2} \right\rangle. \quad \text{(B2)}$$

The expectation value is defined with respect to the distribution Equation (6). One now specialises to the practically relevant situation $T = 6$ and

defines the following function:

$$\begin{aligned} g(x) &= \left\langle \left(1+\frac{\sigma_{\epsilon_i}^2}{\sigma_Y^2}\right)^{-1/2} \right\rangle \Bigg|_{\mu_i=x} \\ &= \int_0^\infty d\eta f_{\chi^2,5}(\eta) \frac{1}{\sqrt{1+\eta\frac{x}{5}}} \\ &= \frac{1}{\Gamma(\frac{5}{2})2^{5/2}} \int_0^\infty d\eta \frac{\eta^{3/2}e^{-\eta^2}}{\sqrt{1+\frac{x}{5}\eta}} \\ &= \frac{25}{6}\sqrt{\frac{5}{2\pi}} x^{-5/2} e^{5/(4x)} \\ &\quad \times \left[K_0\left(\frac{5}{4x}\right) + \left(-1+\frac{2x}{5}\right) K_1\left(\frac{5}{4x}\right) \right]. \end{aligned} \tag{B3}$$

Here, $\Gamma(x)$ denotes the gamma function, and $K_\nu(x)$ denotes the modified Bessel function of the second kind.[23] Next, approximately calculate the eigenvalue $\lambda_{K,\langle C^{\text{var}}\rangle}$ by making the ansatz $u^{(K)}_{i,\langle C^{\text{var}}\rangle} = b_i$

$$\begin{aligned} \sum_{j=1}^{K} \langle C_{ij}^{\text{var}} b_j \rangle &= (1 - ab_i^2)b_i + ab_i \\ &\approx (1 - \frac{a}{K} + a)b_i \end{aligned} \tag{B4}$$

The above approximation is justified because the replacement $b_i^2 \to \frac{1}{K}$ is made in a subleading term ($b_i^2 \ll 1$). Now identify

$$\lambda_{K,\langle C^{\text{var}}\rangle} \approx 1 + a\left(1 - \frac{1}{K}\right). \tag{B5}$$

By using this approximation, the sought-after inverse map G^{-1} has the following component representation:

$$\mu_i = g^{-1}\left(\sqrt{\frac{\langle \lambda_{K,\text{sim}}\rangle - 1}{1-\frac{1}{K}}}\langle u^{(K)}_{i,\text{sim}}\rangle\right). \tag{B6}$$

This study has found that the approximations involved in calculating G^{-1} give rise to errors smaller than 1 per cent. With the analytic expression Equation (B3) for $g(x)$, the inverse function $g^{-1}(x)$ can be calculated numerically without difficulty.