


# MPCR: Multi- and Mixed-Precision Computations Package in R

Mary Lai O. Salvaña<sup></sup>

University of Connecticut,  
Storrs, Connecticut, USA.

Sameh Abdulah<sup></sup>

King Abdullah University  
of Science and Technology,  
Thuwal, Saudi Arabia.

Minwoo Kim<sup></sup>

Pusan National University,  
Busan, South Korea.

David Helmy<sup></sup>

BrightSkies Inc.,  
Alexandria, Egypt.

Ying Sun<sup></sup>

King Abdullah University  
of Science and Technology,  
Thuwal, Saudi Arabia.

Marc G. Genton<sup></sup>

King Abdullah University  
of Science and Technology,  
Thuwal, Saudi Arabia.

---

## Abstract

Computational statistics has traditionally utilized double-precision (64-bit) data structures and full-precision operations, resulting in higher-than-necessary accuracy for certain applications. Recently, there has been a growing interest in exploring low-precision options that could reduce computational complexity while still achieving the required level of accuracy. This trend has been amplified by new hardware such as NVIDIA’s Tensor Cores in their V100, A100, and H100 GPUs, which are optimized for mixed-precision computations, Intel CPUs with Deep Learning (DL) boost, Google Tensor Processing Units (TPUs), Field Programmable Gate Arrays (FPGAs), ARM CPUs, and others. However, using lower precision may introduce numerical instabilities and accuracy issues. Nevertheless, some applications have shown robustness to low-precision computations, leading to new multi- and mixed-precision algorithms that balance accuracy and computational cost. To address this need, we introduce **MPCR**, a novel R package that supports three different precision types (16-, 32-, and 64-bit) and their combinations, along with its usage in commonly-used Frequentist/Bayesian statistical examples. The **MPCR** package is written in C++ and integrated into R through the **Rcpp** package, enabling highly optimized operations in various precisions.

*Keywords:* accuracy, computational statistics, efficient computing, low-precision, mixed-precision, multi-precision.

---

## 1. Introduction

Real numbers in computer architectures are encoded in bits and represented in floating-point format. Released in 1985, the *IEEE 754-1985* governs the standards for floating-point formats and arithmetic (Kahan 1996; Zuras *et al.* 2008). It underwent a revision in 2008, incorporating the addition of 16-bit half-precision, alongside 128-bit quadruple-, 64-bit double-, and 32-bit single-precision formats. The original purpose of introducing the 16-bit half-precision format was for storage. However, it has recently found utility in scientific applications such

as machine learning (Björck *et al.* 2021; Tolliver *et al.* 2022) and linear algebra (Higham *et al.* 2019; Luszczek *et al.* 2019; Scott and Tuma 2023). This trend has increased with the emergence of new hardware that can run low-precision computations substantially faster than full-precision, e.g., Graphics Processing Units (GPUs), Intel CPUs with DL boost, ARM CPUs, Tensor Processing Units (TPUs), Field-Programmable Gate Arrays (FPGAs), and AI hardware accelerators. While employing lower precision can potentially lead to reduced computation times, specific applications demonstrate resilience towards low-precision computations, prompting researchers to explore the creation of innovative multi- and mixed-precision algorithms. These algorithms aim to strike a balance between accuracy and computational efficiency.

In the literature, double-precision used to be the standard numerical format for a wide range of scientific computational workloads such as weather forecasting and climate modeling (Chantry *et al.* 2019; Lang *et al.* 2021), nonlinear equations (Benner *et al.* 2022), numerical linear algebra (Higham and Mary 2022b; Bujanović *et al.* 2023), ordinary differential equations (Hopkins *et al.* 2020), quantum chemistry and physics models (Pederson *et al.* 2022), quantum signal processing models (Ying 2022), neural networks (Hrycej *et al.* 2022), and remote sensing (Nguyen *et al.* 2012; Räss *et al.* 2019; Li and Fotheringham 2020). Recently, there has been a growing interest in reducing the precision of data and computations in these applications and many others. This shift is due to the high cost of double-precision calculations, whereas lower precisions enable hardware acceleration, faster execution times, reduced memory needs, and lower energy consumption. Numerous applications have shown the significant impact of utilizing lower-precision data structures and operations, including in numerical linear algebra (Yamazaki *et al.* 2015a,b; Yang *et al.* 2021; Carson *et al.* 2022), quantum chemistry and quantum physics (Pederson *et al.* 2022), computer simulations of nuclear reactor models (Cherezov *et al.* 2023), computational fluid dynamics (Freytag *et al.* 2022; Lehmann *et al.* 2022), weather forecasting and climate modeling (Rüdisühli *et al.* 2013; Váňa *et al.* 2017; Hatfield *et al.* 2020; Klöwer *et al.* 2020), astronomy (White *et al.* 2023), and deep learning (Murillo *et al.* 2022).

Although lowering the precision offers attractive acceleration, naively reducing the precision can lead to catastrophic round-off errors as the range of the numerical representation is limited, thereby significantly changing model results. Thus, the state-of-the-art in numerical computing is moving away from traditionally specifying a notably reduced precision towards multi- and mixed-precisions to squeeze out more performance while maintaining the accuracy of results (Higham and Mary 2022a). Multi-precision refers to using multiple different floating-point representations for different operations in a single algorithm. Although several precisions are involved, multi-precision is still limited to prescribing a specific precision for each task of an application or algorithm. On the other hand, mixed-precision refers to using a mix of different numerical precisions within a single operation. For example, the state-of-the-art European ocean model called Nucleus for European Modelling of the Ocean (NEMO) demonstrated that from the 942 variables used for ocean simulation, 902 of them could use single-precision and the rest remain double without degrading model accuracy (Tintó Prims *et al.* 2019). Other areas that started to switch to mixed-precision computing are differential equations (Ooi *et al.* 2020), electromagnetic field analysis (Masui and Ogino 2019), and weather and climate (Hatfield *et al.* 2019). Mixed-precision has also been used in solving high-performance computing (HPC) problems, such as those in molecular dynamics, biology, cosmology, and plasma physics. A comprehensive survey on how various HPC applications utilize mixed-

precision can be found in [Netti et al. \(2023\)](#).

Computational statistics is a field that stands to benefit heavily from these innovations in numerical computing. While computational statistics has long been utilizing computational techniques to expedite the use of statistical methodologies that are computationally expensive, the use of low, multi-, and mixed-precision techniques is just gaining momentum in the literature. For instance, [Guivant et al. \(2023\)](#) demonstrated that a Gaussian filter in Kalman filtering can work in low precisions. Their experiments showed that processing times can be accelerated by 50% when the covariance matrix is stored in single or half-precision. [Misra et al. \(2023\)](#) showed that a compiler-inferred fixed-point precision variational inference is 8.15X and 22.67X faster than performing variational inference in single- and double-precision, respectively. [Maddox et al. \(2022\)](#) proposed a conjugate gradients (CG) algorithm used for training Gaussian process models that utilize a mixed-precision approach wherein accuracy-critical computations are cast into full precisions while noncritical parts are performed in lower precisions. They demonstrated up to 3X speedup compared to equivalent models at full precisions. In geostatistics, where the entries of the spatial covariance matrix quantify the dependence between measurements at two locations, [Abdulah et al. \(2019\)](#) proposed a novel Cholesky factorization algorithm that works under a mixed-precision format. The algorithm prescribes double-precision for the diagonal elements while representing the off-diagonal elements in single-precision. They defined a ‘band’ to ascertain the number of diagonal elements to be represented in a double-precision format. Their implementation has relied on the **StarPU** dynamic runtime system for task scheduling on shared- and distributed-memory systems. [Abdulah et al. \(2022\)](#) extended the approach in [Abdulah et al. \(2019\)](#) to include half-precision. They switched the dynamic runtime system to **ParSEC**, which makes on-the-fly decisions on precision conversion. [Cao et al. \(2022\)](#) improved upon the approach in [Abdulah et al. \(2022\)](#) by combining mixed-precision techniques with Tile Low-Rank (TLR) matrix algorithms. Furthermore, they devised a new algorithm that determines the appropriate data structure (dense or TLR) and precision (double, single, or half) for every tile at runtime.

The primary numeric format in the widely used statistical software R ([R Core Team 2023](#)) is double-precision, thereby ensuring a substantial level of accuracy for the calculations performed. Despite the many advantages of using multi- and mixed-precision, such as lower energy usage and reduced storage and data movement costs ([Higham and Mary 2022a](#)), support for precisions lower than double-precision needs to be improved in R. Indeed in R, single-precision type can be introduced using external packages, such as the **float** package ([Schmidt et al. 2022](#)). The **float** package can accommodate single-precision when the system has access to single-precision Basic Linear Algebra Subprograms (BLAS) or LAPACK routines. The package necessitates the underlying BLAS/LAPACK libraries, integrated with the R environment, to support single-precision computation. The package provides support for the majority of the linear algebra operations that are available in R. However, the package design does not readily facilitate easy extension to other precisions, such as half-precision. Another notable R package that supports multi-precision is the **Rmpfr** package which stands for *Multiple Precision Floating-Point Reliable* ([Maechler et al. 2023](#)). The package serves as an interface to the **MPFR** library, a C library for multiple precision floating point arithmetic, which builds on the GNU Multiple Precision Arithmetic C library. The package offers extended precision floating-point representations suitable for precision-demanding tasks such as combinatorics, number theory, and cryptography, where standard floating-point arithmetic can result in unacceptable imprecision. The default precision of **MPFR** objects is 128-bit.

Moreover, the **Rmpfr** package also supports arbitrary precision such that R users can specify their required precision. For example, computing  $\pi$ , which is equal to  $4 \arctan(1)$ , up to 100 decimal places, the following R command can be used: `4 * atan(mpfr(1, 333))`, where the number 1 is encoded with 333-bit, the equivalent of 100 decimal places. A drawback with the **Rmpfr** package is that regular numeric types in R cannot be directly used with **Rmpfr** numbers. When interacting with functions or packages that do not support **Rmpfr**, an explicit conversion to numeric values is required. To interact directly with the GNU Multiple Precision Arithmetic C library (Lucas *et al.* 2023), the R package **gmp** can be used.

This paper introduces a novel R package called **MPCR**, which stands for *Multi- and Mixed-Precision Computational Statistics in R*. **MPCR** is an advanced package designed to provide R users with multi- and mixed-precision data structures and computations. The package harnesses the combined strength of C++ and R, empowering users with high-performance computing capabilities. Specifically tailored for researchers and data scientists working with multi- and mixed-precision arithmetic, **MPCR** is an invaluable tool for achieving efficient and accurate computations. The package supports three different precision types, namely, 16-bit, 32-bit, and 64-bit, as well as their combinations. Central to the package is creating an **MPCR** object with a user-defined precision. Leveraging the underlying BLAS/LAPACK library, **MPCR** facilitates low-precision calculations across various linear algebra operations. Additionally, the package supports mixed-precision calculations by incorporating a tile-based data structure within R. This novel structure enables matrices/vectors to possess different precisions, catering to scientific accuracy requirements. Furthermore, the package allows for parallel linear algebra computations on multicore systems using task-based linear algebra algorithms. The package enables users to generate matrices/vectors of four distinct types: HP-matrix, SP-matrix, DP-matrix, and MP-matrix. HP-, SP-, and DP-matrices are matrices characterized by uniform precision entries. In an HP matrix, elements are stored in half-precision (16-bit), resulting in a smaller memory footprint and facilitating faster algorithms, albeit with a trade-off in accuracy. Each element inside a DP-matrix requires a bigger memory allocation (64-bit), resulting in slower computations but more accurate solutions. An SP-matrix, wherein each element occupies 32-bit memory, offers an intermediate balance between speed and accuracy. Beyond the trio of uniform-precision classifications, the **MPCR** package introduces a mixed-precision (MP-) matrix variant. An MP-matrix is a matrix subdivided into tiles (a term used in scientific computing to refer to blocks) so the tiles can be stored at different precisions. To the best of our knowledge, the **MPCR** package is the first package in R to introduce a tile-based data structure. The **MPCR** package furnishes mixed-precision functionality for linear algebra operations within the R environment.

In the first version of **MPCR**, we provide three tile-based operations for MP-matrices: general matrix multiplication (GEMM), triangular solve (TRSM), and tiled Cholesky factorization (POTRF) (Beaumont *et al.* 2020). We highlight that **MPCR** has considerable potential for expansion to other tile-based matrix operations, allowing R developers to benefit from mixed-precision computation. To assess the performance of an algorithm, it is possible to create different MP-matrix structures. For instance, one can allocate higher precision to the diagonal tiles and lower precision to the off-diagonal tiles. As an illustration of the usage of MP-matrices and MP-operations in various statistical applications, this paper includes examples in Markov Chain Monte Carlo, maximum likelihood estimation in spatial statistics, principal component analysis, and Bayesian inference wherein tiled linear algebra operations are performed on appropriately structured MP-matrices to perform matrix inversion, singular

value decomposition, and the like.

The subsequent sections of the paper are organized as follows: Section 2 gives a background on floating-point arithmetic in modern technology and its realization in hardware. Section 3 briefly reviews tile-based linear algorithms. In Section 4, we delve into the internal design of the package and workflow. In Section 5, we give an overview of the main functions in the package and how to use them with examples and code snippets. Section 6 illustrates the efficiency, performance, and accuracy of **MPCR** objects and functions and compares them with native R objects and functions. In Section 7, we present various Frequentist and Bayesian problems (four in total) and illustrate how to plug-and-play **MPCR** functions to enable faster execution of computationally demanding tasks. We conclude in Section 8.

## 2. Floating-Point Arithmetic in Modern Technology

The most widely used standard for floating-point arithmetic as of 2024 is the *IEEE 754-2008* (Kahan 1996; Zuras *et al.* 2008) and its correction in 2019 (IEEE 2019). This standard significantly improved over its predecessor, *IEEE 754-1985*, and includes a comprehensive set of guidelines that cover almost every aspect of the floating-point theory or simply the approximation rules for real numbers on today’s digital computers. The full name of this standard is *IEEE Standard 754-2019 for Floating-Point Arithmetic*, which is often abbreviated as either *IEEE 754-2019* or simply *IEEE 754*. A floating-point number system is a finite subset  $\mathcal{F}$  of  $\mathbb{R}$ ,  $\mathcal{F} \subset \mathbb{R}$ , which depends on its elements  $(\beta, t, e_{\min}, e_{\max})$ . A number  $x$  in  $\mathcal{F}$  has the form:

$$x = \pm m \times \beta^{e-t+1},$$

where  $t$ ,  $m$ , and  $e$  are integers known as *precision*, *significand* or *mantissa*, and *exponent*, respectively, and  $\beta$  is the base which is 2 for binary (commonly used on all current computers) and 10 for decimal. Here, the significand satisfies  $0 \leq m \leq \beta^t - 1$ , the exponent  $e$  satisfies  $e_{\min} \leq e \leq e_{\max}$ , and the *IEEE* standard requires that  $e_{\min} = 1 - e_{\max}$ . The number is stored in the computer with a format that consists of three fields: a sign bit, exponent bits, and significand bits.

Precision formats that take up more memory (in terms of bits) but can approximate real number values at very high accuracy are regarded as *high* or *full precision*, while those that consume smaller memory with lower accuracy are considered *reduced* or *low precision*. In the *IEEE-754* standard, the double-precision 64-bit floating-point format, denoted FP64, is considered high precision, while the single-precision 32-bit floating-point format, denoted FP32, is regarded as low precision. FP32 uses 1-bit for the sign of the number, 8-bit for the exponent, and 23-bit for the significand and can represent values from  $\pm(2 - 2^{-23}) \times 2^{127}$ . On the other hand, FP64 uses 1-bit for the sign, 11-bit for the exponent, and 52-bit for the significand, and can represent values from  $\pm(2 - 2^{-52}) \times 2^{1023}$ . Another low-precision format is the half-precision 16-bit floating-point, denoted FP16. FP16 uses 1-bit for the sign, 5-bit for the exponent, and 10-bit for the significand and can represent values from  $\pm(2 - 2^{-10}) \times 2^{17}$ . Figure 1 visualizes how a number  $x$  takes up space in a computer’s memory for each precision format.

The FP32 and FP64 are the most widely used formats as a broad range of off-the-shelf general-purpose processors natively supports them. Although FP16 significantly reduces the



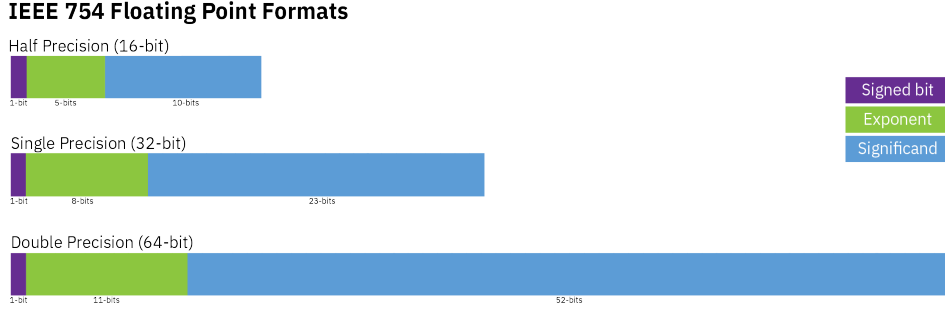


Figure 1: The composition in bits of different precisions based on *IEEE 754-2008* standards.

memory requirements and increases the arithmetic throughput by 2X against FP32 and 4X against FP64, most commercial CPUs do not support FP16. However, recent chips from the leading hardware vendors, including NVIDIA, Intel, and AMD, include FP16 arithmetic in their processing units to enhance performance in applications such as machine learning, gaming, and scientific computing. NVIDIA was one of the first to offer FP16 by enabling it as a storage format in CUDA 7.5 (Sabbagh Molahosseini *et al.* 2012). Since then, FP16 arithmetic has been made available for CUDA-enabled GPUs, starting with the PASCAL architecture (Abdelfattah *et al.* 2020). Other notable hardware that provides support for FP16 is NVIDIA P100 (2016) and V100 (2017), A100 (2020), and H100 (2022) GPUs through tensor cores. One of the standout features of NVIDIA Tensor Cores is their ability to perform operations in mixed-precision by leveraging lower precision (like FP16) for specific calculations to increase performance and efficiency while still leveraging higher precision (like FP32 and FP64) when higher accuracy calculation is mandatory. ARM chips are another example of hardware supporting 16-bit operations through their Thumb instruction set. The instruction set reduces the memory footprint, as the smaller instruction size reduces the amount of code. New ARM chips, like the ARM Cortex series, support a mix of 32-bit (ARM instruction set) and 16-bit (Thumb instruction set) instructions, allowing running an algorithm in multi-precision mode to optimize performance and memory usage (Higham and Pranesh 2019).

Most current hardware chips predominantly support 64-bit and 32-bit floating-point (FP) arithmetic operations. A select few, however, can support 16-bit operations, either through mixed-precision computation, as seen in NVIDIA’s Tensor Cores, or exclusively, like some ARM processors. There has been a significant trend in software development towards adopting partial or full 16-bit computation, particularly in fields such as machine learning and scientific computing. This shift is driving leading hardware manufacturers and even emerging startup companies to enhance their hardware offerings with support for 16-bit FP operations. Additionally, there is a growing anticipation that support for 8-bit FP operations might become more prevalent shortly, reflecting the ongoing evolution in hardware and software domains to balance performance with computational efficiency.

### 3. Tile-based Linear Algebra Algorithms

Linear algebra algorithms underpin the linear algebra operations performed by various computing architectures. These algorithms direct the movement of data and define the tasks needed to perform the operations and arrive at the desired results. The software library LA-

PACK (Anderson *et al.* 1999), and its parallel version, ScaLAPACK (Blackford *et al.* 1996), offer a stable and wide range of linear algebra algorithms necessary for dense linear algebra operations. Additionally, their algorithms were made compatible to parallelism in shared-memory and distributed-memory systems by depending on MPI, OpenMP multithreading, and parallel BLAS and Basic Linear Algebra Communication Subprograms (BLACS) (Luszczek *et al.* 2014). The parallelization, particularly in ScaLAPACK, is carried out using the block-cyclic decomposition technique wherein the matrix is divided into small pieces called *blocks* such that these blocks are then distributed to different MPI processes in a cyclical fashion (Cifani *et al.* 2023). This type of parallelization scheme falls under the traditional bulk synchronous programming model wherein multiple threads will have to wait in idle until the slowest performing thread is finished (Buttari *et al.* 2009; Luszczek *et al.* 2014; Haidar *et al.* 2015). Furthermore, the extent to which linear algebra tasks can be parallelized under this approach depends on the availability of parallel BLAS (Buttari *et al.* 2009).

Tile algorithms were developed to circumvent the limitations inherent to the block-cyclical approach. Tile algorithms propose a major rethinking of the way the matrix is decomposed in order to maximize task parallelism. The parallelization strategy of tile algorithms involves partitioning the matrix into small pieces called *tiles* such that their format allows for a more efficient way to access memory and is more suitable for asynchronous execution combined with dynamic scheduling of tasks across multiple platforms (Haidar *et al.* 2012; Dongarra *et al.* 2014). Tile algorithms can be conveniently expressed using Directed Acyclic Graphs (DAG) with nodes indicating the tasks and the edges specifying data movement and dependencies among tasks (Agullo *et al.* 2009; Buttari *et al.* 2009; Bosilca *et al.* 2012; Haidar *et al.* 2012, 2015; Akbudak *et al.* 2017). Several tile versions of some of the standard linear algebra algorithms, such as, Cholesky factorization (Buttari *et al.* 2009; Haidar *et al.* 2012), LU factorization (Buttari *et al.* 2009; Haidar *et al.* 2012; Dongarra *et al.* 2014), and QR factorization (Buttari *et al.* 2008, 2009; Haidar *et al.* 2012), have been established in the literature.

## 4. The MPCR Package Internal Design and Workflow

Historically, the R community has predominantly utilized 64-bit arithmetic for their computational tasks, aiming for a high level of accuracy that may only sometimes be necessary. In reality, arithmetic with a precision lower than 64-bit can offer the advantages of faster execution and a smaller memory footprint while maintaining the same level of accuracy as 64-bit arithmetic in numerous applications. This work introduces the **MPCR** package, designed to support 64-, 32-, and 16-bit arithmetic. The **MPCR** package also offers multi- and mixed-precision computing capabilities, aiming to optimize computational efficiency without compromising accuracy in many applications.

### 4.1. MPCR Internal Design

The **MPCR** package is designed with both performance and flexibility in mind. We rely on C++ in developing the backend of the package, and we port the code to R through the **Rcpp** package. We use C++ for various reasons: 1) C++ excels in scenarios demanding high performance and efficiency; 2) C++ is ideal for system-level programming that provides more flexibility in managing system resources, such as memory and processing power; 3) C++ has many tools and libraries that can help in building a robust software; 4) C++ can easily be

integrated into different languages, including R and Python. A conceptual view of the software stack for the **MPCR** package is displayed in Figure 2. As can be seen in the figure, every function in **MPCR** is invoked from the R environment, where the **Rcpp** package manages such calls, effectively bridging them to the corresponding C++ functions in the C++ layer.

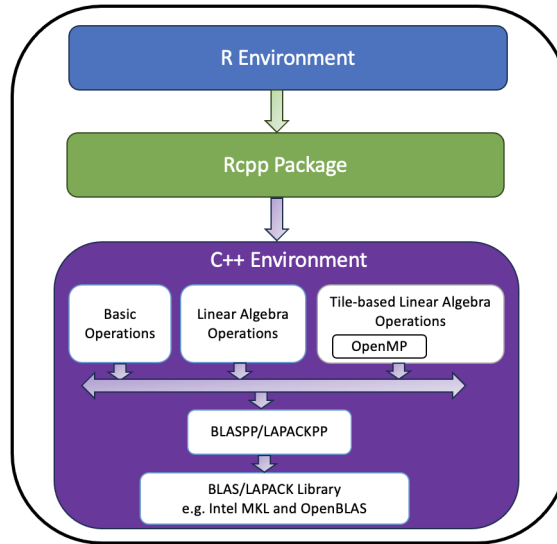


Figure 2: The MPCR package software stack.

The **MPCR** package offers three different types of operations: basic operations, linear algebra operations, and tile-based linear algebra operations. Detailed information about these functionalities is available in the package tutorial, accessible at <https://github.com/stsds/MPCR/blob/main/inst/doc/MPCR-manual.pdf>. The **MPCR** package supports numerous base functions in the R programming language, which can be readily applied to **MPCR**-objects. For instance, when performing the Cholesky factorization (`chol()` in base R) on a single-precision **MPCR**-object, the **MPCR** package utilizes its internal `chol()` function to execute the factorization in single-precision. Such design facilitates integrating the **MPCR** package into any existing R code, enabling single-precision operations with minimal modifications.

The **MPCR** package is designed to offer full extensibility with minimal effort. For instance, the package was built for easy support for integration of new precisions, addition of new linear operations, and inclusion of new tile-based operations. Moreover, the package uses the CMake Build system which enables easy integration for cross-platform and of external libraries. The CMake Build system also allows a full C++ testing environment, enabling a complete CI/CD pipeline for fast and stable development. Furthermore, **MPCR** comes with a fully documented R and C++ code, making it easy for any future development. The package was designed to be fully operational as a separate C++ module to be extended and used in any C++ library. Additionally, **MPCR** uses template C++ functions offering no code redundancy. Lastly, **MPCR** has a fully organized code structure, offering easy and fast code navigation.

## 4.2. Workflow

This section outlines the operational workflow of the **MPCR** package, from a simple call at



the R level and to its execution at the C++ level. Figure 3 presents a high-level overview of the workflow involved in a typical **MPCR** operation.

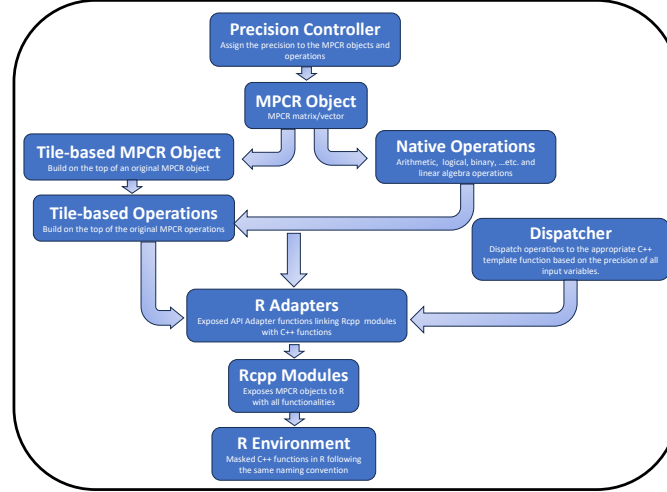


Figure 3: The **MPCR** package workflow.

The process begins with specifying the appropriate precision for an **MPCR** object, which is then conveyed to the C++ code via the **Rcpp** package. The **precision controller** module in the C++ level assigns the correct precision to the **MPCR** objects and their corresponding operations. The next step involves allocating memory for the **MPCR** object, which depends on the structure of the object, e.g., vector or matrix, and the precision, e.g., 32-bit for single-precision and 64-bit for double-precision. Operations at different precision levels can be performed on non-tile-based **MPCR** objects. We refer to the non-tile-based operations collectively as *native MPCR operations*. On the other hand, tile-based **MPCR** objects utilize a different memory allocation approach, with tiles being the fundamental component of any such object. These specialized **MPCR** objects are equipped with tile-based linear algebra operations that can run sequentially or in parallel. Currently, three operations are supported: POTRF, GEMM, and TRSM. Once the object is allocated with particular precision, any call of a C++ function for this object is managed by the *dispatcher* module. This module is responsible for dispatching operations to the appropriate template function based on the precision of the inputs. To link the C++ function with the R functions through the **MPCR** package, the **MPCR R adapters** module is used. This module acts as a mapper between the C++ and R environments. To provide a clearer description of each module, we provide examples of internal codes that illustrate the functionality of each one.

### Code snippet for an **MPCR R adapters** module example at the C++ level

**\*\*MPCR R adapter\*\*** is a module that allows R and C++ to interact. It acts as a bridge between two incompatible interfaces.

```

std::vector <MPCR>
RSVD(MPCR *aInputA, const long &aNu, const long &aNv, const bool &aTranspose)
{
    auto row = aInputA->GetNRow();

```

```

    auto col = aInputA->GetNCol();
    auto nv = aNv;
    auto nu = aNu;

    if (aNv < 0) {
        nv = std::min(row, col);
    }
    if (aNu < 0) {
        nu = std::min(row, col);
    }

    auto precision = aInputA->GetPrecision();
// Allocate three new objects for the SVD output.
    auto d = new MPCR(precision);
    auto u = new MPCR(precision);
    auto v = new MPCR(precision);

    SIMPLE_DISPATCH(precision, linear::SVD, *aInputA, *d, *u, *v, nu, nv,
                    aTranspose)

    std::vector <MPCR> output;
// Converting the output of C++ SVD funtion to a list, to match the shape of
R Return values.
    output.push_back(*d);
    output.push_back(*u);
    output.push_back(*v);

    return output;
}

### Code snippet for an MPCR dispatcher module example

**The Dispatcher** module main task is to choose the right signature for a
function according to the input and output precision. This can be done by the
help of the precision controller module.

**The Precision Controller** module main task is to decide the output precision
according to the input and the promotion strategy, and to decide the right
signature for the dispatcher.

MPCR* RRBind(MPCR *apInputA, MPCR *apInputB)
{
// Getting the precision for object A and B.
    auto precision_a = apInputA->GetPrecision();
    auto precision_b = apInputB->GetPrecision();
// Using the precision controller to decide the output precision.

```

```

    auto output_precision = GetOutputPrecision(precision_a, precision_b);
// Allocating the output MPCR object with the right precision.
    auto pOutput = new MPCR(output_precision);
// Using the precision controller to decide the operation signature for
the dispatcher to use.
    auto operation_comb = GetOperationPrecision(precision_a, precision_b,
                                                output_precision);
// Using the Operation signature to decide what template function to use.
    DISPATCHER(operation_comb, basic::RowBind, *apInputA, *apInputB, *pOutput)
    return pOutput;
}

```

### 4.3. Catch2

*Catch2* is a popular C++ testing framework known for its comprehensive suite of tools that streamline writing and maintaining test cases for C++ code. As a header-only library, it enables C++ developers to easily create and manage their unit tests, offering a straightforward approach to testing.

In the **MPCR** package, *Catch2* incorporates many unit tests, facilitating thorough testing of the functionality of the package as required. For example, integrating new features or enhancements into the package might unexpectedly impact existing code. Running these unit tests via *Catch2* helps ensure software robustness throughout the development process. Furthermore, the **MPCR** package is designed to simplify the expansion of existing test cases, supporting future development and maintaining software reliability over time. Below is an example of using *Catch2* in the **MPCR** package for the singular value decomposition function.

### Code snippets for C++ unit test for the SVD function using *Catch2*.

```

// Input Matrix Values.
    vector <double> values = {1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0,
                             0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0,
                             0, 0, 0, 1, 1, 1};
// Create an MPCR Object.
    MPCR a(values, FLOAT);
// Change MPCR Object to MPCR Matrix.
    a.ToMatrix(9, 4);
// Validate values for testing.
    vector <float> validate_values = {3.464102e+00, 1.732051e+00,
                                     1.732051e+00, 1.922963e-16};
// Allocating three MPCR objects for the SVD output
    MPCR d(FLOAT);
    MPCR u(FLOAT);
    MPCR v(FLOAT);
// This function will dispatch to SVD template function with the right precision.
    SIMPLE_DISPATCH(FLOAT, linear::SVD, a, d, u, v, a.GetNCol(),
                   a.GetNCol())

```

```

// Testing output d size.
    REQUIRE(d.GetSize() == 4);
    auto err = 0.001;
// Setting additional objects to test SVD Mathematically.
    MPCR dd(9, 4, FLOAT);
    for (auto i = 0; i < dd.GetSize(); i++) {
        dd.SetVal(i, 0);
    }
    for (auto i = 0; i < 4; i++) {
        dd.SetValMatrix(i, i, d.GetVal(i));
    }
    vector <double> temp_vals(81, 0);
    MPCR uu(temp_vals, FLOAT);
    uu.ToMatrix(9, 9);
    for (auto i = 0; i < u.GetSize(); i++) {
        uu.SetVal(i, u.GetVal(i));
    }

    MPCR vv = v;
    vv.Transpose();

// Performing  $A = U \Sigma V^H$ 
    MPCR temp_one(FLOAT);
    MPCR temp_two(FLOAT);

    SIMPLE_DISPATCH(FLOAT, linear::CrossProduct, uu, dd, temp_one, false,
                    false)

    SIMPLE_DISPATCH(FLOAT, linear::CrossProduct, temp_one, vv, temp_two,
                    false,
                    false)

    MPCR temp_three(FLOAT);
    SIMPLE_DISPATCH(FLOAT, math::Round, temp_two, temp_three, 1);

    SIMPLE_DISPATCH(FLOAT, math::PerformRoundOperation, temp_three,
                    temp_two, "abs");

// Checking if our SVD function values are valid mathematically
    for (auto i = 0; i < a.GetSize(); i++) {
        REQUIRE(temp_two.GetVal(i) == a.GetVal(i));
    }

```

## 5. Overview of the MPCR Package

To start using the **MPCR** package, it should be installed from the CRAN repository and imported into the R environment by running these commands:

```
R> install.packages("MPCR")
R> library(MPCR)
```

While CRAN hosts the stable **MPCR** package, our GitHub repository contains the latest development features and fixes. To download the package from the GitHub repository:

```
R> library(devtools)
R> install.packages("https://github.com/stsds/MPCR")
R> library(MPCR)
```

### 5.1. Creating MPCR objects

To create an **MPCR** object, i.e., matrix/vector, one invokes the function **new** while specifying the desired **size** and **precision** of the object. The default object created by the function **new** is a zero-vector with a number of elements equal to the **size** argument. To illustrate, the following codes create a single-precision **MPCR** zero-vector with 50 elements:

```
R> MPCR_object <- new(MPCR, 50, "single")
R> MPCR_object
```

MPCR Object: 32-Bit Precision

The function **IsMatrix** can be used to check whether the newly created **MPCR\_object** is a vector or a matrix, e.g.:

```
R> MPCR_object$IsMatrix
```

```
[1] FALSE
```

To display the values contained inside **MPCR\_object**, one can call the function **PrintValues()** as follows:

```
R> MPCR_object$PrintValues()
```

Vector Size : 50

-----

```
[ 1 ]      0      0      0      0      0      0      0
[ 8 ]      0      0      0      0      0      0      0
[ 15 ]     0      0      0      0      0      0      0
[ 22 ]     0      0      0      0      0      0      0
[ 29 ]     0      0      0      0      0      0      0
[ 36 ]     0      0      0      0      0      0      0
[ 43 ]     0      0      0      0      0      0      0
[ 50 ]     0
```

Additionally, `MPCR_object` can be reconfigured as a matrix with a certain number of rows and columns. Suppose one requires `MPCR_object` to be a  $5 \times 10$  matrix. The following codes perform such transformation:

```
R> MPCR_object$ToMatrix(5,10)
R> MPCR_object$IsMatrix
```

```
[1] TRUE
```

```
R> MPCR_object$PrintValues()
```

```
Precision : 32-Bit Precision
Number of Rows : 5
Number of Columns : 10
```

```
-----
[  0  0  0  0  0  0  0  0  0  0  0  ]
[  0  0  0  0  0  0  0  0  0  0  0  ]
[  0  0  0  0  0  0  0  0  0  0  0  ]
[  0  0  0  0  0  0  0  0  0  0  0  ]
[  0  0  0  0  0  0  0  0  0  0  0  ]
```

Elements inside any **MPCR** object can be extracted and their values replaced by indicating the appropriate indices, e.g.:

```
R> MPCR_object[1,1] <- 1
R> MPCR_object[1,2] <- 2
R> MPCR_object[1,3] <- 3
R> MPCR_object[1,4] <- 4
R> MPCR_object[1,5] <- 5
R> MPCR_object$PrintValues()
```

```
Precision : 32-Bit Precision
Number of Rows : 5
Number of Columns : 10
```

```
-----
[  1  2  3  4  5  0  0  0  0  0  0  ]
[  0  0  0  0  0  0  0  0  0  0  0  ]
[  0  0  0  0  0  0  0  0  0  0  0  ]
[  0  0  0  0  0  0  0  0  0  0  0  ]
[  0  0  0  0  0  0  0  0  0  0  0  ]
```

The **MPCR** objects can also be converted to R objects by invoking the **MPCR** function `MPCR.ToNumericMatrix()` or `MPCR.ToNumericVector()` as follows:

```
R> MPCR.ToNumericMatrix(MPCR_object)
```



```

      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    2    3    4    5    0    0    0    0    0
[2,]    0    0    0    0    0    0    0    0    0    0
[3,]    0    0    0    0    0    0    0    0    0    0
[4,]    0    0    0    0    0    0    0    0    0    0
[5,]    0    0    0    0    0    0    0    0    0    0

```

```

R> MPCR_vector <- new(MPCR, 5, "single")
R> MPCR.ToNumericVector(MPCR_vector)

```

```

[1] 0 0 0 0 0

```

Another approach to constructing an **MPCR** object is to convert the usual R object to **MPCR** objects using the function `as.MPCR()`. In the following, we create a  $6 \times 6$  R matrix and convert it into a single-precision  $6 \times 6$  **MPCR** matrix:

```

R> a <- matrix(1:36, 6, 6)
R> MPCR_matrix <- as.MPCR(a,nrow=6,ncol=6,precision="single")

```

```

R> a

```

```

      [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    1    7   13   19   25   31
[2,]    2    8   14   20   26   32
[3,]    3    9   15   21   27   33
[4,]    4   10   16   22   28   34
[5,]    5   11   17   23   29   35
[6,]    6   12   18   24   30   36

```

```

R> MPCR_matrix$PrintValues()

```

```

Precision : 32-Bit Precision
Number of Rows : 6
Number of Columns : 6

```

```

-----
[  1   7  13  19  25  31  ]
[  2   8  14  20  26  32  ]
[  3   9  15  21  27  33  ]
[  4  10  16  22  28  34  ]
[  5  11  17  23  29  35  ]
[  6  12  18  24  30  36  ]

```

## 5.2. Creating MPCR-Tile objects

An **MPCR-Tile** matrix can be created by calling the function `new(MPCRTile, ...)` and supplying the values necessary to configure the desired tile matrix. The required arguments to the `new(MPCRTile, ...)` function are listed as follows:

- `rows`: a scalar value indicating the number of rows in the matrix;
- `cols`: a scalar value indicating the number of columns in the matrix;
- `rows_per_tile`: a scalar value indicating the number of rows in each tile;
- `cols_per_tile`: a scalar value indicating the number of columns in each tile;
- `values`: an R matrix or vector containing all the values that should be in the tile matrix;
- `precisions`: an R matrix or vector of strings, containing the precision type of each tile.

In the following, we demonstrate the usage of `new(MPCRTile, ...)` by creating two tile matrices of different sizes.

Suppose we have a vector `z` with 100 elements, and we want to construct a column tile matrix out of `z` such that it has 5 tiles, with 20 rows in each tile. Moreover, suppose we need the first two tiles to be encoded in double-precision while the rest are in single-precision. The following codes create such a column tile matrix:

```
R> set.seed(1234)
R> n <- 100
R> z <- runif(n, 0, 1)

R> num_tiles = 5
R> prec.z <- matrix(c(rep("double", 2), rep("single", num_tiles - 2)),
+   nrow = num_tiles, ncol = 1)

R> z.tile <- new(MPCRTile, rows = n, cols = 1, rows_per_tile = 20,
+   cols_per_tile = 1, values = z, precisions = prec.z)
R> z.tile
```

```
----- MPCRTile Object -----
Number of Rows : 100
Number of Cols : 1
Number of Tiles : 5
Number of Tiles Per Row : 5
Number of Tiles Per Col : 1
Number of Rows Per Tile : 20
Number of Cols Per Tile : 1
-----
```

To check that the first two tiles are in double-precision while the rest are in single-precision, we can select a specific tile and print out its attributes using the function `MPCRTile.GetTile()` with three arguments, namely, `matrix`, `rowidx`, and `colidx`. The argument `matrix` is simply the tile matrix from which we want to draw out a specific tile. The arguments `rowidx` and `colidx` are, respectively, the row and column indices of the particular tile of interest. The following codes print out the attributes of each of the tiles in the tile matrix `z.tile`:

```
R> MPCRTile.GetTile(z.tile, 1, 1)
```

MPCR Object : 64-Bit Precision

```
R> MPCRTile.GetTile(z.tile,2,1)
```

MPCR Object : 64-Bit Precision

```
R> MPCRTile.GetTile(z.tile,3,1)
```

MPCR Object : 32-Bit Precision

```
R> MPCRTile.GetTile(z.tile,4,1)
```

MPCR Object : 32-Bit Precision

```
R> MPCRTile.GetTile(z.tile,5,1)
```

MPCR Object : 32-Bit Precision

To print the whole `z.tile`, the command `z.tile$MPCRTile.print()` or `print(z.tile)` can be used. Similarly, the values of a specific tile, for instance, the third tile, can be displayed using the following command:

```
R> MPCRTile.GetTile(z.tile,3,1)$PrintValues()
```

Suppose we have a  $100 \times 100$  matrix **M**, and we want to construct a  $5 \times 5$  tile matrix such that only the diagonal tiles are double-precision and the rest are single-precision. The following codes create such a tile matrix:

```
R> library(dplyr)
R> x <- seq(0, 1, length.out = 10)
R> locs <- expand.grid(x, x) %>% as.matrix()
R> n <- nrow(locs)
R> dist_mat <- as.matrix(dist(locs))
R> M <- exp(-dist_mat)

R> num_tiles_rowwise = 5
R> num_tiles_colwise = 5
R> num_tiles_total = num_tiles_rowwise * num_tiles_colwise
R> prec.M <- matrix(rep("single", num_tiles_total), nrow=num_tiles_rowwise,
+   ncol=num_tiles_colwise)
R> diag(prec.M) <- "double"

R> M.tile <- new(MPCRTile, rows=n, cols=n, rows_per_tile=n/num_tiles_rowwise,
+   cols_per_tile=n/num_tiles_colwise, M, prec.M)
R> M.tile
```

```

----- MPCRtile Object -----
Number of Rows : 100
Number of Cols : 100
Number of Tiles : 25
Number of Tiles Per Row : 5
Number of Tiles Per Col : 5
Number of Rows Per Tile : 20
Number of Cols Per Tile : 20
-----

```

### 5.3. Operations on MPCR objects

The **MPCR** package enables multi- and mixed-precision arithmetic. Arithmetic operations on **MPCR** objects can be carried out using the same symbols used to perform those operations on R objects. Moreover, with multi- and mixed-precision arithmetic, one can add or subtract matrices with different precisions such that the resulting **MPCR** object inherits the precision of the **MPCR** object with the highest precision. The following example adds a double-precision  $2 \times 10$  matrix to a single-precision  $2 \times 10$  matrix with a resulting double-precision  $2 \times 10$  matrix.

```

R> s1 <- as.MPCR(1:20,nrow=2,ncol=10,"single")
R> s2 <- as.MPCR(21:40,nrow=2,ncol=10,"double")
R> x <- s1 + s2
R> typeof(x)

```

MPCR Object : 64-Bit Precision

Other usual operations on R vectors and matrices can also be done on **MPCR** vectors and matrices by invoking the conventional function calls in R. For example, the transpose of **MPCR\_matrix** in the previous example can be obtained by `t(MPCR_matrix)` and multiplying two **MPCR** matrices is done using the product operator `%%`, e.g., `s1 %% t(s2)`.

Linear algebra operations are also available and are called as they are in the base R package, such as `chol()`, `chol2inv()`, `solve()`, and `eigen()`. Furthermore, common functions used to evaluate matrices/vectors in R are also made available for **MPCR** objects such as `cbind`, `rbind`, `diag`, `min`, `max`, etc., and can be invoked the same way as the base R package.

### 5.4. Operations on MPCR-Tile objects

The **MPCR** package currently supports the three main tile-based linear algebra operations for tiled matrices, namely, tile-based matrix-matrix multiplication, Cholesky factorization, and solving of a triangular matrix equation.

#### *Tile-Based Matrix-Matrix Multiplication*

The `MPCRTile.gemm()` function performs matrix-matrix multiplication of the form

$$\mathbf{C} = \alpha \mathbf{A} \mathbf{B} + \beta \mathbf{C}. \quad (1)$$

The required arguments to the `MPCRTile.gemm()` function are the following:

- **a**: an **MPCR**-Tile matrix representing matrix **A** in Equation (1);
- **b**: an **MPCR**-Tile matrix representing matrix **B** in Equation (1);
- **c**: an **MPCR**-Tile matrix representing matrix **C** in Equation (1);
- **transpose\_a**: a flag to indicate whether the matrix in argument **a** should be transposed before performing matrix multiplication; takes in the values **TRUE** or **FALSE** (default);
- **transpose\_b**: a flag to indicate whether the matrix in argument **b** should be transposed before performing matrix multiplication; takes in the values **TRUE** or **FALSE** (default);
- **alpha**: a scalar value representing  $\alpha$  in Equation (1);
- **beta**: a scalar value representing  $\beta$  in Equation (1);
- **num\_threads**: an integer that indicates the number of threads to run using **openmp**; the default value is 1, which means serial computations with no parallelization.

The results after performing the matrix-matrix multiplication are stored in the matrix used as an input to the argument **c**. This means that the original data saved into the input matrix to **c** will be overwritten.

Note that the rules of basic matrix-matrix multiplication must also be followed for the tile-based version. This means that the dimensions in terms of tiles of **A**, **B**, and **C** must be compatible. That is, the number of tiles columnwise of **A** must be equal to the number of tiles rowwise of **B** and the resulting dimensions of their product in terms of tiles must be equal to the tile dimensions of **C**.

The following codes demonstrate how to call the `MPCRTile.gemm()` function:

```
R> A <- matrix(1:24, nrow=4, ncol=6)
R> B <- matrix(0, nrow=6, ncol=1)
R> C <- matrix(1, nrow=4, ncol=1)

R> num_tiles_rowwise_A <- 2
R> num_tiles_colwise_A <- 1
R> num_tiles_total_A <- num_tiles_rowwise_A * num_tiles_colwise_A

R> num_tiles_rowwise_B <- 1
R> num_tiles_colwise_B <- 1
R> num_tiles_total_B <- num_tiles_rowwise_B * num_tiles_colwise_B

R> num_tiles_rowwise_C <- 2
R> num_tiles_colwise_C <- 1
R> num_tiles_total_C <- num_tiles_rowwise_C * num_tiles_colwise_C

R> rows_per_tile_A <- nrow(A) / num_tiles_rowwise_A
R> cols_per_tile_A <- ncol(A) / num_tiles_colwise_A

R> rows_per_tile_B <- nrow(B) / num_tiles_rowwise_B
```

```

R> cols_per_tile_B <- ncol(B) / num_tiles_colwise_B

R> rows_per_tile_C <- nrow(C) / num_tiles_rowwise_C
R> cols_per_tile_C <- ncol(C) / num_tiles_colwise_C

R> prec.A <- matrix(rep("single", num_tiles_total_A),
+   nrow=num_tiles_rowwise_A, ncol=num_tiles_colwise_A)
R> prec.B <- matrix(rep("single", num_tiles_total_B),
+   nrow=num_tiles_rowwise_B, ncol=num_tiles_colwise_B)
R> prec.C <- matrix(rep("single", num_tiles_total_C),
+   nrow=num_tiles_rowwise_C, ncol=num_tiles_colwise_C)

R> A.tile <- new(MPCRTile, rows=nrow(A), cols=ncol(A),
+   rows_per_tile=rows_per_tile_A, cols_per_tile=cols_per_tile_A, A, prec.A)
R> B.tile <- new(MPCRTile, rows=nrow(B), cols=ncol(B),
+   rows_per_tile=rows_per_tile_B, cols_per_tile=cols_per_tile_B, B, prec.B)
R> C.tile <- new(MPCRTile, rows=nrow(C), cols=ncol(C),
+   rows_per_tile=rows_per_tile_C, cols_per_tile=cols_per_tile_C, C, prec.C)

```

Before proceeding with the matrix-matrix multiplication, we first print the values in the tile matrix `C.tile` to verify that it is indeed a  $4 \times 1$  column matrix of 1's.

```

R> print(C.tile)

----- MPCRTile Object -----
Number of Rows : 4
Number of Cols : 1
Number of Tiles : 2
Number of Tiles Per Row : 2
Number of Tiles Per Col : 1
Number of Rows Per Tile : 2
Number of Cols Per Tile : 1

-----
[           1           ]
[           1           ]
[           1           ]
[           1           ]

```

Suppose that  $\alpha = 1$  and  $\beta = 0.5$ . The next line of code performs the matrix-matrix multiplication and stores the results in `C.tile`:

```

R> MPCRTile.gemm(A.tile, B.tile, C.tile, transpose_a=F, transpose_b=F,
+   alpha=1, beta=0.5, num_threads=4)

```

Based on the values of `A`, `B`, `C`,  $\alpha$ , and  $\beta$ , we expect the answer to be a  $4 \times 1$  column matrix of 0.5's. To verify that the arithmetic is correct, we check the results by printing the updated values in the tile matrix `C.tile` as follows:



```
R> print(C.tile)
```

```
----- MPCRTile Object -----
```

```
Number of Rows : 4
Number of Cols : 1
Number of Tiles : 2
Number of Tiles Per Row : 2
Number of Tiles Per Col : 1
Number of Rows Per Tile : 2
Number of Cols Per Tile : 1
```

```
-----
[           0.5           ]
[           0.5           ]
[           0.5           ]
[           0.5           ]
```

### *Tile-Based Cholesky Factorization*

The **MPCR** package has a tile-based version of the Cholesky factorization for tile matrices. That is, a positive definite tile matrix  $\mathbf{A}$  can be decomposed into a product of a unique lower triangular tile matrix  $\mathbf{L}$  and its transpose  $\mathbf{L}^\top$  as follows:

$$\mathbf{A} = \mathbf{L}\mathbf{L}^\top. \quad (2)$$

The tile-based Cholesky factorization can be performed by calling the function `chol()` and providing the following required inputs:

- **x**: a positive definite **MPCR**-Tile matrix representing  $\mathbf{A}$  in Equation (2);
- **overwrite\_input**: a flag to indicate whether the resulting lower triangular tile matrix representing  $\mathbf{L}$  in Equation (2) is stored in the tile matrix used as input to the argument **x** above; takes in the values **TRUE** (default) or **FALSE**;
- **num\_threads**: an integer that indicates the number of threads to run using **openmp**; the default value is 1, which means serial computations with no parallelization.

The result after performing `chol()` is a lower triangular tile matrix. Depending on the specification to the argument `overwrite_input`, the resulting lower triangular tile matrix may be stored in the matrix used as the input to the argument **x** or in another user-defined variable. With the  $4 \times 4$  covariance matrix example below, we show how to perform a tile-based Cholesky factorization with **MPCR**:

```
R> library(dplyr)
R> x <- seq(0, 1, length.out = 2)
R> locs <- expand.grid(x, x) %>% as.matrix()
R> n <- nrow(locs)
R> dist_mat <- as.matrix(dist(locs))
```

```

R> M <- exp(-dist_mat)

R> num_tiles_rowwise = 2
R> num_tiles_colwise = 2
R> num_tiles_total = num_tiles_rowwise * num_tiles_colwise

R> rows_per_tile <- nrow(M) / num_tiles_rowwise
R> cols_per_tile <- ncol(M) / num_tiles_colwise

R> prec.M <- matrix(rep("single", num_tiles_total), nrow=num_tiles_rowwise,
+   ncol=num_tiles_colwise)
R> diag(prec.M) <- "double"

R> M.tile <- new(MPCRTile, rows=n, cols=n, rows_per_tile=rows_per_tile,
+   cols_per_tile=cols_per_tile, M, prec.M)
R> M.tile_chol <- chol(M.tile, overwrite=F, num_threads=4)
R> print(M.tile_chol)

----- MPCRTile Object -----
Number of Rows : 4
Number of Cols : 4
Number of Tiles : 4
Number of Tiles Per Row : 2
Number of Tiles Per Col : 2
Number of Rows Per Tile : 2
Number of Cols Per Tile : 2

-----
[      1          0          0          0 ]
[ 0.3678794  0.9298735          0          0 ]
[ 0.3678795  0.1159098  0.9226211          0 ]
[ 0.2431167  0.2994405  0.2641753  0.8839915 ]

```

### *Tile-Based Solving of a Triangular Matrix Equation*

The **MPCR** package supports the tile-based linear algebra routine that solves for  $\mathbf{X}$  in the following triangular system of equations:

$$\mathbf{AX} = \alpha\mathbf{B} \quad \text{or} \quad \mathbf{XA} = \alpha\mathbf{B}, \quad (3)$$

where  $\mathbf{A}$  is any triangular matrix, e.g., the matrix  $\mathbf{L}$  in Equation (2). The routine can be accessed by calling the function `MPCRTile.trsm()` and admitting values to the following parameters:

- a: a triangular **MPCR**-Tile matrix representing  $\mathbf{A}$  in Equation (3);
- b: an **MPCR**-Tile matrix representing  $\mathbf{B}$  in Equation (3);

- **side**: a string that indicates the position of **A** relative to **X**; 'R' for right side or 'L' for left side;
- **upper\_triangle**: a flag that indicates which part of **A** is nonzero; TRUE if **A** is an upper triangular matrix and FALSE otherwise;
- **transpose**: a flag to indicate whether the matrix in argument **a** should be transposed before performing the routine; takes in the values TRUE or FALSE;
- **alpha**: a scalar value representing  $\alpha$  in Equation (3).

The function `MPCRTile.trsm()` returns **X** and is stored in the matrix used as the input to the argument **b**. To illustrate the use of the `MPCRTile.trsm()` function, suppose that the resulting lower triangular tile matrix in the Cholesky factorization in the previous example is **A**. In the following codes, we define **B** as an identity tile matrix and try to solve for **X**:

```
R> B <- diag(4)
R> prec.B <- matrix(rep("single", num_tiles_total), nrow = num_tiles_rowwise,
+   ncol = num_tiles_colwise)
R> B.tile <- new(MPCRTile, rows = n, cols = n, rows_per_tile = rows_per_tile,
+   cols_per_tile = cols_per_tile, B, prec.B)
R> print(B.tile)
```

----- MPCRTile Object -----

```
Number of Rows : 4
Number of Cols : 4
Number of Tiles : 4
Number of Tiles Per Row : 2
Number of Tiles Per Col : 2
Number of Rows Per Tile : 2
Number of Cols Per Tile : 2
```

```
-----
[           1           0           0           0           ]
[           0           1           0           0           ]
[           0           0           1           0           ]
[           0           0           0           1           ]
```

```
R> MPCRTile.trsm(a=M.tile_chol, b=B.tile, side='L', upper_triangle=F,
+   transpose=F, alpha=1)
R> print(B.tile)
```

----- MPCRTile Object -----

```
Number of Rows : 4
Number of Cols : 4
Number of Tiles : 4
Number of Tiles Per Row : 2
Number of Tiles Per Col : 2
```

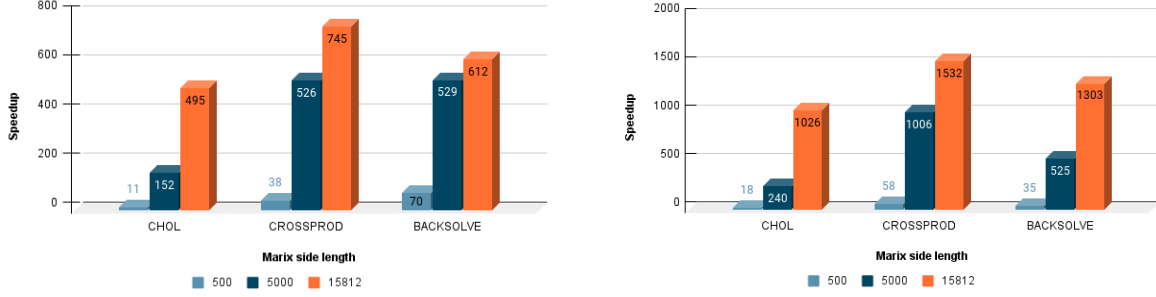


Figure 4: Speedup of double-precision (left) and single-precision (right) linear algebra operations, i.e., `chol()`, `crossprod()`, and `backsolve()` in the **MPCR** package against double-precision operations in R.

Number of Rows Per Tile : 2

Number of Cols Per Tile : 2

```
-----
[      1          0          0          0 ]
[ -0.3956231    1.075415          0          0 ]
[ -0.3490305   -0.1351055    1.083869          0 ]
[ -0.03670389  -0.3239073   -0.3239073    1.131233 ]
```

## 6. Performance Assessment

In this section, we aim to evaluate the performance of single-, double-, and mixed-precisions operations within the **MPCR** package at two different precision levels, i.e., 64-bit and 32-bit. This assessment includes computationally intensive tasks, i.e., non-tile and tile-based linear algebra operations. We focus on linear algebra operations that are time-intensive and those that necessitate modifications for enhanced speed.

Figure 4 illustrates the performance of **MPCR** double- and single-precision operations compared to the R double-precision operations. The left part of the figure shows the speed enhancements achieved by employing **MPCR** double-precision across three different matrix dimensions, specifically 500, 5,000, and 15,812, utilizing three linear algebra procedures: `chol()`, `crossprod()`, and `backsolve()`. The results indicate that **MPCR** double-precision operations can accelerate `chol()`, `crossprod()`, and `backsolve()` by 495X, 745X, and 612X, respectively, compared to the R double-precision operations. The performance improvement is primarily attributed to using highly optimized BLAS/LAPACK libraries like Intel MKL and OpenBLAS, which are automatically used in the **MPCR** package. In contrast, by default, R links to RBLAS, which is not as well optimized and lacks support for single-precision in its current version. The right figure demonstrates the efficiency of **MPCR** single-precision operations relative to R double-precision operations. The illustration reveals that **MPCR** single-precision for `chol()`, `crossprod()`, and `backsolve()` significantly outperforms double-precision operations in R, achieving speedup of 1,026X, 1,532X, and 1,303X, respectively.

The **MPCR** also provides linear scalability with different matrix sizes and consistent speedup when applying single-precision compared to double-precision. Figure 5 shows the execution

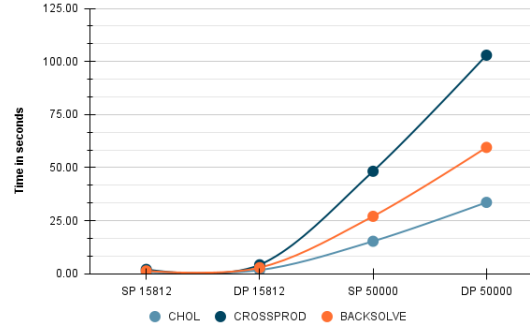


Figure 5: Execution time of three different linear algebra operations, i.e., `chol()`, `crossprod()`, and `backsolve()` using single-precision and double-precision operations with two different matrix sizes 15,812 and 50,000.

times for individual `chol()`, `crossprod()`, and `backsolve()` linear algebra operations with two matrix sizes, 15,812 and 50,000, in single-precision and double-precision formats. Consistent with expectations, a speedup of approximately 2X is shown when precision is lowered from double to single. This increase in efficiency is scalable, as illustrated in the figure.

The **MPCR** package offers two distinct approaches for implementing the Cholesky factorization operation, i.e., `chol()`. This operation can be executed either in the same memory space, referred to as in-place, or using new memory allocations, referred to as out-place. This is an example of performing the Cholesky factorization on a tile-based **MPCR** matrix,

```
R> a <- matrix(c(1.21, 0.18, 0.13, 0.41, 0.06, 0.23,
+ 0.18, 0.64, 0.10, -0.16, 0.23, 0.07,
+ 0.13, 0.10, 0.36, -0.10, 0.03, 0.18,
+ 0.41, -0.16, -0.10, 1.05, -0.29, -0.08,
+ 0.06, 0.23, 0.03, -0.29, 1.71, -0.10,
+ 0.23, 0.07, 0.18, -0.08, -0.10, 0.36), 6, 6)
R> b <- c("float", "double", "float", "float",
+ "double", "double", "float", "float",
+ "double")

R> chol_mat <- new(MPCRTile, 6, 6, 2, 2, a, b)
R> chol_values <- chol(chol_mat, overwrite_input = FALSE, num_thread = 8)
+ # out-place chol()
R> chol(chol_mat, overwrite_input = TRUE, num_thread = 8)
+ # in-place chol()
```

Figure 6 shows the performance when using both calls of the `chol()` operation. As shown, in-place `chol()` outperform out-place `chol()` with different matrix size. The subfigure on the left displays the performance of the single-precision `chol()`, whereas the one on the right illustrates the performance when using double-precision `chol()`.

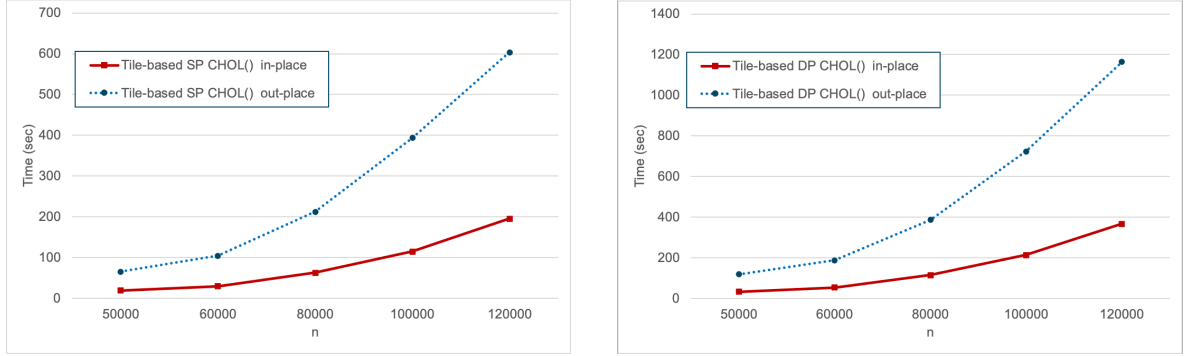


Figure 6: Performance comparison between using in-place and out-place tile-based Cholesky factorization across various matrix sizes.

## 7. Applications

This section offers a range of practical examples demonstrating the use of the **MPCR** package for executing computationally intensive tasks in Frequentist/Bayesian statistical use cases. These examples, drawn from existing applications, illustrate the straightforward integration of different **MPCR** functions. Users dealing with demanding workflows will discover that the multi- and mixed-precision functions provided by **MPCR** significantly enhance performance while preserving the workflow of the original code. This enhancement can dramatically reduce experiment waiting times from hours to minutes or, in some cases, even seconds.

### 7.1. Metropolis-Hastings Algorithm in High Dimensions

The **MPCR** package offers efficient linear algebra operations at different precision levels, as demonstrated in the previous section. One example demonstrating the efficiency of our package is its usage in Markov Chain Monte Carlo (MCMC) simulations, which are essential for many statistical applications such as the Metropolis-Hastings (MH) algorithm.

The MH algorithm is routinely used to generate samples from a target distribution, denoted  $p(\mathbf{z})$ , that may otherwise be difficult to sample from (Metropolis *et al.* 1953; Hastings 1970). The MH algorithm generates a Markov Chain, whose stationary distribution is the target distribution. This means that, in the long run, the samples from the Markov chain will look like the samples from the target distribution, i.e., we use a Markov chain to generate a sequence of vectors, denoted  $(\mathbf{z}_0, \mathbf{z}_1, \dots, \mathbf{z}_I)$ , such that as  $I \rightarrow \infty$ ,  $\mathbf{z}_I \sim p(\mathbf{z})$ .

The Markov Chain vectors are simulated from a distinct distribution called the proposal (conditional) distribution, represented as  $q(\mathbf{z}|\mathbf{z}^{\text{current}})$ . The term “conditional” in the context of proposal distributions in the MH algorithm is simply an indication that the mean of the proposal distribution will be set to the value of the conditioning parameter, which is the current vector in the chain,  $\mathbf{z}^{\text{current}}$ . The proposal distribution is chosen such that it is a distribution from which realizations are easy to simulate and that its realizations can converge to the target distribution. The two most commonly used types of proposal distributions are (a) independent and (b) random-walk. *Independent* proposal distributions are those that do not depend on  $\mathbf{z}^{\text{current}}$ , the most current accepted sample in the chain. This means that  $q(\mathbf{z}|\mathbf{z}^{\text{current}}) = q(\mathbf{z})$ . For instance,  $q(\mathbf{z}|\mathbf{z}^{\text{current}}) = \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . This approach is recommended when the proposal distribution is known to approximate the target distribution well. On the



other hand, proposal distributions under the *random walk* approach are those that depend on  $\mathbf{z}^{\text{current}}$ . For example,  $q(\mathbf{z}|\mathbf{z}^{\text{current}}) = \mathcal{N}_n(\mathbf{z}^{\text{current}}, \mathbf{\Sigma})$  or  $q(\mathbf{z}|\mathbf{z}^{\text{current}}) = \mathcal{U}(\mathbf{z}^{\text{current}} - \epsilon, \mathbf{z}^{\text{current}} + \epsilon)$ , where  $\mathcal{U}(\cdot)$  denotes the uniform distribution (Hastings 1970).

The MH algorithm often struggles in high dimensions especially when incorporating naive proposal distributions such as those listed above (Au and Beck 2001). A more advanced and effective alternative in such scenarios is the Metropolis-Adjusted Langevin Algorithm (MALA) (Xifara *et al.* 2014; Särkkä *et al.* 2021). The key innovation in MALA is the choice of the proposal distribution, the most critical component for the MH algorithm to be successful. The quality of the candidate samples drawn from the independent and random walk proposal distributions earlier are not great since they tend to be rejected by the algorithm. Under MALA, the suggested proposal distribution is

$$q(\mathbf{z}|\mathbf{z}^{\text{current}}) = \mathcal{N}_n \left[ \mathbf{z}^{\text{current}} + \frac{h}{2} \mathbf{M} \nabla \log\{p(\mathbf{z}^{\text{current}})\}, h\mathbf{M} \right], \quad (4)$$

where  $h$  is a user-defined step-size,  $\nabla \log\{p(\mathbf{z})\}$  denotes the gradient of  $\log\{p(\mathbf{z})\}$ , and  $\mathbf{M}$  is a preconditioning matrix. The strength of MALA lies in incorporating the gradient of the target when drawing the next candidate sample (Durmus *et al.* 2017).

When the number of iterations,  $I$ , is large, and the dimensions of the target distribution are large, the MH algorithm can be computationally burdensome to perform. The standard MH algorithm requires computing the likelihood ratio in (6) at every iteration, incurring a computational cost of  $O(n)$  per step, which is prohibitive for large  $n$  (Cornish *et al.* 2019). This section shows how **MPCR** has features that can ameliorate most computational costs surrounding the MH algorithm. First, **MPCR** provides multi-precision support, so vectors and matrices can be stored in double or single-precision. Users can choose precision types for particular vectors and matrices, ultimately reducing the costs of performing the MH algorithm. Second, the **MPCR** package is capable of fast inversion of a high-dimensional covariance matrix and fast vector-matrix operations involved when evaluating the Gaussian log-likelihoods in (6).

The following shows the sequential step in MALA using **MPCR** objects and routines. Additionally, for comparison, we run the same MALA using only R objects.

1. Specify the precision. Choose among **R-Double**, **MPCR-Double**, and **MPCR-Single**.

```
R> precision = 'MPCR-Single'
```

2. Specify the target distribution.

Suppose we want to generate a 2-dimensional (2D) spatial Gaussian random field on a  $120 \times 120$  grid on the unit square with an exponential spatial covariance function model parameterized by  $\boldsymbol{\theta} = (a, \sigma^2)^\top$ , where  $a > 0, \sigma^2 > 0$  are the spatial range and variance parameters, respectively. That is, our target distribution is  $\mathcal{N}_n\{\boldsymbol{\mu}, \mathbf{\Sigma}(\boldsymbol{\theta})\}$ , such that  $n$  is the number of locations, e.g.,  $n = 14,400$ , and we want to simulate an  $n$ -dimensional vector  $\mathbf{Z} = \{Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n)\}^\top \in \mathbb{R}^n$  from the target distribution with  $\boldsymbol{\mu} = \mathbf{0} \in \mathbb{R}^{n \times 1}$ , and  $\mathbf{\Sigma}(\boldsymbol{\theta}) \in \mathbb{R}^{n \times n}$  whose value at index  $(i, j)$  is

$$\text{cov}\{Z(\mathbf{s}_i), Z(\mathbf{s}_j)\} = \sigma^2 \exp\left(-\frac{\|\mathbf{s}_i - \mathbf{s}_j\|}{a}\right). \quad (5)$$

Here,  $Z(\mathbf{s})$  is the observation at location  $\mathbf{s} \in \mathbb{R}^d$  and  $d = 2$ .

- i. Code the mean vector of the target distribution.

```
R> M <- 120
R> n.loc <- M*M
R> locs <- cbind(rep(0:(M-1), M)/(M-1), rep(0:(M-1), each=M)/(M-1))
R> mu = rep(0, nrow(locs)) #target mean vector

R> if(precision %in% c('MPCR-Double', 'MPCR-Single')){
+   mu <- as.MPCR(mu, nrow=n.loc, ncol=1, precision='single')
+ }
```

- ii. Code the covariance matrix of the target distribution.

```
R> dist0 <- as.matrix(dist(locs)) # distance matrix
R> sig <- exp(-dist0/0.5) #target covariance matrix

R> if(precision == 'MPCR-Double'){
+   sig <- as.MPCR(as.matrix(sig), nrow=n.loc, ncol=n.loc,
+     precision='double')
+ }else if(precision == 'MPCR-Single'){
+   sig <- as.MPCR(as.matrix(sig), nrow=n.loc, ncol=n.loc,
+     precision='single')
+ }
```

3. Specify the proposal distribution.

We code the MALA-suggested proposal distribution in Equation (4).

- i. Set values for the parameters in MALA.

```
R> I = 2000
R> h = 0.01
R> pre_M = exp(-dist0/0.05)

R> if(precision == 'MPCR-Double'){
+   pre_M <- as.MPCR(as.matrix(pre_M), nrow=n.loc, ncol=n.loc,
+     precision='double')
+ }else if(precision == 'MPCR-Single'){
+   pre_M <- as.MPCR(as.matrix(pre_M), nrow=n.loc, ncol=n.loc,
+     precision='single')
+ }

R> if(precision %in% c('MPCR-Double', 'MPCR-Single')){
+   pre_M_scaled_inverse <- solve(pre_M$PerformMult(h))
+   L <- t(chol(pre_M$PerformMult(h)))
+ }else{
+   pre_M_scaled_inverse <- solve(h*pre_M)
+   L <- t(chol(h * pre_M))
+ }
```

- ii. Create the functions that will compute  $\nabla \log\{p(\mathbf{z}^{\text{current}})\}$  in Equation (4).

```

R> sig.inv <- solve(sig)
R> gradient <- function(theta){
+   sig.inv %*% theta
+ }

R> gradientStep <- function(theta, t){
+   if(precision %in% c('MPCR-Double', 'MPCR-Single')){
+     theta - (pre_M %*% gradient(theta))$PerformMult(t)
+   }else{
+     theta - t * pre_M %*% gradient(theta)
+   }
+ }

```

4. Create a matrix that will store the values of the accepted samples,  $\mathbf{z}^{\text{current}}$ .

```

R> z_trace <- matrix(, nrow=n.loc, ncol=I)

```

5. Initialize the chain by setting an arbitrary value to  $\mathbf{z}_0 \in \mathbb{R}^n$ .

```

R> set.seed(1234)
R> z_init <- runif(n.loc)
R> if(precision %in% c('MPCR-Double', 'MPCR-Single')){
+   z_init <- as.MPCR(z_init, nrow=n.loc, ncol=1, precision='single')
+ }

```

6. Set the value of  $\mathbf{z}^{\text{current}}$  to  $\mathbf{z}_0$ .

```

R> z_current <- z_init
R> if(precision %in% c('MPCR-Double', 'MPCR-Single')){
+   z_trace[, 1] <- MPCR.ToNumericVector(z_current)
+ }else{
+   z_trace[, 1] <- z_current
+ }

```

7. For  $i = 2, \dots, I$ , iterate through the steps of MALA as follows:

- i. Generate the next vector in the chain  $\mathbf{z}_i \in \mathbb{R}^n$  by simulating a candidate vector  $\mathbf{z}^{\text{candidate}}$  from (4).

```

R> i = 2 #Loop this for i=2,...,I.
R> set.seed(i)
R> x <- rnorm(n.loc)

R> if(precision %in% c('MPCR-Double', 'MPCR-Single')){
+   x <- as.MPCR(x, nrow=n.loc, ncol=1, precision='single')
+ }

R> z_candidate <- gradientStep(z_current, 0.5*h) + L %*% x

```

- ii. Compute the “log acceptance probability”,  $\log A$ , using the formula:

$$\log A = \min[0, \log\{p(\mathbf{z}^{\text{candidate}})\} - \log\{p(\mathbf{z}^{\text{current}})\} + \log\{q(\mathbf{z}^{\text{current}}|\mathbf{z}^{\text{candidate}})\} - \log\{q(\mathbf{z}^{\text{candidate}}|\mathbf{z}^{\text{current}})\}]. \quad (6)$$

```
R> if(precision %in% c('MPCR-Double', 'MPCR-Single')){
+   p_candidate <- MPCR.ToNumericVector((t(z_candidate-mu) %*%
+     sig.inv %*% (z_candidate-mu))$PerformMult(-0.5))
+   p_current <- MPCR.ToNumericVector((t(z_current-mu) %*%
+     sig.inv %*% (z_current-mu))$PerformMult(-0.5))

+   q_current <- MPCR.ToNumericVector((t(z_current-
+     gradientStep(z_candidate,0.5*h)) %*% pre_M_scaled_inverse %*%
+     (z_current-gradientStep(z_candidate,0.5*h)))$PerformMult(-0.5))
+   q_candidate <- MPCR.ToNumericVector((t(z_candidate-
+     gradientStep(z_current,0.5*h)) %*% pre_M_scaled_inverse %*%
+     (z_candidate-gradientStep(z_current,0.5*h)))$PerformMult(-0.5))
+ }else{
+   p_candidate <- -0.5*(t(z_candidate-mu) %*% sig.inv %*%
+     (z_candidate-mu))
+   p_current <- -0.5*(t(z_current-mu) %*% sig.inv %*%
+     (z_current-mu))

+   q_current <- -0.5*(t(z_current-gradientStep(z_candidate,0.5*h))
+     %*% pre_M_scaled_inverse %*% (z_current-gradientStep(z_candidate,
+     0.5*h)))
+   q_candidate <- -0.5*(t(z_candidate-gradientStep(z_current,0.5*h))
+     %*% pre_M_scaled_inverse %*% (z_candidate-gradientStep(z_current,
+     0.5*h)))
+ }

R> log.ratio <- min(0, p_candidate - p_current + q_current - q_candidate)
```

- iii. The acceptance or rejection decision of  $\mathbf{z}^{\text{candidate}}$  as a new member of the chain can be performed using the following criteria:

$$\mathbf{z}_i = \begin{cases} \mathbf{z}^{\text{candidate}}, & \text{if } \log u \leq \log A, \\ \mathbf{z}^{\text{current}}, & \text{otherwise.} \end{cases}$$

Generate a uniformly distributed random number, denoted  $u$ , between 0 and 1.

```
R> if(log(runif(1)) < log.ratio) {
+   z_current <- z_candidate
+ }

R> if(precision %in% c('MPCR-Double', 'MPCR-Single')){
+   z_trace[, i] <- MPCR.ToNumericVector(z_current)
+ }else{
+   z_trace[, i] <- z_current
+ }
```

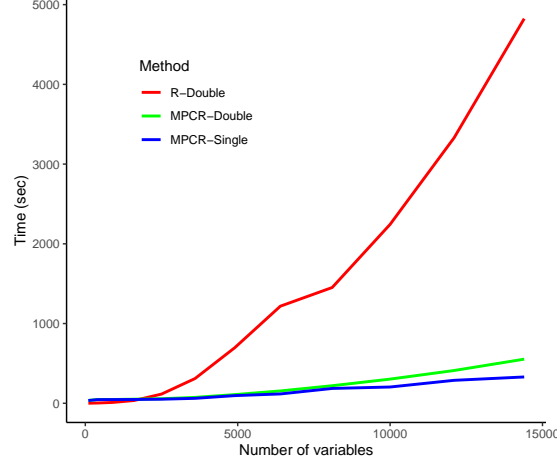


Figure 7: Execution time of MH algorithm using different types of precision for different number of variables.

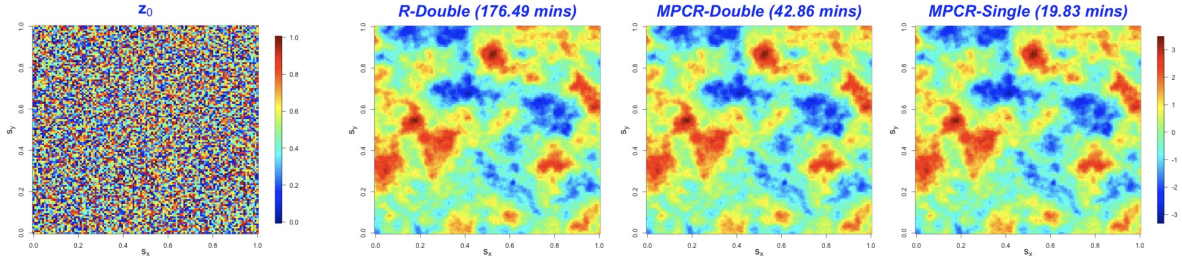


Figure 8: Initial  $\mathbf{z}_0$  and the simulated values after the 2000th iteration using MALA.

Figure 7 shows the execution time when performing the naive MH algorithm with an independent proposal distribution  $\mathcal{U}(\cdot)$  under different execution scenarios, i.e., **R**, **MPCR-Double** and **MPCR-Single**. The plots illustrates the overall speedup that can be obtained using **MPCR** on the simulation tasks under different numerical precisions. Figure 8 shows the 2D spatial fields simulated at the 2000th iteration of MALA. The simulations under different precisions are nearly identical but with a notable difference in execution time. Performing MALA with **MPCR-Single** precision proved to be 9X faster than the naive **R** implementation and 2X faster than **MPCR-Double**.

## 7.2. Maximum Likelihood Estimation of Matérn Covariance Function

Another example demonstrating the advantages of the **MPCR** package is its application in evaluating high-dimensional likelihood functions, which are commonly needed in spatial statistics. Suppose we have a 2D spatial Gaussian random field with 14,400 locations, each associated with a single measurement as shown in Figure 9. Our goal is to model this Gaussian field using the Matérn spatial covariance function:

$$\text{cov}\{Z(\mathbf{s}_i), Z(\mathbf{s}_j)\} = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \mathcal{M}_\nu \left( \frac{\|\mathbf{s}_i - \mathbf{s}_j\|}{a} \right), \quad (7)$$

parameterized by  $\boldsymbol{\theta} = (\nu, a, \sigma^2)^\top$ , where  $\nu > 0, a > 0, \sigma^2 > 0$  are the smoothness, spatial range, and variance parameters, respectively (Guttorp and Gneiting 2006; Matérn 2013). Here,  $\mathcal{M}_\nu(x) = x^\nu \mathcal{K}_\nu(x)$ ,  $\mathcal{K}_\nu(\cdot)$  is the modified Bessel function of the second kind of order  $\nu$ , and  $\Gamma(\cdot)$  is the gamma function.

We can estimate the values of the parameters in the model in (7) via maximum likelihood estimation (MLE), i.e., the best estimates for the model parameters are those that maximize the Gaussian log-likelihood function:

$$l(\boldsymbol{\theta}) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{1}{2} \mathbf{Z}^\top \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1} \mathbf{Z}, \quad (8)$$

where  $|\cdot|$  takes the determinant of  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ . This can be achieved by using non-linear optimizers such as those built-in in R, namely, `optim` and `nlm`. The R package `nloptr`, which serves as the R interface to the `NLopt` library, also provides a suite of different optimization routines that can also be used here. Conventionally, these non-linear optimizers operate in the opposite direction. That is, they work on finding parameters that minimize a given function. Hence, MLE is usually done by minimizing the negative Gaussian log-likelihood.

When  $n$  is large, the MLE becomes impractical due to the extensive execution time and the significant memory required to store the covariance matrix. This is because evaluating the Gaussian log-likelihood function in (8) costs  $O(n^2)$  for storage and  $O(n^3)$  for computation (Abdulah *et al.* 2018) and such costs are incurred at each iteration of the algorithm. This section demonstrates how the MLE operation can be sped up using **MPCR** objects rather than native R objects. Similar to the example in the preceding section, we cast the covariance matrix  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  as an R-Double matrix and as an **MPCR**-Double matrix and compare their corresponding MLE runtimes. MLE can also be performed in lower precisions such as single or half. However, the Cholesky factorization, a key linear algebra operation involved in MLE, is known to suffer badly from round-off errors (Higham and Pranesh 2021; Maddox *et al.* 2022). Hence, casting  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  as **MPCR**-Single may result in  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  being computationally non-positive definite. Nevertheless, despite the sensitivity of some linear algebra operations to lower precisions, we can still reduce the computational costs of MLE by pursuing a mixed-precision arithmetic on  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  such as those proposed in Abdulah *et al.* (2022) and Cao *et al.*

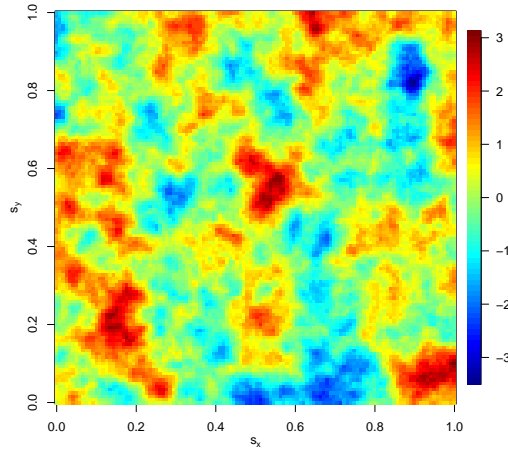


Figure 9: Sample spatial field on a  $120 \times 120$  grid in a unit square generated from a Matérn spatial covariance function model with  $\boldsymbol{\theta} = (1, 0.05, 1)^\top$ .



(2022). Essentially, we can store portions of  $\Sigma(\theta)$  in varying precisions such that  $\Sigma(\theta)$  is kept positive definite but with reduced computational overhead when performing the entire MLE. For our mixed-precision strategy, we cast  $\Sigma(\theta)$  as an **MPCR**-Tile object and employ tile-based linear algebra operations, i.e., tile-based Cholesky factorization. Moreover, we employ the (i) band strategy (Abdulah *et al.* 2022) and the (ii) adaptive precision-aware runtime decision strategy (Cao *et al.* 2022) in determining the appropriate precision for each tile. In the band strategy, we specify the number of tiles closest to the diagonal that will be cast as double-precision, and the rest will be in single-precision. Meanwhile, the adaptive precision-aware runtime decision strategy will rely on the values of the global covariance matrix's Frobenius norm and each tile's Frobenius norm to determine the most appropriate precision for every single tile. Algorithm 1 outlines the steps to performing the adaptive precision-aware runtime decision strategy.

---

**Algorithm 1** Adaptive Precision-Aware Runtime Decision by Cao *et al.* (2022).

---

- 1: Compute the Frobenius norm of  $\Sigma(\theta)$ , i.e.,  $\|\Sigma(\theta)\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n |\Sigma_{ij}(\theta)|^2}$ .
  - 2: Set the value of  $NT$ , the number of tiles in one dimension.
  - 3: Divide  $\Sigma(\theta)$  into tiles, i.e.,  $\Sigma_{ij}(\theta)$ ,  $i = 1, \dots, M_1$  and  $j = 1, \dots, M_2$ , such that  $NT = M_1 M_2$ .
  - 4: Choose two precisions for the mixed-precision strategy and set the values of the variables  $u_{\text{low}}$  and  $u_{\text{high}}$  to their equivalent machine epsilons, e.g.,  $u_{\text{low}} = 2^{-24}$  and  $u_{\text{high}} = 1e^{-8}$ .
  - 5: **for**  $i = 1 : M_1$  **do**
  - 6:     **for**  $j = 1 : M_2$  **do**
  - 7:         Compute the Frobenius norm of tile  $\Sigma_{ij}(\theta)$ .
  - 8:         **if**  $\|\Sigma_{ij}(\theta)\|_F \leq u_{\text{high}} \|\Sigma(\theta)\|_F / (u_{\text{low}} NT)$  **then**
  - 9:             Store tile  $\Sigma_{ij}(\theta)$  in the lower (single) precision.
  - 10:         **else**
  - 11:             Store tile  $\Sigma_{ij}(\theta)$  in the higher (double) precision.
  - 12:         **end if**
  - 13:     **end for**
  - 14: **end for**
- 

In the following, we outline the steps to performing (A) double-precision MLE with R-Double and **MPCR**-Double covariance matrices and (B) mixed-precision MLE involving covariance matrices as **MPCR**-Tile matrices with double- and single-precision.

*A. double-precision*

1. Simulate a sample 2D spatial field from a Matérn spatial covariance function model.

```
R> cov.matern <- function(x, nu, a, sigma_sq){
+   if(nu == 0.5) return(sigma_sq*exp( -x / a))
+   ismatrix <- is.matrix(x)
+   if(ismatrix){nr=nrow(x); nc=ncol(x)}
+   x <- c(x / a)
+   output <- rep(1, length(x))
+   n <- sum(x > 0)
+   if(n > 0) {
```

```

+     x1 <- x[x > 0]
+     output[x > 0] <-
+       (1/((2^(nu - 1)) * gamma(nu))) * (x1^nu) * besselK(x1, nu)
+   }
+   if(ismatrix){
+     output <- matrix(output, nr, nc)
+   }
+   return(sigma_sq * output)
+ }

```

```

R> M <- 120
R> n.loc <- M*M
R> locs <- cbind(rep(0:(M-1), M)/(M-1), rep(0:(M-1), each=M)/(M-1))
R> x <- as.matrix(dist(locs)) # distance matrix
R> theta <- c(1, 0.05, 1) # true parameters
R> cov.R <- cov.matern(x, theta[1], theta[2], theta[3])

```

```

R> # Simulate the spatial field.
R> library(mgcv)
R> set.seed(4)
R> z.R <- rmvn(1, rep(0,M*M), cov.R)

```

2. Create a function that computes the value of the negative Gaussian log-likelihood function. The function `nll` below takes in the argument `pars` which is an R vector object that represents  $\theta$ . Moreover, the parameter values contained in the vector `pars` are transformed to ensure that their values remain within their valid ranges.

```

R> nll <- function(pars, type, precision){
+   nu_param = 2 * 1 / (1 + exp(-pars[1]))
+   a_param = exp(pars[2])
+   sigma_param = exp(pars[3])

+   cat(c(nu_param, a_param, sigma_param), '\n')

+   #Creating the covariance matrix as R-double-precision
+   V <- cov.matern(x, nu_param, a_param, sigma_param)
+   if(type == 'FULL'){
+     if(precision == 'MPCR-Double'){
+       #Casting the covariance matrix as MPCR double-precision
+       cov_mpcr <- as.MPCR(V, nrow=M*M, ncol=M*M, precision='double')
+       L <- chol(cov_mpcr)
+       d <- log(diag(L))
+       log.det.cov <- 2*d$Sum()
+       z.mpcr <- as.MPCR(z.R, nrow=M*M, ncol=1, precision='single')
+       inner_product <- MPCR.trsm(a=L, b=z.mpcr, side='L',
+         upper_triangle=T, transpose=T, alpha=1)
+       inner.prod <- inner_product$SquareSum()
+     }else if(precision == 'R-Double'){

```

```

+       L <- t(chol(V))
+       log.det.cov <- 2*sum(log(diag(L)))
+       z.new <- forwardsolve(L, z.R)
+       inner.prod <- sum(z.new^2)
+     }
+   }else if(type == 'BANDED'){
+     cov.tile <- new(MPCRTile, nr, nc, tr, tc, V, prec.cov.banded)
+     # NOTE: the result of chol is a lower triangular matrix
+     cov.tile_chol <- chol(cov.tile, overwrite = F, num_threads = 4)
+     d <- log(cov.tile_chol$Diag())
+     log.det.cov <- 2*d$Sum()
+     z.tile <- new(MPCRTile, nr, 1, tc, 1, z.R, prec.z)
+     MPCRTile.trsm(a=cov.tile_chol, b=z.tile, side='L',
+       upper_triangle=F, transpose=F, alpha=1)
+     inner.prod <- z.tile$SquareSum()
+   }else if(type == 'ADAPTIVE'){
+     cov.tile <- new(MPCRTile, nr, nc, tr, tc, V, prec.cov)
+     V_norm <- sqrt(sum(V^2))
+     upper_bound <- u_high * V_norm / (nt * u_low)
+     for(ii in 1:tr_total){
+       for(jj in 1:tc_total){
+         test <- MPCRTile.GetTile(cov.tile, ii, jj)
+         tile_norm <- sqrt(test$SquareSum())
+         if(tile_norm < upper_bound){
+           cov.tile$ChangeTilePrecision(ii, jj, "single")
+         }
+       }
+     }
+     # NOTE: the result of chol is a lower triangular matrix
+     cov.tile_chol <- chol(cov.tile, overwrite = F, num_threads = 4)
+     d <- log(cov.tile_chol$Diag())
+     log.det.cov <- 2*d$Sum()
+     z.tile <- new(MPCRTile, nr, 1, tc, 1, z.R, prec.z)
+     MPCRTile.trsm(a=cov.tile_chol, b=z.tile, side='L',
+       upper_triangle=F, transpose=F, alpha=1)
+     inner.prod <- z.tile$SquareSum()
+   }
+   # Computing the negative Gaussian log-likelihood
+   nll <- 0.5*inner.prod + 0.5*log.det.cov + 0.5*n.loc*log(2*pi)
+   return(nll)
+ }

```

3. Perform non-linear optimization. In this example, we use the `nloptr` function with the BOBYQA subroutine (Powell *et al.* 2009) to minimize the log-likelihood function.

```

R> library(nloptr)
R> opts <- list("algorithm" = "NLOPT_LN_BOBYQA", "xtol_rel" = 1e-8,
+   "maxeval" = 1000)

```

```

R> init <- c(-0.3, -1.5, -0.3)
R> fit_R <- nloptr(x0 = init, eval_f = nll, type = 'FULL',
+   precision = 'R-Double', opts = opts)
R> fit_mpcr_double <- nloptr(x0 = init, eval_f = nll,
+   type = 'FULL', precision = 'MPCR-Double', opts = opts)

```

#### B. *Mixed-Precision*

1. Define the tile sizes and the low and high precisions involved in the mixed-precision MLE.

```

R> u_low = 2^(-24)
R> u_high = 1e-8

R> nr <- nc <- n.loc
R> tr <- 2400
R> tc <- 2400
R> tr_total <- n.loc / tr
R> tc_total <- n.loc / tc
R> nt <- tr_total * tc_total

R> # Precisions for z
R> prec.z <- matrix(rep("single", tr_total), tr_total, 1)

R> # Precisions for cov
R> create_prec_banded_mat <- function (prec_matrix, bandwidth) {
+   d <- dim(prec_matrix)
+   outside_band <- .row(d) - bandwidth >= .col(d) |
+     .col(d) - bandwidth >= .row(d)
+   prec_matrix[outside_band] <- 'single'
+   return(prec_matrix)
+ }
R> prec.cov <- matrix(rep("double", nt), n.loc/tr, n.loc/tc)
R> prec.cov.banded <- create_prec_banded_mat(prec.cov, bandwidth=2)

```

2. Perform non-linear optimization.

```

R> fit_tile_adaptive <- nloptr(x0 = init, eval_f = nll,
+   type = 'ADAPTIVE', precision = NULL, opts = opts)
R> fit_tile_banded <- nloptr(x0 = init, eval_f = nll,
+   type = 'BANDED', precision = NULL, opts = opts)

```

Table 1 summarizes the runtimes of different approaches to performing MLE. The results show significant improvement of the mixed-precision strategy over those full precision approaches, with a reduction in the execution time alongside a minimal loss of accuracy.

### 7.3. Principal Component Analysis (PCA)

Another application example is the Principal Component Analysis (PCA) method. PCA is one of the most popular dimensionality-reduction methods in statistics. The goal of PCA is to

Table 1: Summary of MLE results under different types of precision for  $n = 14,400$ . Here, the **nll** column reports the value of the negative Gaussian log-likelihood function when evaluated with the corresponding parameter estimates.

Precision	nll	Parameter Estimates	Execution Time
R-Double	-7,077	$\hat{\theta} = (0.9862619, 0.0513225, 0.9893821)^\top$	48.96 hours
<b>MPCR</b> -Double	-7,077	$\hat{\theta} = (0.9862613, 0.05132348, 0.9894147)^\top$	3.62 hours
<b>MPCR</b> -Mixed (adaptive)	-7,077	$\hat{\theta} = (0.9862476, 0.05126803, 0.9873119)^\top$	3.44 hours
<b>MPCR</b> -Mixed (banded)	-7,077	$\hat{\theta} = (0.9870746, 0.05101309, 0.9812639)^\top$	2.30 hours

construct a new dataset with a relatively small number of independent variables or predictors from a high-dimensional dataset containing a large number of variables or predictors. PCA proceeds by finding a linear combination of the original variables to capture most of the variation observed in the original dataset.

Performing PCA on a space-time dataset is more widely referred to as Empirical Orthogonal Function (EOF) analysis. In EOF analysis, a space-time process  $Z(\mathbf{s}, t)$  can be represented as a linear combination of empirical orthogonal functions (EOFs), i.e.,

$$Z(\mathbf{s}, t) = \sum_{k=1}^K \text{PC}_k(t) \text{EOF}_k(\mathbf{s}),$$

where  $\text{EOF}_k(\mathbf{s})$ , the  $k$ -th EOF evaluated at spatial location  $\mathbf{s}$ , describes the spatial patterns of variability, while  $\text{PC}_k$ , the  $k$ -th principal component evaluated at time  $t$ , describes how the amplitude of the  $k$ -th EOF changes with time. The index  $k$  orders the EOFs such that  $\text{EOF}_1(\mathbf{s})$  corresponds to the spatial pattern that explains the most variance relative to the total variance observed in the data. Likewise,  $\text{EOF}_2(\mathbf{s})$  characterizes the spatial pattern that accounts for the most variance relative to the remaining total variance, and so on.

The EOFs are obtained from the data by computing the spectral decomposition of an empirical covariance matrix or a singular value decomposition (SVD) of the centered data matrix. A matrix  $\mathbf{Z} \in \mathbb{R}^{T \times N}$ , where  $N$  is the number of spatial locations and  $T$  is the number of time points where the observations were made, can be factored using its SVD, i.e.,  $\mathbf{Z} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ , where  $\mathbf{U}$  is a  $T \times T$  orthonormal matrix containing the left singular vectors and  $\mathbf{V}$  is an  $N \times N$  orthonormal matrix containing the right singular vectors. The matrix  $\mathbf{D}$  is a  $T \times N$  diagonal matrix containing the singular values of decreasing magnitude along the main diagonal. The cost of such SVD on a dense space-time data matrix of size  $T \times N$  is  $O(T^2N + N^3)$ , thereby imposing limitations when performing EOF analyses on datasets with large  $T$  and  $N$  (Li *et al.* 2019).

In the following, we show how to use the **MPCR** package to perform EOF analysis on a high-dimensional dataset of 3 hourly averaged zonal wind measurements (in  $m/s$ ) on a regular grid with a nominal  $1^\circ$  horizontal resolution from November 23, 2009 to November 22, 2014, produced by the US National Center for Atmospheric Research (NCAR). We focus the analysis within the region  $[7^\circ N, 85^\circ N] \times [180^\circ W, 20^\circ W]$ . The dataset has 10,707 spatial locations and 14,601 time points.

1. Download the NetCDF file

`b.e21.BHISTcmip6.f09_g17.LE2-1001.001.cam.h3.UBOT.2010010100-2014123100.nc`

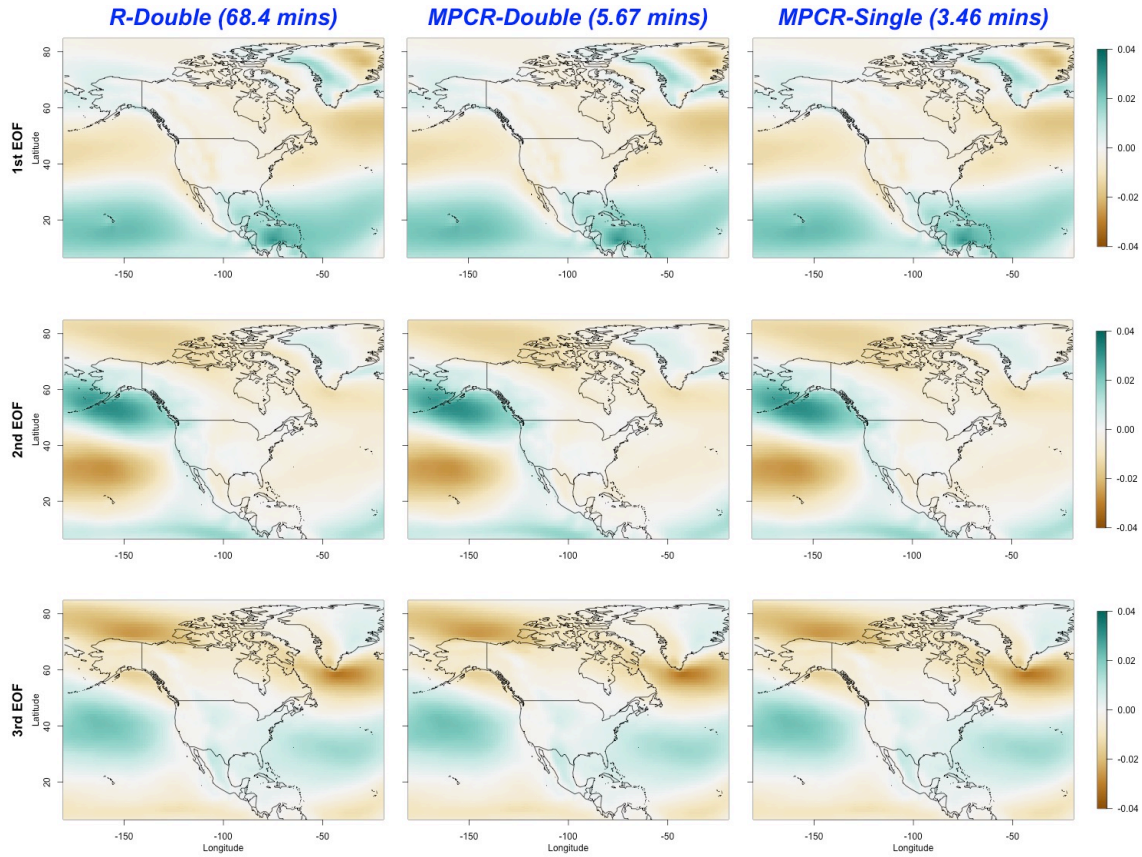


Figure 10: Plots of the first three EOFs of the zonal wind measurements with the corresponding execution time of the SVD using different execution scenarios.

from the link:

[https://www.earthsystemgrid.org/dataset/ucar.cgd.cesm2le.atm.proc.3hourly\\_ave.UBOT.html](https://www.earthsystemgrid.org/dataset/ucar.cgd.cesm2le.atm.proc.3hourly_ave.UBOT.html).

Alternatively, the dataset can be downloaded from this Google Drive link:

[https://drive.google.com/file/d/1rwBUeh80Y\\_wGyhkNKTltdADfsvxudGvP](https://drive.google.com/file/d/1rwBUeh80Y_wGyhkNKTltdADfsvxudGvP).

## 2. Import and process the NetCDF file.

```
R> library(ncdf4)
R> file_name <- "b.e21.BHISTcmip6.f09_g17.LE2-1001.001.cam.h3.UBOT.
+ 2010010100-2014123100.nc"
R> nc <- nc_open(file_name)

R> lat <- ncvar_get(nc,"lat")
R> lon0 <- ncvar_get(nc, "lon")
R> lon <- (lon0+180) %% 360 - 180

R> lat_rng <- c(7,85)
R> lon_rng <- c(-180,-20)
```

```

R> lat_indx <- which(lat >= lat_rng[1] & lat <= lat_rng[2])
R> lon_indx <- which(lon >= lon_rng[1] & lon <= lon_rng[2])

R> #Extract the variable within the region of interest
R> anom_bob <- nvar_get(nc,"UBOT",start = c(lon_indx[1],lat_indx[1],1),
+   count = c(length(lon_indx),length(lat_indx),-1))

```

3. Arrange the data in a matrix

```

R> s2 <- which(!is.na(anom_bob[, ,1]))

R> #Vectorize the array.
R> ubot <- matrix(0,nrow = dim(anom_bob)[3], ncol = length(s2))
R> for(i in 1:dim(anom_bob)[3])
+   ubot[i,] <- anom_bob[, ,i]

```

4. Compute the SVD of the data matrix for 3 precision types.

```

R> #Perform the native R svd() function on an R-Double data matrix.
R> svd_R <- svd(ubot)

R> #Perform the MPCR svd() function on an MPCR-Double data matrix.
R> ubot_mpcr_double <- as.MPCR(ubot, nrow = nrow(ubot), ncol = ncol(ubot),
+   precision = 'double')
R> svd_mpcr_double <- svd(ubot_mpcr_double)

R> #Perform the MPCR svd() function on an MPCR-Single data matrix.
R> ubot_mpcr_single <- as.MPCR(ubot, nrow = nrow(ubot), ncol = ncol(ubot),
+   precision = 'single')
R> svd_mpcr_single <- svd(ubot_mpcr_single)

```

5. Extract the **U**, **V**, and **D** vectors.

```

R> eof_R_u <- svd_R$u
R> eof_R_v <- svd_R$v
R> eof_R_d <- svd_R$d

R> eof_mpcr_double_u <- MPCR.ToNumericMatrix(svd_mpcr_double$u)
R> eof_mpcr_double_v <- MPCR.ToNumericMatrix(svd_mpcr_double$v)
R> eof_mpcr_double_d <- MPCR.ToNumericVector(svd_mpcr_double$d)

R> eof_mpcr_single_u <- MPCR.ToNumericMatrix(svd_mpcr_single$u)
R> eof_mpcr_single_v <- MPCR.ToNumericMatrix(svd_mpcr_single$v)
R> eof_mpcr_single_d <- MPCR.ToNumericVector(svd_mpcr_single$d)

```

6. Visualize the EOFs.

```

R> library(RColorBrewer)
R> library(fields)
R> library(maps)

R> loc <- as.matrix(expand.grid(x=lon[lon_indx],y=lat[lat_indx]))[s2,]
R> coltab <- colorRampPalette(brewer.pal(9,"BrBG"))(2048)

R> #Specify k or the order of the EOF you wish to plot.
R> EOF_INDEX = 1

R> #Map the k-th EOF for R-Double.
R> quilt.plot(loc,eof_R_v[, EOF_INDEX],nx=length(lon[lon_indx]),
+   ny=length(lat[lat_indx]),xlab="",ylab="",col = coltab,
+   zlim = c(-0.04, 0.04))
R> map('world',fill=F, wrap = c(-180, 180), add = TRUE)

R> #Map the k-th EOF for MPCR-Double.
R> quilt.plot(loc,eof_mpcr_double_v[, EOF_INDEX],nx=length(lon[lon_indx]),
+   ny=length(lat[lat_indx]),xlab="",ylab="",col = coltab,
+   zlim = c(-0.04, 0.04))
R> map('world',fill=F, wrap = c(-180, 180), add = TRUE)

R> #Map the k-th EOF for MPCR-Single.
R> quilt.plot(loc,eof_mpcr_single_v[, EOF_INDEX],nx=length(lon[lon_indx]),
+   ny=length(lat[lat_indx]),xlab="",ylab="",col = coltab,
+   zlim = c(-0.04, 0.04))
R> map('world',fill=F, wrap = c(-180, 180), add = TRUE)

7. Compute the percent of variance explained by each EOF.

R> plot(100*(eof_R_d)^2/sum(eof_R_d^2), pch = 16,
+   ylab="Percentage of variance [%]",
+   xlab="", xlim = c(1,20), col = "red")
R> lines(100*(eof_mpcr_double_d)^2/sum(eof_mpcr_double_d^2),
+   col = "green", lwd = 2, lty = 1)
R> lines(100*(eof_mpcr_single_d)^2/sum(eof_mpcr_single_d^2),
+   col = "blue", lwd = 4, lty = 3)

```

The EOFs are the column vectors of  $\mathbf{V}$ , and the PCs are column vectors of  $\mathbf{U}$ . Figure 10 shows a map of the first three EOFs. It can be seen that the EOFs obtained when performing SVD under the three different scenarios are visually identical.

The squares of the singular values in the main diagonal of  $\mathbf{D}$  specify the variance that each EOF explains. Figure 11 plots the percentage of the variance attributable to each of the first twenty EOFs. Like the EOF results, the percent of explained variances are visually identical across different precision types.

#### 7.4. Integrated Nested Single-Half-Arithmetic Laplace Approximation (IN-



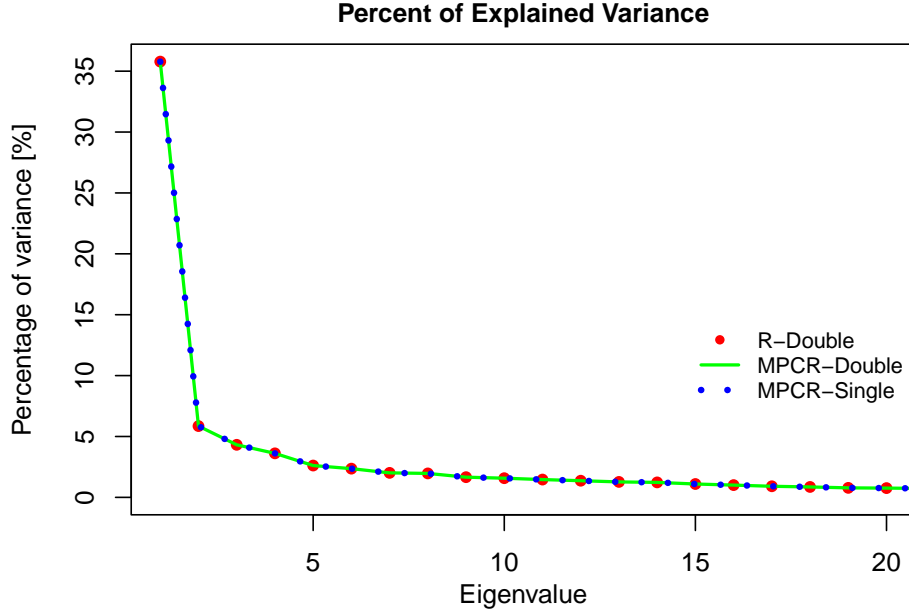


Figure 11: Plots of the percentage of variance explained by the first 20 ordered eigenvalues using different execution scenarios. Under all three scenarios, the first three EOFs explain  $\sim 36\%$ ,  $\sim 6\%$ , and  $\sim 4\%$  of the total variance, respectively.

## SHALA)

Another example to show the capabilities of the **MPCR** package is large-scale Bayesian inference problems, which operate on large precision matrices.

The integrated nested Laplace approximations (INLA) method developed by Rue *et al.* (2009) is a widely known and applied framework for approximate Bayesian inference. INLA is formulated for latent Gaussian models, including but not limited to linear mixed models, spatial, and spatio-temporal models (Rue *et al.* 2017; Bakka *et al.* 2018). INLA has a well-maintained dedicated R package called **R-INLA** (<http://www.r-inla.org>).

A detailed discussion of each step in the INLA method, complete with R code examples, is accessible at the website <https://stefansiebert.net/inla-project/inla-from-scratch>. In the following demonstration, we closely align with their presentation, highlighting how the **MPCR** functions can be seamlessly incorporated into their coding framework.

Consider the following hierarchical model:

- The observations  $y_1, \dots, y_n$  are from a Bernoulli process with parameters  $p_1, \dots, p_n$ , i.e.,

$$p(y_t|p_t) = \begin{cases} p_t, & y_t = 1, \\ 1 - p_t, & y_t = 0, \end{cases}$$

where  $t = 1, 2, \dots, n$ .

- The parameters  $p_1, \dots, p_n$  are assumed to change over time smoothly, and they depend on a latent process  $x_1, \dots, x_n$  through a logit link:

$$p_t = \frac{e^{\beta x_t}}{1 + e^{\beta x_t}}.$$

- The process  $x_1, \dots, x_n$  is a Gaussian process, i.e.,  $(x_1, \dots, x_n)^\top = \{X(t_1), \dots, X(t_n)\}^\top \sim \mathcal{N}_n\{\mathbf{0}, \Sigma(\alpha)\}$ , where  $\Sigma_{ij}(\alpha) = \text{cov}\{X(t_i), X(t_j)\} = \sigma^2 \exp(-|t_i - t_j|/\alpha)$ .
- The hyperparameter  $\alpha$  has a uniform prior  $U(0, 1)$ .

The joint log-density of the observations, latents, and hyperparameters is given by

$$\begin{aligned} \log\{p(\mathbf{y}, \mathbf{x}, \alpha)\} &= \log\{p(\mathbf{y}|\mathbf{x}, \alpha)\} + \log\{p(\mathbf{x}|\alpha)\} + \log\{p(\alpha)\} \\ &\propto \sum_t^n \{\beta x_t y_t - \log(1 + e^{\beta x_t})\} + \frac{1}{2} \log |\mathbf{Q}| - \frac{1}{2} \mathbf{x}^\top \mathbf{Q} \mathbf{x} + \log\{p(\alpha)\}, \end{aligned} \quad (9)$$

where  $\mathbf{Q} = \Sigma^{-1}(\alpha)$  is the inverse covariance matrix or more commonly termed precision matrix. The term “precision” in the context of the precision matrix  $\mathbf{Q}$  should not be confused with the term “precision” used to describe numerical formats, which is the subject of this paper.

Fitting a model using the INLA method has many operations based on linear algebra where the INLA method and the **R-INLA** package proved to be computationally efficient. However, this efficiency results from working only with sparse matrices (Abdul-Fattah *et al.* 2023). In this regard, **MPCR** can be used to overcome the obstacles INLA encounters with dense matrices. Herein, we introduce the Integrated Nested Single-Half-Arithmetic Laplace Approximation (INSHALA) powered by the **MPCR** package. INSHALA is a more computationally efficient version of INLA that operates in low precisions such as 32-bit (single) precision formats. In the model in (9) above, for example, the INSHALA method would store the precision matrix  $\mathbf{Q}$  as an **MPCR**-Single-precision matrix to improve computational performance.

In the following, we show how the INLA-inspired INSHALA methodology works. In particular, we illustrate how to approximate the posterior distribution of  $\alpha$  with **MPCR**-Double and **MPCR**-Single-precision computations. Furthermore, the classical INLA modeling implemented using native R objects is also shown for comparison.

Note that the log posterior distribution of the hyperparameter  $\alpha$  has the following expression:

$$\log\{p(\alpha|\mathbf{y})\} \propto \log\{p(\mathbf{y}, \mathbf{x}, \alpha)\} - \log\{p(\mathbf{x}|\alpha, \mathbf{y})\}, \quad (10)$$

which is derived from another factorization of the joint log-density in (9). Here the value of  $\log\{p(\mathbf{y}, \mathbf{x}, \alpha)\}$  can be computed by a straightforward evaluation of the expression in (9). The value of the conditional log distribution of the latent variables  $\log\{p(\mathbf{x}|\alpha, \mathbf{y})\}$ , on the other hand, is obtained using Laplace approximation. A key step in the INLA method is to define a function  $f(\mathbf{x})$  such that by applying the definition of conditional probability, we can express  $p(\mathbf{x}|\alpha, \mathbf{y})$  as:

$$p(\mathbf{x}|\alpha, \mathbf{y}) = \frac{e^{f(\mathbf{x})}}{\int e^{f(\mathbf{x})} d\mathbf{x}}, \quad (11)$$

and then apply a Laplace approximation to the denominator  $\int e^{f(\mathbf{x})} d\mathbf{x}$ , i.e.,

$$\int e^{f(\mathbf{x})} d\mathbf{x} \approx e^{f(\mathbf{x}_0)} (2\pi)^{n/2} |\mathbf{H}f(\mathbf{x}_0)|^{-1/2}. \quad (12)$$

Here,  $\mathbf{x}_0$  is the mode of  $f(\mathbf{x})$  and  $\mathbf{H}f(\mathbf{x}_0)$  is the Hessian matrix of  $f(\mathbf{x})$  evaluated at the mode. Note that the form in (12) appears as a result of a second-order Taylor expansion of  $f(\mathbf{x})$  around the mode (see Opitz (2017); Flag and Hoegh (2023) for some derivations).

Applying (12) to (11), we get the Laplace approximation of the conditional log distribution  $\log\{p(\mathbf{x}|\alpha, \mathbf{y})\}$  evaluated at mode  $\mathbf{x}_0$  as follows:

$$\log\{p(\mathbf{x}_0|\alpha, \mathbf{y})\} \approx -\frac{n}{2} \log(2\pi) + \frac{1}{2} \log |-\mathbf{H}f(\mathbf{x}_0)|. \quad (13)$$

By substituting (13) to (10), the approximation of the log posterior distribution in (10) is:

$$\log\{p(\alpha|\mathbf{y})\} \propto \log\{p(\mathbf{y}, \mathbf{x}_0, \alpha)\} - \frac{1}{2} \log |-\mathbf{H}f(\mathbf{x}_0)|. \quad (14)$$

In order to compute the posterior distribution of  $\alpha$  using the formula in (14), we need to know the form of  $f(\mathbf{x})$  and consequently its Hessian matrix  $\mathbf{H}f(\mathbf{x})$ . In the example above, a great choice for  $f(\mathbf{x})$  in (11) is such that  $f(\mathbf{x}) \propto \log p(\mathbf{y}, \mathbf{x}, \alpha)$ . From the formulation of  $\log p(\mathbf{y}, \mathbf{x}, \alpha)$  in (9), we write  $f(\mathbf{x})$  as a function of the components that depend on  $\mathbf{x}$  as follows:

$$f(\mathbf{x}) = \sum_t^n \{\beta x_t y_t - \log(1 + e^{\beta x_t})\} - \frac{1}{2} \mathbf{x}^\top \mathbf{Q} \mathbf{x}. \quad (15)$$

By the above form of  $f(\mathbf{x})$ , the gradient vector and Hessian matrix of  $f(\mathbf{x})$  are the following:

$$\nabla f(\mathbf{x}) = \text{vec} \left\{ \beta y_t - \frac{\beta e^{\beta x_t}}{1 + e^{\beta x_t}} \right\} - \mathbf{Q} \mathbf{x}, \text{ and} \quad (16)$$

$$\mathbf{H}f(\mathbf{x}) = -\mathbf{Q} - \text{diag} \left\{ \frac{\beta^2 e^{\beta x_t}}{(1 + e^{\beta x_t})^2} \right\}, \quad (17)$$

where  $\nabla f(\mathbf{x})$  is a column vector of size  $t \times 1$  and  $\mathbf{H}f(\mathbf{x})$  is a  $t \times t$  matrix. Here,  $\text{vec}()$  constructs a column vector while  $\text{diag}()$  builds a diagonal matrix.

Finally, the mode  $\mathbf{x}_0$  of  $f(\mathbf{x})$  can be found iteratively using the steps outlined in Algorithm 2.

In the following, we outline how to perform the approximation of  $\log\{p(\alpha|\mathbf{y})\}$  in (14) via the INSHALA method using the **MPCR** package.

1. Simulate the data.

(a) Simulate  $x_1, \dots, x_n$  from an exponential covariance function model.

```
R> library(mgcv)
R> library(Matrix)
R> library(dplyr)

R> n <- 50~2
R> alpha_true = 0.6
R> sigma_true = 0.1

R> x <- seq(0, n, length.out = n)
R> locs <- cbind(x, 0)
R> dist0 <- as.matrix(dist(locs)) # distance matrix in time
R> cov.R <- sigma_true * exp(-dist0 / alpha_true)
```

---

**Algorithm 2** Solving for the mode  $\mathbf{x}_0$  iteratively.

Source: <https://stefansiebert.net/inla-project/inla-from-scratch>

---

- 1: Set an initial value for  $\mathbf{x}_0$ .
- 2: Taylor expand  $f(\mathbf{x})$  around  $\mathbf{x}_0$  and keep only the terms that depend on  $\mathbf{x}$ , i.e.,

$$\begin{aligned} f(\mathbf{x}) &\approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^\top \mathbf{H}f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) \\ &\propto \{\nabla f(\mathbf{x}_0)^\top - \mathbf{x}_0^\top \mathbf{H}f(\mathbf{x}_0)\} \mathbf{x} + \frac{1}{2} \mathbf{x}^\top \mathbf{H}f(\mathbf{x}_0) \mathbf{x}. \end{aligned}$$

- 3: Find the mode of the second-order Taylor expansion above. This mode serves as a new and improved estimate of the true mode of  $f(\mathbf{x})$ . This can be done by setting its first derivative to zero and solving for  $\mathbf{x}$ , i.e.,

$$\begin{aligned} \nabla f(\mathbf{x}) &\propto \nabla f(\mathbf{x}_0) - \mathbf{H}f(\mathbf{x}_0)\mathbf{x}_0 + \mathbf{H}f(\mathbf{x}_0)\mathbf{x} \\ &= \text{vec} \left\{ \beta y_t - \frac{\beta e^{\beta x_{0,t}}}{1 + e^{\beta x_{0,t}}} \right\} - \mathbf{Q}\mathbf{x}_0 - \left[ -\mathbf{Q} - \text{diag} \left\{ \frac{\beta^2 e^{\beta x_{0,t}}}{(1 + e^{\beta x_{0,t}})^2} \right\} \right] \mathbf{x}_0 \\ &\quad + \left[ -\mathbf{Q} - \text{diag} \left\{ \frac{\beta^2 e^{\beta x_{0,t}}}{(1 + e^{\beta x_{0,t}})^2} \right\} \right] \mathbf{x} \\ &= \text{vec} \left\{ \beta y_t - \frac{\beta e^{\beta x_{0,t}}}{1 + e^{\beta x_{0,t}}} \right\} + \text{diag} \left\{ \frac{\beta^2 e^{\beta x_{0,t}}}{(1 + e^{\beta x_{0,t}})^2} \right\} \mathbf{x}_0 \\ &\quad - \left[ \mathbf{Q} + \text{diag} \left\{ \frac{\beta^2 e^{\beta x_{0,t}}}{(1 + e^{\beta x_{0,t}})^2} \right\} \right] \mathbf{x} = \mathbf{0}, \end{aligned}$$

where the mode of the second-order Taylor expansion of  $f(\mathbf{x})$  is the solution to the equation

$$\mathbf{x} = \left[ \mathbf{Q} + \text{diag} \left\{ \frac{\beta^2 e^{\beta x_{0,t}}}{(1 + e^{\beta x_{0,t}})^2} \right\} \right]^{-1} \left[ \text{vec} \left\{ \beta y_t - \frac{\beta e^{\beta x_{0,t}}}{1 + e^{\beta x_{0,t}}} \right\} + \text{diag} \left\{ \frac{\beta^2 e^{\beta x_{0,t}}}{(1 + e^{\beta x_{0,t}})^2} \right\} \mathbf{x}_0 \right].$$

- 4: Update the value of  $\mathbf{x}_0$  to the value of  $\mathbf{x}$  above, i.e.,  $\mathbf{x}_0 \leftarrow \mathbf{x}$ .
  - 5: Repeat until the distance between  $\mathbf{x}_0$  and  $\mathbf{x}$  is less than a user-defined tolerance, `tol`, i.e.,  $\|\mathbf{x}_0 - \mathbf{x}\|^2 < \text{tol}$ .
- 

```
R> nr <- nrow(cov.R)
R> nc <- ncol(cov.R)

R> set.seed(4)
R> # Simulate x_1, x_2, ..., x_n
R> x_true <- rmvn(1, rep(0, n), cov.R)
```

- (b) Simulate  $p_1, \dots, p_n$  using the logit link.

```
R> beta_true = 10
R> p_true <- 1 / (1 + exp(-beta_true * x_true))
```

(c) Simulate  $y_t$  as Bernoulli( $p_t$ ).

```
R> y <- rbinom(n, 1, p_true)
```

2. Create a function that computes  $\mathbf{Q}$ .

```
R> calc_Q = function(alpha, precision) {
+   cov_R <- cov <- sigma_true * exp( - dist0 / alpha)
+   if(precision == 'MPCR-Double'){
+     cov <- as.MPCR(as.matrix(cov_R), nrow=nrow(cov_R), ncol=ncol(cov_R),
+       precision='double')
+   }else if(precision == 'MPCR-Single'){
+     cov <- as.MPCR(as.matrix(cov_R), nrow=nrow(cov_R), ncol=ncol(cov_R),
+       precision='single')
+   }
+   Q <- solve(cov)
+   return(Q)
+ }
```

3. Create a function that computes the log prior for  $\alpha$ .

```
R> calc_lprior = function(alpha, a=1, b=1) {
+   (a-1) * log(alpha) + (b-1) * log(1 - alpha)
+ }
```

4. Create a function that computes the log joint distribution in (9).

```
R> calc_ljoint = function(y, x, alpha, a=1, b=1, precision) {
+   Q <- calc_Q(alpha, precision)
+   chol_Q <- chol(Q)
+   if(precision %in% c('MPCR-Double', 'MPCR-Single')){
+     d <- log(diag(chol_Q))
+     logdet_Q_half <- d$Sum()
+     x.mpcr <- as.MPCR(x, nrow=length(x), ncol=1, precision='single')
+     inner_product <- chol_Q %*% x.mpcr
+     quad_form <- inner_product$SquareSum()
+   }else{
+     logdet_Q_half <- chol_Q %>% diag %>% log %>% sum
+     quad_form <- crossprod(chol_Q %*% x) %>% drop
+   }
+   res <- sum(beta_true * x * y - log1p(exp(beta_true * x))) +
+     logdet_Q_half - 0.5 * quad_form + calc_lprior(alpha, a, b)
+   return(res)
+ }
```

5. Create a function that computes  $f(\mathbf{x})$  in (15), its gradient (16) and its negative Hessian (17).

```
R> calc_ff = function(x, alpha, precision) {
+   sum(beta_true * x * y - log1p(exp(beta_true * x))) -
```

```

+      0.5 * drop(as.matrix(x %**% calc_Q(alpha, precision) %**% x))
+ }

R> calc_grad_ff = function(x, alpha, precision) {
+   beta_true * y -
+   beta_true * exp(beta_true * x) / (1 + exp(beta_true * x)) -
+   drop(as.matrix(calc_Q(alpha, precision) %**% x))
+ }

R> calc_neg_hess_ff = function(x, alpha, precision) {
+   Q <- calc_Q(alpha, precision)
+   other_term <- diag(beta_true^2 * exp(beta_true * x) /
+     (1 + exp(beta_true * x))^2)
+   if(precision %in% c('MPCR-Double', 'MPCR-Single')){
+     other_term.mpcr <- as.MPCR(other_term, nrow=nrow(other_term),
+       ncol=ncol(other_term), precision='single')
+     return(Q + other_term.mpcr)
+   }else{
+     return(Q + other_term)
+   }
+ }

```

6. Create a function that computes the mode of  $f(\mathbf{x})$ .

```

R> # the function  $g(x) = \log(p(y | x, \theta))$ , its gradient and hessian
R> calc_g0 = function(x) {
+   sum(beta_true * x * y - log1p(exp(beta_true*x)))
+ }

R> calc_g1 = function(x) {
+   beta_true * y - beta_true * exp(beta_true*x) / (1 + exp(beta_true*x))
+ }

R> calc_g2 = function(x) {
+   (-1) * beta_true^2 * exp(beta_true*x) / (1 + exp(beta_true*x))^2
+ }

R> calc_x0 = function(alpha, precision, tol=1e-12) {
+   Q <- calc_Q(alpha, precision)
+   x <- x0 <- rep(0, n)
+   while(1) {
+     g1 <- calc_g1(x)
+     g2 <- calc_g2(x)
+     diag_g <- bandSparse(n=n, k=0, diagonals=list(g2))
+     mode <- (g1 - x0 * g2)
+     if(precision %in% c('MPCR-Double', 'MPCR-Single')){
+       diag_g.mpcr <- as.MPCR(as.matrix(diag_g), nrow=nrow(diag_g),

```

```

+       ncol=ncol(diag_g), precision='double')
+       mode.mpcr <- as.MPCR(mode, nrow=length(x0), ncol=1,
+         precision='single')
+       x <- MPCR.ToNumericVector(solve(Q - diag_g.mpcr) %*% mode.mpcr)
+     }else{
+       x <- drop(solve(Q - diag_g) %*% mode)
+     }
+     if(mean((x-x0)^2 < tol)) {
+       break
+     } else {
+       x0 <- x
+     }
+   }
+   return(x)
+ }

```

7. Create a function that approximates the log posterior in Equation (14) up to an additive constant.

```

R> calc_lpost = function(alpha, precision) {
+   x0 <- calc_x0(alpha, precision)
+   chol_h <- chol(calc_neg_hess_ff(x0, alpha, precision))
+   if(precision %in% c('MPCR-Double', 'MPCR-Single')){
+     d <- log(diag(chol_h))
+     logdet_h_half <- d$Sum()
+   }else{
+     logdet_h_half <- sum(log(diag(chol_h)))
+   }
+   calc_ljoint(y, x0, alpha, precision = precision) - logdet_h_half
+ }

```

8. For each precision type, compute the approximation of the log posterior via the iterative method.

```

R> alpha_vec = seq(0.05, 0.95, len=21)
R> lpost_R = sapply(alpha_vec, calc_lpost, precision = 'R-Double')
R> lpost_mpcr_d = sapply(alpha_vec, calc_lpost, precision='MPCR-Double')
R> lpost_mpcr_s = sapply(alpha_vec, calc_lpost, precision='MPCR-Single')

```

9. Create a function that normalizes the posterior via numerical integration.

```

R> calc_Z = function(alpha_vec, lpost_vec) {
+   nn <- length(alpha_vec)
+   hh <- alpha_vec[2] - alpha_vec[1]
+   ww <- c(1, rep(c(4,2), (nn-3)/2), c(4,1))
+   return(sum(ww * exp(lpost_vec)) * hh / 3)
+ }

```

10. Normalize the posterior.

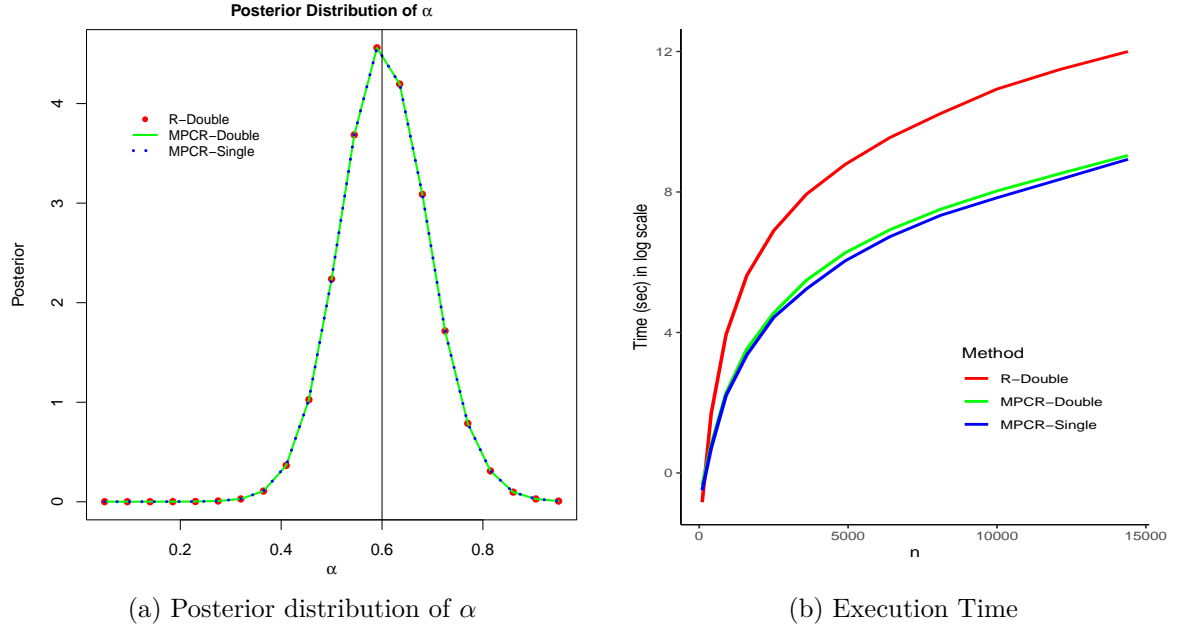


Figure 12: Results of the INSHALA experiments: (a) Approximation of the posterior distribution of  $\alpha$  for  $n = 2500$ . The bold vertical line marks the true value of  $\alpha = 0.6$ . (b) Execution time of computing the posterior distribution for various  $n$  under different types of precision.

```
R> lpost_R = lpost_R - mean(lpost_R)
R> Z_R = calc_Z(alpha_vec, lpost_R)

R> lpost_mpcr_d = lpost_mpcr_d - mean(lpost_mpcr_d)
R> Z_mpcr_double = calc_Z(alpha_vec, lpost_mpcr_d)

R> lpost_mpcr_s = lpost_mpcr_s - mean(lpost_mpcr_s)
R> Z_mpcr_single = calc_Z(alpha_vec, lpost_mpcr_s)
```

We run the aforementioned for different sizes of the precision matrix  $\mathbf{Q}$ . Figure 12a plots the resulting posterior distribution of  $\alpha$  when  $n = 2,500$  for each precision type. From the figure, it can be seen that the INSHALA method yields identical results to that of the INLA method which was obtained using R-double-precision. Figure 12b plots the execution time in log scale when running the INLA and INSHALA procedures. From the figure, it can be seen that for reasonably large values of  $n$ , our proposed INSHALA method yields a non-trivial decrease in execution time while maintaining results identical to those of the INLA method.

## 8. Discussion and Future Work

Employing multi- and mixed-precision computation in scientific applications can significantly speedup the processing of extensive, high-dimensional problems. In this work, we introduce **MPCR**, a package designed to facilitate multi- and mixed-precision computation within the R environment. This package is integrated with well-tuned BLAS/LAPACK libraries, e.g., Intel MKL and OpenBLAS, ensuring optimal performance on available hardware. Herein,



we demonstrate the usage of the package in R through various examples and show how it can enhance the efficiency of existing applications from the execution time perspective by reducing the computation precision while preserving the same accuracy of double precision. The package also provides mixed-precision support of matrix operations that enable faster linear algebra algorithms, e.g., tile-based linear algebra using parallel execution and a mix of precisions, i.e., 32-bit and 64-bit, through leverages OpenMP for parallelization on multicore systems. **MPCR** also offers a robust framework that can support additional precisions, like 128-bit and 16-bit, provided the underlying BLAS/LAPACK library and hardware architecture are compatible. The seamless integration with the optimized BLAS/LAPACK libraries positions **MPCR** as a highly effective alternative to traditional R operations, mainly when R is linked to a less optimized library such as RBLAS.

We plan to extend the **MPCR** package in our future work by integrating GPU support as a high-performance backend. This upgrade will empower R users to execute their linear algebra operations more efficiently. Additionally, we aim to incorporate support for 16-bit precision through tensor core technology available in modern GPU hardware, such as V100, A100, and H100. These advanced technologies can offer a significant speed boost compared to higher precision operations.

## Computational details

The results in this paper were obtained using R 4.3.1 with the **MPCR** 1.0.0 package. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at <https://CRAN.R-project.org/>.

The experiments were conducted on an aarch64-apple-darwin20 (64-bit) platform running under macOS Sonoma 14.0.

## Acknowledgment

This study was funded by the King Abdullah University of Science and Technology (KAUST) in Saudi Arabia. We would like to thank the team at BrightSkies for their technical assistance in the development of the **MPCR** package and to Omar Marzouk and Merna Moawad for their support.

## References

- Abdelfattah A, Anzt H, Boman EG, Carson E, Cojean T, Dongarra J, Gates M, Grützmacher T, Higham NJ, Li S, *et al.* (2020). “A survey of numerical methods utilizing mixed precision arithmetic.” *arXiv preprint arXiv:2007.06674*.
- Abdul-Fattah E, Van Niekerk J, Rue H (2023). “INLA+—Approximate Bayesian inference for non-sparse models using HPC.” *arXiv preprint arXiv:2311.08050*.
- Abdulah S, Cao Q, Pei Y, Bosilca G, Dongarra J, Genton MG, Keyes DE, Ltaief H, Sun Y (2022). “Accelerating geostatistical modeling and prediction with mixed-precision compu-

- tations: A high-productivity approach with parsec.” *IEEE Transactions on Parallel and Distributed Systems*, **33**(4), 964–976.
- Abdulah S, Ltaief H, Sun Y, Genton MG, Keyes DE (2018). “Exageostat: A high performance unified software for geostatistics on manycore systems.” *IEEE Transactions on Parallel and Distributed Systems*, **29**(12), 2771–2784.
- Abdulah S, Ltaief H, Sun Y, Genton MG, Keyes DE (2019). “Geostatistical modeling and prediction using mixed precision tile Cholesky factorization.” In *2019 IEEE 26th international conference on high performance computing, data, and analytics (HiPC)*, pp. 152–162. IEEE.
- Agullo E, Demmel J, Dongarra J, Hadri B, Kurzak J, Langou J, Ltaief H, Luszczek P, Tomov S (2009). “Numerical linear algebra on emerging architectures: The PLASMA and MAGMA projects.” In *Journal of Physics: Conference Series*, volume 180, p. 012037. IOP Publishing.
- Akbudak K, Ltaief H, Mikhalev A, Keyes D (2017). “Tile low rank Cholesky factorization for climate/weather modeling applications on manycore architectures.” In *International Conference on High Performance Computing*, pp. 22–40. Springer.
- Anderson E, Bai Z, Bischof C, Blackford S, Dongarra JDJ, Croz JD, Greenbaum A, Hammarling S, McKenney A, Sorensen D (1999). *LAPACK Users’ Guide*. Third edition. SIAM, Philadelphia, Pennsylvania, USA.
- Au SK, Beck JL (2001). “Estimation of small failure probabilities in high dimensions by subset simulation.” *Probabilistic Engineering Mechanics*, **16**(4), 263–277.
- Bakka H, Rue H, Fuglstad GA, Riebler A, Bolin D, Illian J, Krainski E, Simpson D, Lindgren F (2018). “Spatial modeling with R-INLA: A review.” *Wiley Interdisciplinary Reviews: Computational Statistics*, **10**(6), e1443.
- Beaumont O, Langou J, Quach W, Shilova A (2020). “A makespan lower bound for the tiled cholesky factorization based on alap schedule.” In *Euro-Par 2020: Parallel Processing: 26th International Conference on Parallel and Distributed Computing, Warsaw, Poland, August 24–28, 2020, Proceedings*, pp. 134–150. Springer.
- Benner P, Iannazzo B, Meini B, Palitta D (2022). “Palindromic linearization and numerical solution of nonsymmetric algebraic T-Riccati equations.” *BIT Numerical Mathematics*, **62**(4), 1649–1672.
- Björck J, Chen X, De Sa C, Gomes CP, Weinberger K (2021). “Low-precision reinforcement learning: running soft actor-critic in half precision.” In *International Conference on Machine Learning*, pp. 980–991. PMLR.
- Blackford LS, Choi J, Cleary A, Petitet A, Whaley RC, Demmel J, Dhillon I, Stanley K, Dongarra J, Hammarling S, Henry G, Walker D (1996). “ScaLAPACK: A Portable Linear Algebra Library for Distributed Memory Computers — Design Issues and Performance.” pp. 95–106. URL <http://www.supercomp.org/sc96/proceedings/SC96PROC/DONGARRA/INDEX.HTM>.

- Bosilca G, Bouteiller A, Danalis A, Herault T, Lemarinier P, Dongarra J (2012). “DAGuE: A generic distributed DAG engine for high performance computing.” *Parallel Computing*, **38**(1-2), 37–51.
- Bujanović Z, Kressner D, Schröder C (2023). “Iterative refinement of Schur decompositions.” *Numerical Algorithms*, **92**(1), 247–267.
- Buttari A, Langou J, Kurzak J, Dongarra J (2008). “Parallel tiled QR factorization for multicore architectures.” *Concurrency and Computation: Practice and Experience*, **20**(13), 1573–1590.
- Buttari A, Langou J, Kurzak J, Dongarra J (2009). “A class of parallel tiled linear algebra algorithms for multicore architectures.” *Parallel Computing*, **35**(1), 38–53.
- Cao Q, Abdulah S, Alomairy R, Pei Y, Nag P, Bosilca G, Dongarra J, Genton MG, Keyes DE, Ltaief H, *et al.* (2022). “Reshaping geostatistical modeling and prediction for extreme-scale environmental applications.” In *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–12. IEEE.
- Carson E, Lund K, Rozložník M, Thomas S (2022). “Block Gram-Schmidt algorithms and their stability properties.” *Linear Algebra and its Applications*, **638**, 150–195.
- Chantry M, Thornes T, Palmer T, Düben P (2019). “Scale-selective precision for weather and climate forecasting.” *Monthly Weather Review*, **147**(2), 645–655.
- Cherezov A, Vasiliev A, Ferroukhi H (2023). “Acceleration of Nuclear Reactor Simulation and Uncertainty Quantification Using Low-Precision Arithmetic.” *Applied Sciences*, **13**(2), 896.
- Cifani P, Viviani M, Modin K (2023). “An efficient geometric method for incompressible hydrodynamics on the sphere.” *Journal of Computational Physics*, **473**, 111772.
- Cornish R, Vanetti P, Bouchard-Côté A, Deligiannidis G, Doucet A (2019). “Scalable Metropolis-Hastings for exact Bayesian inference with large datasets.” In *International Conference on Machine Learning*, pp. 1351–1360. PMLR.
- Dongarra J, Faverge M, Ltaief H, Luszczek P (2014). “Achieving numerical accuracy and high performance using recursive tile LU factorization with partial pivoting.” *Concurrency and Computation: Practice and Experience*, **26**(7), 1408–1431.
- Durmus A, Roberts GO, Vilmart G, Zygalakis KC (2017). “Fast Langevin based algorithm for MCMC in high dimensions.” *The Annals of Applied Probability*, **27**(4), 2195–2237.
- Flagg K, Hoegh A (2023). “The integrated nested Laplace approximation applied to spatial log-Gaussian Cox process models.” *Journal of Applied Statistics*, **50**(5), 1128–1151.
- Freytag G, Lima JV, Rech P, Navaux PO (2022). “Impact of Reduced and Mixed-Precision on the Efficiency of a Multi-GPU Platform on CFD Applications.” In *International Conference on Computational Science and Its Applications*, pp. 570–587. Springer.
- Guivant J, Narula K, Kim J, Li X, Khan S (2023). “Compressed Gaussian Estimation under Low Precision Numerical Representation.” *Sensors*, **23**(14), 6406.

- Guttorp P, Gneiting T (2006). “Studies in the history of probability and statistics XLIX on the Matérn correlation family.” *Biometrika*, **93**(4), 989–995.
- Haidar A, Ltaief H, YarKhan A, Dongarra J (2012). “Analysis of dynamically scheduled tile algorithms for dense linear algebra on multicore architectures.” *Concurrency and Computation: Practice and Experience*, **24**(3), 305–321.
- Haidar A, YarKhan A, Cao C, Luszczek P, Tomov S, Dongarra J (2015). “Flexible linear algebra development and scheduling with cholesky factorization.” In *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems*, pp. 861–864. IEEE.
- Hastings WK (1970). “Monte Carlo sampling methods using Markov chains and their applications.” *Biometrika*, **57**(1), 97–109.
- Hatfield S, Chantry M, Düben P, Palmer T (2019). “Accelerating high-resolution weather models with deep-learning hardware.” In *Proceedings of the platform for advanced scientific computing conference*, pp. 1–11.
- Hatfield S, McRae A, Palmer T, Düben P (2020). “Single-precision in the tangent-linear and adjoint models of incremental 4d-var.” *Monthly Weather Review*, **148**(4), 1541–1552.
- Higham NJ, Mary T (2022a). “Mixed precision algorithms in numerical linear algebra.” *Acta Numerica*, **31**, 347–414.
- Higham NJ, Mary T (2022b). “Solving block low-rank linear systems by LU factorization is numerically stable.” *IMA Journal of Numerical Analysis*, **42**(2), 951–980.
- Higham NJ, Pranesh S (2019). “Simulating low precision floating-point arithmetic.” *SIAM Journal on Scientific Computing*, **41**(5), C585–C602.
- Higham NJ, Pranesh S (2021). “Exploiting lower precision arithmetic in solving symmetric positive definite linear systems and least squares problems.” *SIAM Journal on Scientific Computing*, **43**(1), A258–A277.
- Higham NJ, Pranesh S, Zounon M (2019). “Squeezing a matrix into half precision, with an application to solving linear systems.” *SIAM Journal on Scientific Computing*, **41**(4), A2536–A2551.
- Hopkins M, Mikaitis M, Lester DR, Furber S (2020). “Stochastic rounding and reduced-precision fixed-point arithmetic for solving neural ordinary differential equations.” *Philosophical Transactions of the Royal Society A*, **378**(2166), 20190052.
- Hrycej T, Bermeitinger B, Handschuh S (2022). “Training Neural Networks in Single vs Double Precision.” *arXiv preprint arXiv:2209.07219*.
- IEEE (2019). *IEEE Standard for Floating-Point Arithmetic, IEEE Std 754-2019 (Revision of IEEE 754-2008)*. The Institute of Electrical and Electronics Engineers, New York, NY. ISBN 978-1-5044-5924-2. doi:10.1109/IEEESTD.2019.8766229.

- Kahan W (1996). “IEEE standard 754 for binary floating-point arithmetic.” *Lecture Notes on the Status of IEEE*, **754**(94720-1776), 11.
- Klöwer M, Düben P, Palmer T (2020). “Number formats, error mitigation, and scope for 16-bit arithmetics in weather and climate modeling analyzed with a shallow water model.” *Journal of Advances in Modeling Earth Systems*, **12**(10), e2020MS002246.
- Lang ST, Dawson A, Diamantakis M, Dueben P, Hatfield S, Leutbecher M, Palmer T, Prates F, Roberts CD, Sandu I, *et al.* (2021). “More accuracy with less precision.” *Quarterly Journal of the Royal Meteorological Society*, **147**(741), 4358–4370.
- Lehmann M, Krause MJ, Amati G, Sega M, Harting J, Gekle S (2022). “Accuracy and performance of the lattice Boltzmann method with 64-bit, 32-bit, and customized 16-bit number formats.” *Physical Review E*, **106**(1), 015308.
- Li X, Wang S, Cai Y (2019). “Tutorial: Complexity analysis of singular value decomposition and its variants.” *arXiv preprint arXiv:1906.12085*.
- Li Z, Fotheringham AS (2020). “Computational improvements to multi-scale geographically weighted regression.” *International Journal of Geographical Information Science*, **34**(7), 1378–1397.
- Lucas A, Scholz I, Rainer Boehme SJ, Maechler M (2023). **gmp**: *Multiple Precision Arithmetic*. R package version 0.7-2, URL <https://cran.r-project.org/web/packages/gmp>.
- Luszczek P, Kurzak J, Dongarra J (2014). “Looking back at dense linear algebra software.” *Journal of Parallel and Distributed Computing*, **74**(7), 2548–2560.
- Luszczek P, Yamazaki I, Dongarra J (2019). “Increasing accuracy of iterative refinement in limited floating-point arithmetic on half-precision accelerators.” In *2019 IEEE high performance extreme computing conference (HPEC)*, pp. 1–6. IEEE.
- Maddox WJ, Potapczynski A, Wilson AG (2022). “Low-precision arithmetic for fast Gaussian processes.” In *Uncertainty in Artificial Intelligence*, pp. 1306–1316. PMLR.
- Maechler M, Heiberger RM, Nash JC, Borchers HW (2023). **Rmpfr**: *MPFR - Multiple Precision Floating-Point Reliable*. R package version 0.9-3, URL <https://cran.r-project.org/web/packages/Rmpfr>.
- Masui K, Ogino M (2019). “Research on the convergence of iterative method using mixed precision calculation solving complex symmetric linear equation.” *IEEE Transactions on Magnetics*, **56**(1), 1–4.
- Matérn B (2013). *Spatial variation*, volume 36. Springer Science & Business Media.
- Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953). “Equation of state calculations by fast computing machines.” *The Journal of Chemical Physics*, **21**(6), 1087–1092.
- Misra A, Laurel J, Misailovic S (2023). “ViX: Analysis-driven Compiler for Efficient Low-Precision Variational Inference.” In *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 1–6. IEEE.

- Murillo R, Del Barrio AA, Botella G (2022). “The Effects of Numerical Precision In Scientific Applications.” In *2022 Annual Modeling and Simulation Conference (ANNSIM)*, pp. 152–163. IEEE.
- Netti A, Peng Y, Omland P, Paulitsch M, Parra J, Espinosa G, Agarwal U, Chan A, Pattabiraman K (2023). “Mixed precision support in HPC applications: What about reliability?” *Journal of Parallel and Distributed Computing*, **181**, 104746.
- Nguyen H, Cressie N, Braverman A (2012). “Spatial statistical data fusion for remote sensing applications.” *Journal of the American Statistical Association*, **107**(499), 1004–1018.
- Ooi R, Iwashita T, Fukaya T, Ida A, Yokota R (2020). “Effect of mixed precision computing on H-matrix vector multiplication in BEM analysis.” In *Proceedings of the International Conference on High Performance Computing in Asia-Pacific Region*, pp. 92–101.
- Opitz T (2017). “Latent Gaussian modeling and INLA: A review with focus on space-time applications.” *Journal de la Société Française de Statistique*, **158**(3), 62–85.
- Pederson R, Kozłowski J, Song R, Beall J, Ganahl M, Hauru M, Lewis AG, Yao Y, Mallick SB, Blum V, *et al.* (2022). “Large scale quantum chemistry with tensor processing units.” *Journal of Chemical Theory and Computation*, **19**(1), 25–32.
- Powell MJ, *et al.* (2009). “The BOBYQA algorithm for bound constrained optimization without derivatives.” *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, **26**.
- R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Räss L, Kolyukhin D, Minakov A (2019). “Efficient parallel random field generator for large 3-D geophysical problems.” *Computers & Geosciences*, **131**, 158–169.
- Rüdisühli S, Walser A, Fuhrer O (2013). “COSMO in single precision.” *Cosmo Newsletter*, **14**, 5–1.
- Rue H, Martino S, Chopin N (2009). “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **71**(2), 319–392.
- Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK (2017). “Bayesian computing with INLA: a review.” *Annual Review of Statistics and Its Application*, **4**, 395–421.
- Sabbagh Molahosseini A, Sousa L, Emrani Zarandi AA, Vandierendonck H (2012). “Low-precision floating-point formats: From general-purpose to application-specific.” *Approximate Computing*, pp. 77–98.
- Särkkä S, Merkatas C, Karvonen T (2021). “Gaussian Approximations of SDES in Metropolis-Adjusted Langevin Algorithms.” In *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE.
- Schmidt D, Chen WC, Selivanov D (2022). *float: 32-Bit Floats*. R package version 0.3-0, URL <https://cran.r-project.org/package=float>.

- Scott J, Tuma M (2023). “Algebraic preconditioning in low precision works.” *ACM Transactions on Mathematical Software*.
- Tintó Prims O, Acosta MC, Moore AM, Castrillo M, Serradell K, Cortés A, Doblas-Reyes FJ (2019). “How to use mixed precision in ocean models: Exploring a potential reduction of numerical precision in NEMO 4.0 and ROMS 3.6.” *Geoscientific Model Development*, **12**(7), 3135–3148.
- Tolliver E, Pillai V, Jha A, John E (2022). “A Comparative Analysis of Half Precision Floating Point Representations in MACs for Deep Learning.” In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pp. 1–6. IEEE.
- Váňa F, Düben P, Lang S, Palmer T, Leutbecher M, Salmond D, Carver G (2017). “Single precision in weather forecasting models: An evaluation with the IFS.” *Monthly Weather Review*, **145**(2), 495–502.
- White J, Adámek K, Roy J, Dimoudi S, Ransom SM, Armour W (2023). “Bits missing: Finding exotic pulsars using bfloat16 on NVIDIA GPUs.” *The Astrophysical Journal Supplement Series*, **265**(1), 13.
- Xifara T, Sherlock C, Livingstone S, Byrne S, Girolami M (2014). “Langevin diffusions and the Metropolis-adjusted Langevin algorithm.” *Statistics & Probability Letters*, **91**, 14–19.
- Yamazaki I, Tomov S, Dongarra J (2015a). “Mixed-precision Cholesky QR factorization and its case studies on multicore CPU with multiple GPUs.” *SIAM Journal on Scientific Computing*, **37**(3), C307–C330.
- Yamazaki I, Tomov S, Kurzak J, Dongarra J, Barlow J (2015b). “Mixed-precision block Gram Schmidt orthogonalization.” In *Proceedings of the 6th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems*, pp. 1–8.
- Yang LM, Fox A, Sanders G (2021). “Rounding error analysis of mixed precision block Householder QR algorithms.” *SIAM Journal on Scientific Computing*, **43**(3), A1723–A1753.
- Ying L (2022). “Stable factorization for phase factors of quantum signal processing.” *Quantum*, **6**, 842.
- Zuras D, Cowlshaw M, Aiken A, Applegate M, Bailey D, Bass S, Bhandarkar D, Bhat M, Bindel D, Boldo S, *et al.* (2008). “IEEE standard for floating-point arithmetic.” *IEEE Std*, **754**(2008), 1–70.



**Affiliation:**

Mary Lai O. Salvana  
University of Connecticut  
Department of Statistics  
Storrs, CT 06269, USA.  
E-mail: [marylai.salvana@uconn.edu](mailto:marylai.salvana@uconn.edu)  
URL: <https://statistics.uconn.edu/person/mary-lai-salvana>

Sameh Abdulah  
King Abdullah University of Science and Technology  
Extreme Computing Research Center (ECRC)  
Thuwal, 23955-6900, Saudi Arabia.  
E-mail: [sameh.abdulah@kaust.edu.sa](mailto:sameh.abdulah@kaust.edu.sa)  
URL: <https://cemse.kaust.edu.sa/people/person/sameh-abdulah>

Minwoo Kim  
Department of Statistics  
Pusan National University  
Busan, South Korea.  
E-mail: [mwkim@pusan.ac.kr](mailto:mwkim@pusan.ac.kr)

David Helmy BrightSkies Inc.  
Alexandria, Egypt.  
E-mail: [david.helmy@brightskiesinc.com](mailto:david.helmy@brightskiesinc.com)

Ying Sun  
King Abdullah University of Science and Technology  
Extreme Computing Research Center (ECRC) and Statistics Program  
Thuwal, 23955-6900, Saudi Arabia.  
E-mail: [ying.sun@kaust.edu.sa](mailto:ying.sun@kaust.edu.sa)  
URL: <https://cemse.kaust.edu.sa/es>

Marc G. Genton  
King Abdullah University of Science and Technology  
Extreme Computing Research Center (ECRC) and Statistics Program  
Thuwal, 23955-6900, Saudi Arabia  
E-mail: [marc.genton@kaust.edu.sa](mailto:marc.genton@kaust.edu.sa)  
URL: <https://cemse.kaust.edu.sa/stsds>