Geostatistical modeling of financial data: Estimation of large covariance matrices and imputation of missing data

Amelie Hüttner*  Matthias Scherer  Benedikt Gräler

March 7, 2019

We explore how the joint modeling of financial assets can utilize methodologies from geostatistical modeling. The considered approach is essentially based on modeling data as realizations of a (Gaussian) random field. This allows for a parsimonious representation of the dependence structure by means of a covariance function taken to be a function of the distance between observations. A key benefit of this ansatz is the possibility to include new data points, i.e. to consider new companies in financial applications. Consequently, geostatistical modeling has appealing benefits in the contexts of covariance matrix estimation and missing data imputation. We thoroughly discuss the necessary adjustments when applying geostatistical methods to the high-dimensional framework that entails the modeling of financial data, instead of the 2D/3D coordinate space encountered in original applications of the method. We illustrate the two use cases of covariance matrix estimation and missing data imputation on a data set of CDS spreads of constituents of the iTraxx universe.

AMS subject classification: 60G15, 60G60, 62H11, 62H20, 62M30, 91G40, 91G70
JEL subject classification: C21, C51, C53, C55, C58, G10, G17
Keywords: Geostatistics, Gaussian random field, Covariance matrix estimation, Missing data imputation

1 Introduction

The joint modeling of dependent credit spreads is crucial in many applications throughout portfolio optimization, quantitative risk management, and the pricing of portfolio credit derivatives. The challenges encountered are numerous: Credit portfolios often consist of a large number of constituents, which additionally may change over time, and unlike for equity data that is readily available for a large number of firms, data availability and quality is often still an issue for credit derivatives. In the light of these problems, a tractable and parsimonious, yet realistic ansatz for the modeling of the dependence structure that is able to cope with changing constituents and missing data is needed.

*corresponding author: amelie.huettner@tum.de
**Existing approaches:** With regards to the pricing of portfolio credit derivatives, many authors have focused on the joint modeling of default times, see Burtschell et al. (2009); Meissner (2008) (and the references therein) for an overview. However, Cont and Kan (2011) find that the main driver of losses from credit portfolios is not default risk, but rather the risk of a (joint) downturn of the credit spreads. Modeling approaches focusing on this spread risk are, among others, Cont and Kan (2011), who focus on accurate modeling of the stylized facts of CDS spreads / spread returns via time series models for the single-name spread (return) series and a multivariate t-distribution for the residuals, Oh and Patton (2018), who measure systemic risk via CDS spreads utilizing a factor copula model, and Brechmann et al. (2013) and Geidosch and Fischer (2016), who use vine copula models for stress-testing of CDS and risk modeling of credit portfolios, respectively. A common issue in these approaches is the need to re-estimate the models each time a new firm enters (or leaves) the portfolio, which may be computationally costly or even impossible in cases where little to no information on the new firm is available, and may change the modeling of the previously included firms. Cont and Kan (2011) resort to quasi maximum likelihood instead of maximum likelihood estimation to circumvent this problem. Further, a complete record of data is typically required for the estimation of the dependence structure, i.e. in case of a missing observation in one series, the whole data set is truncated. We study a novel approach borrowed from geostatistics that is able to overcome these issues.

**How can financial modeling benefit from spatial approaches?** The transfer of certain spatial statistical methods to applications involving financial data has already been discussed in the literature: Several studies applying spatial models to stock data, e.g. Asgharian et al. (2013); Arnold et al. (2013); Fernandez (2011); Fernández-Avilés et al. (2012); Kou et al. (2016), reveal that spatial dependencies between the underlying entities can provide deeper insights about the dependence structure of stock returns. So far, the main focus of applications lies on contagion and systemic risk. The aspect of spatial dependence inherent in credit spreads has been considered in Blasques et al. (2016) and Keiler and Eder (2013). Both studies employ spatial autoregressive (SAR) models. These models eventually follow a regression approach, where neighboring dependent variables are included in the regression using a spatial weights matrix, which is typically constructed from a distance measure and then row standardized.

In contrast to the studies mentioned above, our approach for incorporating spatial information is borrowed from geostatistics. These techniques were originally designed to forecast data on two- or three-dimensional surfaces from few sample observations and are based on the assumption of an underlying (Gaussian) random field with a low-parametric distance-dependent covariance function, which has appealing implications for the modeling of the dependence structure. There are only few references applying this framework to financial data: Fernández-Avilés et al. (2012) investigate co-movements in stock markets, whereas Arbia and Di Marcantonio (2015) employ such techniques to forecast interest rates.1 We find that for CDS data, geostatistics is a useful tool for estimation of large covariance/correlation matrices, as it supplies a natural estimator which is guaranteed to be positive definite and is able to provide estimators for CDS correlation coefficients of firms with a nonexistent record of CDS spreads. Further, we show that geostatistics supplies a useful tool for missing data imputation which compares favorably to existing

---

1Gaussian random fields have also been applied in interest rate modeling, see Kennedy (1994, 1997).
approaches.

The remainder of the article is organized as follows: Section 2 introduces geostatistical modeling and discusses necessary adjustments when applying it to financial data. Section 3 demonstrates its use for covariance matrix estimation and for the imputation of missing data on a data set of 98 time series of CDS par spreads. Section 4 offers an outlook on possible future research and concludes.

2 Geostatistics

In the following, we provide a brief introduction to the considered modeling approach as applied in a typical problem in geostatistics. The approach is essentially the same for financial data sets, however, complications arise from the necessity of higher-dimensional coordinate systems. These challenges and how they are overcome in the present paper is explained in detail in a later part of this section.

Throughout the article, \( n \) denotes the number of firms, \( d \) denotes the dimension of the considered coordinate space, and \( N \) is the number of observations of the respective time series.

2.1 A short introduction of the method

We give a short introduction to geostatistics, highlighting the necessary adjustments for its use in financial applications. For a detailed overview of the described techniques and some examples for their practical application in geosciences, see, e.g., Cressie (1993).

A typical problem in geosciences is to predict the concentration of a variable of interest at a prespecified location, given measurements of this variable at a set of other locations. This is done by assuming the observations to be taken from a random field

\[
Z = \{Z(s) : s \in \mathbb{R}^d\}, \quad d \leq 3,
\]

which should further satisfy some (weak) stationarity assumptions. Inference is based only on the mean and the covariance function (resp. variogram function as introduced below), which is assumed to be a function of the (absolute value of the) distance vector \( h \) between the observations’ locations. This reflects Tobler’s ‘first law of geography’, cf. Tobler (1970), which essentially states that observations made at close locations are related more strongly than observations made at distant locations.

In the present article, we assume that \( Z \) is a stationary Gaussian random field. This has the convenient implications that \( Z \) is completely determined by its mean and covariance function, and the dependence structure corresponds to a Gaussian copula essentially parameterized by the pairwise distances between the observations. Further, the simple kriging predictor, the best unbiased linear predictor at a new location in case the mean of the field is known, equals the mean of the (Gaussian) predictive distribution, cf. Cressie (1993) and Rasmussen and Williams (2006).

Modeling: It is common practice in geostatistics to express the dependence in terms of the so-called (semi-)variogram \( \gamma(h) \), which is closely related to the covariance function
$C(h)$ and defined as follows:

$$
\begin{align*}
\gamma(h) &:= \frac{1}{2} \mathbb{V}[Z(s + h) - Z(s)], \\
C(h) &:= \text{Cov}(Z(s + h), Z(s)), \\
C(h) &= C(0) - \gamma(h), \quad s, h \in \mathbb{R}^d.
\end{align*}
$$

(1)

If the semivariogram $\gamma$ depends only on the absolute value $h := \|h\|$ of the distance vector, i.e. points on spheres around some center have the same semivariogram value, the model is called isotropic. Isotropy is often an unrealistic assumption, but in many cases a linear transformation of the underlying space restores isotropy in the model:

$$
\gamma(h) = \tilde{\gamma}(\|Ah\|), \quad h \in \mathbb{R}^d, A \in \mathbb{R}^{d \times d},
$$

where $A$ is said linear transformation and $\tilde{\gamma}$ is some isotropic semivariogram model. This is referred to as geometric anisotropy.

In the present article, we assume the field $Z$ to be isotropic.

Just as covariance functions are positive (semi-)definite, semivariogram functions are conditionally negative definite, i.e.

$$
\sum_{i=1}^{n} \sum_{j=1}^{n} a_i \cdot C(s_i - s_j) \cdot a_j \geq 0 \quad \forall n \in \mathbb{N}; s_i, \ldots, s_n \in \mathbb{R}^d; a_1, \ldots, a_n \in \mathbb{R},
$$

$$
\sum_{i=1}^{n} \sum_{j=1}^{n} a_i \cdot 2\gamma(s_i - s_j) \cdot a_j \leq 0 \quad \forall n \in \mathbb{N}; s_i, \ldots, s_n \in \mathbb{R}^d; a_1, \ldots, a_n \in \mathbb{R}, \sum_{i=1}^{n} a_i = 0.
$$

**Estimation:** The semivariogram is estimated from observations by

$$
\hat{\gamma}(h) = \frac{1}{\mathbb{N}(h)} \sum_{N(h)} (Z(s_i) - Z(s_j))^2,
$$

(2)

cf. Cressie (1990, p.69), where $N(h)$ is the set of all sample pairs $s_i, s_j$ separated by distance $h = \|h\|$. In practice, sample pairs are binned to achieve a more stable estimator. Once estimated from data, a valid model is fitted to the empirical variogram. The model ensures that the variogram is indeed conditionally negative definite.

Most parametric isotropic semivariogram models only depend on three parameters $\tau$, $\sigma$, and $\rho$, which define the so-called nugget $\tau^2$, (partial) sill $\sigma^2$, and the range $1/\rho$, see Cressie (1993); Arbia and Di Marcantonio (2015). The nugget captures the variation on scales smaller than the minimum distance between sample locations, as this will not be registered in the estimation of the empirical semivariogram. More formally, it is defined as the limit of $\gamma(h)$ for $h$ approaching zero. The sill is the limit of $\gamma(h)$ as $h$ grows large, i.e. the variance of the field, and the range is the value of $h$ where the sill is first reached.\(^2\)

\(^2\)Some models reach a sill only asymptotically, in which case the range represents the distance where a certain percentage of the sill is reached.
Two examples, the exponential and the Gaussian semivariogram, are given below.

\[
\gamma_{\text{Exp}}(h) = \begin{cases} 
0 & h = 0 \\
\tau^2 + \sigma^2(1 - \exp(-\rho h)) & h \neq 0,
\end{cases}
\]

\[
\gamma_{\text{Gau}}(h) = \begin{cases} 
0 & h = 0 \\
\tau^2 + \sigma^2(1 - \exp(-\rho^2 h^2)) & h \neq 0.
\end{cases}
\]

**Spatial Prediction:** With the fitted model, field values at unobserved locations \( s_{\text{new}} \) can be predicted via kriging, a procedure that essentially estimates a spatial mean. For a random field \( Z \) with known mean, the simple kriging predictor \( p(s_{\text{new}}) \) is defined as the best unbiased linear predictor minimizing the mean-squared prediction error \( E[(Z(s_{\text{new}}) - p(s_{\text{new}}))^2] \).

As \( Z \) is assumed to be Gaussian in the present article, i.e. the observed data are jointly Gaussian, one obtains a predictive distribution at the unobserved locations that is again Gaussian; its mean and variance equal the simple kriging predictor \( p(s_{\text{new}}) \) and the minimal mean-squared prediction error, also referred to as kriging variance, at the unobserved location \( s_{\text{new}} \).

**Remark 2.1** (Assumption of a known mean). *In the present article, we only consider the case where the mean of the field \( Z \) is known. In case it is unknown, it has to be estimated alongside the best unbiased linear predictor, which is referred to as ordinary kriging. For more involved approaches, e.g. more robust approaches when \( Z \) is not Gaussian, see Cressie (1993).*

To summarize, geostatistical modeling involves the following steps:

1. Estimate the sample variogram \( \hat{\gamma} \).
2. Fit a valid variogram model \( \gamma \).
3. Calculate model covariances based on distances and the fitted variogram model.

For financial applications, we have to add the non-trivial preliminary modeling task:

0. Define/select a suitable (financial) distance measure.

In classical geoscience applications, variogram estimation and fitting\(^3\), as well as kriging, can be executed using the gstat package for R, cf. Pebesma (2003). As the function for estimating the sample variogram is tailored to coordinate system dimension \( d \leq 3 \) and expects only a single measurement per location, we have implemented a variant thereof that obtains a distance matrix as input, thus being able to deal with arbitrary coordinate system dimension, and takes the repeated measurements into account for determining the sample variogram. In this context, pairwise complete observations are sufficient for sample variogram estimation. Thus, in case of missing data, we need not exclude days with missing observations completely.

\(^3\)We only consider parametric variogram models in this article.
2.2 Adaption to financial data

The crucial modeling step in translating geostatistical procedures to financial data is the definition of a suitable coordinate system, respectively distance measure. Informally speaking, we have to provide an answer to the question: How ‘far’ is company A from company B? In geosciences, one works with data measured at locations on a plane or sphere, and the association among the measurements is usually decreasing in distance. The Euclidean distance between the measurement locations furnishes a natural measure of closeness. For financial data, no such natural distance is available and it is a pivotal modeling step to design a meaningful one.

After an appropriate distance measure has been found, the question of model-fitting has essentially reduced to the fitting of a valid covariance function to the data. However, the validity of covariance functions is closely tied to the chosen coordinate system, and several parametric models used in geoscience applications are no longer applicable in the higher-dimensional coordinate systems $d > 3$ that we are considering.

2.2.1 Designing appropriate financial distance measures

Distance measures proposed in the literature: Several financial distance measures have already been proposed in previous studies: Fernández-Avilés et al. (2012) construct a financial distance between two countries based on foreign direct investments (FDI). Although it would be reasonable to assume a similar distance (dependence) structure for firms, this measure cannot be adapted to corporates as the relevant data is not publicly available. Asgharian et al. (2013) investigate the performance of several distance measures, namely exchange rate volatility, interest rate volatility, bilateral trade, and geographical distances, with bilateral trade having the highest explanatory power for spatial dependence. Due to the nature of the data set considered in the present study (only European firms, CDS denominated in Euro), it does not make sense to include exchange rate and interest rate based distances, and similar to FDI ties, data on bilateral trade between firms is not publicly available. Fernandez (2011) proposes financial distance measures based on correlations between covariates, namely the ratio of market cap to firm size, the market-to-book ratio, the ratio of total debt to total assets, debt maturity, and dividend yield, where the distance between entities $i$ and $j$ is calculated from their covariate correlation $\rho_{ij}$ following Mantegna and Stanley (2000, Ch.13):

$$d(i, j) = \sqrt{2(1 - \rho_{ij})}. \quad (3)$$

Taking $\rho_{ij}$ to be the sample variant of Pearson’s correlation coefficient, it can be seen from the derivation in Mantegna and Stanley (2000) that this corresponds to the Euclidean distance between the historical covariate vectors. Fernandez (2011) use Spearman’s correlation coefficient instead of Pearson’s, so their distance measure corresponds to the Euclidean distances between the rank-transformed covariate vectors.

Keiler and Eder (2013), one of the few other studies concerning spatial dependence with a focus on credit spread data, consider a financial distance measure based on equity correlations. Keiler and Eder (2013) apply an SAR model and work directly with the (row-standardized) equity correlation matrix as spatial weights matrix, thus the spatial influence is incorporated differently as in geostatistical models.

---

4 Fernandez (2011) use Spearman’s correlation coefficient instead of Pearson’s, so their distance measure corresponds to the Euclidean distances between the rank-transformed covariate vectors.

5 They apply an SAR model and work directly with the (row-standardized) equity correlation matrix as spatial weights matrix, thus the spatial influence is incorporated differently as in geostatistical models.
possibilities. Using it, however, forces us to drop few privately owned companies from our data set.

Our proposed distance measures:  We focus on the following distance measures:

1. Distance based on equity return correlations ($d_E$):
   Similarly to Fernandez (2011), we construct this distance using the Mantegna/Stanley ansatz (3). It corresponds to the Euclidean distance between the equity return vectors, thus the dimension $d$ of the coordinate space is very high, as it equals the length of the return time series. Consequently, many parametric variogram models are no longer valid in this coordinate space, and it is prudent to consider only variogram models that are valid in any dimension $d \in \mathbb{N}$, like the Gaussian or Exponential variogram\(^7\), to make sure one indeed fits a valid variogram model to the sample variogram.

2. Distances based on balance sheet data:
   The coordinate systems resulting from this ansatz are lower-dimensional, the dimension depending on the number of performance figures considered. Like in regression approaches, this might suffer from collinearity in the coordinates. By choosing performance figures with moderate correlation among firms, this issue can be minimized. Another issue, whose impact can be quite severe, is the differing scales of the chosen coordinates: If there is one performance figure which is a lot larger than the others, the Euclidean distance between the vectors of performance figures will be predominantly influenced by this largest figure, thus distorting the spatial influence. At the very least, the coordinates have to be normalized in scale to avoid this issue. Alternatively, one could use the Mahalanobis distance between locations $s_1$ and $s_2$,

\[
d_M(s_1, s_2) := \sqrt{(s_1 - s_2)'\Sigma^{-1}(s_1 - s_2) = d_{\text{Eucl.}}(\Sigma^{-\frac{1}{2}}s_1, \Sigma^{-\frac{1}{2}}s_2),}
\]

where $\Sigma$ is the covariance matrix of the coordinates and $d_{\text{Eucl.}}$ refers to the Euclidean distance. This corresponds to a ‘whitening’ of the data, cf. Kulis (2012).

We consider distances based on balance sheet data for changes in working capital (CWC), EBITDA, net debt, capital expenditures (CapEx), and total revenue, all divided by the respective firm’s market cap to minimize the influence of the firms’ sizes, as well as sector and country categorization, following Markit’s sector categorization supplied for ITRX, resp. Reuter’s information on sector and country. Their correlation matrix is

\[
\begin{pmatrix}
\text{CWC} & \text{EBITDA} & \text{Net Debt} & \text{CapEx} & \text{Total Revenue} \\
1 & -0.3133 & -0.4027 & 0.0737 & 0.2060 \\
-0.3133 & 1 & 0.5873 & -0.9040 & 0.6511 \\
-0.4027 & 0.5873 & 1 & -0.5274 & 0.1809 \\
0.0737 & -0.9040 & -0.5274 & 1 & -0.6697 \\
0.2060 & 0.6511 & 0.1809 & -0.6697 & 1
\end{pmatrix}
\]

\(^6\)A possible solution to this issue would be to use an equally weighted index of equity return data of peer companies as supplied e.g. by Reuters.

\(^7\)To date, these are the only two variogram functions from this class available for variogram fitting in R’s gstat package. This was not a severe limitation, as in our data set, the Gaussian variogram provided a good fit to the sample variograms.
with all coordinates ratios to market cap as indicated above. EBITDA and CapEx are highly correlated, which introduces numerical problems in the computation of the Mahalanobis distance, as the (sector-extended) covariance matrix of coordinates is close to singular. Hence, for the Mahalanobis distance based on balance sheet data and sector information, EBITDA/market cap is excluded from the set of coordinates.\(^8\) To summarize, the considered distance measures are:

- Euclidean distance, considering balance sheet data and sector information as coordinates \(d_{\text{FRS,Eucl.}}\);
- Euclidean distance, considering balance sheet data, sector, and country information as coordinates \(d_{\text{FRSC,Eucl.}}\);
- Mahalanobis distance, considering balance sheet data as coordinates \(d_{\text{FR,M}}\);
- Mahalanobis distance, considering balance sheet data and sector information as coordinates \(d_{\text{FRS,M}}\).

Sector and country information were encoded as 0-1 variables. As for equity data, few firms had to be removed from our data set due to missing balance sheet data. A more severe cut to our data set, however, is the necessary removal of all firms from the financial sector, as for a meaningful characterization of these, different balance sheet data is required.\(^9\)

Figure 1 illustrates the relation between CDS log return correlations and the considered distance measures via hexbin plots. For a meaningful distance measure, a decline of CDS correlation with distance is expected, which is clearly visible in the hexbin plots for \(d_E\), \(d_{\text{FR,M}}\), and \(d_{\text{FRS,M}}\), but less pronounced for \(d_{\text{FRS,Eucl.}}\) and \(d_{\text{FRSC,Eucl.}}\).

**Remark 2.2** (Geometric anisotropy). The issue of different scales of the coordinates is quite similar to the issue of (geometric) anisotropy: Typically, in \(d \in \{2, 3\}\), data-driven approaches, namely the estimation of directional variograms\(^10\), are employed to estimate (geometric) anisotropy parameters. In this context, the sill is assumed to be constant over all directions, only the range varies. This approach, however, requires a certain number of data points for each directional variogram to be feasible, a requirement which is getting increasingly hard to fulfill in higher dimensions. Instead, metric learning, a different data-driven approach borrowed from machine learning applications, cf. Kulis (2012), can be employed to estimate geometric anisotropy in higher dimensions. Indeed, linear metric learning exactly corresponds to a formal estimation of the geometric anisotropy matrix \(A\): Given a set of data points in a \(d\)-dimensional coordinate space and information on a more appropriate distance than the Euclidean, typically in the form of class labels or, simply put, information whether two data points are considered similar or not, linear metric learning approaches infer from the data a linear transformation of the space such that the Euclidean distance in the transformed space more appropriately reflects Tobler’s law.

---

\(^8\)For all other distances, no numerical problems arise, and sample variograms with both the EBITDA and CapEx ratio included differ only marginally from those obtained when one of these ratios is excluded from the set of coordinates.

\(^9\)The balance sheet of a firm in the financial sector differs fundamentally from one of a firm outside the financial sector.

\(^10\)Here, distance vectors \(\mathbf{h}\) are used for sample variogram estimation in Equation (2), instead of norms \(h = \|\mathbf{h}\|\) of distance vectors as in the isotropic case.
Figure 1: Hexbin plots of CDS log return correlations vs. the chosen distance measures. The darker the respective hexagonal field, the more observations it represents.

### 2.2.2 Variogram estimation and the fitting of a valid model

Due to the fact that daily observations of the field are available, and due to the higher-dimensional coordinate spaces, the typical geostatistics procedures have to be tailored to financial data sets. In the following, we explain the issues encountered in variogram estimation and fitting, and explain the necessary adjustments.

**Sample variogram estimation:** In geosciences, one often has only a single observation of the field, or at best a short time series. Financial data, in contrast, is available daily (or at even shorter time intervals), thus many observations of the field are available. These observations are marred by temporal dependence in the single series, which has to be removed if one wants to take only pure spatial dependence into account. We deal with this problem by working with log-returns, which are often considered to be approximately free of spatial dependence, or with the residuals of different time series models fitted to the log-return series. This is explained in greater detail in Section 3.1, where our data set is introduced. Another, more complicated approach is to resort to space-time modeling of the field, which will not be discussed here.

Having removed temporal dependence, we obtain independent observations $Z(s_i, t)$ of the field at location $i$ and time $t$, and we estimate the sample variogram similar as in Equation (2):

$$
\hat{\gamma}(h) = \frac{1}{2|N(h)|N} \sum_{s_i, s_j \in N(h)} \sum_{t=1}^{N} (Z(s_i, t) - Z(s_j, t))^2,
$$

where, as in Equation (2), $N(h)$ refers to the set of firm pairs with (absolute value of) distance (vector) $h = ||h||$ and $N$ is the number of (daily) observations for each firm.
**Variogram model fitting:** When fitting a valid variogram model to the sample variogram, we have two major issues to consider: First, compatibility of the variogram model with the metric used, and second, compatibility of the variogram model with the dimension of the considered coordinate space.

The first issue is discussed in Christakos and Papanicolaou (2000): Essentially, they find that not every variogram model is feasible in combination with any metric. If one decides to use another metric than the Euclidean (e.g. Hamming), one has to check after fitting whether the model is indeed compatible with the chosen distance. We circumvent this issue by focusing on Euclidean distances.\(^\text{11}\)

For the second issue, one has to keep in mind that valid variogram models in dimension \(d_2\) are also valid in dimension \(d_1\) for \(d_1 \leq d_2\), but not vice versa, cf. Cressie (1993, Ch. 2.5.2). If one has chosen a large number of coordinates and a second-order stationary isotropic field, it might be more convenient to work with variogram functions corresponding to (resp. correlation functions from) the well-studied class of correlation functions valid in any dimension, cf. Abrahamsen (1997); Yaglom (1987); Schoenberg (1938). Using \(R\)'s \texttt{gstat} package, this means one is limited to the exponential and Gaussian semivariograms introduced above. Gaussian semivariograms are shied from in geosciences, as these imply very smooth surfaces of the corresponding field, which is considered unrealistic in these applications. However, a-priori there is no reason why this should be an unrealistic behaviour for financial data. Indeed, using the high-dimensional equity return correlation distance, the Gaussian semivariogram fits the data very well.

In some cases one may want to fit a custom variogram function to the sample variogram obtained. A brief overview on how to check if the fitted function is indeed a valid variogram (resp. correlation function) is provided in the Appendix, along with further useful references.

### 3 Applications

The possible applications of the presented method are manifold. In the following, we focus on the problem of covariance matrix parameterization and on the interpolation of missing data. Another possible application, which is not discussed in the present paper, is the prediction of multivariate credit spreads.

#### 3.1 Data set

We illustrate the benefits of incorporating spatial information via geostatistical modeling for covariance matrix estimation and missing data imputation using a data set of CDS spreads, stock prices, and macroeconomic variables referring to the reference entities of the CDS listed in the Markit iTraxx Europe (ITRX IG; investment grade) and iTraxx Europe Crossover (ITRX XO; sub-investment grade) Indices Series 29. The ITRX IG (XO) indices contain the 125 (75) most liquidly traded (sub-) investment grade CDS and exist for maturities 3, 5, 7, and 10 years, with 5 years being the most liquidly traded maturity, similar to single-name CDS. The constituents are adjusted every year in March.

\(^{11}\)Recall that the Mahalanobis distances introduced above are Euclidean distances on a linearly transformed coordinate space, so these do not pose an issue.
and September, and a new on-the-run series is set up. Some of the reference entities are subsidiaries of publicly traded firms which are not traded themselves; in this case, the stock prices and macroeconomic variables of the public parent company are taken.

After removing firms with missing data, our data set comprises a complete record of CDS, stock, and macroeconomic data for 98 firms\textsuperscript{12} in the time period July 25th, 2016 to July 24th, 2018.

For out-of-sample performance evaluation of the proposed approach we use data of Deutsche Post, Henkel, Merck, Air Liquide, Bouygues, Capgemini, and Sodexo.

Remark 3.1. Estimation of the sample variogram requires only pairwise complete observations, as positive definiteness of the covariance (resp. correlation) matrix is ensured by fitting a valid model. Thus, missing CDS data is no knock-out criterion for the application of this method. We only remove time series with missing values to obtain a complete record for comparison in the missing data imputation application. For the coordinates, however, a complete record is required.

Preprocessing CDS data: The geostatistical modeling assumptions require stationary time series; we therefore consider log returns of CDS par spread quotes. Further, we intend to focus on purely spatial dependence. Hence, temporal dependence must be taken care of in some way. Indeed, the presence of temporal dependence in the series cannot be denied: Box tests show that most of the return series are autocorrelated, and heteroskedasticity effects are observed in more than half of the series. We opt for fitting time series models to each log return series. According to Cont and Kan (2011), an AR(1)-GARCH(1,1) is well suited for capturing observed stylized facts of CDS spread return time series.\textsuperscript{13} For comparison, we also carry out our analyses directly on the log returns, as it is a quite common assumption among practitioners that log returns are approximately free of temporal dependence. To comply with the assumption of a Gaussian field, and to be able to directly obtain correlation functions from the fitted variograms, the marginal distributions of each return and residual time series are transformed to the standard normal distribution.

Preprocessing equity data: As mentioned above, privately owned companies have to be removed from the data set. Where possible, stock price quotes were taken from the respective firms’ parent company. Further, 2 firms had to be excluded due to missing equity quotes. The distance measure $d_E$ is constructed from correlations of stock log returns.

\textsuperscript{12}This already serves as an illustration of the data availability issue for CDS: 57 firms had to be removed from the data set due to missing CDS quotes. Further, private companies were removed from the data set as no stock price is available; apart from this, only two firms had to be removed from the data set due to missing equity data. Finally, also companies operating in the financial sector were removed from the data set, as for those different macroeconomic variables are required.

Data obtained via Thomson Reuters Eikon, with CDS quotes from Markit (=MG).

\textsuperscript{13}We also tried fitting ARMA(p,q)-GARCH(1,1) with p,q optimally chosen, as AR(1)-GARCH(1,1) was not always the best-fitting model according to AIC/BIC, as well as ARMA(p,q) with p,q optimally chosen, as for about half of the return series homoscedasticity could not be rejected. However, the results of geostatistical model fitting/covariance matrix estimation/data imputation obtained from the residuals of these models were very similar to those obtained from the residuals of the AR(1)-GARCH(1,1), and are thus not presented in this paper for the sake of brevity.
**Preprocessing balance sheet data:** Balance sheet data is typically reported in yearly intervals. We choose the reported values\(^{14}\) of 2017 as supplied by Thomson Reuters Eikon for the construction of the remaining distance measures.

### 3.2 Covariance/Correlation matrix estimation

Covariance/Correlation matrix estimation is a problem often encountered in finance. It is especially challenging when the number of firms \(n\) is large, as the estimation of \(n(n+1)/2\) parameters (variances and pairwise correlations) is required.

The most natural estimator is the sample covariance matrix of the data set, but this is problematic for several reasons: First, it is singular when the number of observations is small compared to the number of firms considered, \(N < n\), cf. Ledoit and Wolf (2004a). Second, in the case of missing data, one has to make a choice between using only dates with a complete record, which might lead to the \(N < n\) issue, or compute the sample covariance matrix using pairwise complete observations, which may result in an estimator that is not positive definite. Third, for \(n\) large, the sample covariance matrix typically contains a lot of noise compared to the true covariance matrix of the market. Therefore, estimators that impose more structure on the estimated covariance matrix are desirable. At the very least, positive definiteness of the estimator is a necessary requirement. There exists an abundance of literature on this topic, e.g. the series of papers by Ledoit and Wolf (2004a, 2003, 2004b) focusing on shrinkage, Perreault *et al.* (2019) who simplify the covariance matrix by imposing an exchangeable block structure, principal component estimators as e.g. Alexander (2002), factor model estimators as described, e.g., in Fan *et al.* (2008), or Engle (2002)’s DCC model. This list is by no means exhaustive and just intends to give an overview of some of the proposed approaches. None of these approaches, however, is able to cope with missing data. In this case, a popular approach in the context of CDO modeling is to take the equity correlation matrix, cf. RiskMetrics Group (2007), which is inappropriate in the current market conditions, as we will see in the following.

**Natural covariance/correlation matrix estimator from geostatistics:** From Section 2 it can be seen that the geostatistical framework naturally encompasses an estimation procedure for large covariance/correlation matrices via the fitting of a valid variogram function. The resulting estimator has various appealing features: First, it is guaranteed to be positive definite (as the valid variogram is conditionally negative definite). Second, it is a parametric function of distance requiring only few parameters (3 at most in the parametric variogram functions considered here). Third, it can cope with missing data, i.e. given the coordinates, one can predict covariances/correlations of firms for which no data is available. Last but not least it is easy to compute, as it only requires one fit of the variogram function, which is then used for the estimation of all covariances/correlations. It is worth noting that, when estimating covariances via the fitted variogram, all firms have the same variance, which is somewhat unrealistic. This issue can be resolved by standardizing the observations of the field to mean 0 and variance 1, thus obtaining a correlation function from the fitted variogram via Equation (1), and estimating the variances separately for each series.

\(^{14}\)For firms who do not report at the end of December, we take the figure reported at the time closest to Dec. 31st, 2017.
The estimation procedure follows Steps 0.-3. described above: Having selected a (financial) distance measure, one collects the relevant coordinate data plus CDS spreads for a training set. In our case this is the ITRX universe, which consists of the most liquidly traded European CDS, and can hence be considered representative for estimating the correlation function governing the European CDS market. Then one calculates the sample variogram using Equation (2) and fits a valid variogram model, e.g. using R’s gstat package, paying attention to the issues described in Section 2. Finally, using relation (1), one obtains the covariances as functions of the distance.

In our use case, we standardize the field as indicated above, and constrain the nugget and sill parameters of the variogram such that \( C(0) = 1 \) and we obtain correlations as functions of the distance. The fitted variogram models for our chosen distances are given in Table 1 and their fit to the sample variograms with respect to our chosen distances is displayed in Figure 2 for the CDS log return data set. We find that the fitted models for \( d_{FRS, Eucl.} \) and \( d_{FRS, Eucl.} \), as well as the models for \( d_{FR, M} \) and \( d_{FRS, M} \), are very similar for both log returns and AR(1)-GARCH(1,1) residuals.

<table>
<thead>
<tr>
<th>distance</th>
<th>model type</th>
<th>nugget</th>
<th>(partial) sill</th>
<th>range</th>
</tr>
</thead>
<tbody>
<tr>
<td>( d_{E} ), returns</td>
<td>Gaussian</td>
<td>0</td>
<td>1</td>
<td>1.5350</td>
</tr>
<tr>
<td>( d_{FRS, Eucl.} ), returns</td>
<td>Gaussian</td>
<td>0.45</td>
<td>0.55</td>
<td>31.0424</td>
</tr>
<tr>
<td>( d_{FRS, Eucl.} ), returns</td>
<td>Gaussian</td>
<td>0.44</td>
<td>0.56</td>
<td>31.1891</td>
</tr>
<tr>
<td>( d_{FR, M} ), returns</td>
<td>Gaussian</td>
<td>0.42</td>
<td>0.58</td>
<td>16.8436</td>
</tr>
<tr>
<td>( d_{FRS, M} ), returns</td>
<td>Gaussian</td>
<td>0.45</td>
<td>0.55</td>
<td>19.7333</td>
</tr>
<tr>
<td>( d_{E} ), residuals</td>
<td>Gaussian</td>
<td>0</td>
<td>1</td>
<td>1.3868</td>
</tr>
<tr>
<td>( d_{FRS, Eucl.} ), residuals</td>
<td>Gaussian</td>
<td>0.5</td>
<td>0.5</td>
<td>27.2024</td>
</tr>
<tr>
<td>( d_{FRS, Eucl.} ), residuals</td>
<td>Gaussian</td>
<td>0.5</td>
<td>0.5</td>
<td>29.3347</td>
</tr>
<tr>
<td>( d_{FR, M} ), residuals</td>
<td>Gaussian</td>
<td>0.5</td>
<td>0.5</td>
<td>17.8945</td>
</tr>
<tr>
<td>( d_{FRS, M} ), residuals</td>
<td>Gaussian</td>
<td>0.48</td>
<td>0.52</td>
<td>16.2328</td>
</tr>
</tbody>
</table>

Table 1: Fitted variogram models for CDS log returns and AR(1)-GARCH(1,1) residuals for each of our chosen distance measures.

**Performance figures considered for the comparison of models:** We consider the following performance figures:

1. **Matrix-valued loss functions:** The Frobenius loss

\[
\text{trace}((C - \hat{C}^2),
\]

where \( C \) and \( \hat{C} \) are the sample and model covariance matrix, respectively, and the Negative normal log-likelihood

\[
\text{trace}(C\hat{C}^{-1}) - \ln ( \det(C\hat{C}^{-1})) - n,
\]

as Neuberg and Glasserman (2018) find this is a more suitable loss function from a portfolio perspective.

2. **Performance measures addressing the elements of the matrix\(^{15}\):** RMSE and MAPE

\(^{15}\)It is important to remark that MAPE has certain shortcomings as a measure of predictive accuracy, especially when used for model selection, cf. Gneiting (2011); Tofallis (2015). Nevertheless it remains quite popular among practitioners, which is why we include it in our comparison.
Figure 2: Fit of the valid variogram model to the respective sample variogram for the chosen financial distance measures.

for all pairwise correlations and for those of the newly added firms only:

\[
\text{MAPE}_{\text{all}} = \frac{1}{n(n-1)/2} \sum_{k=1}^{n} \sum_{l=k+1}^{n} \frac{\|C_{k,l} - \hat{C}_{k,l}\|}{\|C_{k,l}\|}, \quad C \in \mathbb{R}^{n \times n},
\]

\[
\text{MAPE}_{\text{new}} = \frac{1}{nm + \frac{m(m-1)}{2}} \sum_{k=n+1}^{n+m} \sum_{l=k+1}^{k-1} \frac{\|C_{k,l} - \hat{C}_{k,l}\|}{\|C_{k,l}\|}, \quad C \in \mathbb{R}^{(n+m) \times (n+m)},
\]

\[
\text{RMSE}_{\text{all}} = \sqrt{\frac{1}{n(n-1)/2} \sum_{k=1}^{n} \sum_{l=k+1}^{n} [C - \hat{C}]_{k,l}^2}, \quad C \in \mathbb{R}^{n \times n},
\]

\[
\text{RMSE}_{\text{new}} = \sqrt{\frac{1}{nm + \frac{m(m-1)}{2}} \sum_{k=n+1}^{n+m} \sum_{l=k+1}^{k-1} [C - \hat{C}]_{k,l}^2}, \quad C \in \mathbb{R}^{(n+m) \times (n+m)}.
\]

We compare the sample correlation matrices of our training set of 98 ITRX constituent firms (in-sample) and our training and test set including the 7 new firms listed in Section 3.1 (out-of-sample) to the model correlation matrices obtained from the fitted variograms for our chosen distance measures. For (normal-transformed) CDS log-returns, we include the (normal-transformed) equity correlation matrix for comparison. Both sample correlation matrices are well-defined here, as our training and test sets are without missing data with \(N > n\).

**In-sample comparison:** We find that for CDS log returns, simply taking the corresponding equity correlation matrix is the worst choice, cf. Table 2 and Table 3. Figure 3 reveals...
that this is due to the fact that equity correlations in our data set are systematically much lower than CDS correlations. Considering both matrix-valued and element-wise performance figures, the correlation models parameterized by the equity correlation distance $d_E$ and by the balance sheet ratios and sector-based Mahalanobis distance $d_{FRS,M}$ perform best for CDS log returns. For CDS AR(1)-GARCH(1,1) residuals, the correlation model parameterized by $d_{FRS,E_ucl.}$ performs best according to the Frobenius loss and pairwise correlations RMSE, whereas according to the negative normal log-likelihood and pairwise correlations MAPE, the models parameterized by $d_{FRS,M}$ and $d_{FR,M}$ perform best, respectively.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Equity</td>
<td>890.1386</td>
<td>-</td>
<td>37.9848</td>
<td>-</td>
</tr>
<tr>
<td>$d_E$</td>
<td><strong>73.8306</strong></td>
<td>72.7075</td>
<td>27.5979</td>
<td>24.1303</td>
</tr>
<tr>
<td>$d_{FRS,E_ucl.}$</td>
<td>84.5127</td>
<td>67.4280</td>
<td>28.3136</td>
<td>25.3291</td>
</tr>
<tr>
<td>$d_{FRS,E_ucl.}$</td>
<td>89.3108</td>
<td><strong>67.1137</strong></td>
<td>28.3793</td>
<td>25.3213</td>
</tr>
<tr>
<td>$d_{FR,M}$</td>
<td>98.0467</td>
<td>67.2878</td>
<td>28.6524</td>
<td>25.2926</td>
</tr>
<tr>
<td>$d_{FRS,M}$</td>
<td>79.8826</td>
<td>69.5488</td>
<td><strong>26.5629</strong></td>
<td><strong>23.6696</strong></td>
</tr>
</tbody>
</table>

Table 2: Results of the matrix-valued loss functions (Frobenius loss and negative normal log-likelihood) for CDS log returns and AR(1)-GARCH(1,1) residuals.

<table>
<thead>
<tr>
<th>distance</th>
<th>RMSE returns</th>
<th>RMSE residuals</th>
<th>MAPE returns</th>
<th>MAPE residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equity</td>
<td>0.3060</td>
<td>-</td>
<td>0.5400</td>
<td>-</td>
</tr>
<tr>
<td>$d_E$</td>
<td><strong>0.0881</strong></td>
<td>0.0875</td>
<td><strong>0.1464</strong></td>
<td>0.1523</td>
</tr>
<tr>
<td>$d_{FRS,E_ucl.}$</td>
<td>0.0943</td>
<td>0.0842</td>
<td>0.1620</td>
<td>0.1543</td>
</tr>
<tr>
<td>$d_{FRS,E_ucl.}$</td>
<td>0.0969</td>
<td><strong>0.0840</strong></td>
<td>0.1683</td>
<td>0.1540</td>
</tr>
<tr>
<td>$d_{FR,M}$</td>
<td>0.1016</td>
<td>0.0841</td>
<td>0.1769</td>
<td><strong>0.1519</strong></td>
</tr>
<tr>
<td>$d_{FRS,M}$</td>
<td>0.0917</td>
<td>0.0855</td>
<td>0.1551</td>
<td>0.1570</td>
</tr>
</tbody>
</table>

Table 3: Results of element-wise performance measures for CDS log returns and AR(1)-GARCH(1,1) residuals.

**Out-of-sample comparison:** For the CDS log returns of our new firms, surprisingly the equity correlation matrix outperforms all distance-parameterized model correlation matrices concerning RMSE and MAPE of pairwise correlations. Apparently, for these firms, credit and equity correlations are more similar than for the ITRX constituents, and consequently the geostatistical models overestimate the CDS correlations, cf. Figure 4. The higher CDS correlations of the ITRX constituents may well be caused by being part of the index. The model based on the equity-correlation distance $d_E$ performs best among the geostatistical models in both data sets, CDS log returns and AR(1)-GARCH(1,1) residuals, cf. Table 5. According to both matrix-values loss functions applied to the extended correlation matrices, the geostatistical models parameterized by $d_E$ and $d_{FRS,M}$ perform best for both CDS log returns and AR(1)-GARCH(1,1) residuals, cf. Table 4.
Figure 3: Histograms of the distribution of the entries of \( \text{Cor}_{\text{sample}} - \text{Cor}_{\text{model}} \).

Top: \( \text{Cor}_{\text{sample}} \) is the CDS log return sample correlation matrix and \( \text{Cor}_{\text{model}} \) is the respective model correlation matrix for \( d_E \) and \( d_{\text{FRS,M}} \), and the correlation matrix of (normal-transformed) equity correlation matrix in the rightmost plot. Bottom: \( \text{Cor}_{\text{sample}} \) is the CDS AR(1)-GARCH(1,1) residual sample correlation matrix and \( \text{Cor}_{\text{model}} \) is the respective model correlation matrix for \( d_E \), \( d_{\text{FRS,Eucl.}} \), and \( d_{\text{FR,M}} \).

3.3 Interpolation of missing data

An issue often encountered when working with CDS data is poor data quality. Unlike in equity markets, quotes are not freely available, and often one only obtains data sets with either missing data or the same quote repeated over several days, which is clearly unrealistic. Again, the geostatistical framework offers a nice solution to this problem: Executing Steps 0.-4. as described in Section 2, knowing the distance between firms and the covariance/correlation function of the data set, i.e. the considered market, one can easily interpolate missing values using the kriging technique.

Comparison with existing methods: We compare the performance of the kriging imputation with the performance of a copula-based imputation methods proposed by Di Lascio et al. (2015) and Käärik and Käärik (2009, 2010).

Käärik and Käärik (2009, 2010) assume that the dependence structure of the data set is a Gaussian copula, and they impute the expected value given the observed data. Under the assumption of a Gaussian field, this is very similar to simple kriging as described in Section 2, the only difference being that in geostatistics, the copula is parameterized by distance among the components.

Di Lascio et al. (2015) extend this to a broader selection of copula classes, and instead of imputing the expected value of the conditional distribution given the observed values,

\footnote{In general, this might also be caused by missing liquidity, but our data set comprises the most liquidly traded series by definition of the ITRX index family.}
Table 4: Results of the matrix-valued loss functions for CDS log returns and AR(1)-GARCH(1,1) residuals for new firms only.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Equity</td>
<td>926.2568</td>
<td>-</td>
<td>44.3533</td>
<td>-</td>
</tr>
<tr>
<td>$d_E$</td>
<td><strong>131.9362</strong></td>
<td><strong>112.7960</strong></td>
<td>33.9761</td>
<td>29.2042</td>
</tr>
<tr>
<td>$d_{FRS, Eucl.}$</td>
<td>147.9159</td>
<td>119.4380</td>
<td>32.2188</td>
<td>28.9903</td>
</tr>
<tr>
<td>$d_{FRS, Eucl.}$</td>
<td>157.3199</td>
<td>118.9473</td>
<td>32.3884</td>
<td>28.9790</td>
</tr>
<tr>
<td>$d_{FR, M}$</td>
<td>173.5536</td>
<td>117.4923</td>
<td>32.8869</td>
<td>28.9234</td>
</tr>
<tr>
<td>$d_{FRS, M}$</td>
<td>138.3648</td>
<td>123.9517</td>
<td><strong>30.4045</strong></td>
<td><strong>27.4572</strong></td>
</tr>
</tbody>
</table>

Table 5: Results of element-wise performance measures for CDS log returns and AR(1)-GARCH(1,1) residuals for new firms only.

<table>
<thead>
<tr>
<th></th>
<th>RMSE returns</th>
<th>RMSE residuals</th>
<th>MAPE returns</th>
<th>MAPE residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equity</td>
<td>0.1598</td>
<td>-</td>
<td>0.3483</td>
<td>-</td>
</tr>
<tr>
<td>$d_E$</td>
<td>0.2027</td>
<td><strong>0.1684</strong></td>
<td>0.6289</td>
<td><strong>0.5670</strong></td>
</tr>
<tr>
<td>$d_{FRS, Eucl.}$</td>
<td>0.2118</td>
<td>0.1918</td>
<td>0.6716</td>
<td>0.6692</td>
</tr>
<tr>
<td>$d_{FRS, Eucl.}$</td>
<td>0.2193</td>
<td>0.1915</td>
<td>0.6974</td>
<td>0.6682</td>
</tr>
<tr>
<td>$d_{FR, M}$</td>
<td>0.2311</td>
<td>0.1884</td>
<td>0.7352</td>
<td>0.6558</td>
</tr>
<tr>
<td>$d_{FRS, M}$</td>
<td>0.2034</td>
<td>0.1961</td>
<td>0.6424</td>
<td>0.6841</td>
</tr>
</tbody>
</table>

they impute a value drawn randomly from this conditional distribution. The corresponding code is available in the R package CoImp.

In order to test our model, we deleted observations from our data sets using CoImp’s MCAR (missing completely at random) function and impute these missing values by simple kriging from fitted variograms using the different distances introduced in Section 2.2.1, as well as using the copula-based imputation techniques introduced above. To evaluate the different methods, we use the following performance figures:

$$\text{MAPE} = \frac{1}{M} \sum_{k=1}^{M} \left| \frac{z_k^{\text{imp}} - z_k^{\text{obs}}}{z_k^{\text{obs}}} \right|,$$

$$\text{RMSE} = \sqrt{\frac{1}{M} \sum_{k=1}^{M} (z_k^{\text{imp}} - z_k^{\text{obs}})^2},$$

$$\text{MAE} = \frac{1}{M} \sum_{k=1}^{M} |z_k^{\text{imp}} - z_k^{\text{obs}}|,$$

where $M$ is the number of missing observations and the superscripts ‘imp’ and ‘obs’ refer to imputed and observed values, respectively. As already stated, MAPE has several shortfalls, e.g. being undefined for $z_k^{\text{obs}} = 0$. However, it is a popular performance measure among practitioners and the performance measure of choice in Di Lascio et al. (2015). The missing values with $z_k^{\text{obs}} = 0$ (being less than 0.2% of all missing values) are excluded in the reported MAPE values in Table 6.

In our use case, CoImp identifies the Gaussian copula with unspecified dependence structure as the best-fitting dependence structure among the supplied models$^{17}$ and the Gaussian copula families were Gaussian with constant correlation matrix, Gaussian, Clayton, Gumbel,
Figure 4: Histograms of the distribution of the new entries of Cor_{sample} − Cor_{model}, where Cor_{sample} is the CDS log return sample correlation matrix and Cor_{model} is the respective model correlation matrix for \(d_E\) and \(d_{FRS,M}\), and the correlation matrix of (normal-transformed) equity correlation matrix in the rightmost plot.

<table>
<thead>
<tr>
<th>Distances</th>
<th>MAPE (ret)</th>
<th>RMSE (ret)</th>
<th>MAE (ret)</th>
<th>MAPE (resid)</th>
<th>RMSE (resid)</th>
<th>MAE (resid)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gauss. cop. (cc)</td>
<td>1.603722</td>
<td>0.728327</td>
<td>0.552596</td>
<td>2.502895</td>
<td>0.781716</td>
<td>0.600400</td>
</tr>
<tr>
<td>Gauss. cop. (un)</td>
<td>1.769959</td>
<td>0.730541</td>
<td>0.556732</td>
<td>2.537761</td>
<td>0.787327</td>
<td>0.603779</td>
</tr>
<tr>
<td>CoImp (cc)</td>
<td>2.304826</td>
<td>1.109832</td>
<td>0.887970</td>
<td>3.172209</td>
<td>1.150790</td>
<td>0.918862</td>
</tr>
<tr>
<td>CoImp (un)</td>
<td>2.389993</td>
<td>1.115477</td>
<td>0.894545</td>
<td>3.478642</td>
<td>1.143631</td>
<td>0.915300</td>
</tr>
<tr>
<td>(d_E)</td>
<td>1.646298</td>
<td>0.723373</td>
<td>0.545422</td>
<td>2.617390</td>
<td>0.774820</td>
<td>0.592405</td>
</tr>
<tr>
<td>(d_{FRS,Eucl.})</td>
<td>1.603087</td>
<td>0.727450</td>
<td>0.551777</td>
<td>2.493205</td>
<td>0.780758</td>
<td>0.599383</td>
</tr>
<tr>
<td>(d_{FRS,Eucl.})</td>
<td>1.604094</td>
<td>0.727512</td>
<td>0.551711</td>
<td>2.496190</td>
<td>0.780921</td>
<td>0.599430</td>
</tr>
<tr>
<td>(d_{FR,M})</td>
<td>1.626555</td>
<td>0.727959</td>
<td>0.552468</td>
<td>2.505752</td>
<td>0.780601</td>
<td>0.599547</td>
</tr>
<tr>
<td>(d_{FRS,M})</td>
<td><strong>1.600154</strong></td>
<td><strong>0.723067</strong></td>
<td><strong>0.548492</strong></td>
<td><strong>2.50846</strong></td>
<td><strong>0.775817</strong></td>
<td><strong>0.595213</strong></td>
</tr>
</tbody>
</table>

Table 6: Performance figures for the imputed CDS log returns and AR(1)-GARCH(1,1) residuals from the Gaussian copula approach of Käärik and Käärik (2009, 2010), from CoImp, and from simple kriging using the different distances defined in Section 2.2.1. The labels (cc) and (un) refer to the constant correlation and unspecified dependence structure, respectively.

We find that CoImp performs worst in both data sets according to all performance measures. This, however, is most likely due to the fact that this randomly drawn imputed value is not the optimal point forecast for any of the chosen performance measures, cf. Remark 3.2 below. The imputed values from the geostatistical approach slightly outperform the imputed values from the Gaussian copula approach of Käärik and Käärik (2009, 2010) according to all performance figures, with the forecasts based on \(d_{FRS,M}\) and \(d_E\) performing best for the imputation of log returns, and \(d_E\) and \(d_{FRS,Eucl.}\) performing best for the imputation of AR(1)-GARCH(1,1) residuals, cf. Table 6.

Remark 3.2. (On point forecasts) As illustrated in Gneiting (2011), just comparing some point forecasts by some measures of predictive accuracy is essentially comparing apples to

Frank, and the t copula with degree of freedom fixed to 4. (Imputation was performed with CoImp V. 0.3-1, which does not yet explicitly consider Gaussian copulas other than the constant correlation variant or t copulas, but the package’s authors confirmed that the procedure works just fine for these.)
oranges. He argues for either disclosing the accuracy measure, so that forecasters can give the respective optimal point forecast from their predictive distribution (e.g. the expected value for RMSE), or explicitly asking for a certain functional of the predictive distribution as their point forecast (e.g. the expected value). In this spirit, comparing the performance measures of our forecast and the one made by CoImp has exactly this problem.

**Discussion:** The advantage of the distance-parameterized over the conventionally estimated Gaussian copulas is only marginal for our data set. However, this might be improved substantially for other financial distance measures, or in more diverse data sets. Further, the assumption of a Gaussian field for financial data could be relaxed. The natural next step would be to apply a spatial copula approach as described in Gräler (2014), i.e. use a copula parameterized by distance to model the dependence structure, which can be seen as an approximation to a more complicated field by means of copulas, and also as an extension of the copula-based imputation methods presented in Di Lascio et al. (2015); Käärik and Käärik (2009, 2010). A related ansatz extending vine copulas to incorporate spatial information is described in Erhardt et al. (2015).

### 4 Conclusion and outlook

We discussed the application of geostatistical modeling to financial data sets as well as the necessary adjustments in higher-dimensional coordinate systems. In the light of the presented results, we find that the geostatistical approach is a promising alternative for the joint modeling of dependent credit spreads concerning the estimation of the correlation matrix, especially considering the appealing properties of this approach in the presence of missing data, and the imputation of missing data.

Future research may generalize our results in several ways: In the context of other financial data sets, different distances could be explored. Considering the question of (geometric) anisotropy in higher dimensions, metric learning approaches are worth a closer look. Transcending the sometimes unrealistic assumption of a Gaussian field, one may work with spatially parameterized copulas to model the dependence structure. Finally, space-time models may be explored to be able to model consider temporal and spatial dependence simultaneously.

**Acknowledgement**

The authors thank Ludwig Brummer and Aleksey Min for helpful discussions on the subject of this paper.

**Appendix: Validation of fitted correlation functions**


In the case of second-order stationary fields, this issue is equivalent to validating a fitted correlation function. Here, we follow the approach of Christakos (1984) and validate the estimated variogram against a set of empirical variograms.

---

18 A vine copula is a $d$-dimensional copula constructed solely from bivariate copulas. For an introduction to vine copulas, see, e.g. Aas et al. (2009).
correlation function. For a detailed overview, the interested reader is referred to Abrahamsen (1997), and for background and more details, see, e.g., Schoenberg (1938), Steerneman and van Perlo-ten Kleij (2005), Yaglom (1987).

Bochner’s Theorem gives the following relation to characteristic functions of \( d \)-dimensional random variables (cf. Abrahamsen (1997)):

**Theorem 4.1** (Bochner / Matern). A function \( \rho : \mathbb{R}^d \to \mathbb{R} \) is a stationary correlation function if and only if it is the characteristic function of some \( d \)-dimensional random variable \( X \):

\[
\rho(h) = \mathbb{E}[e^{ih'X}] = \int_{\mathbb{R}^d} e^{ih'x} dF_X(x), \quad h \in \mathbb{R}^d,
\]

where \( F_X \) denotes the distribution function of \( X \).

For stationary isotropic correlation functions this simplifies to

\[
\rho(h) = r(\|h\|) = \int_0^\infty 2^{d-2} \Gamma\left(\frac{d}{2}\right) \frac{J_{d-2}(x\|h\|)}{x\|h\|} \frac{dG(x)}{\|h\|^{d-2}} - \int_0^\infty \Lambda_d(x\|h\|) \frac{dG(x)}{\|h\|},
\]

where \( G(x) = \int_{\|t\| < x} dF_X(x) \). Another representation of stationary isotropic correlation functions due to Schoenberg (1938) is

\[
\rho(\|h\|) = \rho(h) = \int_0^\infty \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi} \Gamma\left(\frac{d-1}{2}\right)} \int_1^1 e^{ih\varphi(1 - \varphi)^{d-3}} d\varphi dF_{\|h\|}(\|y\|).
\]

In case \( F_X \) admits a density \( f \), it is called the spectral density of \( \rho \) and can be obtained as follows for stationary and stationary isotropic correlation functions:

\[
f(x) = (2\pi)^{-d} \int_{\mathbb{R}^d} e^{-ih'x} \rho(h) d^d h,
\]

\[
f(x) = (2\pi)^{-\frac{d}{2}} \int_0^\infty \frac{J_{d-2}(x\|h\|)}{x\|h\|} \frac{dG(x)}{\|h\|^{d-2}} - \int_0^\infty \frac{\Lambda_d(x\|h\|)}{\|h\|} dG(x).
\]

So to verify if a function is indeed a valid stationary (and isotropic) correlation function in \( \mathbb{R}^d \), a simple check is to verify that the corresponding spectral density is indeed a density of some probability measure.

**References**


Ledoit, O. and Wolf, M., A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 2004b, **88**.


