Rosenow, Bernd; Weißbach, Rafael; Altrock, Frank

Working Paper

Modelling correlations in credit portfolio risk II

Technical Report, No. 2007,06

Provided in Cooperation with:
Collaborative Research Center 'Reduction of Complexity in Multivariate Data Structures' (SFB 475), University of Dortmund

Suggested Citation: Rosenow, Bernd; Weißbach, Rafael; Altrock, Frank (2007) : Modelling correlations in credit portfolio risk II, Technical Report, No. 2007,06, Universität Dortmund, Sonderforschungsbereich 475 - Komplexitätsreduktion in Multivariaten Datenstrukturen, Dortmund

This Version is available at:
https://hdl.handle.net/10419/24992

Modelling Correlations in Credit Portfolio Risk II

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

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

[2] *Institut für Wirtschafts- und Sozialstatistik ,*

Universität Dortmund, 44221 Dortmund, Germany

[3] *Credit Risk Management, WestLB AG, 40217 Düsseldorf, Germany*

(Dated: January, 2007)

Abstract

The risk of a credit portfolio depends crucially on correlations between latent covariates, for instance the probability of default (PD) in different economic sectors. Often, correlations have to be estimated from relatively short time series, and the resulting estimation error hinders the detection of a signal. We suggest a general method of parameter estimation which avoids in a controlled way the underestimation of correlation risk. Empirical evidence is presented how, in the framework of the CreditRisk+ model with integrated correlations, this method leads to an increased economic capital estimate. Thus, the limits of detecting the portfolio's diversification potential are adequately reflected.

* The first two authors have contributed equally.

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. Recently, several approaches for describing and modelling concentration risk were discussed [1, 2, 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+ with the method of Bürgisser et al. [1]. 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 minimal model with a reduced dimensionality of the parameter space. A reasonable choice is a parsimonious model with the global default rate as latent factor [6].

Despite the fact that the parameter space of a one-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. We discuss these fluctuations in detail and suggest a bootstrap method which allows to find a level for the parameters that reflects the applicable risk aversion of the individual bank. We exemplify the impact of different conservative estimates on the CreditVaR of a realistic portfolio.

Methodology

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 $T+1$ 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

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

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_i with expectation $E(X_i) = 1$ according to

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

When using CreditRisk+ with a time horizon of one year, one is interested in the relative change of default probabilities. The individual PD, p_A in Eq. 2, is usually taken to depend on the current economic activity at time $t-1$, i.e. it describes a point-in-time rating. The PD for the forthcoming period $t+1$ 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,

$$X_{it} = \frac{\hat{PD}_{it+1}}{\hat{PD}_{it}} + 1 - \frac{1}{T}\sum_{t=1}^{T} \frac{\hat{PD}_{it+1}}{\hat{PD}_{it}} \quad , \tag{3}$$

which is normalized to $\langle X_i \rangle = \frac{1}{T}\sum_{t=1}^{T} X_{it} = 1$ 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

For the illustration of our theoretical concept, we use sector specific time series of insolvency rates for a segmentation of the German economy into $K = 20$ sectors selected by us. We take 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 alike. The data – for a much finer segmentation – are supplied by the federal statistical office of Germany and date unfortunately only from 1994–2000 [13].

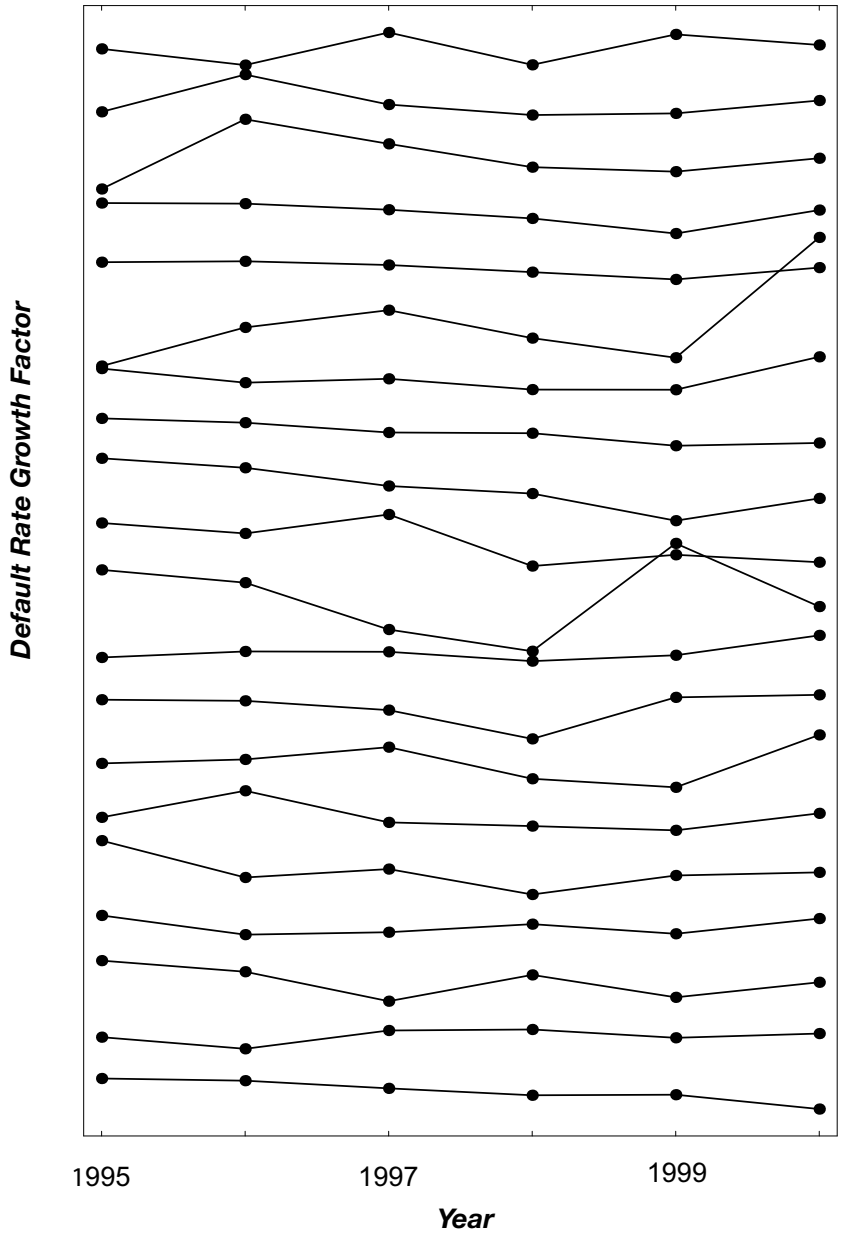

FIG. 1: Default rate growth factors of $K = 20$ German sectors from 1995 to 2000. For clarity, subsequent curves have an offset against each other.

In view of the small sample size, namely $T = 6$, we use a parsimonious one-factor model for the estimation of cross-correlations. As a factor we use relative changes Y_t of the national insolvency rate. The definition of Y_t is analogous to the definition of the X_{it} in Eq. (3). We decompose the sector PDs according to

$$X_{it} = Y_t + \epsilon_{it} \ , \tag{4}$$

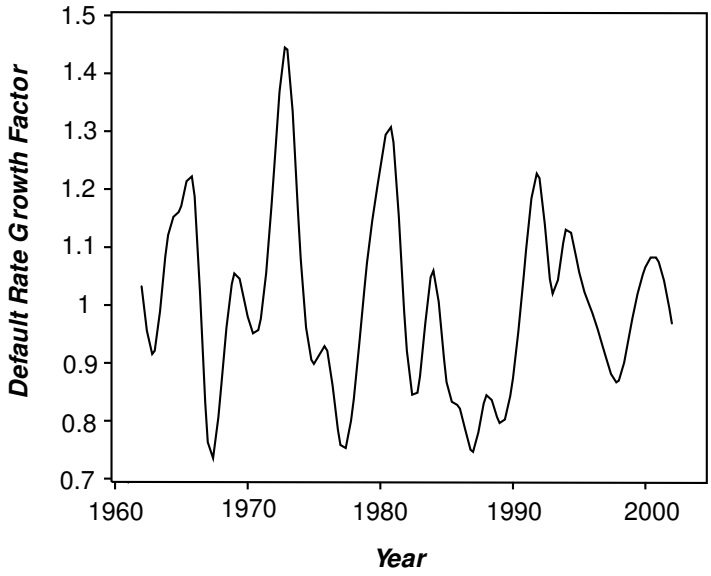

FIG. 2: Relative changes of insolvency rate for the German economy from 1962 to 2003 (until 1994 West Germany).

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

As a consequence, the correlation between the systematic parts of sectors' default rates now is uniformly equal to one. However, this systematic correlation is obscured by the residuals. As Eq. (4) realizes a variance decomposition, it creates a relation between the correlations and volatilities. For reasons of tractability we now make the fundamental assumption that the residuals are uncorrelated among each other and uncorrelated with the factor. Then, one obtains the correlation matrix $\mathsf{C}^{\rm var}$ with elements

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

Here the Kronecker symbol δ_{ij} is one if $i=j$ and zero otherwise. The variance of the residuals ϵ_{it} is denoted by $\sigma_{\epsilon_i}^2$. $\mathsf{C}^{\rm var}$ has an intuitive interpretation: according to Eq. (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 is, the larger is the residual variance of this sector and according to Eq. (5) the correlation coefficients between this sector and other sectors becomes small. C^{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.

Model (4) links the sectorial to the national default rates. Hence, additional to the data for the 20 sectors we use the insolvency rate for the entire German economy, available from 1962 to 2003 (until 1994 West Germany). In order to obtain credible volatility estimates, we need information concerning the stationarity of the time series. The use of relative changes according to Eq. (3) eliminates any linear trend. For a visual assessment of stationarity, we display the time series of sector default rate growth factors $\{X_{it}\}$ in Fig. 1 [14] and the default rate growth factor Y_t for the national economy in Fig. 2. All time series appear to be stationary. As we have only six observations each for the sectorial growth rates, a statistical test for non–stationarity is not feasible. However, for the longer series of national insolvency rates statistical tests are possible, and we test the hypothesis is non–stationarity.

In general, testing theory needs a model for the data and autoregressive models (AR[q]) are common for financial time series:

$$Y_t = a_0 + a_1 Y_{t-1} + \ldots + a_q Y_{t-q} + \eta_t \quad t = p+1, \ldots, T. \tag{6}$$

Here $a_0, \ldots, a_q$ are time-independent parameters, q is called the order of the regression and the innovations η_t represent "white noise", i.e. have expectation $E(\eta_t) = 0$ and variance $Var(\eta_t) = \sigma_\eta^2$. Clearly, if e.g. in an AR[1] model the parameter a_1 is larger than one, the times series is trended, i.e. non mean-stationary, and the volatility of a future Y_t may not be estimated by the empirical volatility of the time series $Y_1, \ldots, Y_{t-1}$. The finding that an a_1 unequal to one indicates a trend generalizes to the rule that the existence of a "unit root" indicates non stationarity. As test for the hypothesis of a unit root - essentially an adoption of the famous t-test - the Dickey-Fuller test was developed in [8] and is now a standard test (see [7, pg.81]). We apply the test to our national insolvency rate data using the SAS macro "dftest". The statistical decision against the hypothesis depends on the error probability α one is willing to risk, the type I error. We will use the common value $\alpha = 5\%$ in the

following. For a given data set, the p-value gives the smallest error rate at which one is able to reject the hypothesis. For a test performed at a level of $\alpha = 5\%$ the decision rule is to reject whenever the p-value is smaller than 5%. Our p-value for the Dickey-Fuller test under the AR(1) model is 0.00079 and enables to reject the trend hypothesis and safely work in a stationarity world. The decision does not change (again at level $\alpha = 5\%$) for larger models, i.e. for orders $q = 2$ to 5.

In addition to testing the mean-stationarity of the time series $\{Y_t\}$, one must assess stationarity of its variance. A finding of clustered volatility would impede the estimation of the current volatility. Again a model is needed and the autoregressive conditional heteroscedastic model (ARCH) is typical for financial time series. The Lagrange-multiplier test for the hypothesis of the absence of ARCH-effects in the volatility (see [3]) does not reject, for orders up to 12 the p-value - using the SAS procedure "autoreg" - is between 0.5 and 0.8. Hence we conclude that the conditional volatility may be considered as constant, in other words, the series is mean-variance stationary.

In the following we estimate both σ_Y^2 and the $\sigma_{\epsilon_i}^2$ with the standard variance estimator, e.g. $\hat{\sigma}_{\epsilon_i}^2 = \frac{1}{T-1}\sum_{t=1}^{T}(\epsilon_{it} - \langle\epsilon_i\rangle)^2$. 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. By using these volatility estimates in Eq. (5), we obtain the canonical correlation estimate $\mathsf{C}^{\rm var}_{\rm canonical}$. 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. The issue becomes relevant for small sample sizes and is investigated in a controlled environment in the next section.

Fluctuations in empirical correlation matrices – a simulation study

In this section, we use the results of Monte Carlo simulations to study the relation between the true cross correlation matrix C and matrices $\mathsf{C}^{\rm sim}$ estimated from time series of length T. We find that the $\{\mathsf{C}^{\rm sim}\}$ differ from C 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 [9, 10].

Assuming that the process Eq. (5) with mutually independent time series Y_t, ϵ_{it}, and ϵ_{jt} is valid, uncertainties in the determination of $\mathsf{C}^{\rm var}_{\rm canonical}$ arise from uncertainties in the

estimation of σ_Y and the σ_{ϵ_i}. As σ_Y is calculated from a long time series including more than forty years of data, its estimation error is negligible in comparison with that of the σ_{ϵ_i} and we set it to zero in the following. For the simulations, we assume normality of the ϵ_{it} due to the increased computational efficiency as compared to the standard assumption of gamma distributed random variables [5]. This gain in efficiency is especially important for the computationally quite demanding iterative calculations described in the next section. We have checked that the deviation between a simulation with normal distributed variables and a simulation with gamma distributed variables is smaller than 3% for the standard deviations defined in Eqs. (8) and (9).

Under the normality assumption of the ϵ_{it} by definition $(\epsilon_{it} - E(\epsilon_i))^2/\sigma^2_{\epsilon_i}$ follows a central χ^2 distribution with one degree of freedom. The sum of T independent χ^2_1 random variables is a χ^2_T variable and the estimation of the mean $E(\epsilon_{it})$ with $\langle \epsilon_i \rangle$ amounts to a reduction of one degree of freedom. Multiplying the ratio $\hat{\sigma}^2_{\epsilon_i}/\sigma^2_{\epsilon_i}$ with $\sigma^2_{\epsilon_i}/\sigma^2_Y$ and application of the density transformation yields that the ratio $\hat{\sigma}^2_{\epsilon_i}/\sigma^2_Y$ follows a χ^2 distribution with $T-1$ degrees of freedom

$$f_i(z) = f_{\chi^2,T-1}\Big(\frac{T-1}{\mu_i} z\Big)\ \frac{T-1}{\mu_i}\ , \tag{7}$$

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

In this section, we study the outcome of model simulations with the help of (7) for the particularly simple hypothetical case where signal Y_t and noise ϵ_{it} have the same volatility, i.e. $\mu_i \equiv 1$, in order to gain qualitative insight into the occurring fluctuations. The corresponding infinite time series correlation matrix $C^{\mathrm{model}}_{ij} = \delta_{ij} - (1-\delta_{ij})/2$ has a largest eigenvalue $\lambda_K = 10.5$ and a corresponding eigenvector $u^{(K)}_i \equiv 1/\sqrt{K}$.

Instead of simulating the time series $\{\epsilon_{it}\}$ and estimating their variance, we remember that for normally distributed $\{\epsilon_{it}\}$ the variance estimator follows a χ^2 distribution. If in addition σ^2_Y is known, then the ratios $\hat{\sigma}^2_{\epsilon_i}/\sigma^2_Y$ are indeed distributed according to Eq. (7). Hence, for each of the 500,000 simulation runs, we i) draw a set of $K = 20$ values for the ratios $\{\hat{\sigma}^2_{\epsilon_i}/\sigma^2_Y\}$ from the χ^2 distribution defined by Eq. (7) with parameters $T = 6$ and $\mu_i \equiv 1$, ii) calculate a matrix $\mathsf{C}^{\mathrm{sim}}$ by inserting the ratios $\{\hat{\sigma}^2_{\epsilon_i}/\sigma^2_Y\}$ in Eq. (5) , iii) and calculate the largest eigenvalue $\lambda_{K,\mathrm{sim}}$ and the corresponding eigenvector $\mathbf{u}^{(K)}_{\mathrm{sim}}$ [15] from this matrix. Averaging over all simulation runs, we finally obtain the probability distribution

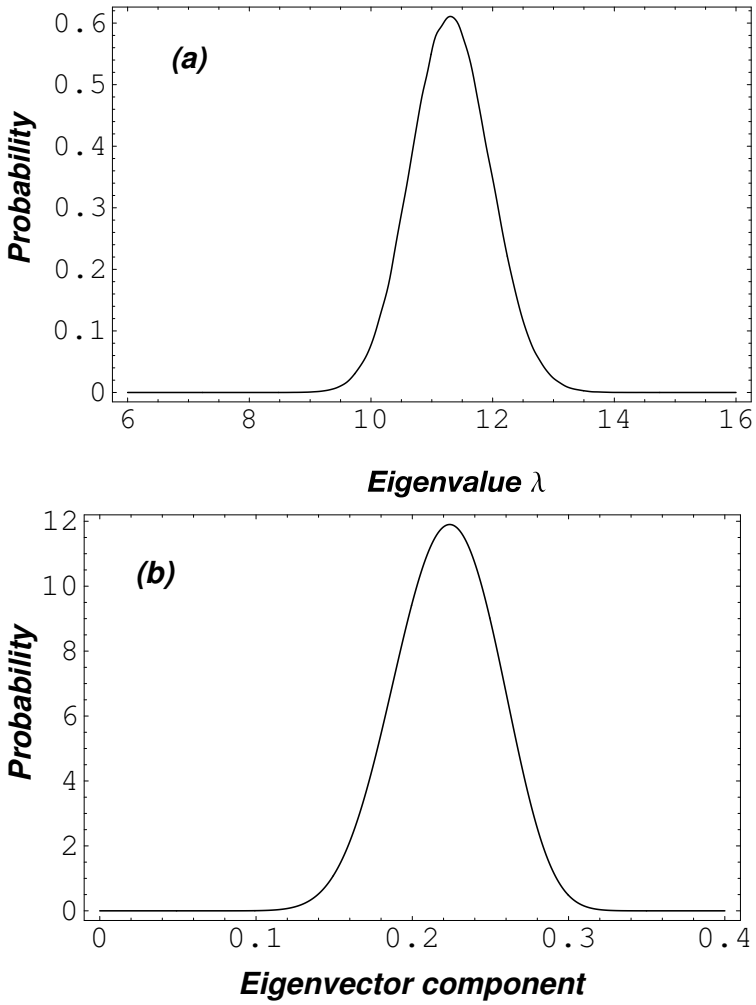

FIG. 3: Distribution of (a) the largest eigenvalue and of (b) all components of the corresponding eigenvector from simulations of the model with $\lambda_{K,\mathrm{model}} = 10.5$.

function (pdf) for both quantities.

The goal of the simulation is to understand i) whether our estimates $\mathsf{C}^{\mathrm{sim}}$ are biased compared to the true correlation matrix $\mathsf{C}^{\mathrm{model}}$, and ii) how large fluctuations from one simulation run to the next are. To answer these questions, we use the fact that by construction all relevant information in the $\{\mathsf{C}^{\mathrm{sim}}\}$ is contained in the largest eigenvalue and the corresponding eigenvector.

We find that both eigenvalue and eigenvector components have broad distributions (see Fig. 3). The distribution of eigenvalues has an average $\langle\lambda_{K,\mathrm{sim}}\rangle = 11.3$ which is significantly larger than the true eigenvalue $\lambda_{K,\mathrm{model}} = 10.5$. Hence, the above described procedure for estimating the correlation matrix is indeed biased towards larger eigenvalues. We quantify

the systematic shift of eigenvalues by the difference $\Delta\lambda = \langle\lambda_{K,\mathrm{sim}}\rangle - \lambda_K$, which is 0.81 for the present simulation.

In addition, one sees from Fig. (3) that there are significant fluctuations around the mean. The magnitude of eigenvalue fluctuations is described by the standard deviation in the simulation

$$\sigma_\lambda = \sqrt{\langle\lambda^2_{K,\mathrm{sim}}\rangle - \langle\lambda_{K,\mathrm{sim}}\rangle^2} \quad . \tag{8}$$

For the distribution shown in Fig. 3 we find $\sigma_\lambda = 0.65$.

There are significant fluctuations of the eigenvector components as well. To quantify them, we calculate the standard deviations

$$\sigma_{u_i} = \sqrt{\langle(u^{(K)}_{i,\mathrm{sim}})^2\rangle - \langle u^{(K)}_{i,\mathrm{sim}}\rangle^2} \; . \tag{9}$$

As all eigenvector components of $\mathsf{C}^{\mathrm{model}}$ are equal, we may aggregate all components of the simulated eigenvectors $\mathbf{u}^{(K)}_{\mathrm{sim}}$, and calculate one common standard deviation $\sigma_u = 0.03$.

Our aim is now to use our knowledge about statistical fluctuations of eigenvalue and eigenvector components to construct a better estimator for $\mathsf{C}^{\mathrm{var}}$. As the matrix $\mathsf{C}^{\mathrm{var}}$ is calculated from a one 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 $\{\sigma^2_{\epsilon_i}\}$ leads to a systematic overestimation of the largest eigenvalue as $\langle\lambda_{K,sim}\rangle = 11.3$ while the true model eigenvalue is $\lambda_K = 10.5$. 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.

As a first step, we want to remove the bias from the estimate $\mathsf{C}^{\mathrm{var}}_{\mathrm{canonical}}$. To achieve this goal, we now take the point of view that its largest eigenvalue $\lambda_{K,\mathrm{canonical}}$ can be interpreted as the expectation value $\langle\lambda_{K,\mathrm{sim}}\rangle$ of a Monte Carlo simulation. We relax the hypothetical assumption $\mu_i \equiv 1\ \forall i$, and use our knowledge of the bias generation to remove the bias. We start from the original volatility estimates that define the set $\{\mu_i\}$. Again, our simulation tells us that using this set for the calibration of our model (4) results in overestimating the correlations, especially in overestimation of the largest eigenvalue of the correlation matrix estimate. We construct a set of smaller model parameters $\{\mu_{i,\mathrm{boot}}\}$ such that $\langle\lambda_{K,\mathrm{sim}}\rangle = \lambda_{K,\mathrm{canonical}}$ and $\langle\mathbf{u}^{(K)}_{\mathrm{sim}}\rangle = \mathbf{u}^{(K)}_{\mathrm{canonical}}$. As the map $G : \{\mu_i\} \to \{\langle\lambda_{K,\mathrm{sim}}\rangle, \langle\mathbf{u}^{(K)}_{\mathrm{sim}}\rangle\}$ is only defined via a Monte Carlo simulation, it cannot easily be inverted. The inversion of G is described

in detail in appendix A. The new parameters are defined by

$$\{\mu_{i,\mathrm{boot}}\} \;=\; G^{-1}\Big(\lambda_{K,\mathrm{canonical}}, \mathbf{u}^{(K)}_{\mathrm{canonical}}\Big)\;. \tag{10}$$

We use the $\mu_{i,\mathrm{boot}}$ as optimal estimators (with respect to estimating the correlation matrix from finite length time series) for the ratios $\sigma^2_{\epsilon_i}/\sigma^2_Y$ in Eq. (5) to derive an unbiased estimate $\mathsf{C}^{\mathrm{var}}_{\mathrm{boot}}$ for the correlation matrix $\mathsf{C}^{\mathrm{var}}$.

The largest eigenvalue of $\mathsf{C}^{\mathrm{var}}_{\mathrm{boot}}$ is $\lambda_{K,\mathrm{boot}} = 11.8$, which is smaller than the previous estimate $\lambda_{K,\mathrm{canonical}} = 12.4$. The difference between the two is due to the systematic eigenvalue shift explained above. The eigenvector $\mathbf{u}^{(K)}_{\mathrm{boot}}$ corresponding to the largest eigenvalue of $\mathsf{C}^{\mathrm{var}}_{\mathrm{boot}}$ is displayed in Fig. 4, it is almost identical to the eigenvector of $\mathsf{C}^{\mathrm{var}}_{\mathrm{canonical}}$.

As a conclusion, even if the generating process for relative PD movements is a simple one–factor model, the empirically found parameters - estimated on basis of the separate univariate times series - can deviate significantly from the theoretical ones. We advocate the point of view that the empirical $\mathsf{C}^{\mathrm{var}}$ 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 [9, 10]. 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 its insolvency. Using the bias corrected correlation estimate $\mathsf{C}^{\mathrm{var}}_{\mathrm{boot}}$ discussed in the last section, the bank risks that the correlations are "accidently" low. The most conservative approach would be to assume all correlations to be 1, i.e. $u^{(K)}_i = 1/\sqrt{K}\;\forall\; i$ and $\lambda_K = K$. But now the model would effectively be a one-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 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. To achieve this goal, we proceed in the following way: we determine parameters $\{\mu_{i,\mathrm{case}}\}$ such that

i) the bias in the largest eigenvalue is removed,

ii) the expectation value $\langle\lambda_{K,\mathrm{sim}}\rangle$ calculated from simulations with parameters $\{\mu_{i,\mathrm{case}}\}$ is by x standard deviations σ_λ larger than the corresponding expectation value calculated based on the parameters $\{\mu_{i,\mathrm{boot}}\}$,

iii) the eigenvector component expectation values $\langle u^{(K)}_{i,\mathrm{sim}}\rangle$ calculated from simulations with parameters $\{\mu_{i,\mathrm{case}}\}$ 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 $\{\mu_{i,\mathrm{boot}}\}$. In contrast to our simulation, now the σ_{u_i} differ from sector to sector.

Having found a set of parameters $\{\mu_{i,\mathrm{case}}\}$ satisfying the above requirements, we use them to calculate conservative estimates $\mathsf{C}_{x\sigma}$ from the formula Eq. (5).

As the requirements i) – iii) cannot be solved directly for the $\{\mu_{i,\mathrm{case}}\}$, we use an iterative routine to determine them. The details of this routine are as follows. In the iterative loop A)–C), we determine the relative size of the $\{\mu_{i,\mathrm{case}}\}$ while keeping their overall size fixed through the requirement $\langle\lambda_{K,\mathrm{sim}}\rangle \equiv \lambda_{K,\mathrm{canonical}}$. We choose $\{\mu_{i,\mathrm{boot}}\}$ as initial values for the $\{\mu_{i,\mathrm{case}}\}$ and iterate the following steps A) to C) of the routine until convergence is reached.

A) We use the parameters $\{\mu_{i,\mathrm{case}}\}$ to calculate $\langle\mathbf{u}^{(K)}_{\mathrm{sim}}\rangle$ and the $\{\sigma_{u_i}\}$ 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

$$\langle u^{(K)}_{i,\mathrm{sim}}\rangle_{\mathrm{ideal}} = \begin{cases} u^{(K)}_{i,\mathrm{canonical}} + x\ \sigma_{u_i} & \text{if } \frac{1}{\sqrt{K}} - u^{(K)}_{i,\mathrm{canonical}} > x\ \sigma_{u_i} \\ u^{(K)}_{i,\mathrm{canonical}} - x\ \sigma_{u_i} & \text{if } u^{(K)}_{i,\mathrm{canonical}} - \frac{1}{\sqrt{K}} > x\ \sigma_{u_i} \\ \frac{1}{\sqrt{K}} & \text{otherwise} \end{cases} \tag{11}$$

As these "ideal values" depend on the $\{\sigma_{u_i}\}$ which in turn are functions of the $\{\mu_{i,\mathrm{case}}\}$, it is not useful to impose the conditions Eq.(11) directly. Instead, we choose an iterative approach and define auxiliary quantities

$$v_i \ = \ \langle u^{(K)}_{i,\mathrm{sim}}\rangle + \eta\Big(\langle u^{(K)}_{i,\mathrm{sim}}\rangle_{\mathrm{ideal}} \ - \ \langle u^{(K)}_{i,\mathrm{sim}}\rangle\Big), \tag{12}$$

which we normalize to unity before proceeding. For our actual calculations, the choice $\eta \approx 0.1$ turned out to be a good compromise between achieving a fast convergence (favors large values of η) and avoiding oscillatory limit cycles of the iterative algorithm (demands small values of η).

C) Next, we calculate a new set of parameters $\{\mu_{i,\text{case}}\}$, which satisfy the equation $\langle \mathbf{u}_{\text{sim}}^{(K)} \rangle = \mathbf{v}$ 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 $G : \{\mu_i\} \to \{\langle \lambda_{K,\text{sim}} \rangle, \langle \mathbf{u}_{\text{sim}}^{(K)} \rangle\}$ is only defined via a Monte Carlo simulation and hence cannot easily be inverted. For the inversion of G se again appendix A.

The new parameters are defined by

$$\{\mu_{i,\text{case}}\} \;=\; G^{-1}\Big(\lambda_{K,\text{canonical}}, \mathbf{v}\Big) \;. \tag{13}$$

In the iterative loop, this new set of parameters is used as input for step A.

To achieve both fast convergence and reliable results, we increase the number N of Monte Carlo simulations in A) from 10^3 to 10^5, as the parameters $\{\mu_{i,\text{case}}\}$ converge to their final values. We stop the iterative routine when the total change resulting from five successive iterations is smaller then one percent. For each value of $x = 1, 2, 3$, we save the vectors $\mathbf{v}_{1\sigma}$, $\mathbf{v}_{2\sigma}$, $\mathbf{v}_{3\sigma}$ from the last iteration cycle and denote them by $\mathbf{v}_{\text{case}}$ when using them in the routine to calculate the overall size of the $\{\mu_{i,\text{case}}\}$.

This iterative routine contains the following steps D)–F), as start values we use the $\{\mu_{i,\text{case}}\}$ from the last iteration of Eq. (13).

D) Via a Monte Carlo simulation, we 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

$$\langle \lambda_{K,\text{sim}} \rangle_{\text{ideal}} \;=\; \lambda_{K,\text{canonical}} + x\,\sigma_\lambda \quad . \tag{14}$$

Again, as σ_λ is a function of the $\{\mu_{i,\text{case}}\}$, it is not useful to enforce the relation Eq. (14) directly. Instead, we define an auxiliary "largest eigenvalue" which contains a small correction

$$\kappa \;=\; \langle \lambda_{K,\text{sim}} \rangle + \eta\Big(\langle \lambda_{K,\text{sim}} \rangle_{\text{ideal}} - \langle \lambda_{K,\text{sim}} \rangle\Big) \quad . \tag{15}$$

F) Using the inversion G^{-1} of the mapping $G : \{\mu_i\} \to \{\langle \lambda_{K,\text{sim}} \rangle, \langle \mathbf{u}_{\text{sim}}^{(K)} \rangle\}$ defined in appendix A, we can now calculate a new set of parameters

$$\{\mu_{i,\text{case}}\} \;=\; G^{-1}\Big(\kappa, \mathbf{v}_{\text{case}}\Big) \;. \tag{16}$$

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

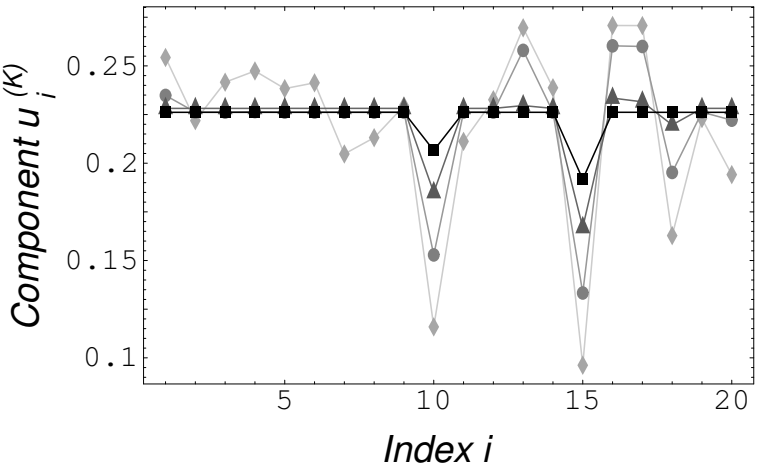

FIG. 4: Comparison between the eigenvector $\mathbf{u}^{(K)}_{\rm boot}$ (diamonds) and the conservative estimates $\mathbf{u}^{(K)}_{1\sigma}$ (circles), $\mathbf{u}^{(K)}_{2\sigma}$ (triangles), and $\mathbf{u}^{(K)}_{3\sigma}$ (squares).

Having derived the sets $\{\mu_{i,\rm case}\}$ which satisfy the conditions $\langle \mathbf{u}^{(K)}_{\rm sim}\rangle = \langle \mathbf{u}^{(K)}_{\rm sim}\rangle_{\rm ideal}$, $\langle \lambda_{K,\rm sim}\rangle = \langle \lambda_{K,\rm sim}\rangle_{\rm ideal}$ to the desired accuracy, we use them in the relation Eq. (4) to derive the "true" infinite time series correlation matrices $\mathsf{C}_{1\sigma}$, $\mathsf{C}_{2\sigma}$, and $\mathsf{C}_{3\sigma}$ for each case. We diagonalize these matrices and calculate their largest eigenvalues $\lambda_{K,1\sigma} = 12.4$, $\lambda_{K,2\sigma} = 13.1$, and $\lambda_{K,3\sigma} = 13.6$. We indeed see that with increasing safety margin the largest eigenvalue grows.

The components of the corresponding eigenvectors $\mathbf{u}^{(K)}_{1\sigma}$, $\mathbf{u}^{(K)}_{1\sigma}$, and $\mathbf{u}^{(K)}_{3\sigma}$ are shown in Fig. (4) together with the components of $\mathbf{u}^{(K)}_{\rm boot}$. 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 components of $\mathbf{u}^{(K)}_{\rm boot}$ fluctuate significantly, the components of $\mathbf{u}^{(K)}_{1\sigma}$ fluctuate less, and $\mathbf{u}^{(K)}_{3\sigma}$ is closest to the null hypothesis of equal components.

Economic implications of the different correlation matrices

In the last section we have described five different estimates for the cross correlation matrix, i.e. $\mathsf{C}^{\rm var}_{\rm canonical}$, $\mathsf{C}^{\rm var}_{\rm boot}$, $\mathsf{C}_{1\sigma}$, $\mathsf{C}_{2\sigma}$, and $\mathsf{C}_{3\sigma}$. To judge the economic implications of these estimates, we study the differences in the loss distribution resulting from these correlation estimations. To do this, we quantify the impact of the different correlation estimates by calculating their influence on CreditVaR and the conditional expectation over the CreditVaR, i.e. the expected shortfall.

The portfolio we 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 bn Euro range with a largest exposure of 750 mn Euro and a smallest exposure of 0.13 mn Euro. The counterpart specific default probability varies between 0.03% and 7%, the expected loss for the total portfolio is 187 mn Euro. Our primary aim 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.

Table I shows the CreditVaR and expected shortfall calculated by using CreditRisk+ and the method of Bürgisser et al. [1], which uses momentfitting (of the first two moments) to integrate correlations: instead of the original set of factors, one uses *one* synthetic factor Z with a variance σ_Z^2 that mimics the portfolio-loss expectation and variance for the correlated factors.

Correlation matrix	CreditVaR	Expected Shortfall
Independence	1.078 (983)	1.209 (1117)
$\mathsf{C}^{\text{var}}_{\text{canonical}}$	1.299 (1186)	1.460 (1348)
$\mathsf{C}^{\text{var}}_{\text{boot}}$	1.283 (1172)	1.441 (1331)
$\mathsf{C}_{1\sigma}$	1.314 (1200)	1.478 (1365)
$\mathsf{C}_{2\sigma}$	1.340 (1223)	1.509 (1392)
$\mathsf{C}_{3\sigma}$	1.366 (1246)	1.539 (1419)
One sector	1.561 (1417)	1.769 (1625)

TABLE I: Analysis of CreditVaR and expected shortfall at level 99.95% (99.90%) for different correlation matrices [in billion Euro]

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 uncertainty, we use the bias-corrected point estimate $\mathsf{C}^{\text{var}}_{\text{boot}}$ as a starting point and add volatilities 1σ, 2σ, and 3σ to the correlation estimate. (The bias correction accounts for a reduction of 16 mn Euro capital as compared to $\mathsf{C}^{\text{var}}_{\text{canonical}}$ on the 99.95% level.) When applying a one–σ estimate, the CreditVaR increases by 31 mn Euro, for the two–σ estimate there is another increase by 27 mn Euro, and using the three–σ estimate the CreditVaR

increases by yet another 26 mn Euro (all at 99.95% level). To put these numbers in perspective, we note that the CreditVaR without including correlations is found to be 1.078 bn Euro, and that the assumption of full correlations among all sectors leads to a CreditVaR of 1.561 bn Euro. The effects on the 99.90% confidence level for the CreditVaR as well as for the expected shortfall are similar.

We believe that 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. Even more important, we expect our conservative method of parameter estimation to provide smooth correlation estimates in the sense that new observations – occurring as times 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, we have 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 like a one-factor model is necessary. However, when using such a model to calculate the corresponding correlation matrix, one typically observes still 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.

We would like 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.

Acknowledgement: We would like to thank A. Müller-Groeling for initiating this project. We thank S. Lösch, C. von Lieres und Wilkau, and A. Wilch for useful discussions. The views expressed here are those of the authors and do not necessarily reflect the opinion of WestLB AG. The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

---

[1] Bürgisser, P., A. Kurth , A. Wagner, and M. Wolf, *Integrating Correlations*, Risk magazine, **12(7)**, 57–60 (1999).

[2] Nagpal, K., and R. Bahar, *Measuring default correlation*, Risk magazine, **14(3)**, 129–132 (2001); *Modelling default correlation*, Risk magazine, **14(4)**, 85–89 (2001).

[3] Engle, R.F., *Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation*, Econometrica, **50**, 987–1007 (1982).

[4] Frey, R., A. McNeil, and M. Nyfeler, *Copulas and credit models*, Risk magazine, **14(10)**, 111–114 (2001).

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

[6] Gordy, M.B., *A risk-factor foundation for ratings-based capital rules*, Journal of Financial Intermediation **12**, 199-232 (2001).

[7] Franses, P.H., *Times Series Models for Business and Economic Forecasting*, Cambridge University Press, 1998.

[8] Dickey, D.A., Fuller, W.A., Distribution of Estimators for Autoregressive Tine Series with Unit Root, *Journal of the American Statistical Association*, **74**, 427-431 (1979).

[9] Laloux, L., P. Cizeau, J.-P. Bouchaud, and M. Potters, *Random Matrix Theory*, Journal of Risk **12**, 69 (1999); see also L. Laloux et al., Physical Review Letters **83**, 1467 (1999).

[10] 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*, Physical Review Letters **83**, 1471 (1999).

[11] Abromowitz, M., and I.A. Stegun eds.,*Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Table*, Dover, New York (1970).

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

[13] Taking a look at the whole sample, the minimum number of entities under default risk per sector was 570, whereas the maximal number was 47858. Averaging over time, the minimum was 648, and the maximum 47084. The median was 2569 and the mean was 10296. For the insolvency frequencies, the time-average minimum for the sectors is 0.1%, the maximum 2.1%, and the median 1.2%.

[14] Throughout the text curly brackets indicate the discrete set that arises when all index combinations are considered in a sensible order, e.g. here for $i = 1, \ldots, K$ and $t = 1, \ldots, T$.

[15] Throughout the text bold characters indicate vectors.

APPENDIX A: 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 $G : \{\mu_i\} \to \{\langle\lambda_{K,\mathrm{sim}}\rangle, \langle\mathbf{u}^{(K)}_{\mathrm{sim}}\rangle\}$ which was defined via Monte Carlo simulations in the section on "Fluctuations in empirical correlation matrices - a simulation study".

To find the inversion algorithm, we first describe an analytic approximation to G. First, we calculate the expectation value $\mathrm{E}(\mathsf{C}^{\mathrm{sim}})$ by averaging the variances in Eq. (5) with respect to the distribution Eq. (7). We have numerically convinced ourselves that the largest eigenvalue $\lambda_{K,\mathrm{E}(\mathsf{C}^{\mathrm{sim}})}$ and corresponding eigenvector $\mathbf{u}^{(K)}_{\mathrm{E}(\mathsf{C}^{\mathrm{sim}})}$ of $\mathrm{E}(\mathsf{C}^{\mathrm{sim}})$ are good approximations (error of the order of one percent) to $\langle\lambda_{K,\mathrm{sim}}\rangle$ and $\langle\mathbf{u}^{(K)}_{\mathrm{sim}}\rangle$ and hence proceed to calculate them. To this end, we introduce the parameterization

$$\mathrm{E}(\mathsf{C}^{\mathrm{sim}}) \;=\; \delta_{ij} \;+\; (1-\delta_{ij})\;\alpha\;\beta_i\;\beta_j\;, \quad \text{with} \quad \sum_{i=1}^{K}\beta_i^2 = 1\;. \tag{A1}$$

The parameters are given by

$$\sqrt{\alpha}\;\beta_i \;=\; \mathrm{E}\Big(\Big(1+\frac{\hat{\sigma}^2_{\epsilon_i}}{\sigma^2_Y}\Big)^{-1/2}\Big)\;. \tag{A2}$$

The expectation value is defined with respect to the distribution Eq. (7). We now specialize to the practically relevant situation $T = 6$ and define a function

$$\begin{aligned} g(x) \;&=\; \mathrm{E}\Big(\Big(1+\frac{\hat{\sigma}^2_{\epsilon_i}}{\sigma^2_Y}\Big)^{-1/2}\Big)\Big|_{\mu_i=x} \\ &=\; \int_0^\infty f_{\chi^2,5}(\eta)\;\frac{1}{\sqrt{1+\eta\frac{x}{5}}}\;d\eta \end{aligned}$$

$$= \frac{1}{\Gamma(\frac{5}{2})2^{5/2}} \int_0^\infty \frac{\eta^{3/2}\ e^{-\eta/2}}{\sqrt{1+\frac{x}{5}\eta}}\ d\eta$$

$$= \frac{25}{6}\sqrt{\frac{5}{2\pi}} x^{-5/2} e^{5/(4x)} \Big[K_0\Big(\frac{5}{4x}\Big) + \Big(-1+\frac{2x}{5}\Big) K_1\Big(\frac{5}{4x}\Big)\Big] \tag{A3}$$

such that $\sqrt{\alpha}\beta_i = g(\mu_i)$. Here, $\Gamma(x)$ denotes the gamma function, and $K_\nu(x)$ denotes the modified Bessel function of the second kind [11]. Next, we approximately calculate the eigenvalue $\lambda_{K,\mathrm{E}(\mathsf{C}^{\mathrm{sim}})}$ by approximating $u^{(K)}_{i,\mathrm{E}(\mathsf{C}^{\mathrm{sim}})} \approx \beta_i$

$$\sum_{j=1}^{K} \mathrm{E}(C^{\mathrm{sim}}_{ij})\ \beta_j \ = \ \Big(1\ -\ \alpha\ \beta_i^2\Big)\beta_i\ +\ \alpha\ \beta_i$$

$$\approx\ \Big(1 - \frac{\alpha}{K}\ +\ \alpha\Big)\beta_i \tag{A4}$$

The above approximation is justified because the replacement $\beta_i^2 \to \frac{1}{K}$ is made in a sub-leading term ($\beta_i^2 \ll 1$). We now identify

$$\langle \lambda_{K,\mathrm{sim}} \rangle\ \approx\ 1\ +\ \alpha\Big(1\ -\ \frac{1}{K}\Big)\ . \tag{A5}$$

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

$$\mu_i\ =\ g^{-1}\Big(\sqrt{\frac{\langle \lambda_{K,\mathrm{sim}} \rangle - 1}{1-\frac{1}{K}}}\ \langle u^{(K)}_{i,\mathrm{sim}} \rangle\Big)\ . \tag{A6}$$

We have convinced ourselves numerically that the approximations involved in calculating G^{-1} give rise to errors of about one percent.