UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

Avaliação de performance da metodologia *CM-generator* através de testes de estresse algorítmico

Maurício dos Anjos da Silva

Ouro Preto-MG
2022

Maurício dos Anjos da Silva

Avaliação de performance da metodologia *CM-generator* através de testes de estresse algorítmico

Monografia de Graduação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do grau de bacharel em Estatística.

Orientador: Anderson Ribeiro Duarte

Coorientador: Helgem de Souza Ribeiro Martins

Ouro Preto

2022

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

S586a Silva, Mauricio dos Anjos da.
Avaliação de performance da metodologia CM-generator através de testes de estresse algorítmico. [manuscrito] / Mauricio dos Anjos da Silva. - 2022.
81 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Anderson Ribeiro Duarte.
Coorientador: Prof. Dr. Helgem de Souza Ribeiro Martins.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Graduação em Estatística .

1. Matrizes. 2. Testes de estresse. 3. Simulação aleatória. I. Duarte, Anderson Ribeiro. II. Martins, Helgem de Souza Ribeiro. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 31

Bibliotecário(a) Responsável: Luciana De Oliveira - SIAPE: 1.937.800

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
COLEGIADO DO CURSO DE ESTATISTICA

FOLHA DE APROVAÇÃO

Maurício dos Anjos da Silva

Avaliação de performance da metodologia *CM-generator* através de testes de estresse algorítmico

Monografia apresentada ao Curso de Estatística da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Estatística

Aprovada em 1º de novembro de 2022

Membros da banca

Dr Anderson Ribeiro Duarte - Orientador (Universidade Federal de Ouro Preto)
Dr Helgem de Souza Martins - Co-orientador (Universidade Federal de Ouro Preto)
Dr Fernando de Souza Bastos (Universidade Federal de Viçosa)
Dr Eduardo Bearzoti (Universidade Federal de Ouro Preto)

Professor Dr. Anderson Ribeiro Duarte, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 01/11/2022

Documento assinado eletronicamente por **Anderson Ribeiro Duarte**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 03/11/2022, às 14:25, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0416908** e o código CRC **67F4311F**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3559-1312 - www.ufop.br

Agradecimentos

Primeiramente a Deus.

A meus pais, pelo incentivo diário mesmo em momentos difíceis, e por todo o carinho e orientação depositados em mim durante todo o processo de construção desse trabalho. Obrigado por entenderem que por vários momentos me faltava tempo, e que nem sempre podiam contar comigo. Por todas as palavras de motivação e por toda força que me deram sou grato, e por acreditarem em mim.

A meus irmãos por todo o carinho, companheirismo e compreensão de sempre, e por estarem comigo nos momentos de maior dificuldade. Obrigado por partilharem a maior dádiva de construir memórias únicas e retirar delas força e inspiração constante para prosseguir.

À minha namorada Larissa, por estar comigo em todos os momentos. Obrigado por me incentivar todos os dias, e por todo apoio e carinho, mesmo em situações difíceis.

A meus avós, por todo carinho que me deram. Obrigado por entenderem quando nem sempre estava por perto, a vocês dedico todo meu carinho e atenção.

Ao meu professor orientador, Dr. Anderson Ribeiro Duarte e ao Coorientador Dr. Helgem de Souza Ribeiro Martins pelo apoio durante todo o processo. Obrigado por todo carinho e compreensão. Obrigado por entenderem minhas dificuldades, e por toda força que me deram para superá-las. Obrigado por todo exemplo e ensinamentos durante toda essa caminhada.

Ao saudoso e honorável professor Spencer. Obrigado por todos os ensinamentos, essencialmente no início de toda essa jornada. Seu esforço e dedicação deixou marcas que jamais serão esquecidas. Por todas as contribuições, ensinamentos e todo o incrível trabalho desenvolvido junto ao departamento. Por essas e outras coisas, suas lembranças sempre serão presentes. Muito obrigado.

Gostaria de agradecer também a meus amigos João e Gabriel, pela amizade e companheirismo. Obrigado pelo carinho, consideração e por toda ajuda que me deram. Obrigado pela paciência, e pela parceria de longas datas.

Meus agradecimentos também aos professores membros da banca de avaliação, por aceitarem o convite e se fazerem membros de avaliação deste trabalho.

A todos vocês, muito obrigado.

Resumo

Procedimentos de simulação de matrizes de correlações aleatórias são essenciais em diversas investigações científicas. Algumas metodologias já foram apresentadas para cumprir este propósito. Em particular, um recente estudo apresentou um método denominado *Custom Matrix generator* (*CM-generator*) para gerar matrizes de correlação que atendem consistentemente às suposições matemáticas. A metodologia pode gerar matrizes de correlação customizadas, tanto em termos de intensidade de correlação quanto em relação à distribuição de proporções entre níveis de intensidade de correlação, para diversas aplicações. O método foi apresentado com algumas investigações acerca de qualidade dos resultados e do tempo de execução através de testes algorítmicos em condições extremas (testes de estresse). Este estudo produz uma investigação mais profunda e com experimentos mais abrangentes. Investigações com este formato aumentam a credibilidade e aplicabilidade de metodologias científicas recentes e inovadoras como o CM-generator.

Palavras-chave: Matrizes de correlação aleatórias; Testes de estresse; Simulação aleatória; Matrizes de correlação personalizadas.

Abstract

Simulation processes for random correlation matrices are paramount in several scientific investigations. Some methodologies have already been presented to fulfill this purpose. In particular, a recent study proposed a method called Custom Matrix generator (CM-generator) to generate correlation matrices that consistently meet the mathematical assumptions. The methodology can generate customized correlation matrices, both in terms of correlation intensity and concerning the distribution of proportions between levels of correlation intensity, for various applications. The method was presented with some investigations about the quality of the results and the execution time through algorithmic tests under extreme conditions (stress tests). This study produces more profound research with more comprehensive experiments. Inquiries with this format increase the credibility and applicability of recent and innovative scientific methodologies such as the CM-generator.

Keywords: Random Correlation Matrices; Stress tests; Random Simulation; Custom Correlation Matrices.

Lista de ilustrações

Lista de tabelas

Lista de *scripts*

Sumário

1 Introdução

Uma matriz de correlação é composta por medidas associativas entre pares de variáveis aleatórias. Nesse sentido, o princípio relacional entre variáveis aleatórias de um determinado experimento aleatório, é disposto de forma agrupada em uma matriz capaz de representar metricamente essa relação por meio das covariâncias e respectivos desvios padrão. Por definição, uma matriz de correlação n-dimensional é simétrica, com todos os seus elementos no intervalo $[-1, 1]$. Além disso, a matriz possui diagonal principal unitária e é positiva semi-definida, ou seja, sem auto-valores negativos.

Alguns procedimentos específicos de análise multivariada demandam matrizes de correlações específicas, que sejam capazes de descrever algum efeito específico de correlacionamento. A possibilidade de produzir, aleatoriamente, matrizes de correlação cujas proporções para determinadas intensidades de correlação sejam controladas, pode ser primordial para a validação de resultados.

A necessidade de geração de matrizes de correlação com características válidas, porém específicas, é recorrente e inerente a contextos práticos muito comuns como estudos em análise multivariada, aplicações no mercado financeiro ou mesmo em procedimentos de teste comuns na área da saúde. De fato, em muitas situações, a geração dos dados reais oriundos do próprio fenômeno de estudo pode ser custosa, arriscada ou mesmo impraticável para testes individuais. As dificuldades podem ser observadas na repetição constante do experimento e na necessidade de novos dados a cada aplicação.

No entanto, a implementação de novos métodos estatísticos está relacionada à verificação por meio de procedimentos não observacionais, ou experimentais. Novas técnicas e procedimentos estatísticos precisam ser submetidos a um conjunto de dados significativamente semelhante aos dados reais, de forma a minimizar vieses que não reflitam a realidade do estudo. Mais que isso, técnicas já implementadas são testadas e melhoradas a todo momento, isso amplia e torna a importância das matrizes de correlação com propriedades válidas ainda mais evidente.

Com isso, diversos métodos computacionais para a geração de matrizes de correlação com propriedades matemáticas válidas foram, e ainda são estudados ao longo das décadas. Em conjuntura, em muitos contextos práticos, as proporções das correlações por níveis de intensidade estão intrinsecamente ligadas à construção do problema.

1.1 Motivação

A análise multivariada busca essencialmente reduzir, agrupar, selecionar ou discriminar elementos de acordo com os objetivos da análise. Porém, mais especificamente em ambientes nos quais há uma grande dimensionalidade da base de dados, sua redução é essencial para uma melhor visibilidade e conclusão sobre a informação subjetiva. A análise fatorial é uma técnica muito comum, que objetiva a criação de fatores não observados a partir das variáveis e suas características. Como resultado, a análise é simplificada pela criação dos fatores que relacionam essas variáveis.

Nesse contexto, há diversas situações que podem se tornar propícias a essa abordagem, dada a quantidade de informação gerada a todo momento sem um tratamento inicial. A fundamentação dessa técnica baseia-se primordialmente nas características individuais e nas correlações observadas entre as variáveis. Dada a importância dos cenários descritos, muitos testes e análises são ponderados por técnicas de análise multivariada. A utilização da matriz de correlação se dá pela necessidade da modelagem da variabilidade da informação de estudo. Contudo, em diversos casos de interesse prático, a execução de testes sequenciados com dados reais é inexequível, ou mesmo significativamente custosa. Com isso, urge a necessidade de métodos computacionais para geração de matrizes de correlação. A ideia principal é centrada em produzir matrizes de correlação que atendam aos preceitos matemáticos, mas suficientemente próximas das necessidades do pesquisador.

Nas últimas décadas, diversas proposições de métodos computacionais para geração de matrizes de correlação foram discutidas, algumas com maior êxito e outras com deficiências bastante específicas.

1.2 Objetivos

Um recente estudo de Duarte, Martins e Oliveira (2022) [1] aborda a geração de matrizes de correlação com propriedades válidas e únicas. O método é capaz de gerar matrizes customizadas, tanto no sentido dos valores a qual assumem as correlações, quanto nas proporções das correlações dentro de uma determinada faixa especificada de intensidade no momento do experimento.

Desse modo, o estudo objetiva executar uma sequência de experimentos bastante abrangentes e demonstrar o comportamento do método mediante essas condições experimentais. Além disso, há a necessidade de testar a consistência desses resultados mediante um cenário de estresse no que tange a dimensionalidade excessiva das matrizes e também customizações com nível de especificidade muito elevado.

1.2.1 Objetivos Gerais

Apresentar diversos experimentos para validação e identificação de possíveis fragilidades no método CM-Generator [1]. Descrever o comportamento do algoritmo quando submetido a diferentes cenários de utilização de forma a evidenciar sua utilização sob condições extremas. Verificar funcionalidades gerais de customização implementadas no método tais como proporção das correlações e limites em cada faixa de correlações estipulada.

1.2.2 Objetivos Específicos

Apresentar comparações de desempenho do método sob condições atípicas de utilização. Estudar a distribuição dos elementos das matrizes em cada dimensão simulada. Estudar a porcentagem de erro sob condições de customização. Construir abordagens de teste sob ambos os limites no intervalo de definição das correlações $[-1, 1]$. Explorar a frequência de observações que acontecem nesses limites, bem como identificar o tempo necessário de execução e compará-lo em ambos os experimentos.

1.3 Contribuições

Com os resultados dos experimentos, será possível concluir e validar a utilidade do método CM-generator mediante cenários descritos em cada tópico ilustrado. Além disso, os experimentos computacionais foram conduzidos para um volume excessivo de simulações, essas repetições contribuem para uma maior consistência dos dados. Tais situações serão detalhadas na apresentação dos diversos resultados experimentais.

2 Fundamentação Teórica

A investigação do problema da geração de matrizes de correlação por meio de mecanismos de simulação computacional é um problema instigante e desafiador. As últimas décadas relatam diversas pesquisas que abordam o tema. Em geral, o desafio central é fornecer uma estratégia computacional para garantir a geração de matrizes aleatórias que atenda aos requisitos necessários para garantir que trata-se de uma matriz de correlação válida. Nesse contexto, será apresentada uma revisão de literatura que tenta delimitar os esforços científicos acerca desse tema.

Chalmers (1975) [2] apresenta um algoritmo com o intuito de gerar matrizes de correlação para um dado vetor de auto-valores fixo. Esse processo é mais viável para matrizes de covariância através da decomposição espectral $V = T\Lambda T'$, em que Λ é a matriz de auto-valores. Porém quando se trata de matrizes de correlação, essa relação não é imediata. Com isso a relação é explicada e comparada à equação geral de um cone, em que o objetivo é encontrar geratrizes para esse cone, no espaço n-dimensional. O estudo garante que é fácil ver que um cone sobre um espaço n-dimensional com $n > 2$ possui infinitas geratrizes.

A descrição de Bendel e Mickey (1978) [3] representa um método de geração de matrizes de correlação com auto-valores especificados. Uma matriz D, com os auto-valores na diagonal principal, entra na primeira iteração do método juntamente com uma matriz ortogonal qualquer gerada por algum método de simulação conveniente. A utilização inicial não garante diagonal unitária, com isso, o mecanismo por trás do método é multiplicar por uma dada rotação, uma matriz de senos e cossenos para que os elementos da diagonal principal sejam fixados na unidade. A matriz rotação, por definição, é ortonormal, ou seja, o produto de um vetor coluna por ele mesmo deve ser 1, e o produto por qualquer outro vetor coluna deve ser 0. Dessa forma, duas matrizes precisam ser geradas. Uma matriz M, que é tomada de forma aleatória, e a matriz P, que é formada pela matriz rotação, em que os valores de θ (ângulos de rotação) são escolhidos de modo a definir uma aplicação que conduza para uma matriz com diagonal unitária.

Davies e Higham (2000) [4] detectaram alguma instabilidade no método de Bendel e Mickey (1978) [3] e apresentaram uma outra aplicabilidade do método baseado na aplicação da metodologia de Drmac (1998) [5]. O método se inicia de uma matriz Q ortogonal gerada de um método qualquer. A matriz deve possuir a propriedade de invariância, em que a probabilidade de seleção de uma matriz Q é a mesma quando ela é deslocada em seu espaço. Com isso, a primeira matriz A, será formada por $A =$

$U\text{diag}(\lambda_i)U^T$. Desse modo, são aplicadas rotações para que cada elemento a_{ii}, se torne unitário. O problema de inconsistência detectado no algoritmo é justamente no primeiro passo, em que a matriz A é definida. Seja a relação $\overline{\gamma_k} = \frac{cku}{1-cku}$, em que c representa um número inteiro próximo a zero. Para essa relação, detectou-se que se $\frac{max(\lambda_i)}{min(\lambda_i)} > \frac{\overline{\gamma_{n^2}}^{-1}}{\sqrt{n}}$, então a matriz de correlação A^* pode ser indefinida. Diante disso, a matriz poderia possuir algum auto-valor menor que 0. A instabilidade desse fato se encontra então na implementação do algoritmo, visto que um erro $\overline{\gamma_k}$ pode ser produzido, e com isso não ser uma matriz de correlação válida por ferir uma das propriedades. A mudança na implementação do algoritmo foi feita a partir da geração de U. Assim, gera-se uma matriz X, $X = U\text{diag}(\sigma_i)V^T$, em que U e V são duas matrizes ortonormais e σ os escalares ordenados escolhidos. Os índices de A, são gerados baseados na norma de cada índice de X.

No mercado financeiro é muito comum a necessidade da geração de matrizes aleatórias em várias aplicações, principalmente com a propriedade de Perron-Frobenius, que afirma que matrizes com entradas positivas possuem todos os seus auto-valores positivos e um deles dominante, capaz de carregar a maior parte da variação dos dados consigo. As matrizes de correlação válidas então possuem as seguintes propriedades: possuir elementos no intervalo $[-1, 1]$, possuir diagonal unitária, ser uma matriz simétrica, ser uma matriz positiva semi-definida. Diante disso, Hüttner e Mai (2019) [6] propõem o algoritmo `randcorr` que inicialmente seleciona uma matriz para seu início e obtém sua decomposição espectral. O algoritmo parte da ideia de aplicar rotações até que a diagonal se transforme em unitária.

Böhm e Hornik (2014) [7] mostram uma abordagem para a geração de matriz de correlação a partir do método da aceitação e rejeição simples. O método inicia com a geração de R matriz identidade, e então gera os valores para a triangular superior a partir de uma distribuição uniforme no intervalo $(-1, 1)$. Para aceitação o método testa se a matriz é positiva definida, caso não seja o algoritmo retorna ao passo inicial. O grande impasse encontrado pelo algoritmo é que conforme aumenta-se o valor de n, o custo computacional aumenta exponencialmente. Muitas tentativas são necessárias antes de aceitar uma matriz. Com isso, como conclusão é possível dizer que o método de aceitação e rejeição simples não é adequado para gerar matrizes de correlação.

Holmes (1991) [8] descreve uma abordagem para geração de matrizes de correlação maiores que 3×3. A proposição parte do fato que a matriz A gerada deve atender $det(A) = 0$. O método permite gerar matrizes de correlação, condicionado ao determinante. Essa abordagem torna-se um pouco inviável à medida que o número de dimensões cresce, pois a expressão que define o intervalo das correlações torna-se exponencialmente maior.

O método proposto por Budden, Hadavas, Hoffman e Pretz (2008) [9] demonstra

um procedimento para geração de matrizes de correlação 4×4. Inicialmente, são descritas as restrições usuais de matrizes de correlação para essa questão. A restrição de ser uma matriz positiva semi-definida é a que mais dificulta o processo, dado que o processo de geração de uma normal multivariada não garante que a matriz será positiva semi-definida, e isso é o suficiente para torná-la inválida em termos práticos. Desse modo, é introduzido o método para geração de matrizes 3×3, com os valores r_{12} e r_{13} escolhidos arbitrariamente, e então o valor r_{23} pode ser obtido pela relação $r_{12}r_{13} - \sqrt{(1-r_{12}^2)(1-r_{13}^2)} \leq r_{23} \leq r_{12}r_{13} + \sqrt{(1-r_{12}^2)(1-r_{13}^2)}$. Isso garante que ela será positiva semi-definida. Essa abordagem apresenta uma extensão, para matrizes 4×4. Como conclusão, várias aplicações são descritas. E fácil ver também que a extensão do método para matrizes de dimensão n é dificultada pelo fato de as expressões crescerem muito quando cresce o valor de n.

Marsaglia e Olkin (1984) [10] apresentam uma primeira abordagem que busca gerar matrizes de correlação aleatórias com uma média especificada. Nesse sentido, o algoritmo proposto considera uma matriz A simétrica, com zeros em sua diagonal, em que $max_i \sum_j \sqrt{a_{ij}^2} < \lambda$. Desse modo, os elementos da diagonal superior da matriz X são gerados de tal forma que o módulo desses elementos sejam menores que os elementos de A, igualando-se os índices. Dada uma matriz C, positiva definida, $R = C + X$. Para os valores de $i < j$, então $R = C + \sqrt{2}\lambda X$. Isso leva então para uma matriz R com auto-valores no intervalo $(1-\lambda, 1+\lambda)$, com λ o menor autovalor de C, então $C + R - I$ possui valor esperado C, e é uma matriz de correlação válida. Um segundo método introduzido, foi garantir que a matriz gerada fosse positiva definida através de sua decomposição TT'. O último exemplo mostra um algoritmo para geração de matrizes de correlação com a obtenção da matriz ortogonal P, semelhante à clássica ortogonalização de Gram-Schmidt. Assim, é necessário entrar com os auto-valores especificados. Apesar da semelhança com o método de Gram-Schmidt, parte dele torna o processo aleatório, o que é um dos objetivos iniciais.

Hardin, Garcia e Golan (2013) [11] descrevem a importância da geração de matrizes de correlação para várias áreas do conhecimento, tais como análise fatorial, modelos de estimação Bayesianos, modelos de classificação, entre outros. Nesse contexto, a geração de matrizes de correlação com propriedades reais é dificultada pelas restrições usuais já abordadas. Um primeiro exemplo é enunciado com base em aplicações genéticas. Esse procedimento estaria inserido em um processo na qual as características genéticas deveriam possuir correlações muito próximas a zero. Outro exemplo enunciado é a estrutura de correlação Toeplitz, utilizada principalmente em modelos de classificação e análise discriminante, mas também usado em alguns modelos de séries temporais. Essa estrutura descreve características fortemente correlacionadas, e em algumas abordagens é conhecido como estrutura auto-regressiva, por levar em conside-

ração a dimensão de índice $k-1$. O último exemplo considera fenômenos naturais, com estudos de caso controle acerca dessas características. Ambos esses casos e estruturas de correlação usam alguma distribuição para a geração de uma matriz de correlação com esse comportamento. Porém, a que se mostra mais eficiente e aplicável utiliza a distribuição normal para gerar os vetores de dados. Com isso, o método é comparado com dados gerados pela distribuição normal, com a inserção de ruídos e o estudo do comportamento desses ruídos. A partir de certo número de simulações o método se torna tão eficiente quanto a distribuição Gaussiana. Os três exemplos foram citados, e o algoritmo foi apresentado em ambos os casos, com a inserção de ruídos sobre os dados seguido de clusterização, a partir de um erro inicial estipulado.

Gosh e Henderson [12] buscam gerar vetores aleatórios de uma determinada distribuição, a partir da distribuição normal. Com isso a distribuição uniforme é usada como intermediária, ou cópula. Os vetores gerados são chamados NORTA (Normal to anything), por justamente fazer o uso da distribuição normal na transformação. Dessa forma, os dados inicialmente são representados por uma distribuição normal multivariada, em que cada vetor linha é uma distribuição marginal também normal univariada. Uma propriedade importante é que, sob normalidade desses vetores, se a correlação $\rho = 0$ então isso implica em independência. A recíproca dessa propriedade não é verdadeira, e isso não é válido para outras distribuições. Com isso, o primeiro passo é gerar uma matriz triangular inferior, tal que ela seja não singular quando fatorada em $M = \Sigma\Sigma'$. Assim, W são os vetores gerados da distribuição normal univariada, e Z é formada por todos os vetores W. X então será cada valor da distribuição normal gerada em Z aplicada na inversa da distribuição desejada. Esse passo é repetido até que sejam transformados todos os elementos de Z. A metodologia de geração ficou nominada método Onion.

Joe (2006) [13] em seu trabalho para geração de matrizes de correlação válidas, utiliza de correlações parciais e múltiplas na apresentação de seus métodos C-vine e D-vine. São introduzidos teoremas sobre a obtenção e utilização dessas medidas de correlação em um espaço d-dimensional. Correlações parciais são medidas de duas variáveis aleatórias na qual, fixado em um espaço plano, têm-se a correlação entre elas. Em outras palavras, em um espaço d-dimensional, com d variáveis $X_1, X_2, X_3, ..., X_d$, a correlação entre X_1 e X_2, com relação às demais $X_3, ..., X_n$, pode ser obtida pela seguinte relação $\rho_{ij} = -\frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}}$ em que C é o cofator da matriz principal para i e j. Para a linha e a coluna fixadas, o cofator é calculado pelo determinante da sub-matriz obtida através da eliminação da linha e da coluna escolhidas, e, em seguida, multiplicando-se esse valor por $(-1)^{(i+j)}$.

Lewandowski, Kurowica e Joe (2009) [14] discutem os métodos de geração Onion [12], C-vine e D-vine [13]. É apresentado um extenso estudo de simulação para

verificar a validade dos métodos propostos. A investigação revela alguma superioridade do método onion, mais rápido quando comparado aos métodos vine.

Diversos outros estudos não incluem proposição de metodologias, mas apresentam aplicações elucidativas para o assunto de geração de matrizes de correlação.

Bouchaud e Potters (2009) [15] buscam descrever o comportamento de matrizes de correlação para aplicações no mercado financeiro. Na abordagem espectral de Marčenko-Pastur o objetivo é descrever como a N-ésima quantidade ou variável se comporta quando comparada a outra variável na T-ésima observação, em que N é o número de medidas de um fenômeno e T suas repetições, ou seja, observações dessas mesmas quantidades em outros momentos. Essa medida pode ser em dias, meses ou outra escala de abordagem definida pelo pesquisador. Com isso, a abordagem de Marčenko-Pastur busca a relação entre a matriz de correlações empírica e a verdadeira, através da relação de normalização $X_{ij} = r_i^t / \sqrt{T}$. São descritas aplicações para os retornos financeiros principalmente no mercado de ações. A análise de componentes principais é associada como uma abordagem essencial para aplicação, dado que esses passos são fortes bases do método. Desse modo, os componentes no mercado de ações fornecem uma sequência de passos correspondentes que fornecem o melhor lucro, com auto-valor definido e variância decrescente ao longo dessa sequência.

Simonian (2010) [16] descreve uma aplicação na área de risco e aplicações financeiras. De fato, em termos práticos, a matriz de correlação gerada precisa atender os critérios mínimos de teste e fazer com que a simulação das variáveis associadas se aproximem o máximo da realidade. Há diversos métodos que possuem o intuito de gerar matrizes de correlação, porém muitos deles não satisfazem a condição de gerar uma matriz positiva semi-definida. Com isso um dos objetivos é tornar uma matriz estimada em sua forma válida. Desse modo, o método apresentado baseia-se na decomposição espectral. Dada uma matriz, encontra-se seus auto-valores e auto-vetores. Assim, na sua matriz de auto-valores, identifica-se os auto-valores menores que zero, e então esses valores são substituídos por 0. Dessa forma, uma nova matriz será produzida ao se multiplicar os auto-valores produzidos com essa transformação pelos respectivos auto-vetores. Essa nova matriz terá seus elementos fixados em 1.

Mittelbach, Matthiesen e Jorswieck (2012) [17] apresentam uma maneira de se amostrar uniformemente matrizes do conjunto das matrizes positivas definidas, e com traço constante. Ao amostrar matrizes através desse método, o traço deve ser igual para todas elas, $tr(A) = a_{11} + a_{22} ... + a_{nn} = \sum_{i=1}^{n} a_{ii}$ com n o número de linhas da matriz A. A ideia do método é utilizar a decomposição de Cholesky e uma reparametrização dos elementos em coordenadas hiperesféricas. A partir dessa parametrização, a matriz U da decomposição de Cholesky $A = U'U$ é obtida. A distribuição da soma dos elementos da diagonal $tr(A)$ é obtida através do método Jacobiano, para voltar à distribuição original.

Esse método também pode ser extendido para matrizes complexas, cuja decomposição é $A = U^*U$. O estudo da distribuição dos vetores gerados é fortemente mencionado, pois uma das restrições é justamente o traço da matriz.

Pourahmadi e Wang (2015) [18] estudam a distribuição de matrizes de correlação aleatórias por meio da parametrização hiperesférica de seus fatores de Cholesky e as distribuições dos ângulos relacionados. A importância desse procedimento no processo de geração de matrizes de correlação de alta dimensão são abordados.

Prôa, O'Higgins e Monteiro (2013) [19] apresentam um estudo de simulação com uma abordagem para o estudo de características genéticas com base nas variáveis já observadas nos parentes anteriores. Essas características dos indivíduos atuais, pode ser modelada através de uma relação com os antepassados. A relação é descrita pela matriz de covariâncias genéticas e fenotípicas relacionadas a essas características. A relação entre essas matrizes é importante, pois em vários casos se torna mais viável utilizar a matriz de covariâncias fenotípicas, ao invés da matriz de covariâncias genéticas.

Hong (1999) [20] discute o erro de modelo causado pela simulação das matrizes de correlação. Em casos reais, o erro de simulação se torna importante se estudar esse comportamento. Dois métodos são abordados, o modelo de Tucker-Koopman-Linn e o algoritmo de Wijsman no intuito de implementar uma aplicação em análise fatorial. Nessa abordagem, a matriz de correlações populacional é particionada em matriz de correlações entre os fatores, e matriz de covariâncias.

Rebonato e Jäckel (2011) [21] descrevem a geração de matrizes de correlação para aplicações financeiras, na área de risco. A importância de realizar simulações reais é que a todo momento há uma variação de risco e mudança nos níveis de interação entre as variáveis. Com isso, e necessário um método para geração de matrizes de correlação válidas, satisfazendo os requisitos mínimos. É apresentado um método que utiliza coordenadas hiperesféricas, juntamente com a decomposição $A = TT'$. E um segundo método proposto, considera a decomposição espectral.

Diversos estudos são apresentados nesta revisão. Por fim, será apresentado o estudo de Duarte, Martins e Oliveira (2022) [1] que é a abordagem para geração aleatória de matrizes de correlação que será objeto desse estudo. O método carrega especificidades únicas que serão apresentadas com mais detalhes nas apresentações metodologicas desse estudo.

3 Aspectos Metodológicos

Diante do propósito especificado, é preponderante apresentar a metodologia CM-generator [1] com maior profundidade de detalhes. Como dito, o método CM-generator é capaz de gerar matrizes de correlação que atendam aos interesses do pesquisador, isso sem desconsiderar as premissas matemáticas que as tornam válidas, excepcionalmente com relação a seus auto-valores. Todavia, essa foi uma das principais fragilidades em vários estudos no decorrer das décadas, fato que levou à reformulações e até mesmo ao abandono de algumas técnicas. A necessidade de aplicações nesse sentido surge pois, independente do tipo de experimento, alguns deles não são explicáveis apenas por uma relação direta ação-reação em que as variáveis são independentes. Por esse motivo, deve ser considerada a relação entre as variáveis aleatórias inerentes ao experimento para explicar determinado fenômeno de interesse prático. Não obstante, vários modelos de teste são caracterizados pela utilização da relação entre as variáveis explicativas do modelo, e, com isso, descrevem a importância de metodologias como o CM-generator.

Suponha um cenário na qual é desejável aplicar testes de avaliação sobre uma nova classe de ativos no mercado financeiro. Porém, de acordo com estudos e testes de correlação anteriores, existem limites especificados para a verdadeira correlação populacional entre as variáveis em questão, supostamente com variáveis como renda média familiar e perfil de investidor. Suponha ainda que estudos anteriores demonstraram que em 95% das observações, a correlação verificada se concentra acima de 0,6. A ideia é então avaliar o comportamento de índices de interesse no mercado, mediante uma nova classe de ativos. Um possível modelo para representação desse cenário, com determinadas faixas específicas de correlação entre as variáveis preditoras, pode ser avaliado mediante essas condições de níveis de correlação. Dado um conhecimento prévio sobre o comportamento das variáveis, a não inserção dessa relação no modelo poderia implicar em vieses não amostrais, em que o enfoque estaria em não compreender o cenário populacional de estudo.

A utilização do método leva em consideração a dimensão, limites de customização e proporção em cada faixa desejada. O controle da intensidade das correlações expande significativamente o ambiente de aplicabilidade acima supracitado. A versatilidade quanto à aplicação de todas essas possibilidades de utilização o torna inovador, mediante os aspectos de customização ainda não explorados.

Uma premissa importante do método é sua simplicidade construtiva. Uma matriz de correlação é gerada a partir de uma amostra aleatória multivariada simulada da distribuição Normal. A ideia central é que exista uma parametrização adequada

para o procedimento de simulação dessa amostra, para que a matriz gerada atenda toda a especificação prévia do pesquisador com base nos parâmetros de customização desejados.

Inicialmente, defina as variáveis aleatórias X, Y e Z tais que:

$$Z \sim N(0,1)$$
$$X \sim N(0,a^2)$$
$$Y \sim N(0,b^2)$$

em que $a, b > 0$. Além disso, considere as variáveis aleatórias U e V tais que:

$$U = X \pm Z \Longrightarrow U \sim N(0, 1+a^2)$$
$$V = Y \pm Z \Longrightarrow V \sim N(0, 1+b^2)$$

De posse dessas variáveis, é possivel determinar as correlações $\rho_{Z,U}$, $\rho_{Z,V}$ e $\rho_{U,V}$, como segue:

$$Cov(Z,U) = Cov(Z, X \pm Z) = Cov(Z,X) \pm Cov(Z,Z)$$

por independência e normalidade, $Cov(Z,X) = 0$. Logo:

$$\begin{aligned} Cov(Z,U) &= Cov(Z, X \pm Z) \\ &= 0 \pm Cov(Z,Z) \\ &= 0 \pm 1 \\ &= \pm 1 \end{aligned}$$

Portanto, $\rho_{Z,U}$ pode ser expressa como:

$$\rho_{Z,U} = \frac{Cov(Z,U)}{\sqrt{Var(Z)}\sqrt{Var(U)}} \Rightarrow \rho_{Z,U} = \pm\frac{1}{\sqrt{1+a^2}} \tag{3.1}$$

De maneira análoga, $\rho_{Z,V}$ pode ser obtida como:

$$\rho_{Z,V} = \frac{Cov(Z,V)}{\sqrt{Var(Z)}\sqrt{Var(V)}} \Rightarrow \rho_{Z,V} = \pm\frac{1}{\sqrt{1+b^2}}. \tag{3.2}$$

A correlação $\rho_{U,V}$ pode então ser obtida da seguinte forma:

$$Cov(U,V) = Cov(X \pm Z, Y \pm Z) = Cov(X,Y) \pm Cov(X,Z) \pm Cov(Z,Y) \pm Cov(Z,Z)$$

A independência entre as variáveis implica em covariância nula, dessa forma:

$$\begin{aligned} Cov(U,V) &= 0 \pm 0 \pm 0 \pm Cov(Z,Z) \\ &= \pm Var(Z) \\ &= \pm 1. \end{aligned}$$

Logo:

$$\begin{aligned} \rho_{U,V} &= \frac{Cov(U,V)}{\sqrt{Var(U)}\sqrt{Var(V)}} \\ &= \pm \frac{1}{\sqrt{1+a^2}\sqrt{1+b^2}}. \end{aligned} \tag{3.3}$$

Suponha agora, a possibilidade de obter números reais positivos a e b para definir limites ℓ_{lower} e ℓ_{upper} que sejam cotas inferiores e superiores para $\rho_{Z,U}, \rho_{Z,V}$ e $\rho_{U,V}$ simultâneamente. Tais cotas satisfariam o seguinte conjunto de inequações:

$$\ell_{lower} < \frac{1}{\sqrt{1+a^2}} < \ell_{upper} \tag{3.4}$$

$$\ell_{lower} < \frac{1}{\sqrt{1+b^2}} < \ell_{upper} \tag{3.5}$$

$$\ell_{lower} < \frac{1}{\sqrt{1+a^2}\sqrt{1+b^2}} < \ell_{upper} \tag{3.6}$$

Duarte, Martins e Oliveira (2022) [1] enunciam então o resultado que segue.

Teorema 3.1. *É possível obter os valores de a e b para satisfazer ambas as inequações* (3.4), (3.5) *e* (3.6) *simultaneamente.*

Prova do Teorema 3.1. *Suponha* $\ell^*_{lower} = \sqrt{\ell_{lower}}$, *e também* $\ell^*_{upper} = \sqrt{\ell_{upper}}$. *É possível verificar que* $\ell^*_{lower} > \ell_{lower}$, *e ainda, verificar que* $\ell^*_{upper} > \ell_{upper}$, *isso decorre da relação* $0 < \ell_{lower} < \ell_{upper} < 1$, *logo:*

$$\ell_{lower} < \ell^*_{lower} < \frac{1}{\sqrt{1+a^2}} < \ell_{upper} < \ell^*_{upper} \tag{3.7}$$

$$\ell_{lower} < \ell^*_{lower} < \frac{1}{\sqrt{1+b^2}} < \ell_{upper} < \ell^*_{upper} \tag{3.8}$$

Dessa maneira, a única restrição que precisa ser satisfeita é $\ell^*_{lower} < \ell_{upper}$, *caso contrário não estariam válidas as inequações* (3.4), (3.5) *e* (3.6). *Para obter a relação para a correlação* $\rho_{U,V}$, *basta multiplicar as inequações* (3.7) *e* (3.8):

$$\ell_{lower} = {\ell^*_{lower}}^2 < \frac{1}{\sqrt{1+a^2}}\frac{1}{\sqrt{1+b^2}} < {\ell^*_{upper}}^2 = \ell_{upper} \tag{3.9}$$

Portanto, para obter os limites para a, basta alguma manipulação algébrica na inequação (3.7)*:*

$$\ell^*_{lower} < \frac{1}{\sqrt{1+a^2}} < \ell^*_{upper} \Longrightarrow \frac{1}{\ell^*_{upper}} < \sqrt{1+a^2} < \frac{1}{\ell^*_{lower}}$$

$$\frac{1}{\ell_{upper}} < 1+a^2 < \frac{1}{\ell_{lower}} \Longrightarrow \frac{1}{\ell_{upper}} - 1 < a^2 < \frac{1}{\ell_{lower}} - 1$$

$$\sqrt{\frac{1}{\ell_{upper}} - 1} < a < \sqrt{\frac{1}{\ell_{lower}} - 1} \tag{3.10}$$

$$I = \left(\sqrt{\frac{1}{\ell_{upper}} - 1}; \sqrt{\frac{1}{\ell_{lower}} - 1}\right) \tag{3.11}$$

De maneira análoga, é possível encontrar o intervalo para b, tal que:

$$\ell^*_{lower} < \frac{1}{\sqrt{1+b^2}} < \ell^*_{upper}$$

Que é o mesmo intervalo apresentado na Equação (3.11)*.*

3.1 Os limites ℓ_{lower} e ℓ_{upper}

É importante observar que a escolha dos valores a e b, restritos ao intervalo definido na equação (3.11), está intimamente ligada com a escolha dos limites ℓ_{lower} e ℓ_{upper} para os valores das medidas de correlação geradas. Na prática, os limites ℓ_{lower} e ℓ_{upper}, em condições de contorno, devem atender ao intervalo $(0,1)$. Porém, para os valores de borda 0 e 1 propriamente ditos, há algumas restrições matemáticas. Esses limites devem estar bem definidos mesmo em uma sequência de experimentos completamente aleatória, uma vez que as correlações dependem dos valores de a e b, como exposto nas equações (3.1), (3.2) e (3.3). Nesse sentido, considerando o limite superior do intervalo para escolha dos valores a e b, têm-se que:

$$\lim_{\ell_{lower}\to 0} \sqrt{\frac{1}{\ell_{lower}} - 1} \to \infty$$

e, com isso, ℓ_{lower} não deve assumir o valor 0 de fato, mas valores suficientemente próximos a 0 são viáveis. Diante disso, não é possível definir $\ell_{lower} = 0$, por restrições matemáticas. Para o limite superior do intervalo, sabe-se que:

$$\lim_{\ell_{upper} \to 1} \sqrt{\frac{1}{\ell_{upper}} - 1} = 0$$

Nesse caso, não há diretamente uma restrição matemática para o limite superior. Apesar disso, o valor resultante do intervalo $a = 0$ indica que a variável gerada em determinado experimento não teria variabilidade, seria degenerada em um único valor, pois a representa seu desvio padrão. De mesmo modo, é possível então, tomar valores suficientemente próximos a 1, mas não iguais a 1.

Visto de outra forma, quando $\ell_{lower} \to 0$ e $\ell_{upper} \to 1$, o intervalo definido na equação (3.11) converge para o intervalo $\mathbb{R}^+$, ou seja, qualquer valor de desvio padrão seria viável. Note que, do ponto de vista lógico, a escolha do valor do desvio padrão pode ser suficientemente grande, porém um número real, e suficientemente pequena porém diferente de 0. De posse disso, os limites próximos ao intervalo $(0, 1)$ são sensíveis a pequenas mudanças em sua estrutura. Uma vez que têm-se valores para tais limites, uma pequena mudança pode representar um aumento significativo nos valores dos intervalos para a e b. Com isso, um dos objetivos do método é controlar também essa sensibilidade.

3.2 Mecanismo de Geração das Matrizes de Correlação

A descrição anterior permite a construção de uma abordagem para geração aleatória de matrizes de correlação. A metodologia descrita parte da geração de variáveis aleatórias para que, posteriormente, seja possível compor um vetor aleatório que será utilizado para obter as correlações desejadas. Os passos de execução são descritos de modo a considerar que a geração dessas variáveis independa do mecanismo de geração de variáveis aleatórias. Inicialmente, a metodologia será apresentada para a geração de matrizes de correlação sem quaisquer restrições prévias. Em segundo momento, será conduzida uma descrição a partir da definição dos limites para os valores das correlações, ou seja, já com algum nível de customização. Por fim, o método completamente customizado, que considera além dos limites de customização, diferentes níveis limitados (intervalos) de correlações e ainda, as proporções de correlações para cada um desses níveis.

3.2.1 Método Geral para Geração de Matriz de Correlações

Considere uma amostra aleatória com n observações da distribuição normal padrão $Z \sim N(0,1)$, para compor as coordenadas do vetor $\vec{Z}$. Seja ainda X, uma matriz com k vetores representados pelas variáveis aleatórias $X_1, X_2, ..., X_k$, com $X_i \sim N(0, a_i^2)$. Por fim, determine a matriz U, da seguinte maneira:

$$U = X + Z = [U_1, U_2, ...U_i...U_k]$$

cada elemento U_i da matriz U possui por definição n elementos, em que $U_i \sim N(0, 1 + a_i^2)$, como descrito na equação 3.1. Desse modo, a correlação ρ_{U_i,U_j} é tal que:

$$\rho_{U_i,U_j} = \pm \frac{1}{\sqrt{1 + a_i^2}\sqrt{1 + a_j^2}}$$

como descrito na Equação 3.3. As variáveis contidas nas matrizes Z e X são obtidas por mecanismos de simulação computacional. Destarte, todas as pressuposições para garantir a validade matemática da matriz de correlação gerada são satisfeitas por construção, afinal trata-se de uma matriz de correlação amostral.

Nesse sentido, o algoritmo requer o tamanho da dimensão da matriz desejada, e as medidas de variabilidade necessárias. A matriz U então é gerada a partir dos procedimentos descritos. E sua respectiva matriz de correlação (note que U é composta por k amostras aleatórias de tamanho n) representa a matriz de interesse.

3.2.2 Método para Geração de Matriz de Correlações em Intervalo pré-definido

Uma extensão ao método é representada pela definição dos limites de correlação conforme a escolha do usuário. O ganho e a flexibilidade de aplicação inserida a esse contexto a partir da definição dos limites de borda aumentam significativamente. Com isso, a ideia é condicionar os valores a e b de acordo com as definições das Inequações (3.4), (3.5), (3.6). Dessa maneira, é possível definir os limites para escolha dos desvios padrão condicionados aos limites especificados pelo pesquisador de modo a garantir a validade das propriedades relatadas anteriormente.

Outra propriedade inserida nesse contexto, é a capacidade de definir se o limite especificado trata-se de um limite inferior, superior ou ambos, em que as correlações concentram-se em um intervalo pré-definido.

Caso tenha sido escolhido apenas um limite inferior ℓ_{lower}, os valores dos desvios padrão serão obtidos no intervalo de borda inferior maior que zero e borda superior menor que $\sqrt{\frac{1}{\ell_{lower}} - 1}$. De mesma forma, caso tenha sido definido apenas um limite

superior ℓ_{upper}, os valores dos desvios padrão serão obtidos no intervalo de borda inferior maior que $\sqrt{\frac{1}{\ell_{upper}} - 1}$ e borda superior ilimitada. Em um cenário cujas correlações precisam estar em um intervalo, ambos os limites ℓ_{lower} e ℓ_{upper} são especificados para fornecer o intervalo de escolha dos desvios padrão descrito na Equação (3.11).

3.2.3 Método para Geração de Matriz de Correlações Customizadas

De posse das descrições supracitadas, o nível máximo de customização pode ser alcançado. A situação de diferentes níveis de intensidade de correlação, em que cada um desses níveis é definido por um intervalo limitado. Além disso, com as proporções de correlações pertencentes a cada um desses níveis pré-fixada pelo usuário do método. Esse, seria o nível máximo de customização inerente ao método.

Suponha o interesse na obtenção de uma matriz de correlações tal que 30% das correlações estejam entre $0 < |\rho| < 0{,}3$, 50% estejam entre $0{,}3 < |\rho| < 0{,}7$ e 20% sejam superiores a 0,7. Isso implica que 40% das correlações devem ser geradas de modo que os valores de desvios padrão adotados satisfaçam os valores de $\ell_{lower} = 0$ e $\ell_{upper} = 0{,}3$. De forma análoga, 50% precisam ser gerados de modo a considerar que os desvios padrão utilizados sejam obtidos para $\ell_{lower} = 0{,}3$ e $\ell_{upper} = 0{,}7$. Por fim, 20% devem ser gerados tal que os desvios padrão considerem os limites $\ell_{lower} = 0{,}7$ e $\ell_{upper} = 1{,}0$. Apesar disso, apenas estas escolhas não garantem que a matriz obtida segue as referências de customização especificadas previamente. Isso sob a condição de mais de um intervalo para os limites de customização. Mais que isso, certamente o resultado obtido não será correspondente às expectativas do usuário.

Esta situação decorre do fato de que para dois ou mais intervalos distintos, invariavelmente existirá na matriz, pelo menos uma correlação que não atende nenhum dos dois intervalos certamente. Essa medida de correlação é decorrente de duas variáveis aleatórias geradas com desvios padrão obtidos em intervalos distintos, o que não produz uma correlação que atenda aos limites descritos nas inequações do Teorema 3.1 para os dois intervalos simultaneamente. O objetivo então é obter uma matriz inicial e, a partir dela, modificar (através de alguma estratégia otimizadora) as proporções iniciais e gerar uma nova matriz. Esse processo deve então ser repetido até se obter uma matriz suficientemente próxima da estrutura desejada. Com isso, à medida em que é realizado esse procedimento repetidas vezes, a matriz gerada converge para a matriz desejada.

O método de otimização utilizado possui conceitos bastante semelhantes ao bastante difundido algoritmo *Simulated Annealing*. Através de um vetor de probabilidades em cada faixa, inserido pelo usuário, a cada iteração a matriz convergirá na direção da matriz desejada com uma precisão desejada, a partir de uma solução inicial. Considere então $\Gamma = (\gamma_1, \gamma_2, ..., \gamma_k)$ o vetor de proporções desejadas inicialmente. Gere então uma matriz P_1 a partir dessas proporções. Seja também $\Theta = (\theta_1, \theta_2, ..., \theta_k)$ o vetor com as

proporções obtidas na matriz gerada no passo atual. O intuito é obter a máxima e mínima diferença entre as coordenadas dos dois vetores, e as respectivas posições dessas diferenças. Uma perturbação no vetor Γ nas posições de máxima e miníma diferenças será implementada. Será então obtido um novo vetor de proporções iniciais para uma nova matriz de correlações ser gerada. Com isso, a alteração em Γ consiste em adicionar ou subtrair as diferenças encontradas. Defina também os valores N e K, em que N é o número máximo de passos de execução, e K o máximo valor aceitável para a máxima diferença encontrada entre coordenadas dos vetores.

O fluxograma apresentado na Figura 1 descreve as etapas do procedimento de otimização.

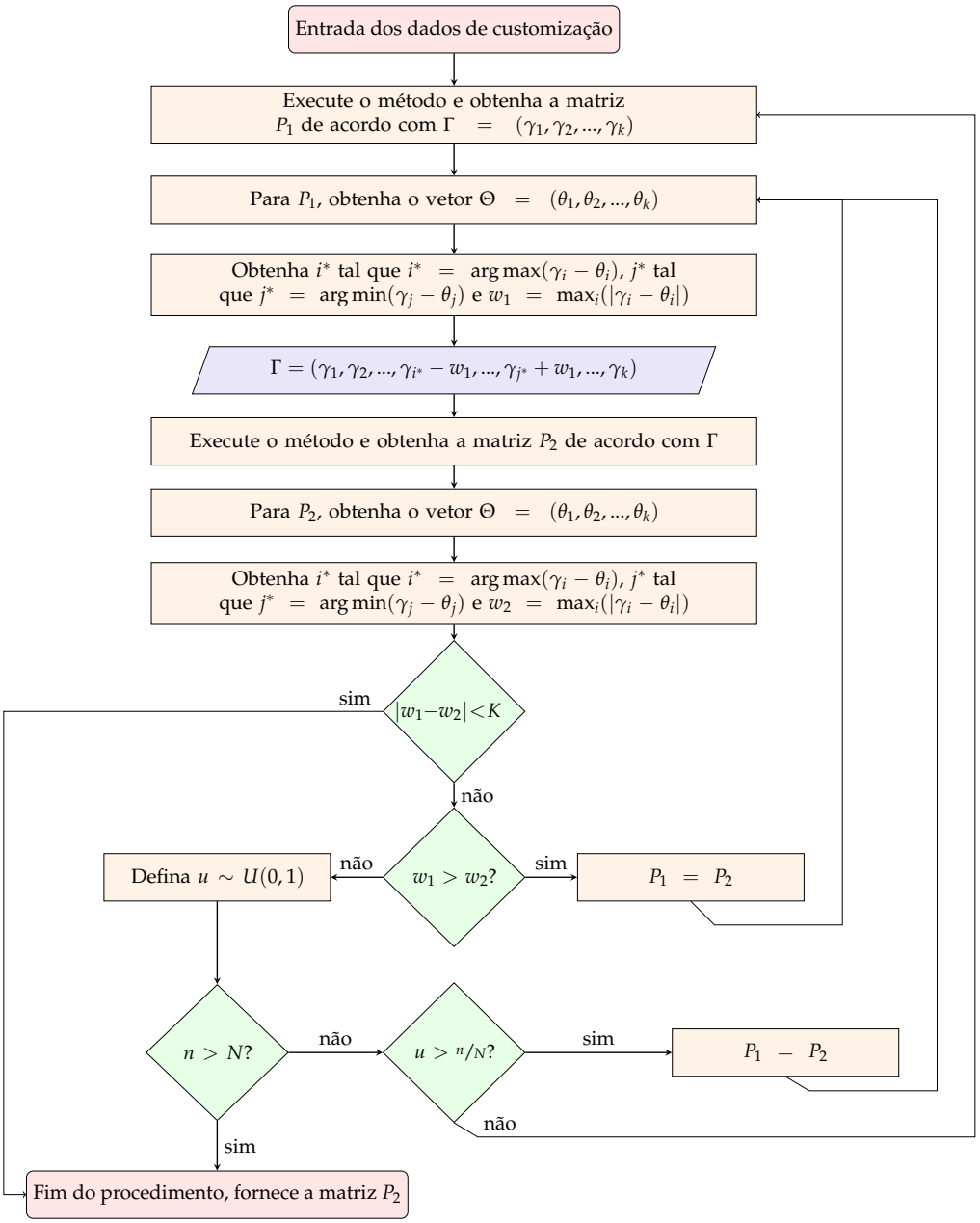


Figura 1 – Fluxograma de execução do método.

De posse dessas considerações, é notório que as alterações em Γ a cada passo garantem que a matriz converge para a matriz desejada. As condições de parada então são ligadas ao máximo entre os elementos de Γ e Θ, em que é definido o valor w_i e comparado com o resultado obtido no passo anterior.

4 Resultados Alcançados

Uma vez estabelecida e detalhada toda a metodologia, um vasto conjunto de experimentos foi conduzido para verificar o comportamento do método. O interesse é formular condições experimentais para executar testes em condições extremas, sejam elas condições de dimensionalidade excessiva ou também de especificidades na customização das matrizes geradas.

Os códigos utilizados para gerar as matrizes descritas nesse estudo foram implementados no software estatístico R [22]. As implementações produzem as situações específicas de estresse algorítmico usando a pacote gencor. Durante o procedimento, o pacote gencor [23] fornece uma implementação do CM-generator e está disponível no CRAN (*The Comprehensive R Archive Network*) através do endereço eletrônico `https://cran.r-project.org/web/packages/gencor/index.html`. Todos os experimentos computacionais foram conduzidos em um dispositivo Intel(R) Core(TM) i7-4790, 3.60 GHz, executando, Windows 10 Pro 64 bits, com 16,00 GB de memória RAM.

Os experimentos estão divididos em três grupos, experimentos sem especificações nas correlações, experimentos com limites mínimos ou máximos para as correlações e por fim, experimentos completamente customizados. Dentro de cada um destes grupos, existem sub-configurações experimentais que serão detalhadas.

A geração de matrizes de correlação sem quaisquer especifidades é relevante para o estudo do desempenho do método com relação a esses cenários. Por outro lado, aplicar testes comparativos de desempenho inerentes a cenários de customização em diversos níveis a depender do interesse do pesquisador são também de grande validade. Como apresentado anteriormente, mesmo em cenários sem nenhuma especificação os limites ℓ_{lower} e ℓ_{upper} precisam estar bem definidos para não atingir as extremidades do intervalo, como explicitado anteriormente. Com isso, os valores utilizados foram $\ell_{lower} = 0{,}01$ e $\ell_{upper} = 0{,}9999$.

Desse modo, os experimentos apresentados dividem-se em três cenários. O primeiro deles é um ambiente sem nenhum nível de customização. Posteriormente, é introduzido algum nível de customização sob a introdução de limites superior ou inferior preestabelecido. Nas matrizes customizadas, como o objetivo é gerar uma matriz mais próxima da matriz desejada quanto possível, seja em questão de aproximação do percentual em cada faixa nos limites predefinidos, seja na permutação de cada percentual ao longo do intervalo $I = (0, 1)$. Por fim, o último cenário em que as matrizes são geradas sob mesmas proporções e também é definido sob os maiores

níveis de customização. Cada experimento foi subdividido para matrizes com menores dimensões e maiores.

4.1 Experimentos sem Especificações nas Correlações

O primeiro conjunto de experimentos é um ambiente sem nenhum nível de customização. Apenas a dimensão das matrizes é especificada. Foram geradas 100 matrizes para cada dimensão no intervalo entre 3 e 500.

Os experimentos sem especificidades são importantes para avaliar o desempenho do método, como enunciado. Estes experimentos iniciais foram realizados para visualizar a estrutura da distribuição das correlações geradas e a exigência de tempo computacional. A Figura 2 mostra a distribuição das correlações geradas para todas as matrizes do experimento

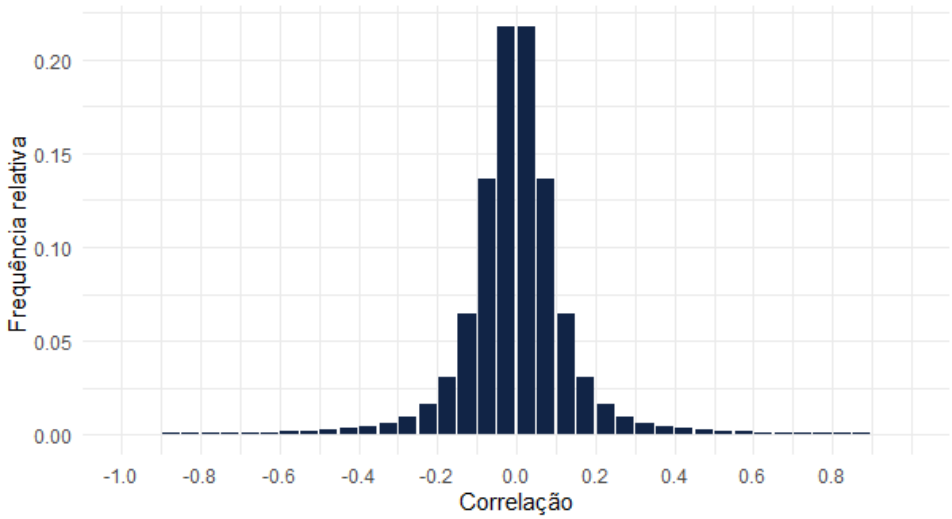
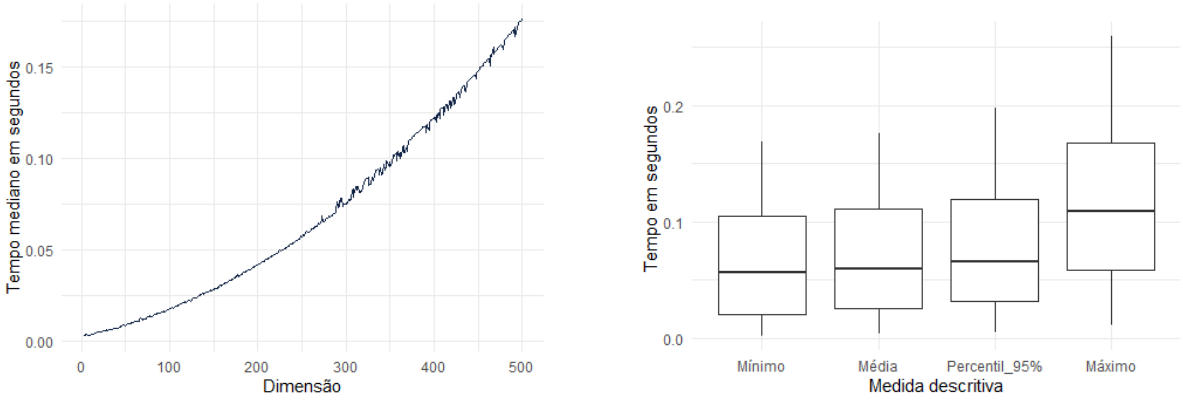


Figura 2 – Histograma para as correlações geradas nas dimensões 3 a 500 para matrizes sem especificações prévias.

As correlações geradas possuem distribuição simétrica em torno de zero. De fato, esse comportamento é esperado, isso se deve à natureza de construção do método, em que as matrizes são geradas a partir da distribuição normal e com os desvios padrão tomados uniformemente através dos limites especificados. Estas são considerações que seguem as descrições metodológicas anteriores.

A Figura 3 explora as informações de tempo computacional. Através da Figura 3(a) é possível ver que o tempo mediano de execução cresce com o aumento da dimensão da matriz gerada, como previsto. Este crescimento é quase linear e mesmo para dimensões bastante elevadas, como 500, o tempo mediano é inferior a 0,20 segundos.



(a) Tempo de execução por dimensão.

(b) Boxplot do tempo de execução.

Figura 3 – Análise gráfica dos tempos de execução do procedimento computacional.

O tempo máximo de simulação possui maior variabilidade como é identificado na figura 3(b). Entretanto, mesmo para situações extremas, o máximo ainda apresentou tempos relativamente pequenos. Os tempos de simulação aumentam para maiores dimensões, como é possível observar na Figura 3(a). Para uma melhor compreensão desse comportamento, foi conduzido um experimento com dimensões mais elevadas. Foram geradas 100 matrizes para cada dimensão de teste, e consideradas as dimensões $1000, 2000, 3000, \ldots, 10000$, com um salto de 1000 dimensões. Nesse segundo momento são verificadas matrizes por dimensão, e sob mesmas configurações de especificidade, porém sob condições mais extremas de dimensionalidade. A Figura 4 apresenta a distribuição das correlações geradas para todas as matrizes do experimento para a situações de estudo com alta dimensionalidade.

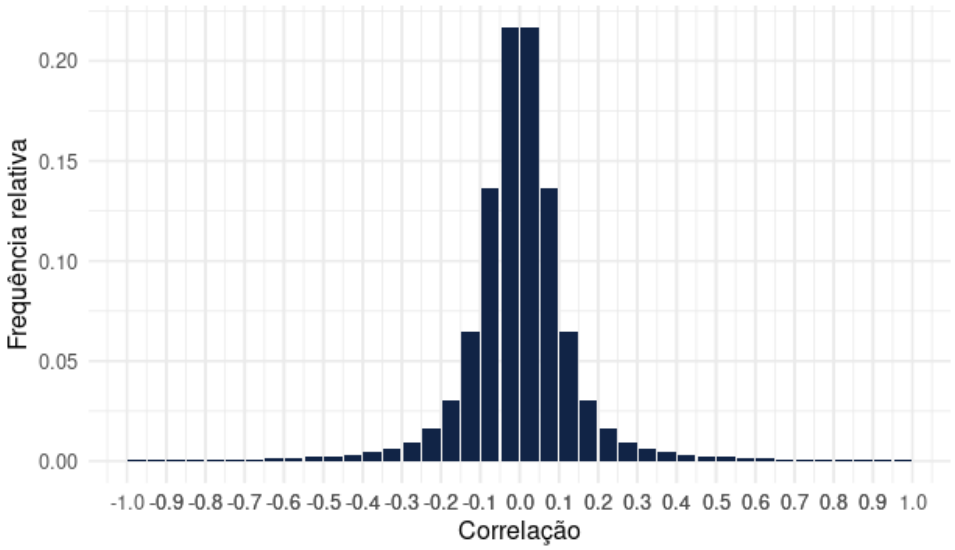


Figura 4 – Histograma para as correlações geradas nas dimensões $1000, 2000, 3000, \ldots,$ 10000 para matrizes sem especificações prévias.

Mesmo para dimensões mais elevadas, o comportamento da distribuições das

correlações permanece inalterado. Novamente as correlações geradas possuem distribuição simétrica em torno de zero. Além disso, o peso das caudas da distribuição é bastante semelhante ao verificado na Figura 2.

A Figura 5 ilustra as informações de tempo computacional. É notório que o tempo de simulação aumenta consideravelmente conforme aumenta-se a dimensão da matriz, porém com o mesmo padrão de curvatura exposto na análise anterior. Nesse sentido, esses fatos indicam que, mediante a esses cenários, o aumento na dimensão da matriz não representa uma perturbação que comprometa o consumo de tempo computacional na execução do método, e o tempo representa apenas uma função da dimensão descrita pelo custo de execução. O maior tempo computacional registrado nesse experimento ocorreu, como previsto, na maior dimensão (10000) com 53 segundos.

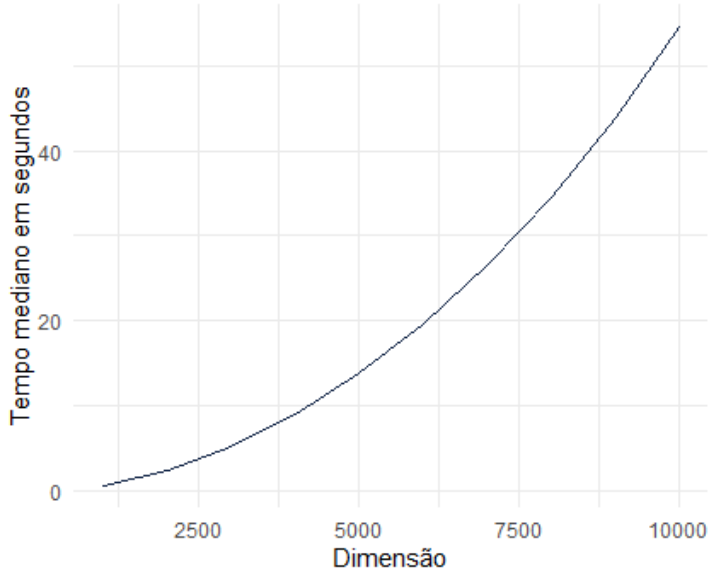


Figura 5 – Tempo de execução por dimensão para experimento sem especificidades e alta dimensionalidade.

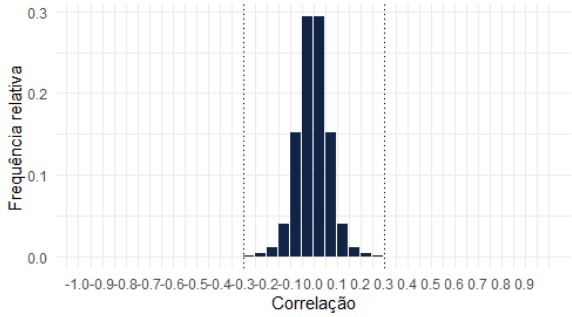
4.2 Experimentos com Limites Mínimos ou Máximos para as Correlações

4.2.1 Experimentos com Limite Superior

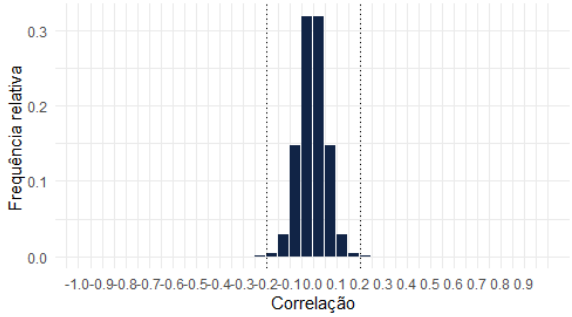
Após a descrição dos experimentos para matrizes gerais, sem nenhum nível de customização, é notória a consistência do método em gerar matrizes sob essas condições. Em seguida, é de interesse prático a experimentação a partir de algum nível de customização. Destarte, os passos a seguir são executados a partir da escolha do limite superior ℓ_{upper} para as correlações geradas. Essa escolha faz com que os desvios padrão sejam limitados inferiormente e ilimitados superiormente.

Dessa maneira, a abordagem desse experimento foi avaliada para os limites $\ell_{upper} = (0{,}05; 0{,}1; 0{,}2; 0{,}3)$. Para cada limite especificado, foram simuladas matrizes com dimensões de 3 a 500, e 100 replicações para cada dimensão. O procedimento seguinte recorre às mesmas características anteriores, a menos da cota superior.

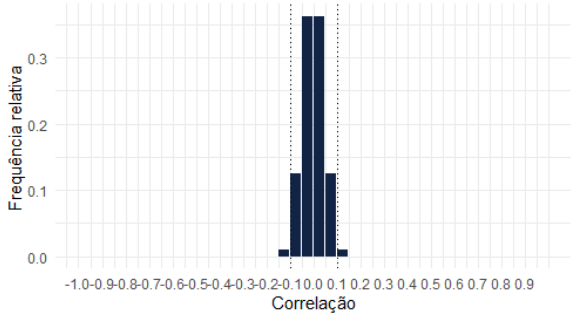
A Figura 6 apresenta estes resultados. É possível perceber que as frequências observadas são influenciadas pelo limite especificado. As correlações que excedem o limite estipulado, representam o percentual de valores que não se encaixam nas especificações prévias. É importante relembrar que uma condição de confirmação do Teorema 3.1 é garantir que $\ell^*_{lower} < \ell_{upper}$, quanto menor for o valor estipulado para ℓ_{upper}, mais restritiva se torna essa condição, daí a ocorrência desses pequenos escapes. Por esse motivo, é relevante a representação das frequências relativas descritas pelos elementos dentro e fora do limite, em que será possível visualizar o desempenho do método nessa abordagem.



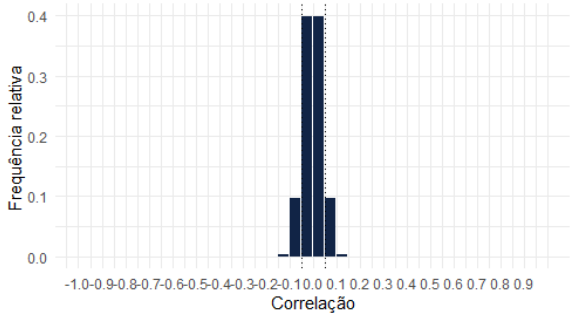
(a) Histograma para $\ell_{upper} = 0{,}3$.



(b) Histograma para $\ell_{upper} = 0{,}2$.



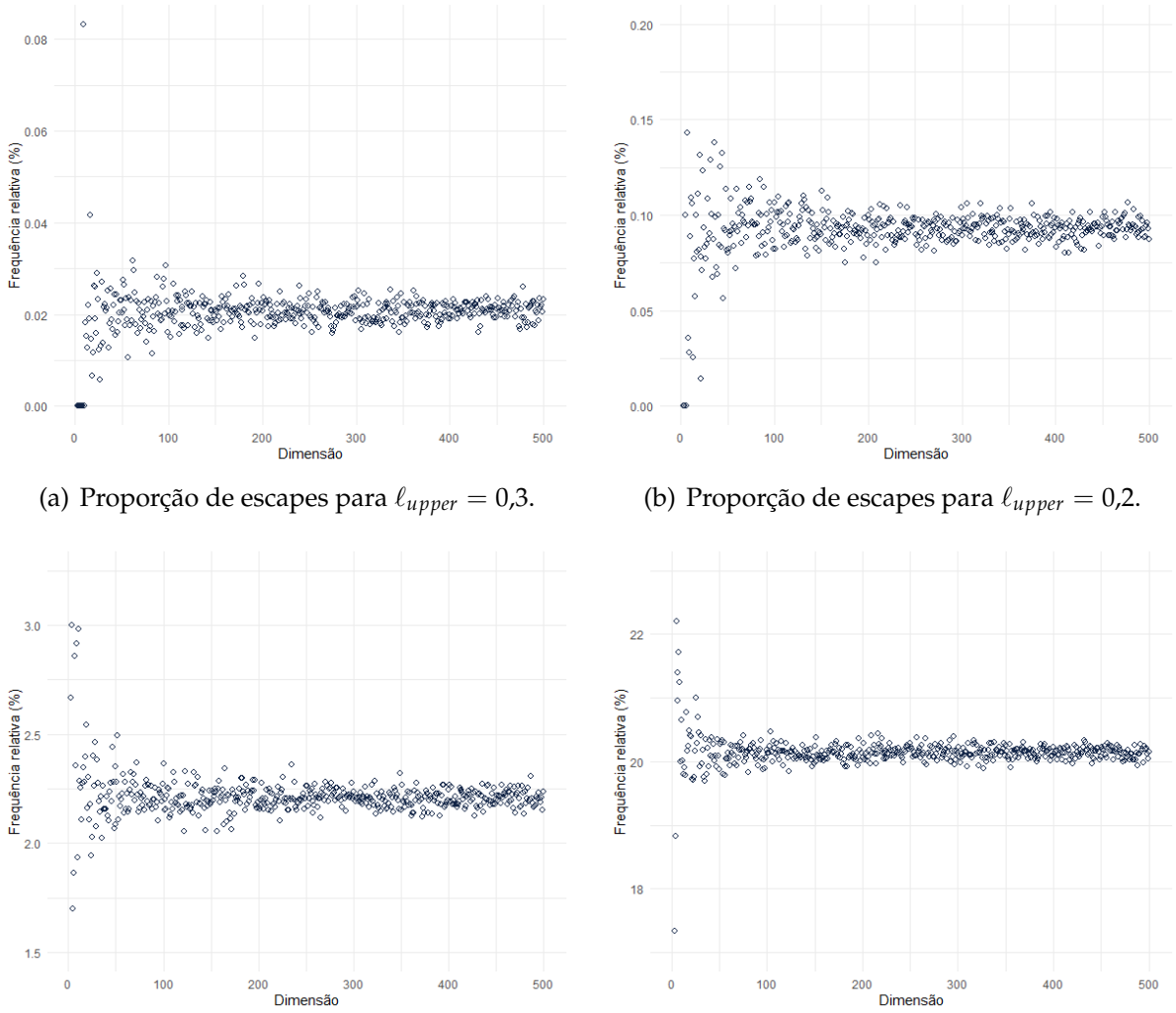
(c) Histograma para $\ell_{upper} = 0{,}1$.



(d) Histograma para $\ell_{upper} = 0{,}05$.

Figura 6 – Histograma para as correlações geradas com limites superiores estabelecidos previamente.

A Figura 7 ilustra como ocorreram os escapes ao longo das rodadas de simulação para as diversas dimensões testadas. É possível verificar uma variabilidade maior para dimensões pequenas. Porém essa variabilidade desaparece rapidamente e o percentual médio de escapes fica estável, ou seja, unicamente dependente da escolha de ℓ_{upper}.



(a) Proporção de escapes para $\ell_{upper} = 0{,}3$.

(b) Proporção de escapes para $\ell_{upper} = 0{,}2$.

(c) Proporção de escapes para $\ell_{upper} = 0{,}1$.

(d) Proporção de escapes para $\ell_{upper} = 0{,}05$.

Figura 7 – Percentual de valores fora do limite superior especificado para cada dimensão avaliada.

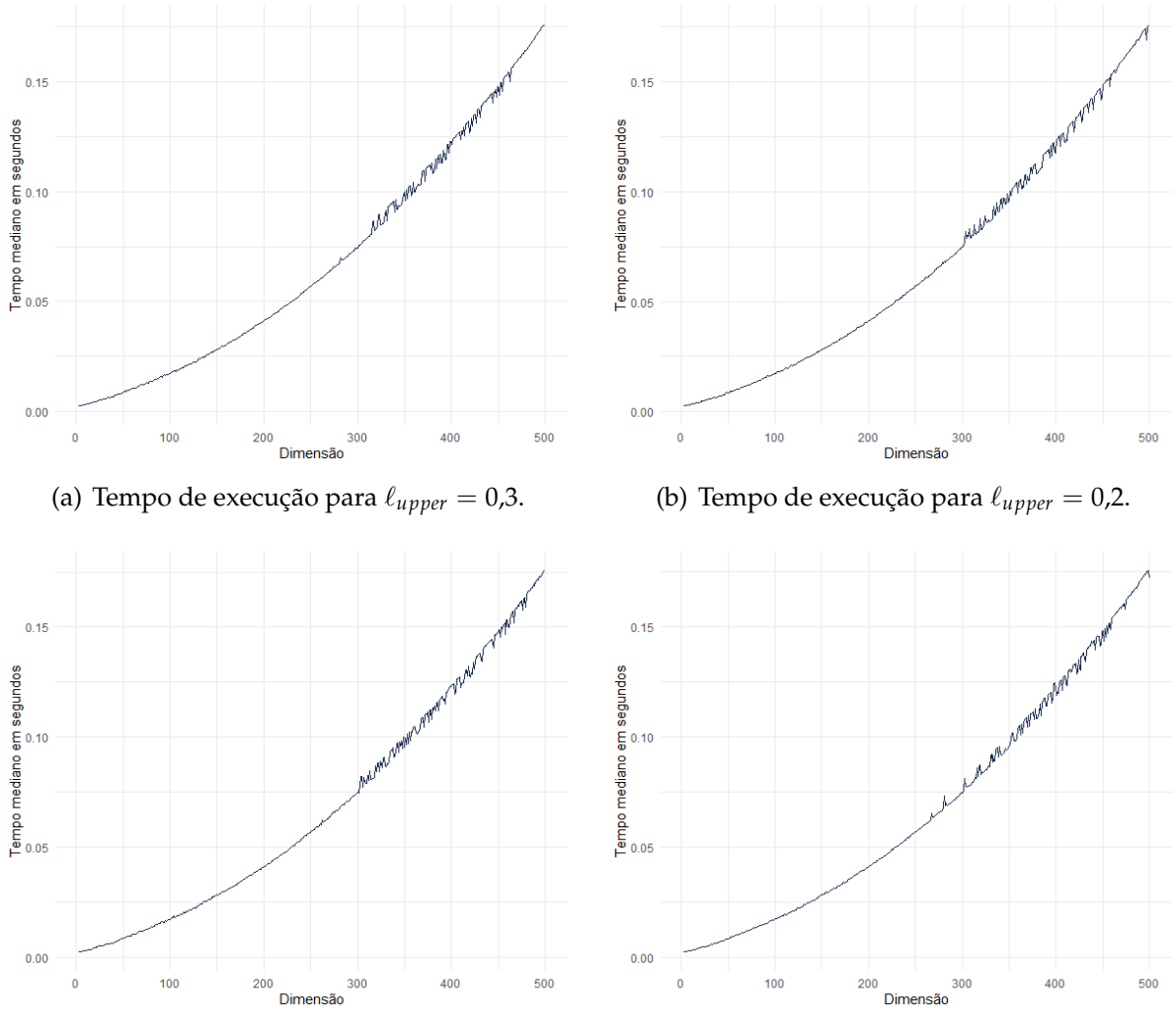
A qualidade de aproximação obtida é bastante satisfatória, para ℓ_{upper} até 0,2, a quantidade de correlações abaixo de ℓ_{upper} foi sempre superior à 99,9%. Estes valores foram menores para as condições mais extremas de ℓ_{upper}. A Tabela 1 apresenta a proporção média entre as 100 replicações.

Tabela 1 – Proporção média de correlações que satisfazem e não satisfazem a cota superior ℓ_{upper}.

ℓ_{upper}	dentro do limite (%)	fora do limite (%)
0,3	99,9791	0,0209
0,2	99,9048	0,0952
0,1	97,7772	2,2228
0,05	79,8839	20,1161

Além da análise anterior, é preciso verificar possíveis impactos no tempo computacional de execução do procedimento. Para tanto, duas informações são bastante

relevantes: se a escolha de ℓ_{upper} afeta os tempos computacionais e se o crescimento do tempo computacional com o aumento dimensional permanece com a mesma estrutura do experimento anterior.A Figura 8 ilustra o comportamento do tempo computacional.



(a) Tempo de execução para $\ell_{upper} = 0{,}3$.

(b) Tempo de execução para $\ell_{upper} = 0{,}2$.

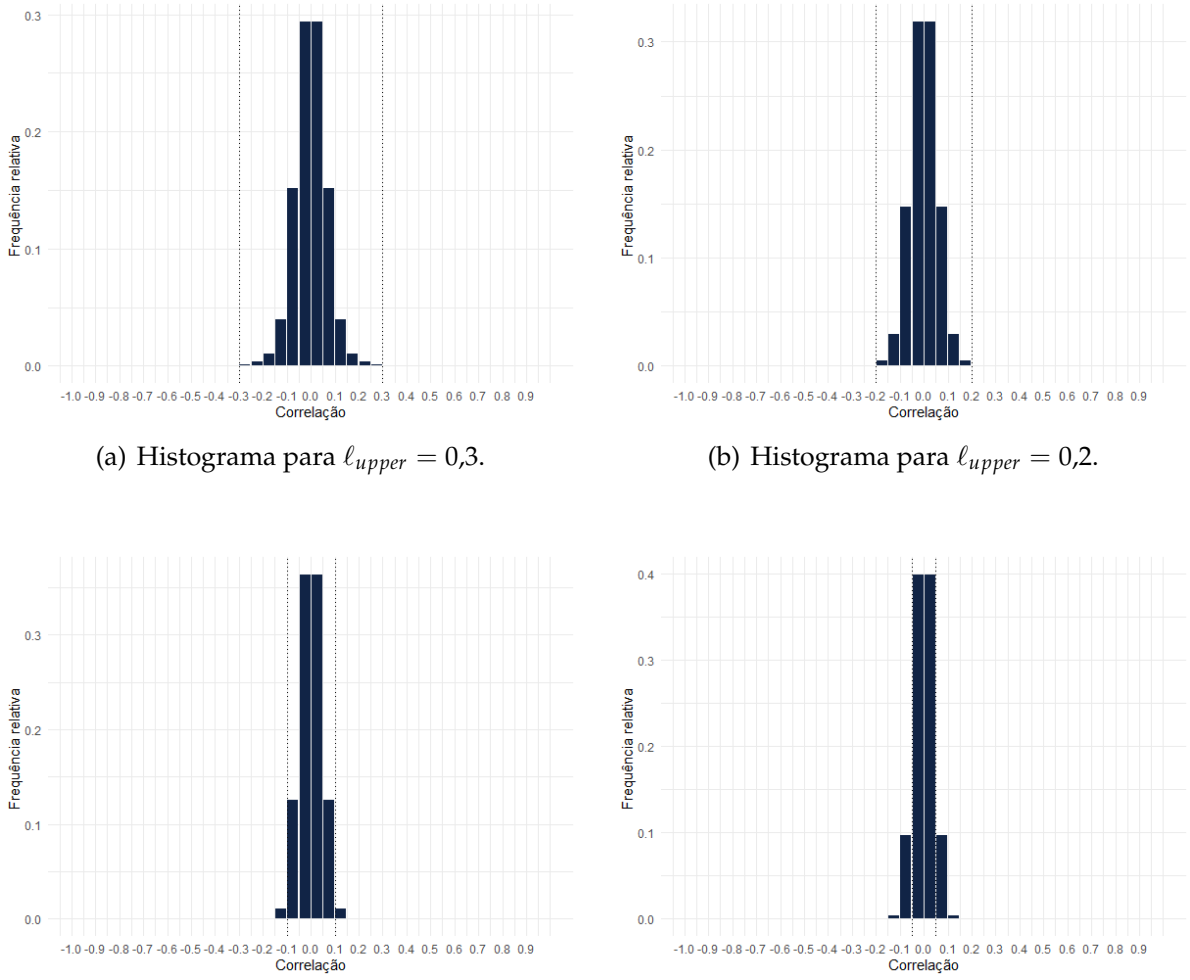
(c) Tempo de execução para $\ell_{upper} = 0{,}1$.

(d) Tempo de execução para $\ell_{upper} = 0{,}05$.

Figura 8 – Tempo de execução computacional para matrizes com limite superior de correlação especificado.

Novamente, como executado para o experimento sem qualquer tipo de especificação, foi conduzido um experimento para dimensões maiores. O mesmo experimento anterior, com $\ell_{upper} \in \{0{,}05; 0{,}1; 0{,}2; 0{,}3\}$, foi executado para gerar matrizes de dimensões $1000, 2000, ..., 10000$, sendo avaliadas 100 matrizes para cada dimensão.

A Figura 9 ilustra o comportamento dos valores de correlação para cada limite estipulado. A execução em condições de elevada dimensionalidade apresentam uma condição de "*stress*"para a metodologia, extremamente importante para a verificação de possíveis anomalias de funcionamento. Assim como verificado anteriormente, ao fixar ℓ_{upper} algumas medidas de correlação escapam dessa especificação.



(a) Histograma para $\ell_{upper} = 0{,}3$.

(b) Histograma para $\ell_{upper} = 0{,}2$.

(c) Histograma para $\ell_{upper} = 0{,}1$.

(d) Histograma para $\ell_{upper} = 0{,}05$.

Figura 9 – Histograma para as correlações geradas com limites superiores estabelecidos previamente para matrizes nas dimensões $1000, 2000, 3000, \ldots, 10000$.

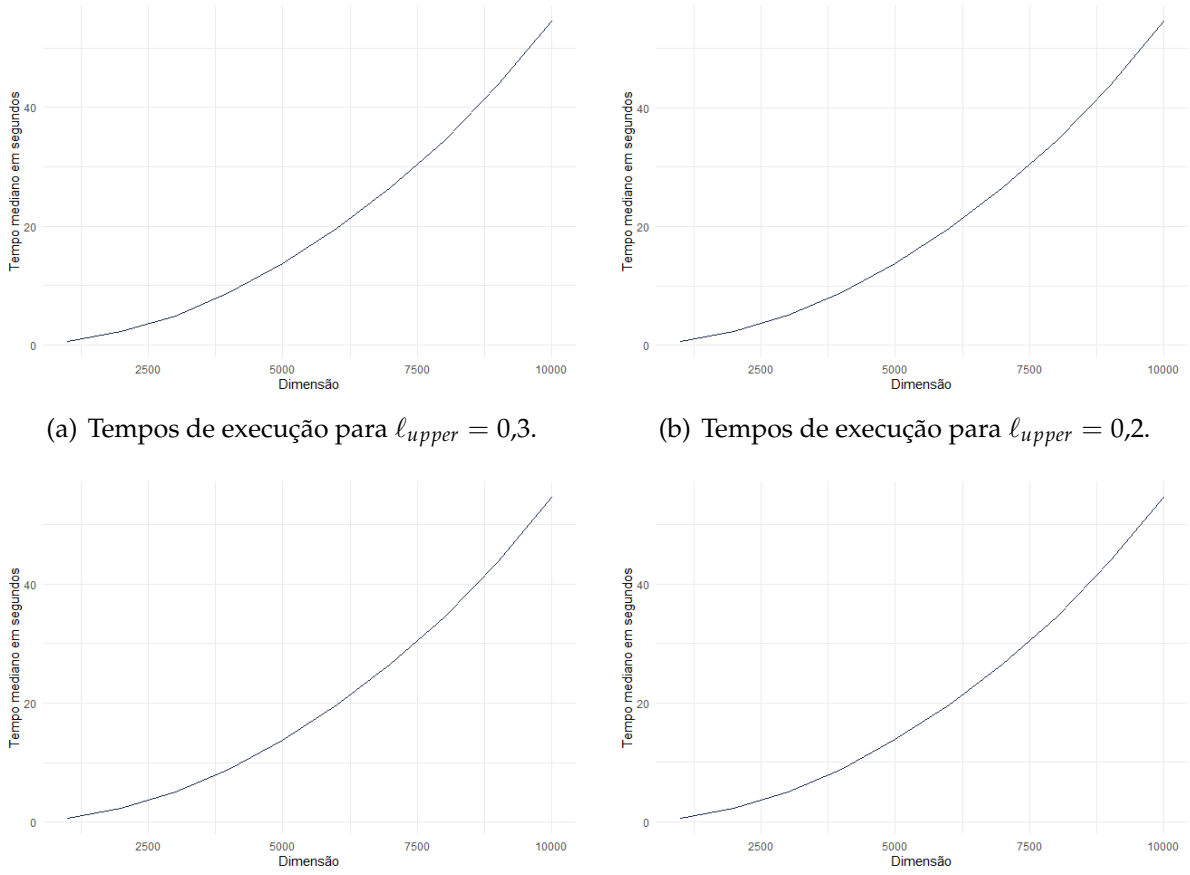
Os resultados apresentados anteriormente se mantiveram para dimensões maiores. A relação das frequências de cada execução é representada pela Tabela 2. As relações de proporção são muito semelhantes às verificadas na abordagem anterior, o que faz crer que o fator dimensional não comprometa o procedimento.

Tabela 2 – Proporção média de correlações que satisfazem e não satisfazem a cota superior ℓ_{upper} para alta dimensionalidade.

ℓ_{upper}	dentro do limite (%)	fora do limite (%)
0,3	99,9785	0,0215
0,2	99,9059	0,0941
0,1	98,8794	1,1206
0,05	79,8338	20,1662

Como abordado anteriormente, é preciso verificar possíveis impactos no tempo computacional de execução do procedimento para dimensões mais elevadas. A Figura

10 ilustra o comportamento do tempo computacional.



(a) Tempos de execução para $\ell_{upper} = 0{,}3$.

(b) Tempos de execução para $\ell_{upper} = 0{,}2$.

(c) Tempos de execução para $\ell_{upper} = 0{,}1$.

(d) Tempos de execução para $\ell_{upper} = 0{,}05$.

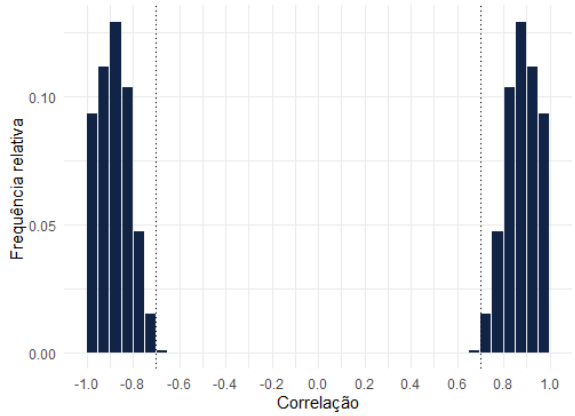
Figura 10 – Tempos de execução para as correlações geradas com limites superiores estabelecidos previamente para matrizes nas dimensões $1000, 2000, 3000, \ldots,$ 10000.

4.2.2 Experimentos com Limite Inferior

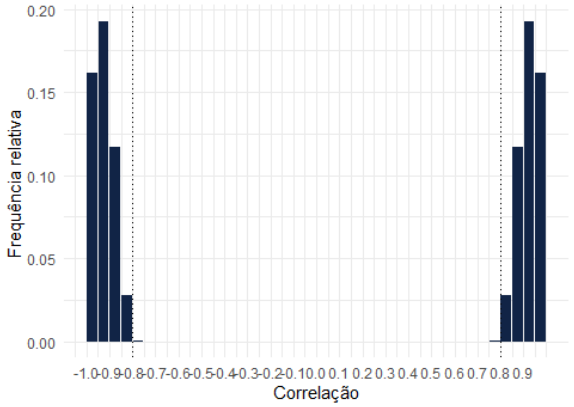
Assim como há grande relevância na observação dos resultados quando ℓ_{upper} é preestabelecido, a mesma observação é válida quando ℓ_{lower}, o limite inferior de customização, é previamente especificado. Por outro lado, para este estudo é conveniente tratar cada um dos limites separadamente. Por esse motivo, cenários completamente customizados serão tratados posteriormente.

Como demonstrado anteriormente na equação (3.11), quando o limite inferior ℓ_{lower} é introduzido previamente os valores dos desvios padrão são limitados inferiormente por um valor maior do que 0 e limitados superiormente por $\sqrt{\frac{1}{\ell_{lower}} - 1}$.

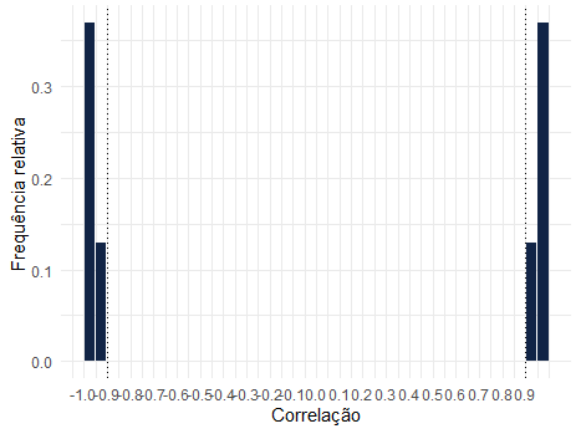
O mesmo procedimento foi executado de modo que o limite inferior fosse preestabelecido. Foram utilizados os limites $\ell_{lower} \in \{0{,}7; 0{,}8; 0{,}9; 0{,}95\}$. Foram geradas matrizes de dimensões 3 a 500, e 100 matrizes para cada dimensão. A Figura 11 ilustra as frequências percentuais observadas nesse contexto, com ℓ_{lower} estabelecido previamente.



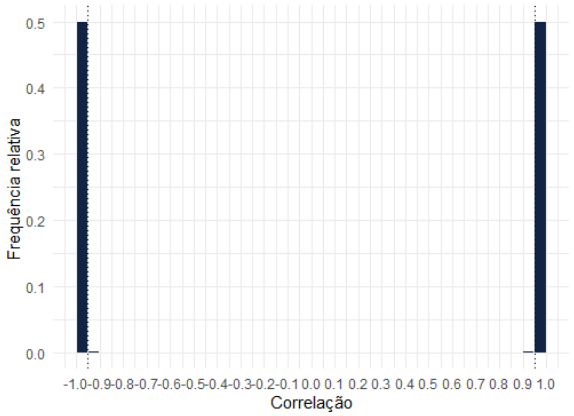
(a) Histograma para $\ell_{lower} = 0{,}7$.



(b) Histograma para $\ell_{lower} = 0{,}8$.



(c) Histograma para $\ell_{lower} = 0{,}9$.

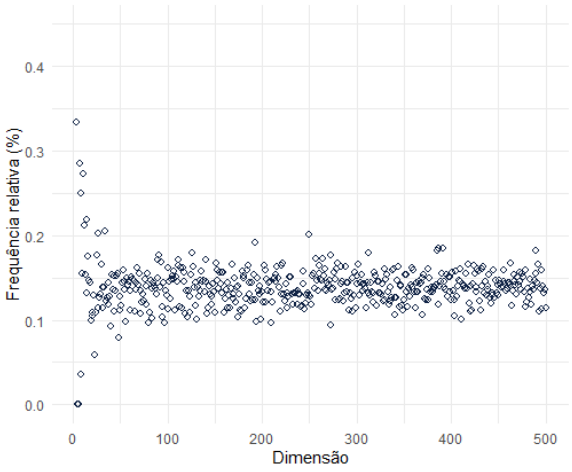


(d) Histograma para $\ell_{lower} = 0{,}95$.

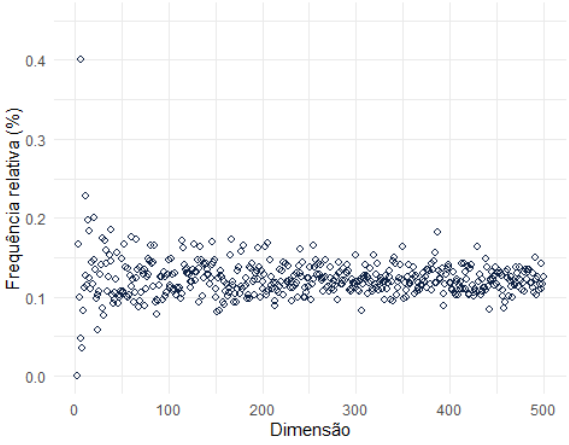
Figura 11 – Histograma para as correlações geradas com limites inferiores estabelecidos previamente.

Os valores que satisfazem a cota estipulada se encaixam nas especificações prévias. Conforme a equação (3.11), uma vez estabelecido o limite inferior, os valores dos desvios padrão são limitados inferiormente por um valor maior que zero e ilimitados superiormente. Pela ilustração apresentada na Figura 11 e em conformidade com os resultados anteriores, o número de elementos de correlações fora das especificações é muito baixo.

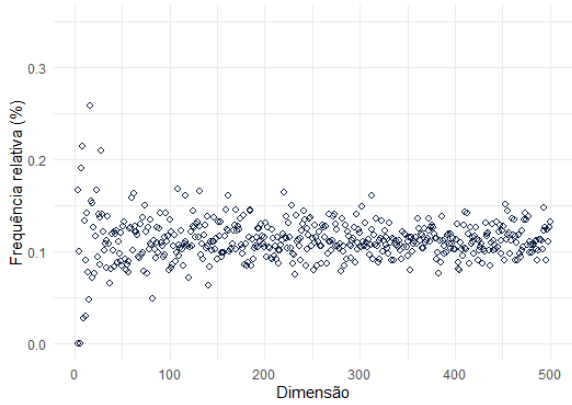
A Figura 12 ilustra um cenário parecido com a descrição da Figura 7. Diante de todas as correlações simuladas, o percentual de valores fora do limite estabelecido diminui em variabilidade e mantém novamente uma média constante, à medida que cresce a dimensão das matrizes simuladas. Conforme diminui-se o valor do limite ℓ_{lower}, observa-se um aumento no percentual de correlações que escapam para fora do limite estipulado.



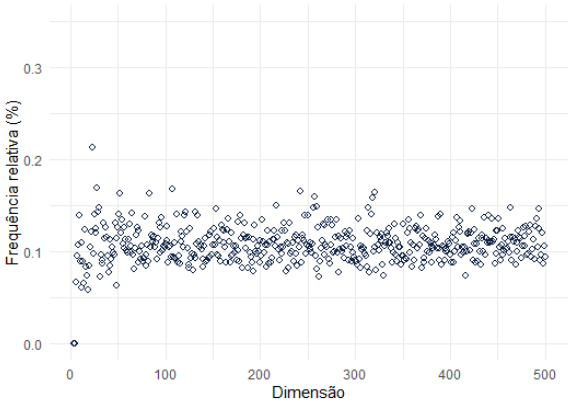
(a) Proporção de escapes para $\ell_{lower} = 0{,}7$.



(b) Proporção de escapes para $\ell_{lower} = 0{,}8$.



(c) Proporção de escapes para $\ell_{lower} = 0{,}9$.



(d) Proporção de escapes para $\ell_{lower} = 0{,}95$.

Figura 12 – Percentual de valores fora do limite inferior especificado para cada dimensão avaliada.

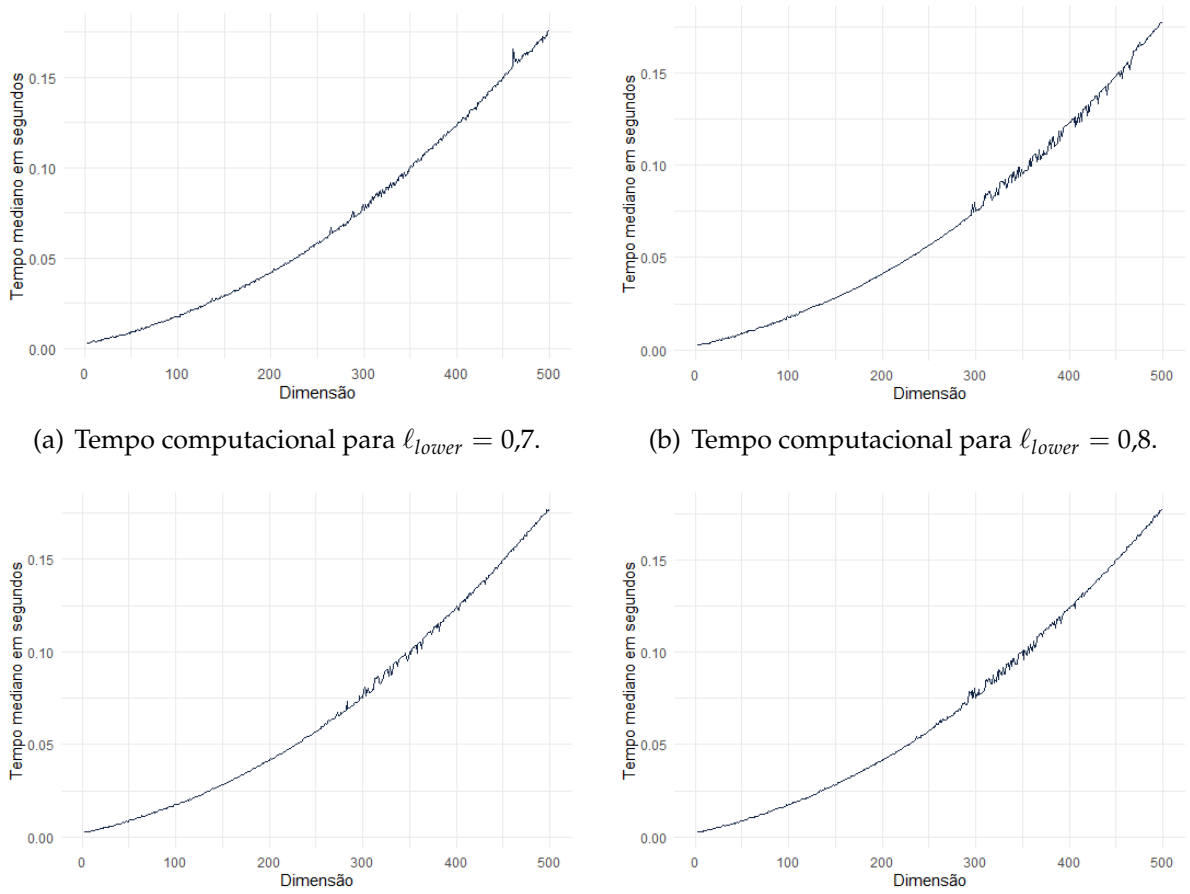
A relação das frequências de cada execução é representada pela Tabela 3 que descreve o índice percentual de valores dentro e fora do limite especificado. Novamente as relações de proporção são semelhantes às verificadas nos experimentos anteriores, o que confirma que o fator dimensional não compromete o procedimento.

Tabela 3 – Proporção média de correlações que satisfazem e não satisfazem a cota inferior ℓ_{lower}

ℓ_{upper}	dentro do limite (%)	fora do limite (%)
0,7	99,8605	0,1395
0,8	99,8786	0,1214
0,9	99,8889	0,1121
0,95	99,8913	0,1087

O índice proporcional de correlações fora e dentro das especificações prévias descrito na Tabela 3 são coerentes. Conforme aumenta-se o valor de ℓ_{lower} o intervalo para os desvios padrão é menor, isso estabelece uma relação inversamente proporcional.

Embora haja um índice percentual de valores fora do limite especificado previamente, os valores encontrados são bastante expressivos e satisfatórios, uma vez que o menor percentual de valores que não excederam a cota estipulada ainda se encontra superior a 99,8%. Para verificar impactos no tempo computacional de execução do procedimento, a Figura 13 ilustra o tempo computacional quando ℓ_{lower} é preestabelecido.



(a) Tempo computacional para $\ell_{lower} = 0{,}7$.

(b) Tempo computacional para $\ell_{lower} = 0{,}8$.

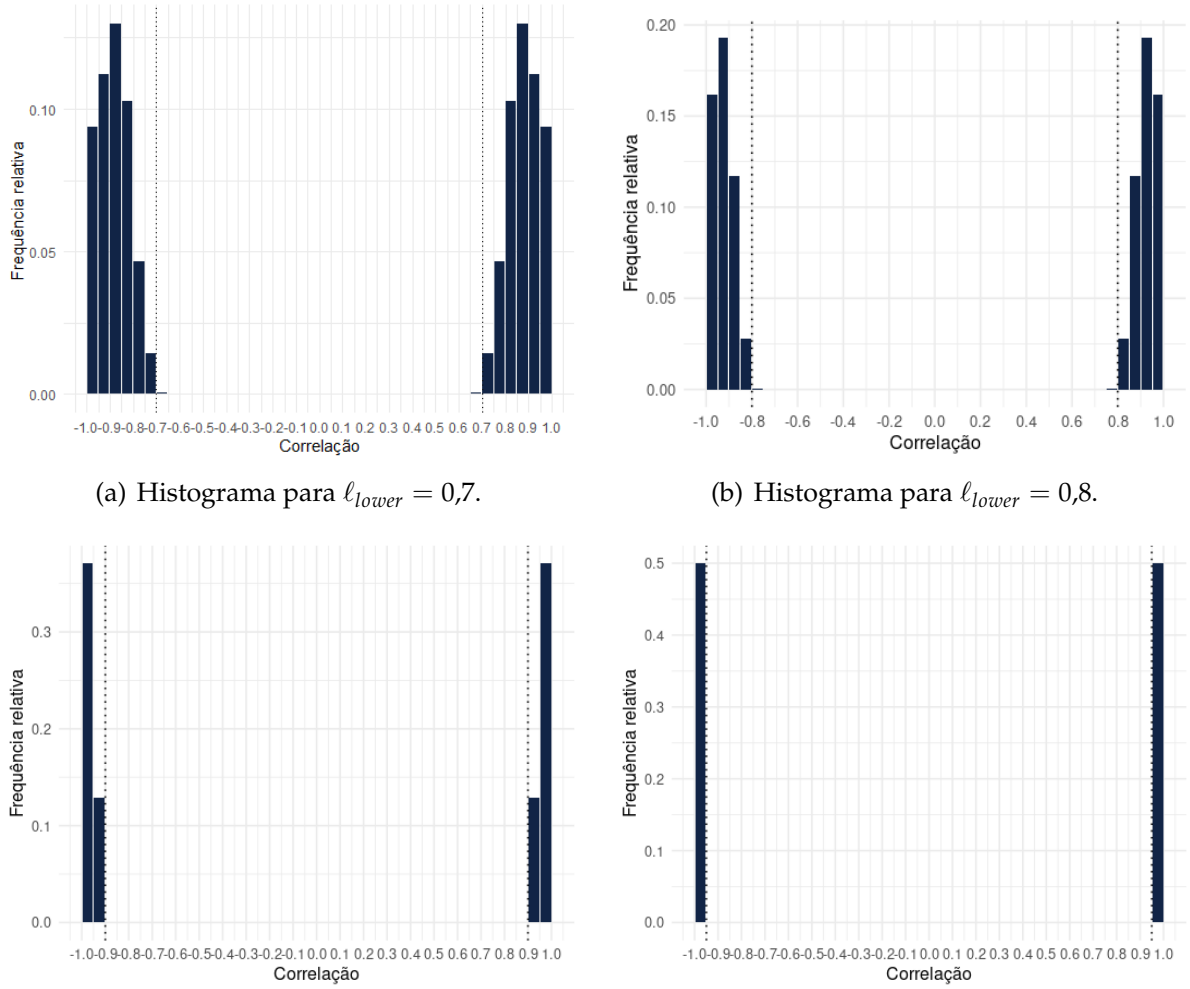
(c) Tempo computacional para $\ell_{lower} = 0{,}9$.

(d) Tempo computacional para $\ell_{lower} = 0{,}95$.

Figura 13 – Tempo computacional para correlações com o limite inferior estabelecido previamente.

Por meio dos tempos computacionais representados na Figura 13 é notório que os tempos são similares aos tempos representados na Figura 8, com um valor mediano (em segundos) gasto próximo a 0,2 para a dimensão 500. Semelhantemente ao resultado obtido para ℓ_{upper} especificado, a escolha do limite inferior não altera o tempo computacional.

Ainda conforme a abordagem anterior, é válida a avaliação das correlações para a alta dimensionalidade, quando ℓ_{lower} corresponde ao limite preestabelecido. Nesse sentido, foram simuladas matrizes de dimensão de $1000, 2000, ..., 10000$, e 100 matrizes para cada dimensão, para os mesmos limites ℓ_{lower} indicados no experimento anterior. Estes resultados podem ser visualizados através da Figura 14.



(a) Histograma para $\ell_{lower} = 0{,}7$.

(b) Histograma para $\ell_{lower} = 0{,}8$.

(c) Histograma para $\ell_{lower} = 0{,}9$.

(d) Histograma para $\ell_{lower} = 0{,}95$.

Figura 14 – Análise gráfica das frequências das correlações para a alta dimensionalidade, com ℓ_{lower} preestabelecido.

Pela análise da Figura 14 fica evidente que os resultados expressos configuram índices percentuais de correlação muito próximos às frequências apresentadas na Figura 11. Portanto, a alta dimensionalidade é um fator que não expressa uma perturbação significativa ao método sob esses cenários, em que o valor de ℓ_{lower} é preestabelecido previamente.

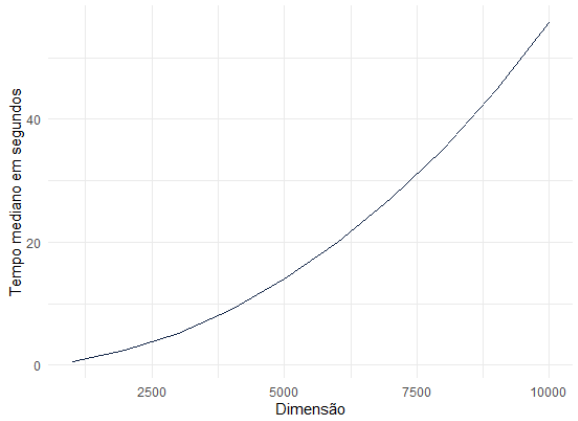
Mais que isso, é possível notar que os valores de correlação que não se encaixam às especificações detectados pelo experimento seguem o mesmo princípio identificado na Figura 11, em que o percentual de correlações fora aumenta com a redução do limite ℓ_{lower}, por aumentar o desvio padrão da distribuição simulada.

A Tabela 4 expressa uma média das frequências observadas para as correlações obtidas sob alta dimensionalidade e ℓ_{lower} preestabelecido, para valores dentro e fora do limite especificado.

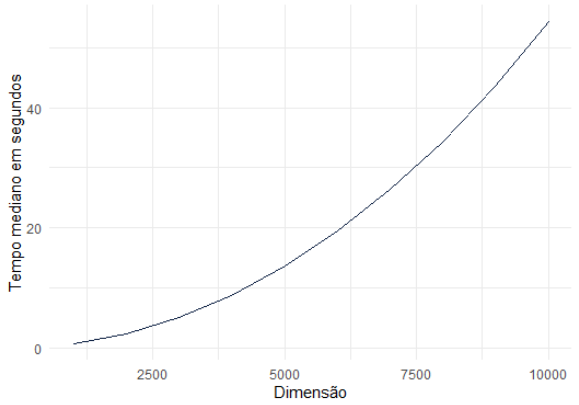
Tabela 4 – Proporção média de correlações que satisfazem e não satisfazem a cota inferior ℓ_{lower} para a alta dimensionalidade

ℓ_{lower}	Dentro do limite (%)	Fora do limite (%)
0,7	99,8776	0,1224
0,8	99,8848	0,1152
0,9	99,8929	0,1071
0,95	99,885	0,1150

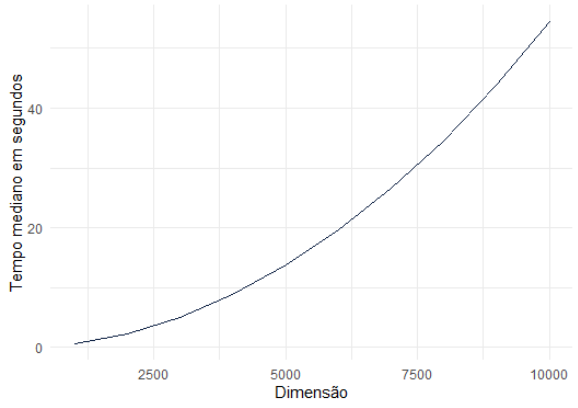
Para a alta dimensionalidade, as frequências de correlações se mantiveram acima de 99,8% o que demonstra também um resultado bastante consistente. A Figura 15 ilustra o tempo computacional do experimento.



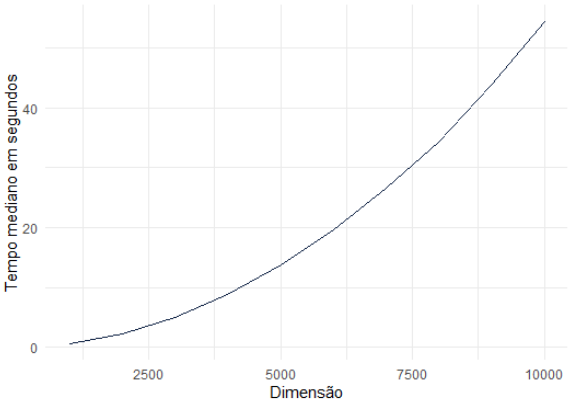
(a) Tempo computacional para $\ell_{lower} = 0{,}7$.



(b) Tempo computacional para $\ell_{lower} = 0{,}8$.



(c) Tempo computacional para $\ell_{lower} = 0{,}9$.



(d) Tempo computacional para $\ell_{lower} = 0{,}95$.

Figura 15 – Tempo computacional para correlações com o limite inferior estabelecido previamente para a alta dimensionalidade.

Os tempos de simulação ilustrados na Figura 15 demonstram novamente a mesma situação ilustrada nas Figuras 8, 10 e 13, em que os limites de customização não interferem no tempo computacional utilizado. É relevante ainda uma comparação com os resultados dos tempos computacionais descritos na Figura 5, em que os mesmos resultados são obtidos porém sem nenhum nível de customização. Isso indica novamente

que a inserção do limite inferior de customização não representa uma perturbação à execução do método, como identificado para o limite superior.

4.3 Experimentos Completamente Customizados para a Geração de Matrizes de Correlação

Após a descrição dos experimentos para matrizes gerais e com o limite superior e inferior preestabelecidos separadamente, é necessária uma abordagem para cenários completamente customizados. Neste caso, é possível inserir não somente mais que um limite para as correlações, mas também proporções associadas a cada faixa especificada.

O procedimento de otimização utilizado é ilustrado na figura 1, em que são demonstrados os passos de execução e como a inserção de um vetor de proporções permite que o resultado convirja para a matriz desejada, ponderado pelo número de execuções desejado e um valor K correspondente ao erro de aceitação entre a aproximação feita e as proporções inseridas.

A metodologia experimental abrange então a observação das correlações sobre os limites especificados, e suas respectivas frequências sobre cada faixa desejada. Além disso, o fator dimensional, custo computacional e performance do método são delimitadores para avaliar o cenário completamente customizado. O número de repetições do procedimento de otimização utilizado foi arbitrariamente fixado em 1000 repetições, com uma precisão fixada em $K = 0{,}03$, em que ambas informações representam os critérios de parada associados ao processo de otimização. Desse modo, após a execução, o procedimento foi repetido até a obtenção de uma precisão de no mínimo 0,03 (máxima diferença entre as proporções inseridas e as proporções encontradas na matriz naquele passo), ou até o fim das 1000 repetições inseridas previamente.

Foram simuladas matrizes de dimensão 3 a 100, com 30 matrizes para cada dimensão. O número reduzido de réplicas em relação aos experimentos anteriores se justifica devido ao maior tempo de execução do procedimento customizado. Entretanto, mesmo o número reduzido de réplicas é suficiente para fornecer evidências sobre o desempenho da técnica. Os limites especificados formam o conjunto $(0{,}2; 0{,}7)$, o que implica dizer que as correlações deverão estar contidas nas três faixas seguintes: $(-0{,}2; 0{,}2)$, $(-0{,}7; -0{,}2) \cup (0{,}2; 0{,}7)$ e $(-1; -0{,}7) \cup (0{,}7; 1)$.

Os experimentos foram conduzidos também para permutações do vetor de proporções iniciais Γ sugerido para testes. O intuito foi verificar o desempenho do método quando submetido a diferentes proporções em diferentes faixas de correlação. O objetivo após todas as execuções é fornecer uma matriz que possua uma aproximação para as proporções em cada faixa, de acordo com os parâmetros K e N, associados à

precisão da aproximação.

Uma verificação das frequências das correlações obtidas através de simulação para cada vetor de proporções Γ, permite uma compreensão do desempenho do método sob nível máximo de customização. Aliado à isso, uma análise de tempo computacional para execução também é bastante relevante nesse estudo. O vetor de proporções iniciais Γ foi introduzido ao experimento como $\Gamma = (0{,}5; 0{,}495; 0{,}005)$. Sob essas condições, o resultado esperado inclui 50% das correlações entre $(-0{,}2; 0{,}2)$, depois 49,5% das correlações no intervalo $(-0{,}7; -0{,}2) \cup (0{,}2; 0{,}7)$, e por fim, 0,5% das correlações no intervalo $(-1; -0{,}7) \cup (0{,}7; 1)$.

É importante observar que este experimento é uma configuração extrema, um experimento com estratégia de contaminação. Na prática, todas as correlações estão contidas dos dois primeiros intervalos e apenas uma contaminação é imposta ao terceiro intervalo. A mesma estratégia foi replicada alterando o intervalo que receberá a contaminação. Foram geradas matrizes também com o vetor $\Gamma = (0{,}5; 0{,}005; 0{,}495)$, e ainda $\Gamma = (0{,}005; 0{,}5; 0{,}495)$. O resultado com esse nível de customização permite uma análise detalhada dos resultados, pois as repetições com as mesmas proporções porém em faixas diferentes permitem avaliar o desempenho do método sob esse contexto de *stress* de execução. A Figura 16 ilustra as frequências das correlações obtidas para a primeira configuração, com $\Gamma = (0{,}5; 0{,}495; 0{,}005)$.

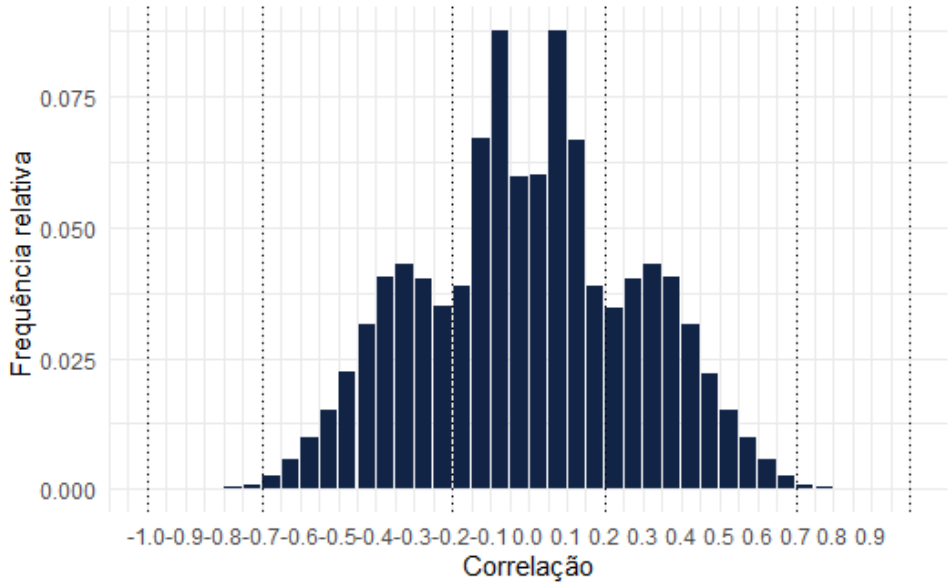
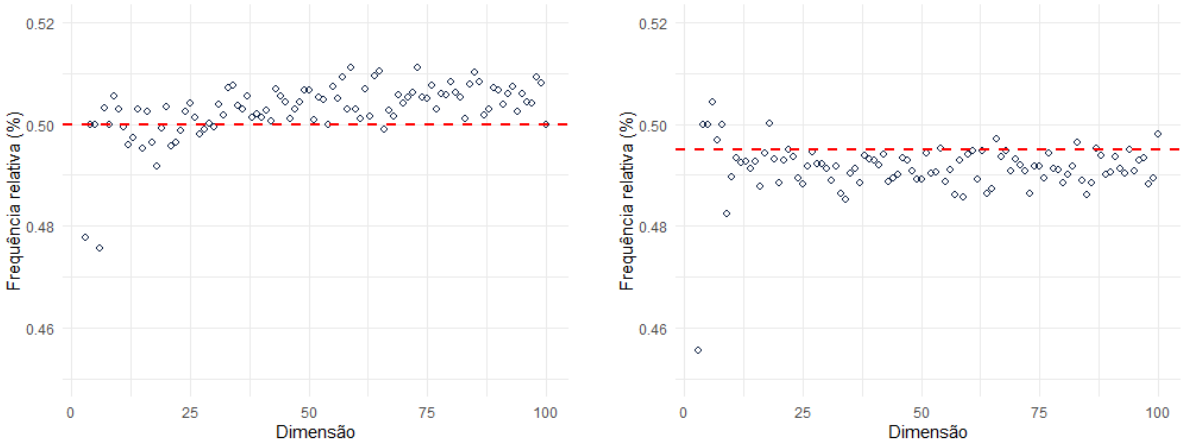


Figura 16 – Histogramas de frequência para as correlações geradas sob nível máximo de customização e $\Gamma = (0{,}5; 0{,}495; 0{,}005)$.

O efeito proporcional da contaminação no intervalo $(-1; -0{,}7) \cup (0{,}7; 1)$ fica bastante evidente. As frequências relativas por dimensão, para os dois intervalos não contaminados podem ser observadas na Figura 17.



(a) Intervalo $(-0{,}2; 0{,}2)$. (b) Intervalo $(-0{,}7; -0{,}2) \cup (0{,}2; 0{,}7)$.

Figura 17 – Frequências relativas observadas nas faixas de correlação não contaminadas para $\Gamma = (0{,}5; 0{,}495; 0{,}005)$.

Já a Figura 18 apresenta as frequências relativas por dimensão, para o intervalo contaminado. Para as dimensões menores é possível observar uma maior variabilidade, porém este efeito é rapidamente dissipado. Para dimensões maiores, a contaminação sempre apresentou proporção de ocorrência bastante condizente com o vetor Γ proposto. Visto de uma outra forma, o método apresenta ótimo comportamento para gerar matrizes que atendam o perfil de contaminação para o intervalo de correlações que estejam em valores elevados (maiores que 0,7).

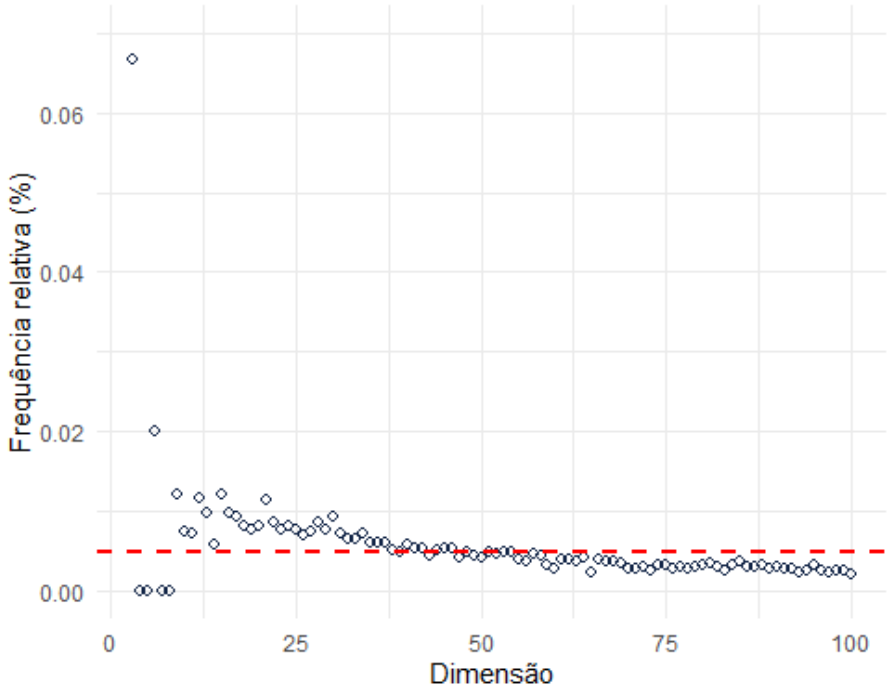


Figura 18 – Frequências relativas observadas na faixa de correlação contaminada para $\Gamma = (0{,}5; 0{,}495; 0{,}005)$.

Um segundo formato para esse experimento posicionou o efeito de contaminação no intervalo $(-0{,}7; -0{,}2) \cup (0{,}2; 0{,}7)$, ou seja, para correlações em patamar médio. A Figura 19 ilustra as frequências das correlações obtidas para o vetor de proporções $\Gamma = (0{,}5; 0{,}005; 0{,}495)$.

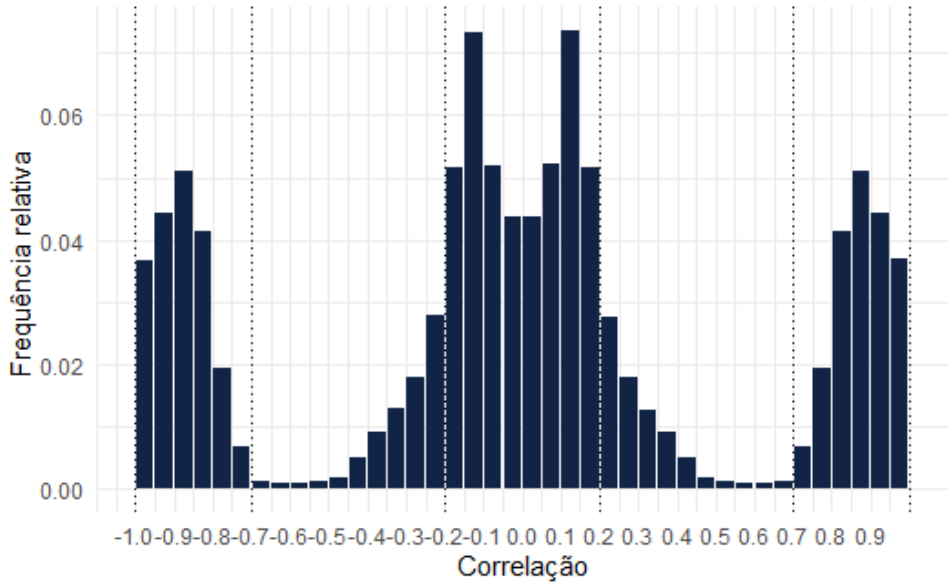
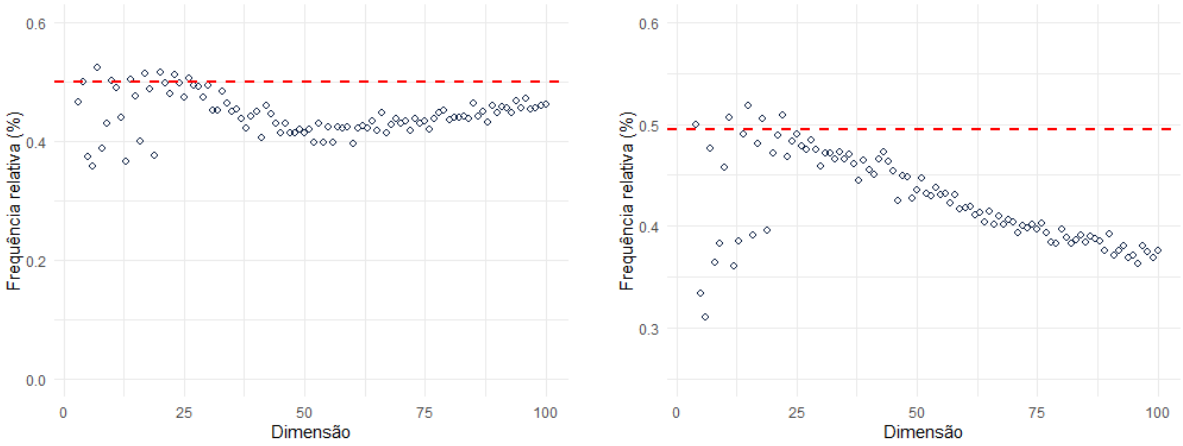


Figura 19 – Histogramas de frequência para as correlações geradas sob nível máximo de customização e $\Gamma = (0{,}5; 0{,}005; 0{,}495)$.

Novamente o efeito proporcional da contaminação é notório, porém a sensação visual é de um volume maior que os níveis de estabelecidos em Γ. As frequências relativas por dimensão, para os dois intervalos não contaminados são apresentados na Figura 20.

Na Figura 21 são ilustradas as frequências relativas por dimensão, para o intervalo contaminado. Novamente, para as dimensões menores, ocorre uma maior variabilidade que se dissipa com o crescimento dimensional das matrizes. Neste experimento, a contaminação sempre esteve em patamares superiores aos estipulados nas proporções definidas pelo vetor Γ.

É possível verificar que, com o aumento dimensional, a frequência relativa observada apresenta uma queda em direção do valor estipulado, mas não uma queda suficiente para alcançar o objetivo prévio. Visto de outra forma, o método apresenta uma fragilidade para gerar matrizes que atendam o perfil de contaminação para o intervalo de correlações médias. Este resultado não é completamente inesperado, isso decorre de trabalhar com uma proibição de gerar correlações numa faixa bastante ampla, e de valores de correlação tão usuais. Seria razoável supor que este poderia ser um obstáculo difícil de transpor.



(a) Intervalo $(-0{,}2; 0{,}2)$. (b) Intervalo $(-1; -0{,}7) \cup (0{,}7; 1)$.

Figura 20 – Frequências relativas observadas nas faixas de correlação não contaminadas para $\Gamma = (0{,}5; 0{,}005; 0{,}495)$.

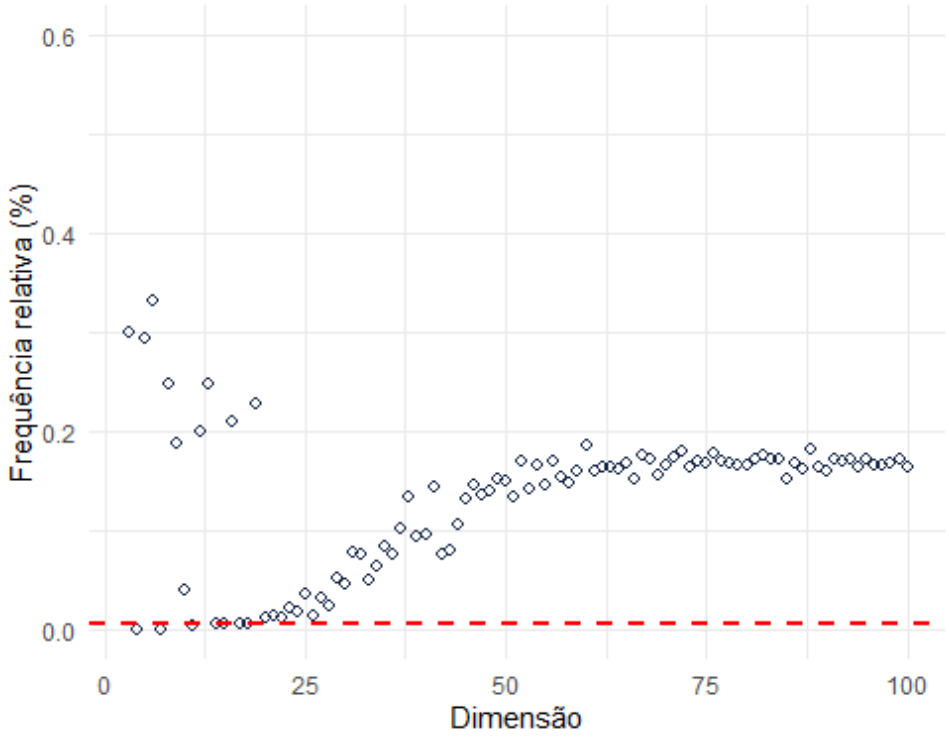


Figura 21 – Frequências relativas observadas na faixa de correlação contaminada para $\Gamma = (0{,}5; 0{,}005; 0{,}495)$.

É razoável supor que se o parâmetro N (quantidade de rodadas de otimização) da estratégia de otimização for aumentado, a performance do método melhorasse. O método tende a convergir com a precisão desejada desde que submetido ao esforço computacional necessário para este propósito. É importante lembrar que este é um teste em condições extremas, diferente da geração de matrizes em condições de contorno mais usuais. Essa situação reforça a validade de uma investigação baseada nestas condições extremas aqui apresentadas.

Por fim, este estudo apresenta um terceiro formato para o experimento. Agora, adicionou-se o efeito de contaminação no intervalo $(-0{,}2; 0{,}2)$, ou seja, para correlações mais baixas. A Figura 22 mostra as frequências relativas das correlações simuladas obtidas para o vetor de proporções $\Gamma = (0{,}005; 0{,}5; 0{,}495)$. De fato, o setor contaminado revela frequências baixas de correlação como era esperado. Assim como nas configurações anteriores, o efeito proporcional da contaminação é visual para os níveis fixados em Γ. Como para os casos anteriores, uma análise associada a cada intervalo é mais informativa.

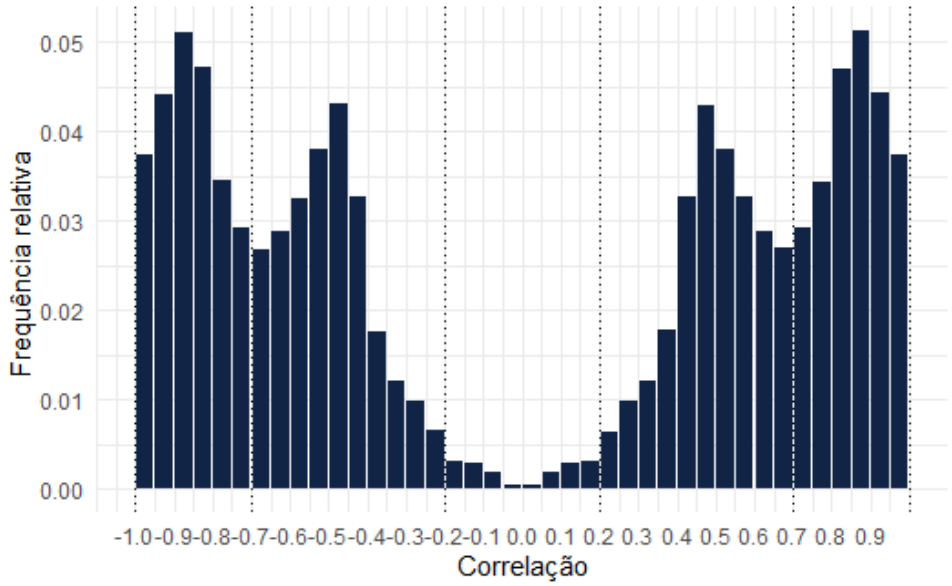
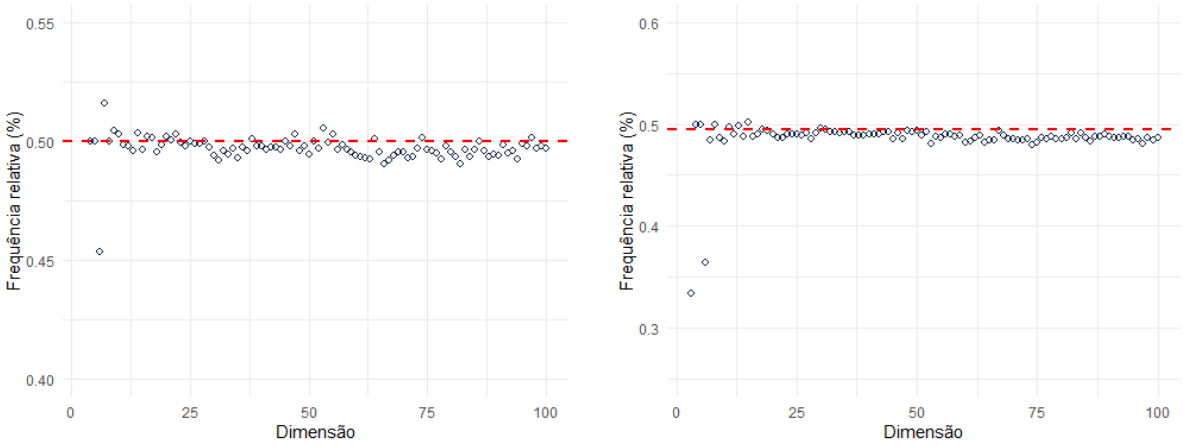


Figura 22 – Histogramas de frequência para as correlações geradas sob nível máximo de customização e $\Gamma = (0{,}005; 0{,}5; 0{,}495)$.

As frequências relativas por dimensão, para os dois intervalos não contaminados são observadas através da Figura 23. É possível notar que o comportamento dos resultados voltou a ser parecido com o primeiro experimento, com $\Gamma = (0{,}5; 0{,}495; 0{,}005)$. Para os dois intervalos, as frequências relativas obtidas através dos resultados simulados se aproximam da configuração do vetor $\Gamma = (0{,}005; 0{,}5; 0{,}495)$.

Já na Figura 24 são ilustradas as frequências relativas por dimensão, para o intervalo contaminado. Novamente, para as dimensões menores ocorre uma maior variabilidade que se dissipa com o crescimento dimensional. Neste experimento, a contaminação sempre esteve em patamares superiores aos estipulados no vetor Γ.

Para uma análise numérica mais pormenorizada, a Tabela 5 descreve o percentual médio encontrado para todas as dimensões, por faixa de correlação e por cada especificação do vetor de proporções iniciais Γ. A sequência de proporções em que os valores reais mais se distanciaram foi $\Gamma = (0{,}5; 0{,}005; 0{,}495)$. Com um percentual de



(a) Intervalo $(-0{,}7; -0{,}2) \cup (0{,}2; 0{,}7)$. (b) Intervalo $(-1; -0{,}7) \cup (0{,}7; 1)$.

Figura 23 – Frequências relativas observadas nas faixas de correlação não contaminadas para $\Gamma = (0{,}005; 0{,}5; 0{,}495)$.

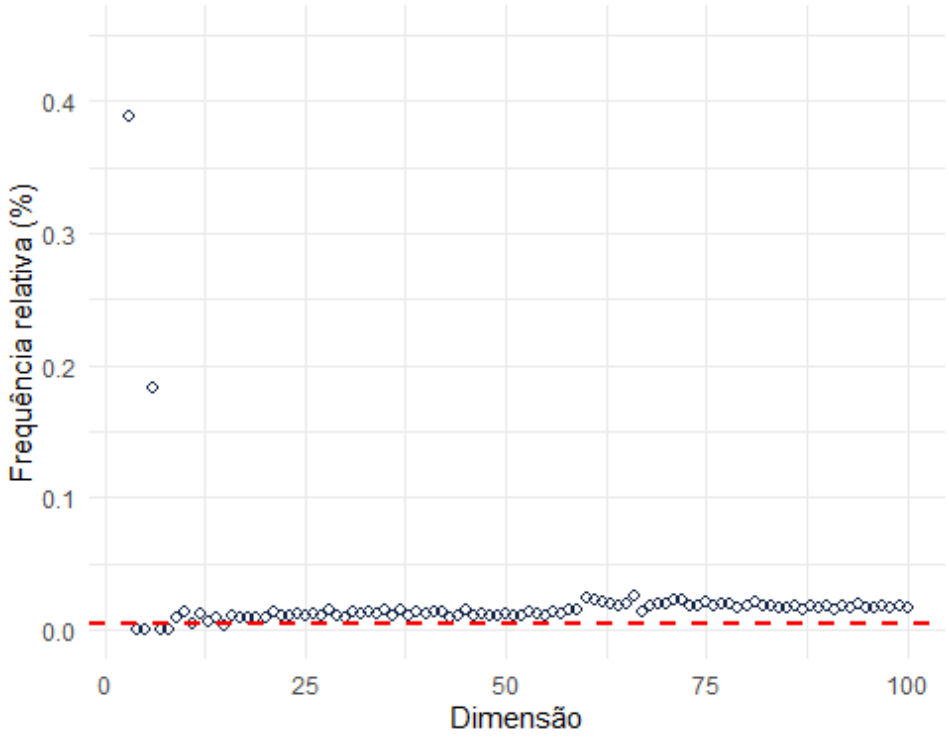


Figura 24 – Frequências relativas observadas na faixa de correlação contaminada para $\Gamma = (0{,}005; 0{,}5; 0{,}495)$.

escape de aproximadamente 5,8%, tem-se um percentual de 2,8% abaixo do nível de tolerância de $0,03$. Para a faixa de correlações $(-0{,}7; -0{,}2) \cup (0{,}2; 0{,}7)$ o percentual de escape foi de $15,34\%$. Finalmente, a última faixa $(-1; -0{,}7) \cup (0{,}7; 1)$ compreende um percentual de escape de $6,5\%$.

É válido ressaltar que o fator dimensional da matriz está intrinsecamente ligado à proporção desejada sob cada faixa e ao resultado do método. Sob uma proporção

Tabela 5 – Proporção média para todas as matrizes completamente customizadas.

Faixa de correlação	Proporções iniciais Γ		
	$(0{,}5; 0{,}495; 0{,}005)$	$(0{,}5; 0{,}005; 0{,}495)$	$(0{,}005; 0{,}5; 0{,}495)$
$(-0{,}2; 0{,}2)$	0,5050643	0,4419262	0,0168656
$(-0{,}7; -0{,}2) \cup (0{,}2; 0{,}7)$	0,4914743	0,1584262	0,4962580
$(-1; -0{,}7) \cup (0{,}7; 1)$	0,0034614	0,3996476	0,4868764

baixa como $0,005$, matrizes de dimensão menor possuem uma aproximação menos eficiente, em que apenas um elemento faz grande diferença na proporção final.

Os percentuais de correlações verificados através das Figuras 18, 21 e 24 que retratam a situação de contaminação apresentam as funcionalidades da metodologia para confrontar situações de notória adversidade. Particularmente o vetor de proporções $\Gamma = (0{,}5; 0{,}005; 0{,}495)$, ressalta a condição mais extrema. Como mencionado anteriormente, acontece uma demanda por elevado volume de iterações do mecanismo de otimização. Ainda assim, os resultados não são plenamente satisfatórios. É razoável supor que seja necessário um maior número de iterações preestabelecidas através do parâmetro N.

Essa dificuldade pode também ser observada através de uma análise do tempo computacional consumido. A necessidade de muitas rodadas do procedimento de otimização amplia sobremaneira o tempo computacional. A Figura 25 ilustra o tempo mediano de processamento para o vetor de proporções $\Gamma = (0{,}5; 0{,}005; 0{,}495)$.

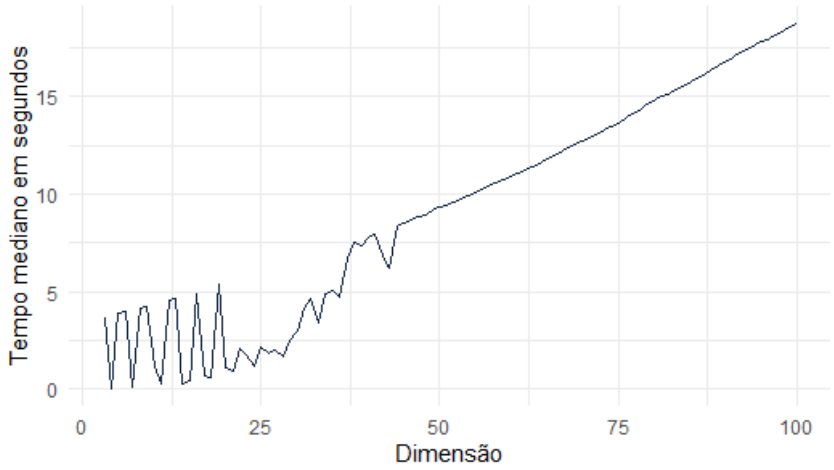
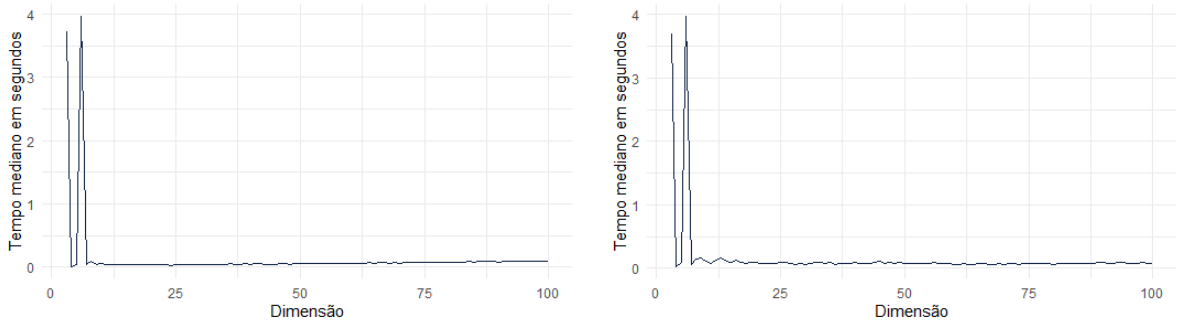


Figura 25 – Tempo computacional exigido para geração de correlações com $\Gamma = (0{,}5; 0{,}005; 0{,}495)$.

O tempo computacional apresentou um crescimento aparentemente linear com a dimensão. Não é proibitivo, porém não é o desejável nessa condição. Por outro lado, para as demais situações, a análise de tempo computacional revela o desempenho bastante promissor da técnica. De toda forma, vale ressaltar que mesmo para a dimensão 100, o

tempo mediano ficou abaixo de 20 segundos. A Figura 26 apresenta os resultados para as demais configurações do vetor Γ.



(a) $\Gamma = (0{,}5; 0{,}495; 0{,}005)$. (b) $\Gamma = (0{,}005; 0{,}5; 0{,}495)$.

Figura 26 – Tempo computacional exigido para geração de correlações com outros vetores Γ.

Para as duas outras configurações do vetor Γ, os tempos medianos ficam limitados pela fronteira dos 4 segundos. A medida que a dimensão cresce não ocorre aumento do tempo computacional, a maior flexibilidade para distribuir as correlações mantém os tempos em limiares bem baixos. O tempo mediano, a partir da dimensão 10, torna-se praticamente constante, e semelhante em ambos os vetores Γ preestabelecidos.

5 Considerações Finais

O objetivo principal do método *CM-generator* é ser capaz de gerar matrizes de correlação de quaisquer dimensões e com valores variados para as medidas de correlação obtidas. Essa metodologia é recente, inovativa e hábil para o propósito, porém carecia de testes mais profundos acerca de sua qualidade e adaptabilidade para condições extremas.

Diante disso, o desenvolvimento do presente estudo possibilitou compreender o comportamento do método *CM-generator* mediante cenários extremos de utilização, bem como a identificação de possíveis fragilidades sob esses cenários, dado o contexto inovador e de grande aplicabilidade da metodologia.

A execução dos testes aleatórios permitiu avaliar a estrutura da distribuição das correlações quando nenhum parâmetro é especificado, e identificar, por meio de comparação, fatores que unicamente interferem na performance inerente ao processo.

Os testes com o limite superior e inferior preestabelecidos permitiram uma avaliação com um nível acima de customização. Isso possibilitou uma análise sob condição apenas dessa mudança. Esse ponto de investigação foi eficiente ao garantir a boa adaptabilidade do método para uma condição específica de geração de matrizes de correlação.

Finalmente, a condução dos testes de contaminação, completamente customizados, permitiram uma avaliação do comportamento do método sob cenários mais extremos de aplicação, em que é possível estabelecer previamente faixas para as correlações, e estabelecer proporções em cada uma delas. A avaliação por meio do tempo computacional de simulação, permitiu uma comparação do desempenho do método a partir de cada nível de customização.

Sob esses resultados, o desempenho do método então é condicionado a vários fatores como os limites especificados para os desvios padrão da distribuição simulada, o número de repetições de otimização e a precisão da aproximação desejada. Desse modo, além de todo o contexto de customização, é possível que a aplicação seja ponderada pelo usuário por recursos como tempo, poder computacional a ser utilizado e precisão, descrevendo uma grande flexibilidade de utilização.

O transcorrer dessa investigação permitiu que algumas falhas de implementação na versão disponível do pacote `gencor` fossem identificadas e corrigidas. Essa é uma contribuição de suma importância e que transcende o nível usual de resultados esperados em um trabalho de conclusão de curso de graduação.

Além dos resultados alcançados, este estudo deixa aberto um conjunto de possíveis caminhos de continuidade. A condução de um volume ainda maior de execuções para os contextos de alta dimensionalidade. Essa abordagem pode fornecer mais informações de qualidade e tempo computacional. Além disso, experimentos com outras configurações dos intervalos de partição das correlações e novas configurações do vetor de proporções Γ podem elucidar mais itens acerca do experimento completamente customizado. A partir dos resultados aqui dispostos, é possível também analisar os limites inferior e superior para níveis ainda maiores ou menores de correlação. Para esse procedimento, é importante um ambiente computacional parecido com a descrição dos experimentos, ou superior.

Referências

[1] Duarte, Anderson Ribeiro, Helgem de Souza Ribeiro Martins e Fernando Luiz Pereira Oliveira: *CM-generator: a methodology for generating customized correlation matrices*. (submitted paper), páginas 1–22, 2022. Citado 5 vezes nas páginas 2, 3, 10, 11 e 13.

[2] Chalmers, Collin P: *Generation of correlation matrices with a given Eigen-Structure*. Journal of Statistical Computation and Simulation, 4(2):133–139, 1975. Citado na página 5.

[3] Bendel, Robert e M. Ray Mickey: *Population correlation matrices for sampling experiments*. Communications in Statistics - Simulation and Computation, 7(2):163–182, 1978. Citado na página 5.

[4] Davies, Philip I. e Nicholas J. Higham: *Numerically stable generation of correlation matrices and their factors*. BIT Numerical Mathematics, 40(4):640–651, 2000. Citado na página 5.

[5] Drmac, Zlatko: *Accurate computation of the product-induced singular value decomposition with applications*. SIAM journal on numerical analysis, 35(5):1969–1994, 1998. Citado na página 5.

[6] Hüttner, Amelie e Jan Frederik Mai: *Simulating realistic correlation matrices for financial applications: correlation matrices with the Perron–Frobenius property*. Journal of Statistical Computation and Simulation, 89(2):315–336, 2019. Citado na página 6.

[7] Böhm, Walter e Kurt Hornik: *Generating random correlation matrices by the simple rejection method: Why it does not work*. Statistics & Probability Letters, 87:27–30, 2014. Citado na página 6.

[8] Holmes, Robin B.: *On random correlation matrices*. SIAM Journal on Matrix Analysis and Applications, 12(2):239–272, 1991. Citado na página 6.

[9] Budden, Mark, Paul Hadavas e Lorrie Hoffman: *On the generation of correlation matrices*. Applied Mathematics E-Notes, 8:279–282, 2008. Citado na página 6.

[10] Marsaglia, George e Ingram Olkin: *Generating Correlation Matrices*. SIAM Journal on Scientific and Statistical Computing, 5(2):470–475, 1984. Citado na página 7.

[11] Hardin, Johanna, Stephan Ramon Garcia e David Golan: *A method for generating realistic correlation matrices*. The Annals of Applied Statistics, 7(3):1733–1762, 2013. Citado na página 7.

[12] Ghosh, Soumyadip e Shane G Henderson: *Behavior of the NORTA method for correlated random vector generation as the dimension increases*. ACM Transactions on Modeling and Computer Simulation (TOMACS), 13(3):276–294, 2003. Citado na página 8.

[13] Joe, Harry: *Generating random correlation matrices based on partial correlations*. Journal of Multivariate Analysis, 97(10):2177–2189, 2006. Citado na página 8.

[14] Lewandowski, Daniel, Dorota Kurowicka e Harry Joe: *Generating random correlation matrices based on vines and extended onion method*. Journal of Multivariate Analysis, 100:1989–2001, 2009. Citado na página 8.

[15] Bouchaud, Jean Philippe e Marc Potters: *Financial applications of random matrix theory: a short review*. arXiv preprint arXiv:0910.1205, 2009. Citado na página 9.

[16] Simonian, Joseph: *The most simple methodology to create a valid correlation matrix for risk management and option pricing purposes*. Applied Economics Letters, 17(18):1767–1768, 2010. Citado na página 9.

[17] Mittelbach, Martin, Bho Matthiesen e Eduard A Jorswieck: *Sampling uniformly from the set of positive definite matrices with trace constraint*. IEEE transactions on signal processing, 60(5):2167–2179, 2012. Citado na página 9.

[18] Pourahmadi, Mohsen e Xiao Wang: *Distribution of random correlation matrices: Hyperspherical parameterization of the Cholesky factor*. Statistics & Probability Letters, 106:5–12, 2015. Citado na página 10.

[19] Prôa, Miguel, Paul O'Higgins e Leandro R Monteiro: *Type I error rates for testing genetic drift with phenotypic covariance matrices: a simulation study*. Evolution: International Journal of Organic Evolution, 67(1):185–195, 2013. Citado na página 10.

[20] Hong, Sehee: *Generating correlation matrices with model error for simulation studies in factor analysis: A combination of the Tucker-Koopman-Linn model and Wijsman's algorithm*. Behavior Research Methods, Instruments, & Computers, 31(4):727–730, 1999. Citado na página 10.

[21] Rebonato, Riccardo e Peter Jäckel: *The most general methodology to create a valid correlation matrix for risk management and option pricing purposes*. Available at SSRN 1969689, 2011. Citado na página 10.

[22] R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. `https://www.R-project.org/`. Citado na página 21.

[23] Martins, Helgem Souza Ribeiro e Anderson Ribeiro Duarte: *gencor: Generate Customized Correlation Matrices*, 2022. `https://CRAN.R-project.org/package=gencor`, R package version 1.0.2. Citado 2 vezes nas páginas 21 e 53.

[24] Wickham, Hadley: *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016, ISBN 978-3-319-24277-4. `https://ggplot2.tidyverse.org`. Citado 2 vezes nas páginas 53 e 57.

[25] Wickham, Hadley, Jim Hester e Jennifer Bryan: *readr: Read Rectangular Text Data*, 2022. `https://CRAN.R-project.org/package=readr`, R package version 2.1.3. Citado na página 53.

[26] Auguie, Baptiste: *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. `https://CRAN.R-project.org/package=gridExtra`, R package version 2.3. Citado na página 53.

[27] Wickham, Hadley, Romain François, Lionel Henry e Kirill Müller: *dplyr: A Grammar of Data Manipulation*, 2022. `https://CRAN.R-project.org/package=dplyr`, R package version 1.0.10. Citado na página 57.

[28] Wickham, Hadley: *Reshaping Data with the reshape Package*. Journal of Statistical Software, 21(12):1–20, 2007. `http://www.jstatsoft.org/v21/i12/`. Citado na página 57.

Apêndices

APÊNDICE A – *Scripts* de execução dos experimentos

Durante a execução dos experimentos, foram utilizados também os pacotes `ggplot2` [24], `readr` [25] e `gridExtra` [26] além do pacote com a implementação da metodologia, o pacote `gencor` [23]. O *script* que segue foi utilizado para a execução dos experimentos aleatórios.

```
for (i in c(3:500)){
  times=c()
  data_graph=c()
  x_cut_values=c()
  for (j in 1:100){
    x =gencor(i)
    aux_matrix = upper.tri(x$Matrix, diag = F)
    x1=x$Matrix[aux_matrix]
    x_cut = cut(x1, breaks = c(seq(-1, 1, 0.05)),
                labels = c(seq(1, 40, 1)))
    x_cut_values= c(x_cut_values,(table(x_cut)))
    times = c(times, x$Runtime)
  }
  lab=c(seq(1, 40,1))
  data_values_cut= data.frame(x_cut_values, rep(lab, 100))
  write.table(times,
              file=paste("output_general_times_",
                                 i, ".txt", sep = ""))
  write.table(data_values_cut,
              file=paste("output_general_values_",
                                       i, ".txt", sep = ""))
  times=c()
  x_cut_values=c()

  cat("\014")
  print(paste("EXPERIMENTO 1 - ",
              "Dimensao ", i,
              " - Simulacao: ", j))
}
```

***Script* de execução A.1 –** Experimentos aleatórios

O *script* a seguir, foi utilizado na geração de matrizes com limite superior. A definição do limite inferior ou superior pode ser especificada pelo parâmetro `method`

contido na função `gencor`. A representação abaixo ilustra a geração de matrizes com limite superior $\ell_{upper} = 0{,}01$.

```
for (i in c(3:500)){
  times=c()
  data_graph=c()
  x_cut_values=c()
  for (j in 1:100){
    x =gencor(i, method = "low", lim_low = 0.01)
    aux_matrix = upper.tri(x$Matrix, diag = F)
    x1=x$Matrix[aux_matrix]
    x_cut = cut(x1, breaks = c(seq(-1, 1, 0.05)),
                labels = c(seq(1, 40, 1)))
    x_cut_values= c(x_cut_values,(table(x_cut)))
    times = c(times, x$Runtime)
  }
  lab=c(seq(1, 40,1))
  data_values_cut= data.frame(x_cut_values, rep(lab, 100))
  write.table(times, file=paste("output_general_times_",
                                i, ".txt", sep = ""))
  write.table(data_values_cut, file=paste("output_general_values_",
                                          i, ".txt", sep = ""))
  times=c()
  x_cut_values=c()
  cat("\014")
  print(paste("EXPERIMENTO 3 (Parte 1) - ",
              "Dimensao ", i,
              " - Simulacao: ", j))
}
```

***Script* de execução A.2 –** Experimentos com limite inferior e superior

O *script* que segue, representa as frequências relativas das correlações geradas em cada matriz e o tempo de execução da dimensão i e simulação j, para a execução das simulações para o experimento completamente customizado. As faixas utilizadas são: $(-0{,}2; 0{,}2)$, $(-0{,}7; -0{,}2) \cup (0{,}2; 0{,}7)$ e $(-1; -0{,}7) \cup (0{,}7; 1)$, e foram inseridas no parâmetro `custom lim`. As proporções iniciais $\Gamma = (0{,}5; 0{,}495; 0{,}005)$ foram inseridas no parâmetro `custom prop`, contidos na função `gencor`.

```
for (i in c(3:100)){
  times=c()
  data_graph=c()
  x_cut_values=c()
  for (j in 1:30){
    x =gencor(i, method = "custom",
              custom_lim = c(0.2, 0.7),
              custom_prop = c(0.5, 0.495,0.005))
    aux_matrix = upper.tri(x$Matrix, diag = F)
    x1=x$Matrix[aux_matrix]
    x_cut = cut(x1, breaks = c(seq(-1, 1, 0.05)),
                labels = c(seq(1, 40, 1)))
    x_cut_values= c(x_cut_values,(table(x_cut)))
    times = c(times, x$Runtime)
  }
  lab=c(seq(1, 40,1))
  data_values_cut= data.frame(x_cut_values,
                              rep(lab, 30))

  write.table(times,
              file=paste("output_general_times_",
                                 i, ".txt", sep = ""))
  write.table(data_values_cut,
              file=paste("output_general_values_",
                         i, ".txt", sep = ""))
  times=c()
  x_cut_values=c()

  cat("\014")
  print(paste("EXPERIMENTO 7 (Parte 1) - ",
              "Dimensao ", i, " - Simulacao: ", j))
}
```

***Script* de execução A.3 –** Experimentos completamente customizados

APÊNDICE B – *Script* R para os gráficos gerados

Os gráficos foram gerados a partir dos pacotes `ggplot2` [24], `dplyr` [27] e `reshape2` [28]. O *script* R abaixo contém a configuração utilizada para visualização dos gráficos a partir dos arquivos de simulação.

```
library(gencor)
library(ggplot2)
library(readr)
library(gridExtra)
library(reshape2)
library(dplyr)

#Lendo e sumarizando os tempos gerados
times_x = c()
media = c()
percentil_95 = c()
maximo = c()
minimo = c()
mediana = c()

#Para matrizes de 3 a 500
for (i in c(3:500)){
  x2 = read.table(paste(paste("output_general_times_",
                              i, ".txt", sep = "")))
  times_x=rbind(times_x,x2)
  media = c(media, mean(x2$x))
  percentil_95=c(percentil_95, quantile(x2$x, .95))
  maximo=c(maximo, max(x2$x))
  minimo=c(minimo, min(x2$x))
  mediana = c(mediana, median(x2$x))
}

data_line_times = data.frame(media, percentil_95,
                             minimo, maximo, mediana)

#Gerando o grafico para os tempos
ggplot(data_line_times) +
  aes(x = c(3:500), y = `mediana`) +
  geom_line(colour = "#112446") +
  labs(
    x="Dimensao",
```

```
37      y = "Tempo mediano em segundos"
38    ) +
39    theme_minimal()
40
41  #Para matrizes de dimensoes de 3 a 500
42  for (i in c(3:500)){
43    x2 = read.table(paste(paste("output_general_times_",
44                                i, ".txt", sep = "")))
45    times_x=rbind(times_x,x2)
46    media = c(media, mean(x2$x))
47    percentil_95=c(percentil_95, quantile(x2$x, .95))
48    maximo=c(maximo, max(x2$x))
49    minimo=c(minimo, min(x2$x))
50    mediana = c(mediana, median(x2$x))
51  }
52  data_line_times = data.frame(media, percentil_95,
53                               minimo, maximo, mediana)
54
55  #Gerando o grafico para os tempos
56  ggplot(data_line_times) +
57    aes(x = c(3:500), y = `mediana`) +
58    geom_line(colour = "#112446") +
59    labs(
60      x="Dimensao",
61      y = "Tempo mediano em segundos"
62    ) +
63    theme_minimal()
64
65  #Para matrizes de dimensoes de 1000 a 10000
66  for (i in seq(1000, 10000, 1000)){
67    x2 = read.table(paste(paste("output_general_times_",
68                                i, ".txt", sep = "")))
69    times_x=rbind(times_x,x2)
70    media = c(media, mean(x2$x))
71    percentil_95=c(percentil_95, quantile(x2$x, .95))
72    maximo=c(maximo, max(x2$x))
73    minimo=c(minimo, min(x2$x))
74    mediana = c(mediana, median(x2$x))
75  }
76
77  data_line_times = data.frame(media, percentil_95,
78                               minimo, maximo, mediana)
79
80  #Gerando o grafico para o tempo mediano
81  ggplot(data_line_times) +
82    aes(x = seq(1000, 10000, 1000), y = `mediana`) +
83    geom_line(colour = "#112446") +
```

```
84    labs(
85      x="Dimensao",
86      y = "Tempo mediano em segundos"
87    ) +
88    theme_minimal()
89
90  times_dimension = c()
91  for (j in c(3:500)){
92    times = c(rep(j, 100))
93    times_dimension = c(times_dimension, times)
94  }
95
96  times_x$Dimensao = times_dimension
97
98
99  #Gerando o grafico para os tempos
100 ggplot(times_x) +
101   aes(x = times_dimension, y = x) +
102   geom_line(shape = "circle", size = 1.5, colour = "#112446") +
103   labs(
104     y = "Tempo (em segundos)",
105     title = "Tempos de simulacao por dimensao - Experimento 7",
106     subtitle = "Contaminacao 3"
107   ) +
108   theme_minimal()
109
110 #Gerando o boxplot(Usado no Experimento aleatorio)
111 data_box = data.frame(minimo, media, percentil_95, maximo)
112 colnames(data_box)=c("Minimo", "Media", "Percentil_95%", "Maximo")
113 data_boxplot = melt(data_box)
114 data_boxplot %>% ggplot() +
115   geom_boxplot(aes(x=variable, y=value))+
116   labs(x="Medida descritiva",  y="Tempo em segundos ")+
117   theme_minimal()
118
119 #Gerando o histograma do experimento
120 x1= c()
121 for (i in c(3:500)){
122   x2 = read.table(paste(paste("output_general_values_",
123                               i, ".txt", sep = "")))
124   x2=aggregate(x2$x_cut_values~x2$rep.lab..100.,
125                data = x2, sum)
126   colnames(x2)=c("Interval", "x_cut_values")
127   x1=rbind(x1,x2)
128 }
129
130
```

```
#transformando os dados e variavel dos intervalos para fator
data_hist= aggregate(x1$x_cut_values~x1$Interval,
                     data = x1, sum)
data_hist$`x1$Interval`=as.factor(data_hist$`x1$Interval`)
data_hist$Intervalos = c(seq(-0.975, 1, 0.05))
colnames(data_hist) = c("Labels", "x_cut", "Intervalos")
data_hist$Proporcoes = prop.table(data_hist$x_cut)

#Criando o grafico para correlacoes obtidas <0.01
ggplot(data_hist) +
  aes(x = (data_hist$Intervalos), y = data_hist$Proporcoes) +
  geom_col(fill = "#112446") +
  geom_vline(xintercept = c(-1, -0.7, -0.2, 0.2, 0.7, 1),
             linetype="dotted")+
  labs(x = "Correlacao", y = "Frequencia relativa ",
       subtitle = "") +
  ggtitle("")+
  scale_x_continuous(breaks = round(seq(min(data_hist$Intervalos),
                                        max(data_hist$Intervalos),
                                        by = 0.1),1)) +
  theme_minimal()

#Gerando o grafico de frequencias relativas por dimensao
contador = 0
falha_por_dimensao = c()
sum(data_hist$x_cut)


#Procedimento do experimento 7 - proporcao 0,7 a 1
for (i in 3:100){
  percentual = (sum(x1$x_cut_values[(1+contador*40):(6+contador*40)])+
                sum(x1$x_cut_values[(35+40*contador):(40+40*contador)]))/
                sum(x1$x_cut_values[(1+contador*40):(40+contador*40)])
  marca_dimensao = i
  falha_por_dimensao=c(falha_por_dimensao, percentual)
  contador = contador+1
}

#Procedimento do experimento 7 - proporcao 0,2 a 0,7
for (i in 3:100){
  percentual = (sum(x1$x_cut_values[(7+contador*40):(16+contador*40)])+
                sum(x1$x_cut_values[(25+40*contador):(34+40*contador)]))/
                sum(x1$x_cut_values[(1+contador*40):(40+contador*40)])
  marca_dimensao = i
  falha_por_dimensao=c(falha_por_dimensao, percentual)
  contador = contador+1
}
```

```
178
179 #Procedimento do experimento 7 - proporcao 0,2
180 for (i in 3:500){
181   percentual = ((sum(x1$x_cut_values[(7+40*contador):(34+40*contador)]))/
182                   sum(x1$x_cut_values[(1+contador*40):(40+contador*40)]))*
                        100
183   marca_dimensao = i
184   falha_por_dimensao=c(falha_por_dimensao , percentual)
185   contador = contador+1
186 }
187
188 Dimensao = seq(3, 500, 1)
189 dados_falha_por_dimensao = data.frame(Dimensao, falha_por_dimensao )
190 dados_falha_por_dimensao$Percentual = dados_falha_por_dimensao$falha_por_
      dimensao
191 #Gerando o grafico de frequencias relativas por dimensao
192 ggplot(dados_falha_por_dimensao) +
193   aes(x = Dimensao, y = Percentual) +
194   geom_point(
195     shape = "circle filled",
196     size = 1.95,
197     colour = "#112446"
198   ) +
199   labs(y = "Frequencia relativa (%)") +
200   ylim(c(0, 0.45))+
201   theme_minimal()
```

***Script* de execução B.1 –** Geração de gráficos