Bridging Gaussian Markov Random Fields and Copulas: A Fast Algorithm for Efficient Gaussian Copula Density Computation with Matérn-like Precision Matrices

Brynjólfur Gauti Guðrúnar Jónsson*

University of Iceland,

E-mail: brynjolfur@hi.is

Abstract

Gaussian Markov Random Fields (GMRFs) have long been a powerful tool for modeling spatial and temporal dependencies in various fields. Similarly, copulas have proven invaluable for modeling complex dependency structures in multivariate data. However, the combination of these two approaches - using GMRFs within copula models - has historically been computationally inefficient, limiting their joint application to smaller datasets or simpler models. This work presents a novel algorithm that overcomes these limitations, allowing for fast and efficient computation of Gaussian Copula densities using GMRF precision structures. By bridging the gap between GMRFs and copulas, this method opens up new possibilities for analyzing large-scale, high-dimensional spatial and spatio-temporal data with complex dependency structures.

Introduction

Review

Gaussian Markov Random Fields (GMRFs) and copulas are two powerful statistical tools, each offering unique strengths in modeling complex data structures. GMRFs excel in capturing spatial and temporal dependencies, particularly in fields such as environmental science, epidemiology, and image analysis. Their ability to represent local dependencies through sparse precision matrices makes them computationally attractive for high-dimensional problems. Copulas, on the other hand, provide a flexible framework for modeling multivariate dependencies, allowing separate specification of marginal distributions and their joint behavior.

The Gaussian copula, in particular, has gained popularity due to its interpretability and connection to the multivariate normal distribution. However, combining GMRFs with copulas has historically been computationally challenging, limiting their joint application to smaller datasets or simpler models.

Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be a multivariate random vector with marginal distribution functions F_i for $i = 1, 2, \dots, n$. The joint distribution function of $\mathbf{X}$ can be written as:

$$F_{\mathbf{X}}(\mathbf{x}) = C(F_1(x_1), F_2(x_2), \dots, F_n(x_n)),$$

where C is the Gaussian copula defined by the GMRF precision matrix $\mathbf{Q}$. The Gaussian copula C is given by:

$$C(u_1, u_2, \dots, u_n) = \Phi_{\mathbf{Q}}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \dots, \Phi^{-1}(u_n)),$$

where $\Phi_{\mathbf{Q}}$ is the joint cumulative distribution function of a multivariate normal distribution

with mean vector $\mathbf{0}$ and precision matrix $\mathbf{Q}$, and Φ^{-1} is the inverse of the standard normal cumulative distribution function.

A critical requirement for the precision matrix $\mathbf{Q}$ governing the GMRF copula C is that $= \mathbf{Q}^{-1}$ should have a unit diagonal, i.e. the marginal variance is equal to one everywhere.. This ensures it operates on the same scale as the transformed data, $\Phi^{-1}(u_i)$. However, this can be challenging as GMRFs are typically defined in terms of precision matrices that often imply non-unit marginal variances.

This paper presents a novel algorithm that bridges the gap between GMRFs and copulas, allowing for fast and efficient computation of Gaussian copula densities using GMRF precision structures. Our method focuses on creating a Matérn-like precision matrix $\mathbf{Q}$ with unit marginal variance and efficiently computing the multivariate Gaussian copula density of $\mathbf{Z} = \Phi^{-1}(\mathbf{u})$, where $u_i \sim \text{Uniform}(0, 1)$, $i = 1, \ldots, n$.

The key innovation lies in leveraging the special structure of the precision matrix:

$$\mathbf{Q} = \mathbf{Q}_1 \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{Q}_1,$$

where $\mathbf{Q}_1$ is the precision matrix of a standardized one-dimensional AR(1) process and $\otimes$ denotes the Kronecker product. By employing efficient eigendecomposition techniques, our method avoids explicit formation and inversion of the large precision matrix $\mathbf{Q}$, making it particularly suitable for high-dimensional spatial data. In additions to the exact method, we mention approximations to $\mathbf{Q}$ using circuland and folded circulant matrices.

Problem Formulation

Consider a spatial field on a regular $n \times n$ grid. Our objective is to compute the Gaussian copula density efficiently for this field. This computation involves:

1. Specifying a precision matrix $\mathbf{Q}$ that represents the spatial dependence structure.

2. Ensuring the implied covariance matrix $= \mathbf{Q}^{-1}$ has unit diagonal elements.
3. Computing the density for large spatial fields in a computationally efficient manner.

Precision Matrix Structure

We use a Matérn-like precision matrix structure defined as:

$$\mathbf{Q} = \tilde{\mathbf{Q}}^{(+\mathbf{1})}, \quad \nu \in \{0, 1, 2\}$$
$$\tilde{\mathbf{Q}} = \mathbf{Q}_1 \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{Q}_1,$$

where $\mathbf{Q}_1$ is the precision matrix of a standardized one-dimensional AR(1) process, and $\otimes$ denotes the Kronecker product.

Eigendecomposition Method

The properties of Kronecker products and sums are utilized to compute the eigendecomposition of $\mathbf{Q}$. This approach facilitates:

1. Calculation of $\log|\mathbf{Q}|$.
2. Computation of quadratic forms $\mathbf{x}^T\mathbf{Q}\mathbf{x}$.
3. Determination of the diagonal elements of $= \mathbf{Q}^{-1}$ without explicit matrix inversion.

Unit Marginal Variance Adjustment

To ensure unit marginal variances, which is necessary for copula applications, we employ a scaling method for the precision matrix.

Approximation Methods

We introduce circulant and folded circulant approximations to $\mathbf{Q}$. These approximations offer potential computational advantages, which we analyze in terms of efficiency and accuracy trade-offs.

Methods

Gaussian Copula Density Computation

The Gaussian copula density for a random vector $\mathbf{U} = (U_1, ..., U_n)$ with $U_i \sim \text{Uniform}(0, 1)$ is given by:

$$c(\mathbf{u}) = |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2}\mathbf{z}^T(\mathbf{Q} - \mathbf{I})\mathbf{z}\right)$$

where $\mathbf{z} = (z_1, ..., z_n)$ with $z_i = \Phi^{-1}(u_i)$, $\mathbf{Q}$ is the precision matrix, and $\mathbf{I}$ is the identity matrix.

The log-density can be expressed as:

$$\log c(\mathbf{u}) = \frac{1}{2}\log|\mathbf{Q}| - \frac{1}{2}\mathbf{z}^T\mathbf{Q}\mathbf{z} + \frac{1}{2}\mathbf{z}^T\mathbf{z}$$

Our goal is to efficiently compute this log-density for large spatial fields.

Precision Matrix Structure

We define the precision matrix $\mathbf{Q}$ as:

$$\mathbf{Q} = (\mathbf{Q}_1 \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{Q}_1)^{(\nu+1)}, \quad \nu \in \{0, 1, 2\}$$

where $\mathbf{Q}_1$ is the precision matrix of a one-dimensional AR(1) process. This matrix is then scaled so that it implied unit marginal variances.

Computation Process

Step 1: Eigendecomposition of $\mathbf{Q}_1$

We first compute the eigendecomposition of $\mathbf{Q}_1$:

$$\mathbf{Q}_1 = \mathbf{V}\ \mathbf{V}^T$$

where $\mathbf{V}$ is the matrix of eigenvectors and is the diagonal matrix of eigenvalues. Then, the eigendecomposition of $\mathbf{Q}$ is:

$$\mathbf{Q} = (\mathbf{V} \otimes \mathbf{V})(\ \otimes \mathbf{I} + \mathbf{I} \otimes\ )^{(\nu+1)}(\mathbf{V} \otimes \mathbf{V})^T.$$

This means that the eigenvectors of $\mathbf{Q}$ are $\mathbf{v}_j \otimes \mathbf{v}_i$ and the corresponding eigenvalues are $\lambda_i + \lambda_j$.

Step 2: Computation of Marginal Standard Deviations

Using the eigendecomposition of $\mathbf{Q}_1$, we compute the marginal standard deviations and store them in a vector , i.e. $\sigma_i = \sqrt{\Sigma_{ii}}$:

$$= \sqrt{\sum_{i,j} \frac{(\mathbf{v}_j \otimes \mathbf{v}_i)^2}{(\lambda_i + \lambda_j)^\nu}}$$

where $\mathbf{v}_i$ are the eigenvectors and λ_i are the eigenvalues of $\mathbf{Q}_1$.

Step 3: Scaling the Eigendecomposition

We scale the eigendecomposition of $\mathbf{Q}$ using the marginal standard deviations:

$$\mathbf{DQD} = \mathbf{D}(\mathbf{V} \otimes \mathbf{V})(\ \otimes \mathbf{I} + \mathbf{I} \otimes\ )^{\nu}(\mathbf{V} \otimes \mathbf{V})^T\mathbf{D}$$
$$= (\tilde{\mathbf{V}} \otimes \tilde{\mathbf{V}})(\tilde{}\ \otimes \mathbf{I} + \mathbf{I} \otimes \tilde{}\ )(\tilde{\mathbf{V}} \otimes \tilde{\mathbf{V}})^T$$

where $\mathbf{D}$ is a diagonal matrix with $D_{ii} = \sigma_i$.

Step 4: Efficient Computation of Log-Density

Using this scaled eigendecomposition, we efficiently compute:

1. Log-determinant: $\log|\mathbf{DQD}| = \sum_{i,j} \log(\tilde{\lambda}_i + \tilde{\lambda}_j)$
2. Quadratic form: $\mathbf{z}^T\mathbf{DQDz} = \sum_{i,j}(\tilde{\lambda}_i + \tilde{\lambda}_j)y_{ij}^2$, where $y_{ij} = (\tilde{}\mathbf{v}_j \otimes \tilde{}\mathbf{v}_i)^T\mathbf{z}$

This approach allows for efficient computation of the Gaussian copula density without explicitly forming the full $n^2 \times n^2$ precision matrix $\mathbf{Q}$ or its inverse .

Circulant and Folded Circulant Approximations

While the eigendecomposition method provides an exact solution, it can be computationally expensive for very large spatial fields. To address this, we introduce circulant and folded circulant approximations that offer potential computational advantages.

Circulant Matrices

A circulant matrix C is a special kind of matrix where each row is a cyclic shift of the row above it. It can be fully specified by its first row or column, called the base c:

$$C = \begin{pmatrix} c_0 & c_1 & c_2 & \cdots & c_{n-1} \\ c_{n-1} & c_0 & c_1 & \cdots & c_{n-2} \\ c_{n-2} & c_{n-1} & c_0 & \cdots & c_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_1 & c_2 & c_3 & \cdots & c_0 \end{pmatrix} = (c_{j-i \mod n})$$

The key property of circulant matrices is that they are diagonalized by the Discrete Fourier Transform (DFT) matrix. If F is the DFT matrix, then for any circulant matrix C:

$$C = F\Lambda F^H$$

where Λ is a diagonal matrix of eigenvalues and F^H is the conjugate transpose of F.

Block Circulant Matrices

For two-dimensional spatial fields, we use block circulant matrices with circulant blocks (BCCB). An Nn × Nn matrix C is block circulant if it has the form:

$$C = \begin{pmatrix} C_0 & C_1 & C_2 & \cdots & C_{N-1} \\ C_{N-1} & C_0 & C_1 & \cdots & C_{N-2} \\ C_{N-2} & C_{N-1} & C_0 & \cdots & C_{N-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_1 & C_2 & C_3 & \cdots & C_0 \end{pmatrix} = (C_{j-i \mod N})$$

where each C_i is itself a circulant n × n matrix.

Computational Advantages

The circulant structure allows for efficient computation using the Fast Fourier Transform (FFT):

1. Matrix-vector multiplication: For a circulant matrix C with base c and a vector v

$$Cv = \sqrt{n}\text{DFT}(\text{DFT}(c) \odot \text{IDFT}(v)),$$

2. Matrix inverse: The base of C^{-1} is given by

$$\frac{1}{n}\text{IDFT}(\text{DFT}(c)^{-1}).$$

Approximations for Q1

Let Q_1 be the precision matrix of a one-dimensional AR(1) process with n observations. The exact form of Q_1 is:

$$\mathbf{Q}_1 = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

Circulant Approximation

The circulant approximation to Q_1, denoted as $\mathbf{Q}_1^{(circ)}$, is:

$$\mathbf{Q}_1^{(circ)} = \frac{1}{1-\rho^2}\begin{bmatrix} 1+\rho^2 & -\rho & 0 & \cdots & 0 & -\rho \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\rho & 0 & 0 & \cdots & -\rho & 1+\rho^2 \end{bmatrix}$$

This approximation treats the first and last observations as neighbors, effectively wrapping the data around a circle.

Folded Circulant Approximation

The folded circulant approximation, $\mathbf{Q}_1^{(fold)}$, is based on a reflected version of the data. We

double the data by reflecting it, giving us the data $x_1, \ldots, x_n, x_n, \ldots, x_1$. We then model this doubled data with a $2n \times 2n$ circulant matrix. If written out as an $n \times n$ matrix, it takes the form:

$$\mathbf{Q}_1^{(fold)} = \frac{1}{1-\rho^2} \begin{bmatrix} 1-\rho+\rho^2 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1-\rho+\rho^2 \end{bmatrix}$$

This approximation modifies the first and last diagonal elements to account for the reflection of the data. As x_1 now is the first and last data point, then we avoid the circular dependence from the regular circulant approximation.

Extension to the Full Q Matrix

For a two-dimensional spatial field on an $n \times n$ grid, we construct the full precision matrix Q using a Kronecker sum:

$$\mathbf{Q} = (\mathbf{Q}_1 \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{Q}_1)^{(\nu+1)}$$

where $\otimes$ denotes the Kronecker product, I is the $n \times n$ identity matrix, and ν is a smoothness parameter.

When we approximate Q_1 with a circulant matrix, this Kronecker sum results in a block-circulant matrix with circulant blocks (BCCB). To see this, let's consider the case where $\nu = 0$ for simplicity:

$$\mathbf{Q} = \mathbf{Q}_1 \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{Q}_1$$

Now, let Q_1 be approximated by a circulant matrix C with base vector c = [c0, c1, …, cn-1]. Then:

$$\mathbf{Q}_1 \approx \mathbf{C} = \begin{pmatrix} c_0 & c_1 & \cdots & c_{n-1} \\ c_{n-1} & c_0 & \cdots & c_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ c_1 & c_2 & \cdots & c_0 \end{pmatrix}$$

The Kronecker product C I results in a block matrix where each block is a scalar multiple of I:

$$\mathbf{C} \otimes \mathbf{I} = \begin{pmatrix} c_0\mathbf{I} & c_1\mathbf{I} & \cdots & c_{n-1}\mathbf{I} \\ c_{n-1}\mathbf{I} & c_0\mathbf{I} & \cdots & c_{n-2}\mathbf{I} \\ \vdots & \vdots & \ddots & \vdots \\ c_1\mathbf{I} & c_2\mathbf{I} & \cdots & c_0\mathbf{I} \end{pmatrix}$$

Similarly, I C results in a block matrix where each block is a copy of C:

$$\mathbf{I} \otimes \mathbf{C} = \begin{pmatrix} \mathbf{C} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C} \end{pmatrix}$$

The sum of these two matrices is a block-circulant matrix with circulant blocks:

$$\mathbf{Q} \approx \mathbf{C} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{C} = \begin{pmatrix} \mathbf{B}_0 & \mathbf{B}_1 & \cdots & \mathbf{B}_{n-1} \\ \mathbf{B}_{n-1} & \mathbf{B}_0 & \cdots & \mathbf{B}_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_1 & \mathbf{B}_2 & \cdots & \mathbf{B}_0 \end{pmatrix}$$

where each Bi is a circulant matrix. Specifically, B0 = c0I + C, and for i 0, Bi = ciI.

This BCCB structure allows us to use 2D FFT for efficient computations. The base matrix c for this BCCB structure is:

$$\mathbf{c} = \begin{bmatrix} 2+2\rho^2 & -\rho & 0 & \cdots & 0 & -\rho \\ -\rho & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\rho & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

This base matrix c captures the structure of the precision matrix Q and allows for efficient computation of eigenvalues using the 2D Fast Fourier Transform (FFT), enabling rapid calculation of the log-determinant and quadratic forms needed for the Gaussian copula density.

Results

To evaluate the computational efficiency of our proposed methods, we conducted a benchmarking study comparing the exact solution with the circulant and folded circulant approximations. We measured the time taken to evaluate the density of a Gaussian copula for various grid sizes, ranging from 100x100 to 40000x40000.

Computational Efficiency

Table 1 presents the results of a benchmark comparing the time it takes to evaluate the gaussian copula density described above. For each grid size, we report the computation time for the exact method and the two approximations, along with the speed-up factor relative to the exact method. Each calculation was performed twenty times and the medians are shown in the table.

Table 1: Table 1. Benchmarking how long it takes to evaluate the density of a Mátern(ν)-like field with correlation parameter ρ, scaled to have unit marginal variance.

Q_size	Exact	Circulant		Folded	
	Time	Time	Speed-Up	Time	Speed-Up
100x100	194.69 s	33.89 s	5.75x	44.77 s	4.35x
400x400	309.61 s	37.45 s	8.27x	122.78 s	2.52x
900x900	790.30 s	85.65 s	9.23x	155.47 s	5.08x
1600x1600	1.96ms	98.71 s	19.86x	243.64 s	8.05x
3600x3600	8.66ms	151.70 s	57.1x	484.07 s	17.89x
10000x10000	71.81ms	343.07 s	209.31x	1.36ms	52.76x
19600x19600	246.40ms	667.58 s	369.09x	2.89ms	85.32x
40000x40000	997.32ms	1.44ms	694.46x	7.69ms	129.69x

The results demonstrate significant computational gains from both approximation methods, with the efficiency advantage increasing for larger grid sizes. Key observations include:

1. **Scalability**: Both approximation methods show superior scalability compared to the exact method. As the grid size increases, the speed-up factor generally increases, indicating that the approximations become increasingly advantageous for larger spatial fields.

2. **Circulant Approximation Performance**: The circulant approximation consistently outperforms the folded circulant approximation in terms of speed. For the largest grid size (40000x40000), it achieves a remarkable 694.46x speed-up over the exact method.

3. **Folded Circulant Approximation**: While not as fast as the circulant approximation, the folded circulant method still offers substantial speed improvements, reaching a 129.69x speed-up for the 40000x40000 grid.

4. **Trade-off Considerations**: The choice between the circulant and folded circulant approximations may depend on the specific requirements of the application. While the circulant approximation is faster, the folded circulant method may offer better accuracy, particularly near the edges of the spatial field.

These results underscore the practical value of our approximation methods, especially for large-scale spatial analyses where computational efficiency is crucial. The substantial speed-ups achieved, particularly for larger grids, demonstrate the potential of these methods to enable the analysis of much larger spatial datasets than previously feasible with exact methods.