JOSINO JOSÉ BARBOSA

DATA-DRIVEN CLUSTER ANALYSIS METHOD: UMA NOVA METODOLOGIA PARA DETECÇÃO DE *OUTLIERS* EM DADOS MULTIVARIADOS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

Orientador: Fernando Luiz Pereira de Oliveira

Coorientador: Anderson Ribeiro Duarte

VIÇOSA - MINAS GERAIS
2021

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
B238d
2021

Barbosa, Josino José, 1985-
Data-driven Cluster Analysis Method : uma nova metodologia para detecção de outliers em dados multivariados / Josino José Barbosa. – Viçosa, MG, 2021.
77 f. : il. ; 29 cm.

Orientador: Fernando Luiz Pereira de Oliveira.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Outlier (Estatística). 2. Análise multivariada. 3. Análise de agrupamentos. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.5

JOSINO JOSÉ BARBOSA

DATA-DRIVEN CLUSTER ANALYSIS METHOD: UMA NOVA METODOLOGIA PARA DETECÇÃO DE *OUTLIERS* EM DADOS MULTIVARIADOS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

APROVADA: 20 de abril de 2021.

Assentimento:

Josino José Barbosa
Autor

Fernando Luiz Pereira de Oliveira
Orientador

Dedico este trabalho à minha esposa Ana Regina e à minha filha Beatriz.

Agradecimentos

Agradeço aos meus pais, Geralda e José, pelo amor, carinho, dedicação e por sempre me incentivar nos estudos, mesmo diante de todas as dificuldades.

Aos meus irmãos, Josãne e Josias, pelo apoio, incentivo e por tudo que fizeram por mim.

À minha esposa, Ana Regina, pelo amor, paciência e apoio nos momentos difíceis.

À minha filha, Beatriz, pelo carinho e por me proporcionar tantas alegrias.

Aos professores do Programa de Pós-Graduação em Estatística Aplicada e Biometria da UFV, pelos ensinamentos que contribuíram para a minha formação acadêmica.

Ao secretário do Programa de Pós-Graduação em Estatística Aplicada e Biometria, Júnior Pires, pelo empenho e dedicação em nos atender.

Aos professores Anderson Ribeiro Duarte e Fernando Luiz Pereira de Oliveira, pela confiança depositada durante o desenvolvimento do doutorado. Obrigado pela orientação, dedicação, incentivo e por todo o aprendizado.

Aos professores José Ivo Ribeiro Júnior, Marcelo Ângelo Cirillo e Rivert Paulo Braga Oliveira, por terem aceitado o convite para participar da banca e pelas contribuições apresentadas.

Aos colegas de pós-graduação pela excelente convivência e conhecimento compartilhado, em especial aos amigos André, Mateus, Marcelo e Maurício.

Ao amigo Helgem pelo apoio desde a graduação, pela companhia nas viagens para Viçosa e pela parceria neste doutorado.

À FAPEMIG, CNPq, CAPES, UFV e UFOP, pelo apoio financeiro para o desenvolvimento deste trabalho.

Agradeço especialmente a Deus por ter me proporcionado alcançar mais essa conquista e por ter colocado todas estas pessoas no meu caminho.

O sucesso é uma consequência e não um objetivo.

Gustave Flaubert

Resumo

BARBOSA, Josino José, D.Sc., Universidade Federal de Viçosa, abril de 2021. ***Data-driven Cluster Analysis Method*: uma nova metodologia para detecção de *outliers* em dados multivariados**. Orientador: Fernando Luiz Pereira de Oliveira. Coorientador: Anderson Ribeiro Duarte.

Metodologias para identificação de *outliers* multivariados são de grande importância em análise estatística. Observações aberrantes podem revelar informações relevantes para variáveis sob investigação. Aplicações estatísticas sem uma prévia identificação de possíveis valores extremos podem apresentar resultados controversos e induzir decisões equivocadas. Além disso, em diversos contextos, os *outliers* são pontos de grande interesse prático e sua identificação torna-se o principal objetivo. Diante disso, esse estudo tem por objetivo propor uma nova técnica de detecção de *outliers* multivariados baseada em análise de agrupamentos. A técnica considera informações inerentes ao próprio banco de dados e também informações de conhecimento prévio do pesquisador acerca das populações sob investigação. A avaliação da metodologia foi conduzida através de calibração e comparação com três métodos de detecção já difundidos por meio de dados simulados. A investigação comparativa considera duas técnicas de detecção baseadas na clássica distância de Mahalanobis e uma técnica também baseada em análise de agrupamentos. As medidas de sensibilidade, especificidade e acurácia são utilizadas para aferir a qualidade dos métodos, assim como uma análise quanto ao tempo computacional necessário para a execução dos procedimentos. Além disso, os métodos foram empregados num conjunto de dados reais. A nova técnica proposta revelou uma notória superioridade em relação às demais, tanto na qualidade de detecção de *outliers* através dos dados simulados, quanto na adequabilidade na aplicação do conjunto de dados reais.

Palavras-chave: *Outliers* multivariados. Simulação. Análise de agrupamentos. DDCAM.

Abstract

BARBOSA, Josino José, D.Sc., Universidade Federal de Viçosa, April, 2021. **Data-driven Cluster Analysis Method: a new methodology for detection outliers in multivariate data**. Advisor: Fernando Luiz Pereira de Oliveira. Co-advisor: Anderson Ribeiro Duarte.

Methodologies for identifying multivariate outliers are of great importance in statistical analysis. Aberrant observations can reveal relevant information for variables under investigation. Statistical applications without prior identification of possible extreme values can present controversial results and induce wrong decisions. In addition, in several contexts, outliers are points of great practical interest and their identification becomes the main objective. Therefore, this study aims to propose a new technique for detecting multivariate outliers based on cluster analysis. The technique considers information inherent to the database itself and also information of the researcher's prior knowledge about the populations under investigation. The evaluation of the methodology was carried out through calibration and comparison with three detection methods already disseminated through simulated data. The comparative investigation considers two detection techniques based on the classic Mahalanobis distance and one technique also based on cluster analysis. Sensitivity, specificity and accuracy measures are used to assess the quality of the methods, as well as an analysis of the computational time required to perform the procedures. In addition, the methods were used on a real data set. The proposed new technique revealed a notorious superiority in relation to the others, both in the quality of detecting outliers through the simulated data, and in the suitability in the application of the real data set.

Keywords: Multivariate Outliers. Simulation. Cluster analysis. DDCAM.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A análise estatística de conjuntos de dados, sejam esses conjuntos de pequeno ou grande volume de dados, sempre requer um estudo descritivo cuidadoso e cauteloso. O referido cuidado deve ser a prévia para a utilização de quaisquer técnicas estatísticas mais sofisticadas. Em particular, a presença de observações aberrantes, atípicas, usualmente mencionadas como *outliers* pode deteriorar sobremaneira a investigação estatística.

A avaliação sobre a presença de *outliers* em dados univariados não é completamente trivial do ponto de vista teórico. Essa investigação se torna ainda mais sofisticada quando na busca por *outliers* multivaridados, ou seja, *outliers* presentes em conjuntos de dados com duas ou mais variáveis analisadas simultaneamente.

Estudos sobre identificação de *outliers* podem beneficiar diversas áreas de atuação e, consequentemente, gerar uma gama de aplicações em situações práticas. Alguns exemplos de aplicações específicas em problemas de detecção de *outliers* são apresentados:

Exemplo 1.1 *Garantir a qualidade do produto, bem como controlar o desempenho do sistema, são tarefas importantíssimas em processos industriais. Com os avanços tecnológicos e o aumento da capacidade de produção, o controle de desempenho do sistema e a qualidade de produção está cada vez mais complicado. Dessa forma, o uso de ferramentas baseadas nos próprios dados do processo torna-se uma solução atraente para o controle e monitoramento do processo. Em situações práticas, técnicas de detecção de outliers podem ser utilizadas para controlar e monitorar processos industriais, identificar possíveis falhas de produção ou desempenho, bem como auxiliar na calibração de ferramentas para este fim.*

Exemplo 1.2 *A deteção de outliers tem sido bastante utilizada em aplicações em redes de sensores. Segundo De et al. (2013), redes de sensores consistem em dispositivos distribuídos espacialmente que se comunicam por meio de rádio sem fio e detectam cooperativamente as condições físicas ou ambientais. Eles fornecem um alto grau de visibilidade dos processos físicos ambientais. As redes de sensores também são muito úteis em cenários catastróficos ou de emergência, como inundações, incêndios, vulcões, campos de batalha, onde a participação humana é muito perigosa e as redes de infraestrutura são impossíveis ou muito caras. Portanto, padrões anormais em bancos de dados oriundos destas redes podem ser classificados como possíveis outliers e sua identificação pode ser muito útil para a tomada de decisões.*

Exemplo 1.3 *Diversos problemas ambientais têm origem na utilização dos solos. A má utilização deste importante recurso natural pode provocar alterações climáticas, perda de biodiversidade, poluição das águas, dos solos e do ar. A contaminação dos solos traz riscos diretos e indiretos para a saúde dos seres vivos, sejam eles devido ao uso para habitação, lazer, agricultura, dentre outros. Uma forma viável e prática de se fazer um levantamento geral da qualidade dos solos é através da coleta e medição geoquímica de amostras. É razoável supor que terras potencialmente contaminadas se comportam como outliers em um banco de dados geoquímico, pois normalmente apresentam características deferentes de solos normais. Esses locais precisam ser identificados e investigados para uma possível correção, a fim de minimizar seus riscos potenciais.*

A relevância de procedimento de detecção de *outliers* é uma motivação abrangente para estabelecer objeto de estudo. Os diversos métodos já apresentados, em geral, são validados através de técnicas de simulação. Dados são simulados com ou sem a presença de valores aberrantes gerados através do mecanismo de simulação e o procedimento de detecção em estudo é utilizado. Na maioria dos estudos, os dados simulados para testes são concebidos através da geração de populações normais multivariadas contaminadas, por meio da mistura de distribuições normais multivariadas através de simulações de Monte Carlo.

1.1 Justificativa e Motivação

A justicativa para este estudo está na possibilidade de melhorias do método de identificação de *outliers* multivariados proposto por Barbosa et al. (2018). Nesta abordagem, Barbosa et al. (2018) utilizou o método de análise de agrupamento k-médias com o objetivo de agrupar os indivíduos semelhantes. Para concluir se um determinado grupo de indivíduos é um grupo de *outliers* multivariados ou não, utilizou como critério, uma medida baseada no desvio-padrão amostral (s) das distâncias entre os centroides dos grupos e a mediana dos dados. Dessa forma, se a distância euclidiana entre o centroide de um grupo e a mediana dos dados for superior a $2,5 \times s$, os indivíduos deste grupo são classificados como *outliers*.

A principal deficiência apresentada pelo método de Barbosa et al. (2018) está no fato do número de grupos do agrupamento k-médias ter uma escolha completamente *ad-hoc* de $k = n/10$. É intuitivo supor que tal escolha tende a se comportar bem em alguns cenários, mas pode ser completamente deficitária em outros. A discussão deste entrave gera a motivação central para este estudo, verificar o efeito de diferentes escolhas do valor k. O estudo apresentado por Barbosa et al. (2020) buscou elucidar este efeito. Neste estudo, foram realizados experimentos com a metodologia original que adota $k = n/10$ e uma segunda metodologia que escolhe o valor k que maximiza a acurácia média em uma sequência de procedimetos de detecção de *outliers* em dados simulados. Vale destacar que essa metodologia não tem sentido prático para dados reais, uma vez que não há como escolher previamente o valor k que maximizará a acurácia no processo de

detecção dos *outliers*. Logo, inicialmente, pretendeu-se mostrar que é possível melhorar o desempenho do método proposto por Barbosa et al. (2018), para posteriormente empregar técnicas que visem encontrar escolhas mais adequadas para o valor k. No capítulo 4, serão apresentados resultados experimentais que indicam que o método original pode, de fato, passar por estratégias de melhorias para se tornar mais adaptativo aos conjuntos de dados sob investigação. Uma avaliação experimental efevetiva dos possiveis efeitos será discutida.

1.2 Objetivos

O objeto principal deste estudo é propor uma nova técnica de detecção de *outliers* multivariados que apresente melhorias significativas em relação ao método proposto por Barbosa et al. (2018), através de estratégias de otimização e aprendizado através dos próprios dados. Os objetivos secundários são:

i. estabelecer um simulador de dados da distribuição normal multivariada contaminada baseado em matrizes de correlações mais realístas;

ii. determinar medidas adequadas para avaliar a qualidade de técnicas de detecção de *outliers* multivariados;

iii. comparar, de acordo com as medidas determinadas, a nova metodologia aqui proposta com três métodos de detecção relevantes: o método proposto por Barbosa et al. (2018) e dois métodos baseados em distância de Mahalanobis, via estimador robusto MVE (*Minimum Volume Ellipsoid*) e via estimador robusto MCD (*Minimum covariance determinant*);

iv. apresenatar uma aplicação da medotologia proposta em um conjunto de dados reias.

1.3 Organização da Tese

Esta tese encontra-se organizada a partir da seguinte estrutura:

- capítulo 1: seção introdutória que delimita o problema de interesse e motiva o estudo;
- capítulo 2: é apresentada uma revisão do estado da arte em relação ao problema de detecção de *outliers* em dados multivariados;
- capítulo 3: detalha o método proposto por Barbosa et al. (2018) e apresenta toda a discussão metodológica de aprimoramento e concepção do novo método aqui proposto;

- capítulo 4: são apresentados os mecanismos utilizados na geração dos dados simulados, as medidas utilizadas para aferir a qualidade dos procedimentos de detecção dos *utliers* multivariados, os resultados e discussões acerca das simulações executadas, uma comparação entre os métodos em estudo e os reultados da aplicação no conjunto de dados reias;

- capítulo 5: são apresentadas conclusões sobre a metodologia proposta quanto a sua eficiência e também possíveis limitações. Além disso, discutem-se propostas de continuidade de pesquisa e demais aspectos relevantes.

Referências Bibliográficas

Barbosa, J. J., Duarte, A. R. e Martins, H. S. R. (2020). A performance evaluation in multivariate outliers identification methods. Ciência & Natura, 42:1–14.

Barbosa, J. J., Pereira, T. M. e Oliveira, F. L. P. (2018). Uma proposta para identificação de outliers multivariados. Ciência & Natura, 40:1–8.

De, D., Song, W.-Z., Xu, M., Shi, L. e Tan, S. (2013). Advances in real-world sensor network system. Em Advances in Computers, volume 90, capítulo: 1, páginas 1–90. Elsevier.

Capítulo 2

Revisão Bibliográfica

2.1 Valores *outliers*

De acordo com Hawkins (1980), uma observação ou subgrupo de informações aparentemente inconsistentes, se comparadas ao restante de algum conjunto de dados, são definidos como *outliers*. Na concepção de Barnett e Lewis (1994) observações que suscitam suspeitas por se encontrarem muito desviadas em relação ao restante do conjunto de dados podem ser provenientes de geração através de algum mecanismo distinto da natureza usual do dados.

O conceito de *outlier* multivariado remete à distância em subespaços $k-$dimensionais definidos pelas variáveis envolvidas na investigação. Segundo Jolliffe e Cadima (2016) alguma observação pode configurar um *outlier* multivariado mesmo sem configurar um *outlier* univaridado quando analisadas cada uma das variáveis individualmente.

Um grande volume de estudos de diversas aplicações utilizam metodologias para a detecção de *outliers* multivariados. Em sua maioria, as aplicações que são apresentadas partem de dados criados em processos produtivos. Quando esses processos, de forma anômala, apresentam comportamentos incomuns, fora de uma previsibilidade usual, valores *outliers* são gerados, como mencionado por Aggarwal (2017). Essa constatação ilustra uma relevante importância de análise de valores *outliers* multivariados. Informções de grande relevância sobre características raras, porém existentes, acerca dos dados sob investigação estão presentes e impactam os processos de geração de dados.

Estabelecer estratégias para detecção dessas características incomuns leva a uma extensa gama de aplicações de interesse prático. Aggarwal (2017) ilustra diversas aplicações dessa natureza como: verificações de fraudes do sistema financeiro e operações de crédito; procedimentos para diagnósticos médicos; sensores de detecção de ocorrências de eventos; entre muitos outros.

2.2 Algumas técnicas para identificação de *outliers* em dados multivariados

Diversos estudos com proposição de metodologia para detecção de *outliers* multivariados têm sido desenvolvidos, alguns deles baseados na identificação de valores aberrantes através da distância de Mahalanobis. Rousseeuw e Zomeren (1990) apresentam um método baseado no estimador robusto MVE (*Minimum volume ellipsoid*). O estimador do MVE pode ser obtido com a utilização do elipsoide de menor volume capaz de cobrir pelo menos k pontos do conjunto amostral, em que $n/2 < k < n$.

Filzmoser (2005) e Filzmoser et al. (2005) apresentam um método baseado no estimador robusto MCD (*Minimum covariance determinant*). O estimador MCD é determinado por um subconjunto de tamanho h, em que $n/2 < h < n$, que minimiza o determinante da matriz de covariâncias amostral, calculado apenas sob os h pontos. A estimativa de locação é a média destes pontos, enquanto a estimativa de dispersão é proporcional à sua matriz de covariância, em que a escolha do tamanho de h determina a robustez do estimador.

Atkinson e Riani (2002) apresentam um método iterativo baseado em análises gráficas através do método *Forward Search*. Este método é utilizado para a construção de modelos robustos, que incorpora um corte flexível baseado nos dados, para a detecção de *outliers* multivariados e estruturas suspeitas. Ao começar com pequenos subconjuntos de dados, as observações que estão próximas do modelo ajustado são adicionadas às observações usadas na estimativa dos parâmetros. À medida que esse subconjunto cresce, estimativas de parâmetros, estatísticas de teste e medidas de ajuste, são monitoradas. Mais detalhes podem ser observados em Atkinson e Riani (2004); Atkinson et al. (2010).

Filzmoser et al. (2008) propõem um método hábil para identificação de *outliers* multivariados em conjuntos de dados de alta dimensão. Trata-se de uma estratégia de implementação simples baseada em um procedimento de re-escalonagem de dados por meio da mediana (MED) e do desvio absoluto da mediana (MAD). Este algoritmo utiliza propriedades simples de componentes principais para identificar *outliers* multivariados e é capaz de analisar a situação dos dados comumente encontrados em certas aplicações biológicas nas quais o número de dimensões é várias ordens de magnitude maior do que o número de observações.

Berton et al. (2010) apresentam um método baseado em redes complexas. A técnica utiliza uma medida de distância de caminhada aleatória acoplada a um índice de dissimilaridade entre pares de vértices para identificação de *outliers* multivariados. O método determina uma visão de toda a rede para cada nó e infere que *outliers* multivariados são aqueles nós cujas visualizações diferem significativamente da maioria dos nós. Normalmente, o *outlier* multivariado é detectado através da aplicação de um critério específico, por exemplo, os mais distantes do nó central. Consequentemente, apenas um tipo de *outliers* multivariados que satisfazem os critérios predefinidos pode ser determinado. Por outro lado, o método in-

corpora informações locais e globais da rede devido ao recurso de passeio aleatório e pode fornecer resultados de detecção de *outliers* multivariados mais gerais.

O estudo de Valadares et al. (2012) apresenta uma análise através de detecção de *outliers* para dados multivariados de rede de sensores sem fio. Através de simulação, são considerados os cenários específicos de uma rede de sensores, foi realizada uma comparação entre três métodos gerais para a identificação dos *outliers* multivariados: *Minimum Volume Ellipsoid* (MVE), *Minimum Covariance Determinant* (MCD) e *Max-Eigen Difference* (MED). Ao final, os métodos MVE e MCD foram considerados mais eficientes que o método MED nas aplicações consideradas, principalmente para grandes volumes de dados.

Veloso e Cirillo (2016) apresentam uma metodologia para identificação de *outliers* multivariados baseada em componentes principais com amostras corrigidas por distância do tipo qui-quadrado de Pearson e Yates. Através de simulação de Monte Carlo, foi construído um teste de significância para indicar os componentes principais que melhor discriminam as discrepâncias. Ao final do estudo, conclui-se que o mais adequado é a seleção dos componentes principais pelo teste de significância por meio da distância qui-quadrado de Pearson.

Outro estudo baseado em componentes principais é apresentado por Filzmoser et al. (2009). Nesta abordagem, foram utilizados dados de composição com *outliers* multivariados, que necessitam de uma transformação especial antes de se aplicar ferramentas de análise de dados multivariados. Neste sentido, foi utilizado pelos autores a transformação logratio isométrica antes de se aplicar a análise de componente principal (PCA) robusta.

Um procedimento para identificar *outliers* multivariados por meio da distribuição cumulativa de extremos em um modelo de resposta Gama foi proposto por Resende et al. (2017). Neste trabalho, para validar o método, os autores utilizaram cenários de simulação definidos pela combinação de diferentes amostras, taxa de contaminação e distribuições com diferentes graus de assimetria. Além disso, probabilidades relacionadas a erros de classificação e precisão foram obtidas por meio de simulações de Monte Carlo.

Kutsuna e Yamamoto (2017) propõem um método para detecção de valores extremos por meio de diagramas de decisão binária e o método *leave-one-out*. A densidade *leave-one-out* é proposta como uma nova medida para detectar *outliers* multivariados, que é definida como uma razão entre o número de elementos de dados dentro de uma região e o volume da região, depois que uma observação específica é removida. Kutsuna e Yamamoto (2017) mostraram que a densidade *leave-one-out* pode ser avaliada de forma muito eficiente em um conjunto de regiões ao redor de cada observação em um determinado conjunto de dados através de diagramas de decisão binários. A complexidade de tempo deste método é quase linear em relação ao tamanho do conjunto de dados, enquanto que a precisão de detecção de *outliers* multivariados ainda é comparável à de outros métodos.

Zhu et al. (2017) propõem um método para detectar trajetórias anômalas, em dispositivos equipados com GPS, com a ajuda do conjunto de

dados da trajetória histórica e das rotas populares. As anormalidades espaciais e temporais são levadas em consideração simultaneamente para melhorar a precisão da detecção. Este método utiliza um algoritmo de detecção de *outlier* multivariado em tempo real, que contém uma etapa de pré-processamento *off-line* e outra etapa de detecção *on-line*. Na etapa de pré-processamento *off-line*, o índice de transferência dependente do tempo (TTI) e o gráfico de transferência dependente do tempo (TTG) são construídos de acordo com o conjunto de dados de trajetória histórica. Em seguida, na etapa de detecção *on-line*, o TTI e o TTG são usados para acelerar o progresso da detecção.

Luo et al. (2018) introduzem um método de triagem de *outliers* multivariados baseado em variogramas para vetores de imagens médicas. O variograma é uma ferramenta geoestatística poderosa para caracterizar a dependência espacial de processos estocásticos. Uma vez que a correlação espacial de vetores de deslocamento inválidos, que são considerados como vetores de *outliers* multivariados, tende a se comportar de maneira diferente dos vetores de deslocamento normais, eles podem ser identificados de forma eficiente no variograma.

Van Zoest et al. (2018) apresentam um método de detecção de *outliers* multivariados em redes de sensores de qualidade do ar urbano, baseado em uma classificação espaço-temporal, com foco nas concentrações horárias de NO_2. Neste método, as observações em um intervalo de um ano são divididas em 16 classes espaço-temporais. Essas classes refletem contexto urbano versus estações de tráfego urbano, dias de semana versus fins de semana e quatro períodos por dia. Para cada classe espaço-temporal, os *outliers* multivariados são detectados através da média e o desvio-padrão da distribuição normal subjacente à distribuição normal truncada das observações de NO_2.

O trabalho apresentado por Barbosa et al. (2018) propõe uma técnica de detecção de *outliers* multivariados executada através de análise de agrupamentos. A geração dos dados foi realizada atráves de simulação via Método de Monte Carlo e a técnica de mistura de distribuições normais multivariadas. Nesta abordagem, utilizou-se o método de análise de agrupamento *k*-médias, com o objetivo de agrupar os indivíduos semelhantes. Para concluir se um determinado grupo de indivíduos é um grupo de *outliers* multivariados ou não, utilizou-se como critério uma medida baseada no desvio-padrão amostral (s) das distâncias entre os centroides dos grupos e a mediana dos dados. Dessa forma, se a distância euclidiana entre o centroide de um grupo e a mediana dos dados for superior a $2,5 \times s$, os indivíduos deste grupo são classificados como *outliers* multivariados.

Wang et al. (2019) introduzem um novo modelo de detecção de *outlier* multivariado que usa um gráfico virtual, denominado *Virtual Outlier Score* (VOS). As informações locais são combinadas com as conexões implícitas na representação gráfica do conjunto de dados original. Este modelo constrói um gráfico de similaridade através dos vizinhos semelhantes de cada objeto e introduz um acoplamento de nó virtual com uma coleção de arestas virtuais para gerar um gráfico virtual. Um processo de passeio aleatório de Markov customizado é então executado no gráfico virtual for-

temente conectado sob o princípio de que um potencial *outlier* multivariado deve ter mais peso para ser visitado. Após atingir o equilíbrio, o vetor de distribuição estacionário é utilizado para deduzir a pontuação do *outlier* multivariado virtual.

Wang e Mao (2019) apresentam um esquema de detecção de *outliers* multivariados que pode ser usado diretamente para monitoramento ou controle de processo industriais. Com base na regressão do processo gaussiano tradicional, foram desenvolvidos algoritmos de detecção, dos quais a função média, função de covariância, função de verossimilhança e método de inferência, são especialmente concebidos.

O estudo de Wahid e Rao (2019) traz um algoritmo de detecção de *outliers* multivariados baseado em distância através do clássico algoritmo de otimização *Particle Swarm Optimization* (PSO). Neste algoritmo, atribui-se um grau periférico a cada ponto de dados através da soma das distâncias entre os pontos e seu conjunto vizinho mais próximo. Em seguida, o PSO é usado para detectar subespaços periféricos onde podem existir *outliers* multivariados.

Lu et al. (2020) propõem um algoritmo de detecção de nós *outliers* multivariados em redes de sensores sem fio com base no processamento de sinal gráfico. Inicialmente, de acordo com as características de posição do sensor, foi estabelecido um modelo de sinal gráfico denominado *K-Nearest Neighbors*. Em seguida, um teste estatístico foi construído com base na razão de suavidade do sinal do gráfico antes e depois da filtragem *low-pass*. Por fim, o julgamento da existência de nós *outliers* multivariados foi realizado por meio do teste estatístico e um limite de decisão.

Lejeune et al. (2020) apresentam uma abordagem para a identificação de *outliers* multivariados em dados funcionais multivariados. Essa abordagem é realizada através de um método pelo qual diferentes características periféricas são capturadas com base em funções de mapeamento de geometria diferencial. Nesse sentido, são extraídas características de forma que refletem a periferia de uma curva com um alto grau de interpretabilidade.

Kamalov e Leung (2020) propõem um novo algoritmo de detecção de *outliers* multivariados baseado na análise de componentes principais e estimativa de densidade de kernel. Este método foi concebido para enfrentar os desafios de lidar com dados de alta dimensão. Dados originais são projetados em um espaço menor e a estrutura inata dos dados é utilizada para calcular pontuações anormais para cada observação do conjunto de dados.

O estudo de Tabjula et al. (2021) apresenta uma análise de *outlier* multivariado para detecção de defeito através de amostragem esparsa no monitoramento de saúde estrutural de onda guiada. Neste estudo, os autores propõem uma abordagem estatística baseada na detecção de *outliers* multivariados para identificar e localizar defeitos em placas compostas. É utilizado um número menor de pontos de detecção em comparação com outras técnicas de imagem convencionais. As etapas principais envolvidas nesta abordagem são a seleção esparsa aleatória dos pontos de detecção, seguida por um processo de detecção de *outliers* multivariados de duas etapas com base no limiar e no cálculo do desvio absoluto mediano.

Dentro deste vasto conjunto de técnicas, algumas serão abordadas com

maior profundidade neste estudo. Para tanto, um ferramental específico, já bem fundamentado na literatura, precisa também ser revisado. Tal revisão será abordada nas seções subsequentes do presente capítulo revisional.

2.3 Distribuição Normal

A primeira aparição da distribuição normal (como uma aproximação da distribuição binomial) apareceu em um panfleto datado de 12 de novembro de 1733 de Abraham De Moivre. O panfleto estava em latim, em 1738, Abraham De Moivre apresentou os resultados do referido panfleto em idioma inglês em De Moivre (1738). A distribuição normal foi posteriormente redescoberta por Laplace (1776) e Gauss (1809). Também conhecida como distribuição gaussiana, a distribuição normal é a mais importante das distribuições de probabilidade, pois serve de base teórica para a estatística inferencial. Outro fator de importância se deve ao teorema central do limite, pois ele garante que a média dos dados converge para uma distribuição normal conforme o número de dados aumenta, mesmo que os dados originais não sejam distribuídos segundo uma normal.

A distribuição normal tem como parâmetros a média e o desvio-padrão. A média indica a posição central da distribuição e o desvio-padrão a dispersão dos dados. Uma variável aleatória contínua X tem distribuição normal se sua função densidade de probabilidade for dada pela equação 2.1:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \qquad -\infty < x < \infty. \tag{2.1}$$

em que μ é a média e σ é o desvio-padrão da distribuição.

A Figura 2.3.1 apresenta o gráfico da distribuição normal. Conforme pode ser visto, o gráfico tem o formato semelhante a um sino e depende dos valores dos parâmetros μ e σ.

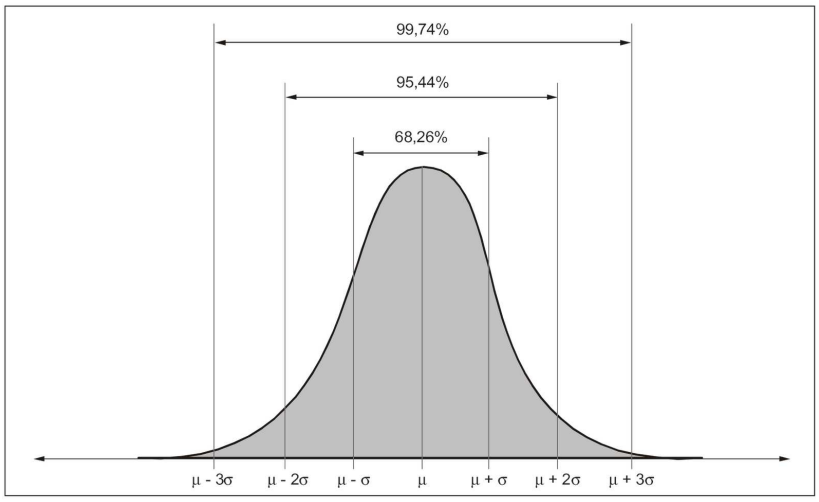

Figura 2.3.1: Representação gráfica da Distribuição Normal.

2.4 Distribuição Normal Multivariada

Conforme menciona Ferreira (2011), a distribuição normal multivariada é uma generalização da distribuição normal univariada para $p \geqslant 2$ dimensões, ou seja, quando se tem duas ou mais variáveis aleatórias conjutamente em estudo. Ela desempenha um papel muito importante nos estudos da análise multivariada, pois representa uma aproximação adequada de distribuições populacionais e dados experimentais. Além disso, ela é capaz de descrever conjuntos de variáveis aleatórias de valores reais correlacionados e pode ser utilizada em várias áreas como engenharia, psicologia, economia, entre outras.

Seja um vetor aletório $\boldsymbol{X}' = [X_1, X_2, \ldots, X_p] \in \mathbb{R}^p$. É dito que $\boldsymbol{X}'$ tem distribuição normal multivariada se sua função densidade de probabilidade é dada pela equação 2.2:

$$f(x_1, x_2, \ldots, x_p) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left[-1/2(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] \tag{2.2}$$

em que $\boldsymbol{\mu}$ é o vetor de médias, $\boldsymbol{\Sigma}$ a matriz de covariâncias e p é o índice da dimensão da distribuição normal p-variada e indica o número de variáveis em estudo.

A partir da distribuição normal univariada, o termo $\left(\frac{X-\mu}{\sigma}\right)^2$ é generalizado para $(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$, que é definida como a distância de Mahalanobis. Quando $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, ou seja, tem distribuição normal multivariada, então cada um dos elementos de $\boldsymbol{X}$ tem distribuição normal univariada.

Quando $\boldsymbol{X}$ é um vetor tal que cada um dos elementos de tem distribuição normal univariada então $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ tem distribuição normal multivariada, então cada um dos elementos de tem distribuição normal univariada.

A Figura 2.4.1 apresenta o gráfico da distribuição normal multivariada para o caso bidimensional, ou seja, para $p = 2$ e ainda com $\boldsymbol{\Sigma}$ igual a matriz identidade.

2.5 Simulação de dados via Distribuição Normal Multivariada Contaminada

Em estudos que envolvem simulação, muitas vezes é necessário simular dados que contenham valores *outliers* multivariados. Em geral, um *outlier* multivariado é uma observação que tem uma pequena probabilidade de ser gerada aleatoriamente. Consequentemente, *outliers* multivariados são geralmente gerados através de uma distribuição diferente, ou através do conhecimento da distribuição geradora de dados para construir valores de dados discrepantes. Por exemplo, a partir de uma distância especificada do centro da distribuição. Essa construção pode ser realizada através da

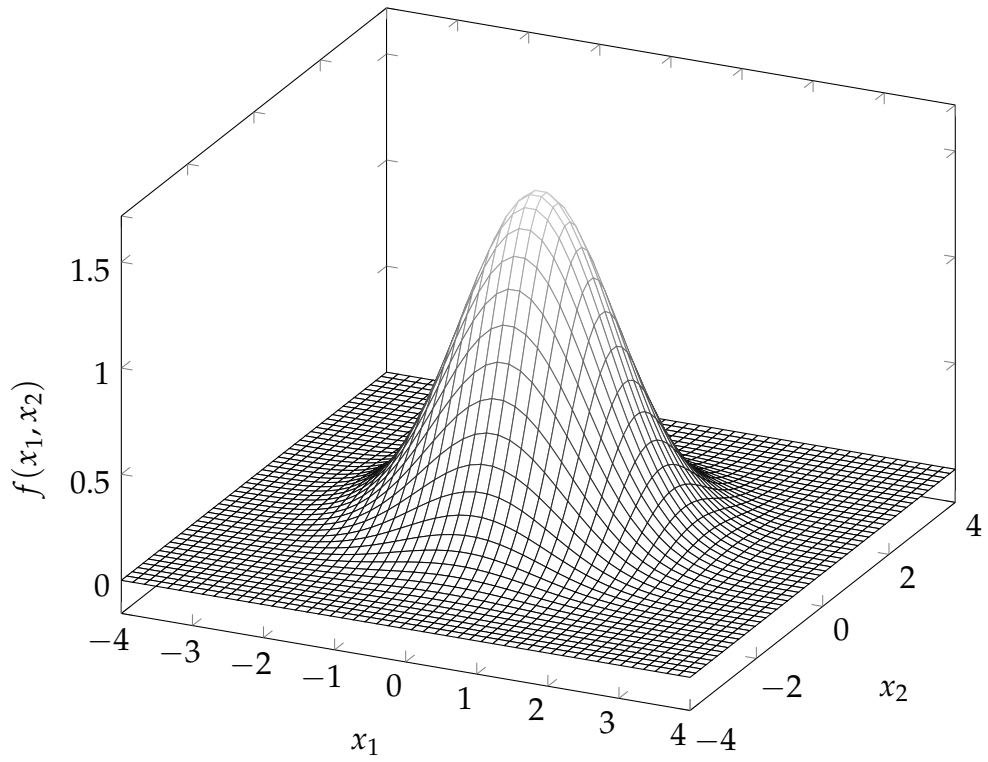

Figura 2.4.1: Representação gráfica da Distribuição Normal Bivariada.

mistura de distribuições normais multivariadas.

A referida mistura gera populações cuja distribuição é usualmente conhecida como distribuição normal multivariada contaminada. Para tanto, considere um vetor aleatório $\boldsymbol{X}' = [X_1, X_2, \ldots, X_p] \in \mathbb{R}^p$, com distribuição normal multivariada contaminada, sua função densidade de probabilidade será dada pela equação 2.3:

$$\begin{aligned} f(x_1, x_2, \ldots, x_p) = (1-\delta) 2\pi^{-p/2} \left|\boldsymbol{\Sigma}_1\right|^{-1/2} exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right] \\ +(\delta) 2\pi^{-p/2} \left|\boldsymbol{\Sigma}_2\right|^{-1/2} exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)'\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right] \end{aligned} \tag{2.3}$$

em que $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ é observação multivariada de $\boldsymbol{X}'$, $(1-\delta)$ é a probabilidade de que o processo seja realizado por $\mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, δ é a probabilidade que o processo seja realizado por $\mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, $\boldsymbol{\Sigma}_i$ é uma matriz positiva definida, $\boldsymbol{\mu}_i$ é o vetor de médias com $i = 1, 2$ e $0 \leqslant \delta \leqslant 1$.

Os procedimentos usuais para estudos que demandam simulações com distribuições dessa natureza partem da pré-fixação de dois vetores de médias distintos $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$ e de uma matriz de covariância $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ tais que:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\mu}_2 = \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \\ \mu \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \sigma_{1,2} & \sigma_{1,3} & \dots & \sigma_{1,p} \\ \sigma_{2,1} & 1 & \sigma_{2,3} & \dots & \sigma_{2,p} \\ \vdots & & \ddots & & \vdots \\ \sigma_{p-1,1} & \sigma_{p-1,2} & \dots & 1 & \sigma_{p-1,p} \\ \sigma_{p,1} & \sigma_{p,2} & \dots & \sigma_{p,p-1} & 1 \end{bmatrix}$$

em que $\mu \in \mathbb{R}$ com $\mu \neq 0$, $\sigma_{i,j} = \rho \ \forall i \neq j$ e $\rho \in [-1, 1]$ é o coeficiente de correlação entre as variáveis.

Apesar de se tratar de um procedimento usual de simulaçao para execução de testes dessa natureza, trata-se de um procedimento muito pouco realista. Um cenário de distribuições multivariadas com média fixa em todas as coordenadas e correlação constante para todas as combinações par a par não parece nada condizente com uma situação realista. Os procedimentos usuais são executados dessa forma dada a carência de mecanismos eficazes para a simulação de matriz de correlação independentes de conjuntos de dados reais. O presente estudo não tem o propósito de questionar a validade de investigações anteriores, mas sim, estabelecer uma análise mais próxima de um cenário real. Uma estratégia mais detalhada de simulação de dados para esse propósito será abordada no capítulo 4.

2.6 Análise de Agrupamentos

De acordo com Malhotra (2012), a análise de agrupamentos, também conhecida com análise de *clusters*, é uma técnica utilizada para classificar objetos em grupos relativamente homogêneos, chamados de agrupamentos ou conglomerados. Uma das características da análise de agruamentos é que os objetos tendem a ser semelhantes entre si dentro dos grupos, mas diferentes quando comparados a outros agrupamentos.

Conforme mencionado por Hair et al. (2009), as características de cada objeto são combinadas em uma medida de semelhança, que pode ser de similaridade ou dissimilaridade. Essa medida é calculada para todos os pares de objetos, isso possibilita a comparação entre os objetos e a associação das observações semelhantes por meio da análise de agrupamento. Uma forma de representar a similaridade entre objetos é através de medidas de distância que indicam a proximidade entre os mesmos. Os métodos de agrupamento são divididos em duas classes: os métodos hierárquicos e os não-hierárquicos.

Segundo Everitt et al. (2001), os métodos hierárquicos são técnicas relativamente simples nas quais os dados são particionados (ou agrupados) sucessivamente e produzem uma representação hierárquica dos agrupamentos. Os métodos hierárquicos tem como principal característica um algoritmo capaz de fornecer mais de um tipo de partição dos dados, em que um grupo pode ser mesclado a outro em determinado passo do algoritmo. Uma das vantagens é que esses métodos não exigem que já se tenha um número inicial de agrupamentos, porém são considerados inflexíveis,

uma vez que não se pode trocar um elemento de grupo. Eles podem ser classificados em dois tipos: aglomerativos e divisivos. Nos métodos aglomerativos, todos os elementos começam separados e são posteriormente agrupados em etapas, um a um, até que se tenha um único grupo com todos os elementos. O número ideal de grupos é escolhido dentre todas as opções através de algum critério específico. Já nos métodos divisivos todos os elementos começam juntos em um único grupo e são posteriormente separados, um a um, até que cada elemento seja seu próprio agrupamento. Assim como nos métodos aglomerativos, o número ótimo de grupos é escolhido dentre todas as possíveis combinações.

Os métodos não-hierárquicos, também conhecidos como métodos de particionamento, são caracterizados pela necessidade de se definir à priori o número de grupos, para produzir uma partição em um número fixo de classes. Estes métodos tem por objetivo determinar a classificação dos n indivíduos em k grupos que otimize algum critério de homogeneidade interna e heterogeneidade externa. Uma das vantagem dos métodos não-hierárquicos é que eles são mais flexíveis, uma vez que os elementos podem ser trocados de grupo durante a execução do algoritmo. Além disso, podem ser utilizados em bases de dados maiores, pois são bem mais rápidos computacionalmente que os métodos hierárquicos. Como desvantagens, destacam-se a necessidade do número de grupos ser escolhido a priori, o que pode levar a conclusões equivocadas caso o número de grupos não seja o ideal e no fato do algoritmo ser, em geral, sensível às condições iniciais, isso possibilita gerar resultados diferentes a cada execução.

2.6.1 Método Ward

Criado por Ward Jr (1963), o método Ward, também conhecido como método de variância mínima, é um método hierárquico aglomerativo que busca partições que minimizem a perda relacionada a cada agrupamento. Essa perda pode ser quantificada pela equação 2.4:

$$W = \sum_{k} \sum_{i \in k} (\boldsymbol{x}_{ik} - \bar{\boldsymbol{x}}_k)^t (\boldsymbol{x}_{ik} - \bar{\boldsymbol{x}}_k) \tag{2.4}$$

em que $\boldsymbol{x}_{ik}$ é a i-ésima observação pertencente ao k-ésimo agrupamento e $\bar{\boldsymbol{X}}_k$ é o vetor de médias do k-ésimo agrupamento. O método Ward busca produzir grupos de tal forma que a união dos indivíduos produza o menor incremento possível no valor de W.

2.6.2 Método K-médias

Formalizado por MacQueen (1967), o K-médias é um método de partição não-hierárquico que particiona o conjunto de dados através de uma técnica iterativa. Neste método, o número de *clusters* k é definido a prioi e os n indivíduos são agrupados nos k *clusters*. Este é um dos métodos de agrupamento mais conhecidos e utilizados atualmente. O método utiliza a soma de quadrados residual (RSS) de modo que se tenha homogeneidade dentro dos agrupamentos e heterogeneidade entre os grupos. Segundo Bussab

(1990), quanto menor for a soma de quadrados residual, mais homogêneos serão os elementos dentro de cada grupo e melhor será a partição. A soma de quadrados residual pode ser calculada pela equação 2.5:

$$RSS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\boldsymbol{x}_{ij} - \boldsymbol{c}_j)'(\boldsymbol{x}_{ij} - \boldsymbol{c}_j) \tag{2.5}$$

em que $\boldsymbol{x}_{ij}$ é o vetor correspondente a i-ésima observação pertencente ao j-ésimo *cluster*, $\boldsymbol{c}_j$ é o vetor de médias (centroide) do j-ésimo *cluster*, k é o número de *clusters* e n_i é o número de observações do j-ésimo *cluster*.

De acordo com Ferreira (2011), a implementaçaõ do algoritmo do método K-médias pode ser realizada a partir dos seguintes passos:

1. selecione aleatoriamente k elementos como centroides iniciais, nominados $(c_1, c_2, \ldots, c_k)$, alternativamente, a escolha dos centroides iniciais pode ser realizada de alguma outra forma;
2. calcule as distâncias entre cada elemento e os k centroides e classifique os elementos no grupo mais próximo;
3. calcule o centroide de cada grupo;
4. repita os passos 2 e 3 até que não ocorram mais mudanças nos centroides.

2.7 A Distância de Mahalanobis

A identificação de candidatos a *outliers* multivariados através da distância de Mahalanobis pode ser realizada baseada em quantis teóricos da distribuição χ^2 segundo Rousseeuw e Zomeren (1990). O uso da distância de Mahalanobis, aqui nominada $\mathcal{MD}$, é sugerido por diversos autores como uma medida bastante adequada em estratégias para detecção de *outliers* multivariados. Pode-se definir a distância de Mahalanobis amostral a partir da equação 2.6:

$$\mathcal{MD}_i = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'\boldsymbol{S}^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})} \tag{2.6}$$

em que $\boldsymbol{x}_i$ é i-ésima observação amostral do conjunto de dados multivariados $\boldsymbol{X}$, ainda, $\bar{\boldsymbol{x}}$ é o vetor de médias amostrais do conjunto $\boldsymbol{X}$, e $\boldsymbol{S}$ é a matriz de variâncias e covariâncias amostrais.

O interesse central é verificar, dentre as observações amostrais, quais possuem a distância quadrática de Mahalanobis ($\mathcal{MD}_i^2$) superior ao quantil $1 - \alpha$ da distribuição $\chi^2_{(p)}$, em que os graus de liberdade p representam o número de variáveis consideradas, segundo Rousseeuw e Zomeren (1990).

Dentre as mais variadas medidas de distância, em especial, a de Mahalanobis é extremamente sensível à presença de *outliers* multivariados. Valores extremos, ou grupo de valores aberrantes, podem influenciar severamente essas métricas para distâncias. Portanto, um questionamento imediato é: "como uma distância facilmente influenciada por *outliers* multivaria-

dos, pode ser capaz de identificá-los?". Basta tratar as partes mais sensíveis dessa medida, a média e a matriz de variâncias, tais medidas são calculadas de forma robusta, ou seja, técnicas robustas de estimação devem ser utilizadas. Dentre os estimadores robustos, pode-se destacar o *Minimum Covariance Determinant*, aqui nominado MCD e o *Minimum Volume Ellipsoid*, aqui dito MVE.

2.8 Estimadores robustos MCD e MVE

Proposto por Rousseeuw (1984), o elipsoide de volume mínimo (MVE) é um estimador baseado no elipsoide de menor volume capaz de cobrir pelo menos k pontos do conjunto amostral $\boldsymbol{X}$, em que $n/2 < k < n$, e seu valor de decomposição é essencialmente $(n-k)/n$. Segundo Hampel et al. (2011), o valor de decomposição é a fração de *outliers* multivariados que, quando excedido, levará a estimativas completamente enviesadas. O MVE possui ponto de ruptura 0,5 e é equivariante, ou seja:

$$\begin{aligned} T(\boldsymbol{x}_1\boldsymbol{A}+\boldsymbol{b},...,\boldsymbol{x}_n\boldsymbol{A}+\boldsymbol{b}) &= T(\boldsymbol{x}_1,...,\boldsymbol{x}_n)\boldsymbol{A}+\boldsymbol{b} \quad \text{ e} \\ C(\boldsymbol{x}_1\boldsymbol{A}+\boldsymbol{b},...,\boldsymbol{x}_n\boldsymbol{A}+\boldsymbol{b}) &= \boldsymbol{A}'C(\boldsymbol{x}_1,...,\boldsymbol{x}_n)\boldsymbol{A}, \end{aligned} \tag{2.7}$$

em que $\boldsymbol{b}$ é um vetor de dimensão p e $\boldsymbol{A}$ é uma matriz p-dimensional não singular, como descrito por Rousseeuw e Zomeren (1990). Este estimador robusto é definido pelo par $(\boldsymbol{T},\boldsymbol{C})$ em que $\boldsymbol{T}$ é um vetor $p-$dimensional da média amostral e $\boldsymbol{C}$ é a sub-matriz quadrada de dimensão p da matriz de covariância amostral. A distância robusta definida para o MVE é dada pela equação 2.8:

$$\mathcal{RD}_i = \sqrt{(\boldsymbol{x}_i - T(\boldsymbol{X}))'\ C(\boldsymbol{X})^{-1}\ (\boldsymbol{x}_i - T(\boldsymbol{X}))}. \tag{2.8}$$

O algoritmo básico de reamostragem para aproximar o MVE, denominado MINVOL, foi proposto por Rousseeuw e Leroy (1987). Este algoritmo considera um subconjunto de teste de $p+1$ observações e calcula sua média e matriz de covariância. O elipsoide correspondente é então inflado ou desinflado para conter exatamente k observações. Esse procedimento é repetido várias vezes, e o elipsoide de menor volume é retido. Para conjuntos de dados pequenos é possível considerar todos os subconjuntos de tamanho $p+1$, mas para conjuntos de dados maiores os subconjuntos de teste são obtidos aleatoriamente.

A função `cov.mve`, disponível no pacote `MASS` do software R Core Team (2021), calcula o estimador robusto MVE e utiliza em seu algoritmo padrão um valor de k igual a parte inteira de $(n+p+1)/2$. Além disso, o número de subconjuntos padrão usado para estimativas iniciais pode ser até 5.000 subconjuntos.

Também proposto por Rousseeuw (1984), o determinante mínimo da covariância estimada (MCD) está entre os métodos de mais vasta utilização para a construção de estimadores robustos, isso por se tratar de um

algoritmo computacionalmente rápido, de acordo com Rousseeuw e Driessen (1999). O estimador MCD é determinado por um subconjunto de tamanho h, em que $n/2 < h < n$, que minimiza o determinante da matriz de covariâncias amostral, calculado apenas sob os h pontos. A estimativa de locação é a média destes pontos, enquanto a estimativa de dispersão é proporcional à sua matriz de covariância, em que a escolha do tamanho de h determina a robustez do estimador. Com um compromisso entre robustez e eficiência, Filzmoser et al. (2005) utilizaram um valor de $h \approx 0,75n$, em que n é o tamanho amostral.

O valor de decomposição do estimador MCD é de aproximadamente $(n-h)/n$, ou seja, igual ao MVE. Entretanto, o MCD apresenta vantagens sobre o MVE. De acordo com Butler et al. (1993), a eficiência estatística do MCD é melhor porque ele é assintoticamente normal, enquanto que o MVE tem uma taxa de convergência mais baixa.

Segundo Rousseeuw e Driessen (1999), distâncias robustas com base no MCD são mais precisas do que aquelas baseadas no MVE e, portanto, mais adequadas para expor *outliers* multivariados. Além disso, Rousseeuw e Driessen (1999) afirmam que seu algoritmo MCD supera em muito o MVE em termos de eficiência estatística e velocidade de computação.

O cálculo do estimador MCD pode ser realizado através da função `covMcd`, disponível no pacote `robustbase` do software R Core Team (2021), cuja a implementação usa o algoritmo *FAST MCD* de Rousseeuw e Driessen (1999) para aproximar o estimador do determinante da covariância mínima.

Além destes conceitos, o método de detecção de *outliers* multivariados através de análise de agrupamentos Barbosa et al. (2018) precisa ser apresentada com mais detalhes para subsidiar este estudo. A partir deste ponto a referida técnica será nominada CAM (*Cluster Analysis Method*).

2.9 CAM: um método para identificação de *outliers* em dados multivariados

Seja $\mathcal{P}$ um conjunto de dados multivariados sob investigação, no qual a presença de *outliers* seja um fato. Considere a utilização do procedimento de análise de agrupamentos $k-$médias, com interesse em estabelecer agrupamentos por semelhança entre os indivíduos de $\mathcal{P}$. O método de agrupamento $k-$médias é iniciado pela escolha de k elementos como centroides. O procedimento usual é que tal escolha seja completamente aleatória, mas não existe perda de generalidade ao supor a proposição de algum critério específico prévio para tal escolha.

O procedimento executado através de escolha completamente aleatória de centroides é capaz de produzir partições com diversos formatos. Não existem garantias de que duas realizações do mesmo procedimento, para o mesmo conjunto de dados, forneçam exatamente a mesma partição. Por outro lado, a proposição de algum critério específico de escolha de centroides é capaz de fornecer mais estabilidade ao procedimento e, portanto, aos

resultados alcançados, através da utilização da técnica de agrupamento.

A aleatoriedade no procedimento de escolha dos centroides pode ser facilmente controlada. A escolha de centroides se baseia em um procedimento de simulação de variáveis aleatórias, ou seja, um sorteio aleatório. Uma semente pré-fixada para a inicialização do processo de geração de números "pseudo" aleatórios é suficiente para contornar tal efeito de aleatoriedade indesejado. Esse cuidado é suficiente no intuito de garantir que para o mesmo conjunto de dados, em duas realizações distintas, sempre será obtida a mesma partição.

Para uma partição obtida através do procedimento $k-$médias, é possível mensurar a distância euclidiana entre o centroide de algum agrupamento com respeito à mediana do conjunto completo de dados. Quanto mais distante o centroide de um determinado agrupamento estiver, em relação à mediana do conjunto completo de dados, maior o potencial para que os elementos do agrupamento do referido centroide sejam *outliers*.

A proposição original CAM para detecção de *outliers* multivariados levou em conta o desvio-padrão (s_c) entre os centroides de agrupamentos e a mediana ($\tilde{X}$) do conjunto completo de dados. Agrupamentos cuja distância euclidiana entre seu centroide e a mediana ultrapassavam a cota de $2{,}5 \times s_c$ foram considerados agrupamentos de *outliers*. Este estudo trabalhará com a cota $\phi \times s_c$, ou seja, a versão prévia da metodologia CAM utiliza $\phi = 2{,}5$. Investigações sobre o efeito do valor ϕ serão abordadas no capítulo 4.

Para um cenário de pontos pertencentes ao $\mathbb{R}^2$ uma visualização se torna viável e bastante útil para a compreensão da técnica. Suponha o cenário de um conjunto de dados dividido em cinco agrupamentos, a Figura 2.9.1 ilustra essa condição.

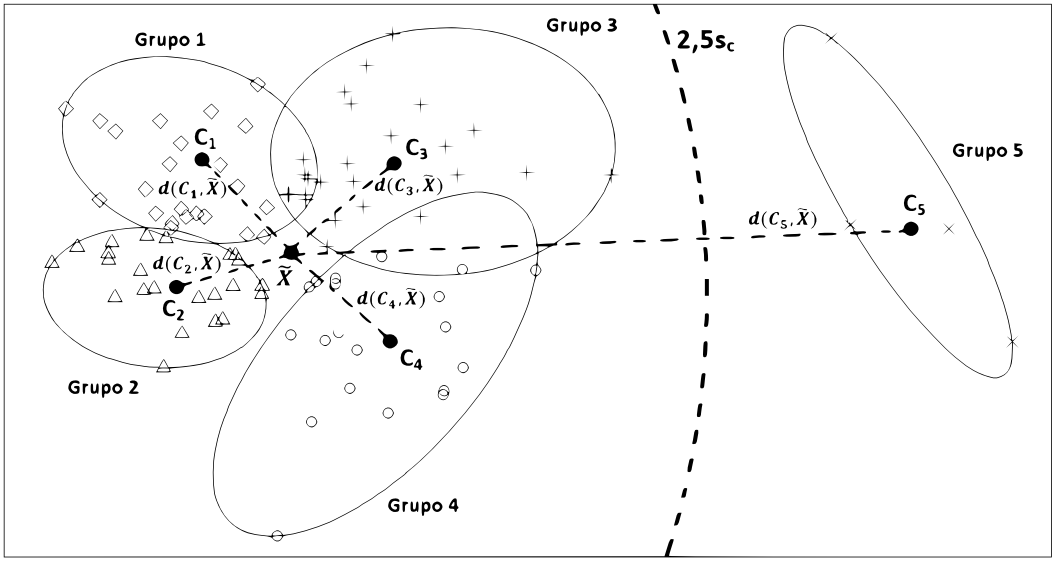

Figura 2.9.1: Visualização gráfica do método de identificação de *outliers* por análise de agrupamentos.

A figura 2.9.1 mostra claramente o agrupamento 5 posicionado com distância euclidiana superior à $2{,}5 \times s_c$ com respeito à mediana do conjunto de dados, representada na imagem por $\tilde{X}$. Por outro lado, os demais

agrupamentos estão suficientemente próximos da mediana com respeito ao critério $2{,}5 \times s_c$. Logo os elementos do agrupamento 5 são classificados pelo método CAM como valores *outliers*.

A técnica é dependente da escolha do valor k, quantidade de agrupamentos. O estudo de Barbosa et al. (2018) apresenta uma escolha completamente *ad-hoc* de $k = n/10$. É intuitivo supor que tal escolha tende a se comportar bem em alguns cenários, mas em contrapartida, ser completamente deficitária em outros. A discussão deste entrave gera um objetivo secundário discutido por Barbosa et al. (2020): verificar o efeito de diferentes escolhas do valor k. Essa análise será melhor explorada no capítulo 4, que apresenta resultados experimentais.

Referências Bibliográficas

Aggarwal, C. C. (2017). An Introduction to Outlier Analysis, páginas 1–34. Springer International Publishing.

Atkinson, A. C. e Riani, M. (2002). Forward search added-variable t-tests and the effect of masked outliers on model selection. Biometrika, 89(4):939–946.

Atkinson, A. C. e Riani, M. (2004). The forward search and data visualisation. Computational Statistics, 19(1):29–54.

Atkinson, A. C., Riani, M. e Cerioli, A. (2010). The forward search: Theory and data analysis. Journal of the Korean Statistical Society, 39(2):117–134.

Barbosa, J. J., Duarte, A. R. e Martins, H. S. R. (2020). A performance evaluation in multivariate outliers identification methods. Ciência & Natura, 42:1–14.

Barbosa, J. J., Pereira, T. M. e Oliveira, F. L. P. (2018). Uma proposta para identificação de outliers multivariados. Ciência & Natura, 40:1–8.

Barnett, V. e Lewis, T. (1994). Outliers in Statistical Data. John Wiley & Sons.

Berton, L., Huertas, J., Araújo, B. e Zhao, L. (2010). Identifying abnormal nodes in complex networks by using random walk measure. Em IEEE Congress on Evolutionary Computation, páginas 1–6. IEEE.

Bussab, W. O. (1990). Introdução à análise de agrupamentos. ABE.

Butler, R., Davies, P. e Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. The Annals of Statistics, páginas 1385–1400.

De Moivre, A. (1738). The doctrine of chances: A method of calculating the probability of events in play. W. Pearson, second edition.

Everitt, B., Landau, S. e Leese, M. (2001). Cluster Analysis. Arnold, 4th edition.

Ferreira, D. F. (2011). Estatística multivariada. Editora UFLA, 2 edition.

Filzmoser, P. (2005). Identification of multivariate outliers: a performance study. Austrian Journal of Statistics, 34(2):127–138.

Filzmoser, P., Garrett, R. e Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. Computers & Geosciences, 31(5):579–587.

Filzmoser, P., Hron, K. e Reimann, C. (2009). Principal component analysis for compositional data with outliers. Environmetrics: The Official Journal of the International Environmetrics Society, 20(6):621–632.

Filzmoser, P., Maronna, R. e Werner, M. (2008). Outlier identification in high dimensions. Computational Statistics & Data Analysis, 52(3):1694–1711.

Gauss, C. F. (1809). Theoria motus corporum coelestium in sectionibus conicis solem ambientium auctore Carolo Friderico Gauss. sumtibus Frid. Perthes et IH Besser.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. e Tatham, R. L. (2009). Análise multivariada de dados. Bookman Editora.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. e Stahel, W. A. (2011). Robust statistics: the approach based on influence functions, volume 196. John Wiley & Sons.

Hawkins, D. M. (1980). Identification of Outliers, volume 11. Chapman and Hall.

Jolliffe, I. T. e Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065):20150202.

Kamalov, F. e Leung, H. H. (2020). Outlier detection in high dimensional data. Journal of Information & Knowledge Management, 19(01).

Kutsuna, T. e Yamamoto, A. (2017). Outlier detection using binary decision diagrams. Data Mining and Knowledge Discovery, 31(2):548–572.

Laplace, P. (1776). Mémoires de mathématique et de physique présentés à l'académie royale des sciences par divers savans et lûs dans ses assemblées. L'Académie Royale des Sciences par divers Savans et lûs dans ses Assemblées.

Lejeune, C., Mothe, J., Soubki, A. e Teste, O. (2020). Shape-based outlier detection in multivariate functional data. Knowledge-Based Systems.

Lu, G., Zhou, L., Lyu, S., Shi, C. e Su, K. (2020). Outlier node detection algorithm in wireless sensor networks based ongraph signal processing. Journal of Computer Applications, 40(3):783–787.

Luo, J., Frisken, S., Machado, I., Zhang, M., Pieper, S., Golland, P., Toews, M., Unadkat, P., Sedghi, A., Zhou, H., Mehrtash, A., Preiswerk, F., Cheng, C., Golby, A., Sugiyama, M. e Wells III, W. M. (2018). Using the variogram for vector outlier screening: application to feature-based image registration. International Journal of Computer Assisted Radiology and Surgery, 13(12):1871–1880.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Em Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, páginas 281–297. Oakland, CA, USA.

Malhotra, N. K. (2012). Pesquisa de marketing: uma orientação aplicada. Bookman Editora.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Resende, M., Brighenti, C. R. G. e Cirillo, M. Â. (2017). Procedure to identify outliers through cumulative distribution of extremes in a gamma response model. Communications in Statistics-Simulation and Computation, 46(9):6937–6946.

Rousseeuw, P. J. (1984). Least median of squares regression. Journal of the American statistical association, 79(388):871–880.

Rousseeuw, P. J. e Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41(3):212–223.

Rousseeuw, P. J. e Leroy, A. M. (1987). Robust regression and outlier detection. John Wiley & Sons.

Rousseeuw, P. J. e Zomeren, B. C. V. (1990). Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, 85(411):633–639.

Tabjula, J. L., Kanakambaran, S., Kalyani, S., Rajagopal, P. e Srinivasan, B. (2021). Outlier analysis for defect detection using sparse sampling in guided wave structural health monitoring. Structural Control and Health Monitoring.

Valadares, F. G., Aquino, A. L. L. e Rabelo, R. A. (2012). Detecção de outliers multivariados em redes de sensores sem fio. Em XLIV Simpósio Brasileiro de Pesquisa Operacional. SBPO.

Van Zoest, V., Stein, A. e Hoek, G. (2018). Outlier detection in urban air quality sensor networks. Water, Air, & Soil Pollution, 229(4):111.

771 Veloso, M. V. S. e Cirillo, M. A. (2016). Principal components in the discrimination of outliers: A study in simulation sample data corrected by pearson's and yates´ s chisquare distance. Acta Scientiarum. Technology, 774 38(2):193–200.

Wahid, A. e Rao, A. C. S. (2019). A distance-based outlier detection using particle swarm optimization technique. Em Information and 777 Communication Technology for Competitive Strategies, páginas 633–643. Springer.

Wang, B. e Mao, Z. (2019). Outlier detection based on gaussian process with 780 application to industrial processes. Applied Soft Computing, 76:505–516.

Wang, C., Liu, Z., Gao, H. e Fu, Y. (2019). Vos: A new outlier detection model using virtual graph. Knowledge-Based Systems, 185.

783 Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American statistical association, 58(301):236–244.

Zhu, J., Jiang, W., Liu, A., Liu, G. e Zhao, L. (2017). Effective and effici-786 ent trajectory outlier detection based on time-dependent popular route. World Wide Web, 20(1):111–134.

Capítulo 3

Material e Métodos

Este capítulo apresenta o principal produto desta tese, o desenvolvimento teórico que fundamenta e respalda a metodologia DDCAM proposta. Para este fim, a técnica de detecção de *outliers* multivariados executada através de análise de agrupamento (Barbosa et al., 2018), mencionada e detalhada no capítulo anterior é de suma importância na construção da metodologia apresentada nesta tese.

3.1 DDCAM (*Data-driven Cluster Analysis Method*): uma nova metodologia para identificação de *outliers* em dados multivariados

O foco central desse estudo está na proposição desta nova metodologia para detecção de valores *outliers* multivariados. A nova técnica parte de duas premissas que serão ajustadas com base nas informações do pesquisador que utiliza o método, e também das próprias informações inerentes ao conjunto de dados em estudo. Essa idéia construtiva do novo método justifica a nomenclatura *Data-driven*.

Inicialmente, suponha avaliar um conjunto composto por n observações de dados $p-$variados, ou seja, existem p variáveis unidimensionais para cada observação do conjunto de dados. Admita ainda que, a população da qual se extraiu a amostra de dados, siga o modelo de probabilidades $\mathcal{D}_p$, um modelo $p-$variado. Seria adequado supor que observações *outliers*, são possíveis observações muito pouco prováveis de ocorrer para o modelo de probabilidades $\mathcal{D}_p$, e ainda, que talvez se ajustem melhor a algum modelo alternativo de probabilidades $\mathcal{D}_p^*$.

Sem perda de generalidade, admita que os modelos probabilísticos $\mathcal{D}_p$ e $\mathcal{D}_p^*$ sejam iguais a menos do seu vetor de médias. Claro, no cenário real as diferenças podem ser superiores a isto, entretanto tal consideração não afetará esta construção metodológica. Para observações em uma única amostra, mas com elementos provenientes das duas distribuições, uma pergunta bastante relevante vem a tona. Qual a distância entre os vetores de médias faria o pesquisador considerar factível supor que as observações que se ajustam melhor ao modelo $\mathcal{D}_p^*$ são candidatos à valores extremos,

isso quando comparadas aos valores oriundos da distribuição $\mathcal{D}_p$? Usualmente essa distância é representada em quantidades de desvios-padrão, este estudo vai estabelecer que tal quantidade é representada pela letra grega τ.

É bastante razoável admitir que o valor de τ desvios-padrão que seja adequado para representar a referida distância é influenciado sobremaneira pela natureza dos dados sob análise. Um valor τ pode ser adequado para algumas variáveis, mas completamente inadequado para outras. Desta forma, o conhecimento do pesquisador acerca dos dados sob investigação é de extrema relevância para estabelecer a cota τ adequada nesse tipo de estudo.

Considere uma amostra $\mathcal{X}$ de n observações p-variadas sob investigação, que possa ser subdividida em $\mathcal{X}_1$ com n_1 observações, para n_1 definido pela parte inteira de $(1-\xi)\times n$, e $\mathcal{X}_2$ com $n-n_1$ observações, para $\xi \in (0,1)$. Seja $\bar{\boldsymbol{X}}_1$, o vetor de médias de $\mathcal{X}_1$ e $\bar{\boldsymbol{X}}_2$ o vetor de médias de $\mathcal{X}_2$. E ainda, que a j-ésima coordenada de $\bar{\boldsymbol{X}}_2$ seja diferente da coordenada em $\bar{\boldsymbol{X}}_1$ por mais que uma distância pré-fixada, suficientemente grande, e que este efeito se repita em todas as p coordenadas. É razoável supor que dados em $\mathcal{X}_2$ sejam avaliados como observações *outliers* independentemente do valor ξ?

Em outras palavras, se por exemplo, 50% dos dados amostrais estiverem no subconjunto $\mathcal{X}_1$ e 50% no subconjunto $\mathcal{X}_2$, ainda assim pareceria razoável admitir que metade dos valores amostrais são observações *outliers*? A resposta parece claramente ser por negar tal afirmação. Na prática, parece completamente adequado admitir a existência de uma quantidade máxima aceitável, de elementos em $\mathcal{X}_2$, que sejam admitidos como observações *outliers*.

Diante desse raciocínio, os valores de τ (distância mínima em desvios-padrão entre médias), k (quantidade de agrupamentos utilizados), ϕ (quantidade de desvios-padrão no processo de agrupamento para estabelecer agrupamentos de observações *outliers*) e ξ (proporção máxima admissível de *outliers*) são valores que serão utilizados na concepção da metodologia DDCAM.

O valor τ deve ser pré-fixado pelo usuário da metodologia de acordo com seu conhecimento prévio acerca das variáveis sob investigação. O conhecimento do investigador sobre as variáveis em estudo é bastante relevante, um valor τ geral, inerente somente ao método, com certeza seria uma grande perda de flexibilidade para adaptação em diferentes conjuntos de dados.

O valor de k afeta sobremaneira a estratégia de agrupamento. Critérios específicos para escolha de k são de extrema relevância, como mencionado por Barbosa et al. (2020). Uma inovadora proposição de critério de escolha mais adequada para o númuro de agrupamentos k será discutida posteriormente.

Já o valor ϕ é também inerente à estratégia de agrupamento, a proposição original do metodo CAM utilizou $\phi = 2{,}5$. Algumas investigações sobre o efeito e possíveis alterações no valor de ϕ serão abordadas no capítulo 4.

Por fim, o valor ξ será utilizado com intuito de se estimar a taxa de mistura δ que será utilizada no procedimento. O valor ξ pode ser pré-fixado pelo usuário do método, ou então estimado através dos próprios dados. Será apresentado aqui propostas distintas para o estimador $\hat{\delta}$, ou seja, para a taxa de mistura entre as distribuições dos dados comuns e *outliers*. Os referidos estimadores serão utilizados posteriormente nos experimentos numéricos simulados.

O procedimento DDCAM será executado em 4 estágios prévios, denominados: Estimação do valor δ, Processo de refinamento - I, Processo de refinamento - II, Busca pelo valor adequado k e, posteriormente, a execução do procedimento *Cluster Analysis Method* com a parametrização estabelecida pelos estágios anteriores.

3.1.1 Estimação do valor δ

O objetivo de estimar δ corresponde ao problema de delimitar qual o volume de *outliers* parece adequado para atender a razoabilidade dos dados. Em outras palavras, a proporção de elementos supostamente provenientes da distribuição dos valores *outliers*. Para tanto, o método DDCAM, inicialmente parte da premissa de estabelecer o valor ϕ para critério de $\phi \times s_c$ de distância do agrupamento em relação à mediana para estabelecer *outliers* constituído na metodologia CAM.

Além disso, como mencionado anteriormente, em uma amostra de dados multivariados $\mathcal{X}$, dividida em duas sub-amostras $\mathcal{X}_1$ e $\mathcal{X}_2$, é razoável supor que as observações de $\mathcal{X}_2$ sejam avaliadas como *outliers* quando as médias amostrais de $\mathcal{X}_1$ e $\mathcal{X}_2$ distam para alguma cota suficientemente grande. O que dizer sobre essa cota suficientemente grande? Tal distância, por razões óbvias, deve ser cotada em quantidade de desvios-padrão. Neste caso, será representado por η, ou seja, a distância em questão será expressa por $(\eta \times \sigma_j)$, em que σ_j seja o desvio-padrão da população na j-ésima variável.

Assim como a discussão sobre τ, o valor de η precisa ser previamente definido. Existe uma associação clara entre os valores de η e τ. Note que τ representa a distância entre vetores de médias de população para admissibilidade de que uma das populações seja de possíveis *outliers*. Já o valor η representa a distância entre coordenadas de vetores de médias amostrais usada para estimar a proporção de possíveis *outliers* na amostra sob investigação. Um raciocínio mais simplista faz o método escolher automaticamente para η o valor arbitrado pelo usuário para τ. Este estudo apresentará experimentos para verificar se tal escolha é adequada para o método proposto.

Uma proposição para um estimador adequado $\hat{\delta}$ pode ser construída através de uma análise prévia univariada. Sejam $\bar{X}_i$ e s_i, respectivamente, a média e o desvio-padrão amostrais pertencente ao vetor $(x_{1i}, x_{2i}, \ldots, x_{ni})$ de n observações da i-ésima variável sob investigação. Um estimador $\hat{\xi}_i$ será definido como a proporção de observações do vetor $(x_{1i}, x_{2i}, \ldots, x_{ni})$ que está a uma distância maior que η desvios padrão amostrais s_i da média $\bar{X}_i$.

A construção definitiva do estimador $\hat{\delta}$ será através de uma função das estimativas univariadas $\hat{\xi}_i$. Um raciocínio básico propõe o estimador da média, aqui nominado $\hat{\delta}_{\text{média}}$, dado pela equação 3.1:

$$\hat{\delta}_{\text{média}} = \frac{\sum_{i=1}^{p} \hat{\xi}_i}{p} \quad . \tag{3.1}$$

Uma segunda proposta tem uma versão mais conservadora, produz o estimador $\hat{\delta}_{\text{máximo}}$, dado pela equação 3.2, que toma o máximo entre as estimativas univariadas. O conservadorismo está associado em sempre majorar a estimativa $\hat{\delta}_{\text{média}}$, isso conduz para admintir um volume maior de possíveis valores *outliers*.

$$\hat{\delta}_{\text{máximo}} = \max_{\substack{1 \leqslant i \leqslant p \\ i \in \mathbb{N}}} (\hat{\xi}_i) \quad . \tag{3.2}$$

3.1.2 Processo de refinamento - I

Considere uma amostra $\mathcal{X}$ de n observações p-variadas sob investigação e que as quantidades τ, η e $\hat{\delta}$ estão definidas. Para o procedimento de agrupamentos descrito para a abordagem CAM de detecção de *outliers* seria necessário estabelecer previamente o número de agrupamentos k. Seja $\mathcal{K}$ o conjunto dos possíveis valores k para número de grupos possíveis utilizados nos agrupamentos investigados. Nesse cenário, $\mathcal{K} = \{2, \ldots, k_{max}\}$. A primeira concepção metodológica aqui é, como definir o valor de k_{max}. A concepção inicial foi $k_{max} = {}^{n}/_{\log(n)}$. A proposta de dividir a quantidade total de dados por $\log(n)$ é inspirada na clássica fórmula de Sturges (1926) que sugere suavizar n através do logaritmo na busca pela quantidade mais adequada para o número de classes em tabelas de distribuição de frequências.

A proposta será de construir o agrupamento com $k = 2$ e verificar se o menor (em quantidade de elementos) grupo (subconjunto) possui no máximo $\hat{\delta} \times n$ elementos (aqui, por flexibilidade será adotado o arredondamento para o menor inteiro maior que $\hat{\delta} \times n$). Se essa condição é aceita, o agrupamento com $k = 2$ é um agrupamento válido, caso contrário, o agrupamento com $k = 2$ será dito inválido e não será investigado. O procedimento descrito será repetido para $k = 3$ e assim subsequencialmente até $k = k_{max}$. Ao final desse processo, haverá um conjunto $\mathcal{K}$ refinado, aqui nominado $\mathcal{K}_1$. Note que $\mathcal{K}_1 \subseteq \mathcal{K}$.

3.1.3 Processo de refinamento - II

De posse desse primeiro refinamento, um segundo refinamento é proposto. Para cada valor $k \in \mathcal{K}_1$, será verificado se existe algum agrupamento cujo centroide dista da mediana mais que $\phi \times s_c$ e que **simultaneamente** possua no máximo $\hat{\delta} \times n$ elementos. Se essa condição não for satisfeita, esse valor k também deverá ser desprezado da investigação. Assim, será produzido

um segundo conjunto de valores k ainda mais refinado, denominado $\mathcal{K}_2$ com $\mathcal{K}_2 \subseteq \mathcal{K}_1 \subseteq \mathcal{K}$.

Durante os dois processos refinamentos, se $\mathcal{K}_1 = \emptyset$, então o método conclui que não existem observações *outliers*. Analogamente se $\mathcal{K}_1 \neq \emptyset$, mas $\mathcal{K}_2 = \emptyset$ então o método também conclui que não existem valores *outliers* no conjunto de dados investigado.

3.1.4 Busca pelo valor adequado k

Para situações em que $\mathcal{K}_2 \neq \emptyset$ deverá se estabelecer o critério para escolher o valor $k \in \mathcal{K}_2$ mais adequado para o procedimento de detecção de *outliers*. Isto é, na verdade, buscar a escolha k que tende a garantir melhores resultados em medidas específicas para qualificar o procedimento de detecção. Este estudo adota as medidas de sensibilidade, especificidade e acurácia como medidas bastante adequadas para tal avaliação. A escolha do valor mais adequado para k proposta no método DDCAM se baseia no critério de informação Bayesiano (BIC), bastante difundido para seleção entre proposta de modelos.

O critério de informação Bayesiano (BIC), também conhecido como critério de informação de Schwarz (1978), é uma medida muito utilizada para seleção de modelos, que visa penalizar modelos mais complexos, com um número excessivo de parâmetros e evitar a ocorrência de sobreajuste. Este critério baseia-se na função de verossimilhança e o modelo escolhido é aquele com menor valor de BIC. Conforme mencionado por Hastie et al. (2009), uma fórmula genérica para o cálculo do valor de BIC é dada pela equação 3.3:

$$BIC = -2\log(L) + \log(n) \times df \tag{3.3}$$

em que L é o valor maximizado da função de verossimilhança do modelo, n é o número de observações em estudo e df é o número de parâmetros estimados pelo modelo. Ramsey et al. (2008) utilizou o valor BIC para escolha do número de clusters no algoritmo k-médias a partir da equação 3.4:

$$BIC(k) = \frac{1}{\hat{\sigma}^2} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\boldsymbol{x}_{ij} - \boldsymbol{c}_j)'(\boldsymbol{x}_{ij} - \boldsymbol{c}_j) + \log(n) \times k \times p, \tag{3.4}$$

em que $\boldsymbol{x}_{ij}$ é o vetor correspondente a i-ésima observação pertencente ao j-ésimo cluster, $\boldsymbol{c}_j$ é o vetor de médias (centroide) do j-ésimo cluster, n é o número de observações, k é o número de clusters, p é o número de variáveis no conjunto de dados multivariados e $\hat{\sigma}$ é uma estimativa do desvio-padrão dos dados amostrais. O estimador usual nesta situação é dado pela equação 3.5:

$$\hat{\sigma}^2 = \left(\frac{1}{n \times p}\right) \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{X}})' (\boldsymbol{x}_i - \bar{\boldsymbol{X}}), \tag{3.5}$$

em que $\boldsymbol{x_i}$ é a i-ésima observação p-variada e $\bar{\boldsymbol{X}}$ é o vetor de médias amos-

trais.

No algoritmo k-médias, a padronização dos dados é indicada quando as variáveis não são homogêneas, isso evita que variáveis com alta grandeza escalar tenham peso exagerado na execução do agrupamento. A padronização prévia dos dados segue o formato $[(x_i - \bar{X})/s]$, isso para todas as variáveis que são coordenadas das observações em que s é a estimativa de desvio-padrão da respectiva variável.

Segundo Mohamad e Usman (2013), realizar a padronização dos dados antes da aplicação do algoritmo k-médias leva à obtenção de um resultado de agrupamento de melhor qualidade, eficiência e precisão. Para dados já padronizados, a expressão $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\boldsymbol{x}_{ij} - \boldsymbol{c}_j)'(\boldsymbol{x}_{ij} - \boldsymbol{c}_j)$ representa a soma de quadrados residual (RSS) e o valor do critério de informação Bayesiano pode ser reescrito pela equação 3.6:

$$BIC(k) = RSS + \log(n) \times k \times p \ . \tag{3.6}$$

Dessa forma, quando $\mathcal{K}_2 \neq \emptyset$, o valor k será k^*, um valor que minimiza o critério de informação Bayesiano restrito aos valores k que satisfazem as condições de $\mathcal{K}_2$, ou seja, dado pela equação 3.7:

$$k^* = \arg\min_{k \in \mathcal{K}_2} BIC(k) \ . \tag{3.7}$$

É importante observar que o critério de informação Bayesiano sofre influência da estrutura do agrupamento e também das próprias variáveis do conjunto de dados. Esse fator sugere a própria informação contida no conjunto de dados para fazer a calibração da escolha mais adequada de k para a metodologia DDCAM.

Conforme mencionado anteriormente, o método $k-$médias é usualmente iniciado pela escolha aleatória de k elementos como centroides e que tal escolha é capaz de produzir partições com diversos formatos. Portanto, para fornecer mais estabilidade ao procedimento, utilizou-se o método de agrupamento Ward para a geração dos centroides iniciais do método $k-$médias. Esse cuidado é suficiente para garantir que, para o mesmo conjunto de dados, em duas realizações distintas, sempre será obtida a mesma partição.

Uma vez estabelecido através dos estágios anteriores o valor mais adequado para k^*, o procedimento DDCAM é agora executado para esta escolha de valor k específica, e não mais para uma escolha completamente *ad-hoc* de valor k como na proposição original do método CAM. A Figura 3.1.1 apresenta um fluxograma com o esquema de execução para o método DDCAM.

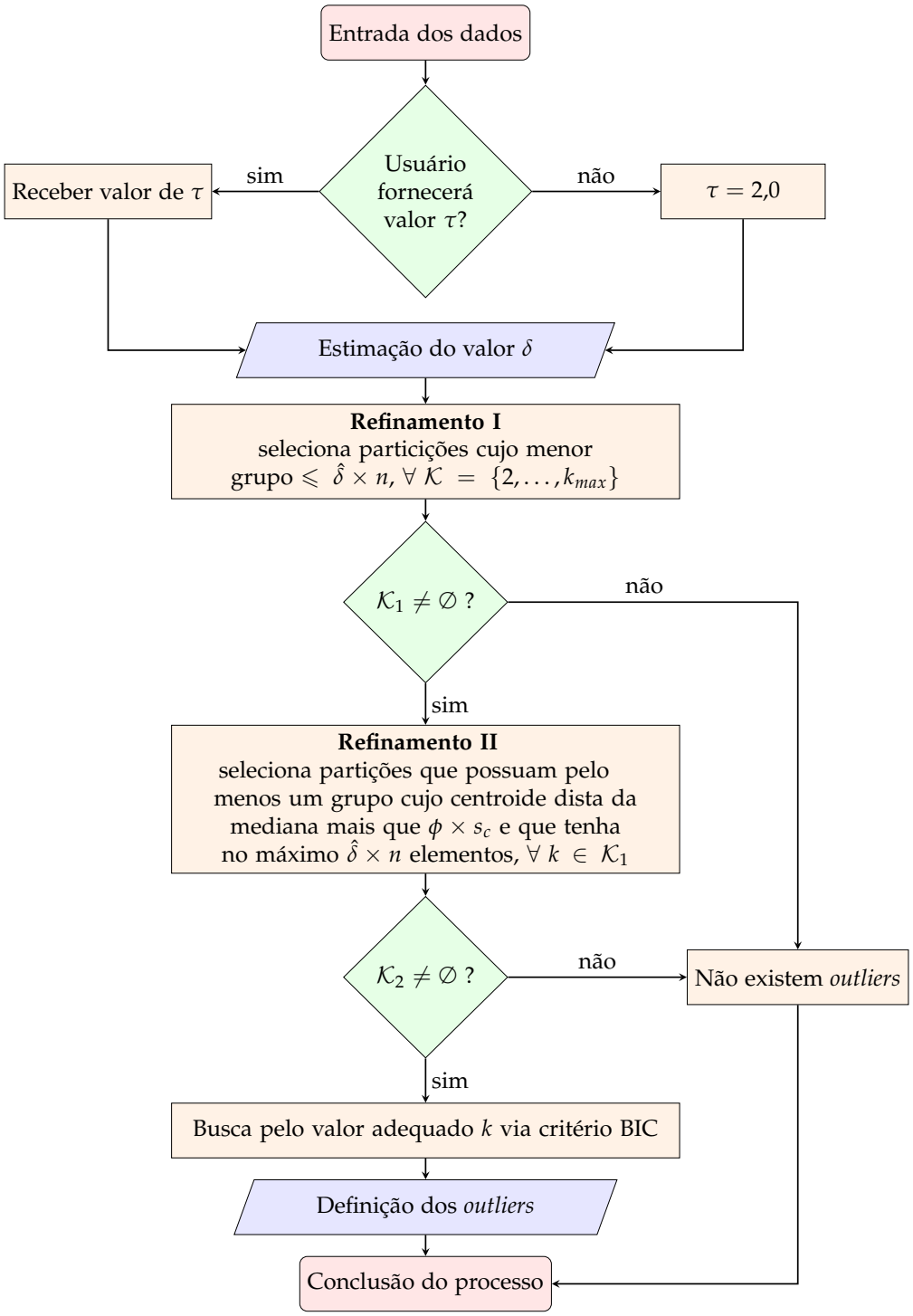

Figura 3.1.1: Fluxograma de execução do método DDCAM.

Referências Bibliográficas

1023 Barbosa, J. J., Duarte, A. R. e Martins, H. S. R. (2020). A performance evaluation in multivariate outliers identification methods. Ciência & Natura, 42:1–14.

Barbosa, J. J., Pereira, T. M. e Oliveira, F. L. P. (2018). Uma proposta para identificação de outliers multivariados. Ciência & Natura, 40:1–8.

Hastie, T., Tibshirani, R. e Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Mohamad, I. B. e Usman, D. (2013). Standardization and its effects on k-means clustering algorithm. Research Journal of Applied Sciences, Engineering and Technology, 6(17):3299–3303.

Ramsey, S. A., Klemm, S. L., Zak, D. E., Kennedy, K. A., Thorsson, V., Li, B., Gilchrist, M., Gold, E. S., Johnson, C. D., Litvak, V., Garnet Navarro, G., Roach, J. C., Rosenberger, C. M., Rust, A. G., Yudkovsky, N., Aderem, A. e Shmulevich, I. (2008). Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics. PLoS Computational Biology, 4(3):e1000021.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2):461–464.

Sturges, H. A. (1926). The choice of a class interval. Journal of the American Statistical Association, 21(153):65–66.

Capítulo 4

Resultados e Discussão

4.1 Dados simulados

Os procedimentos usuais para estudos que demandam simulações com distribuições multivariadas, dessa natureza, partem da pré-fixação de dois vetores de médias distintos $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$ e de uma matriz de covariância $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ tais que:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\mu}_2 = \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \\ \mu \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 1 & \sigma_{1,2} & \sigma_{1,3} & \dots & \sigma_{1,p} \\ \sigma_{2,1} & 1 & \sigma_{2,3} & \dots & \sigma_{2,p} \\ \vdots & & \ddots & & \vdots \\ \sigma_{p-1,1} & \sigma_{p-1,2} & \dots & 1 & \sigma_{p-1,p} \\ \sigma_{p,1} & \sigma_{p,2} & \dots & \sigma_{p,p-1} & 1 \end{bmatrix}$$

em que $\mu \in \mathbb{R}$ com $\mu \neq 0$, $\sigma_{i,j} = \rho \ \forall i \neq j$ e $\rho \in [-1, 1]$ é o coeficiente de correlação entre as variáveis.

Como dito anteriormente, este procedimento de simulação é bastante usual para execução de testes dessa natureza. Trata-se de um procedimento pouco realista, a situação de distribuições multivariadas com vetor de médias fixo com todas as coordenadas iguais e ainda com matriz de correlações constantes para todas as combinações par a par. Diversos estudos anteriores foram executados dessa forma, em virtude da lacuna quanto à existência de mecanismos eficazes para a geração de matriz de correlação que sejam completamente independentes de conjuntos de dados reais. Como dito anteriormente, essa investigação não tem a pretenção de invalidar investigações anteriores, mas sim, estabelecer uma análise mais realista. A seguir, será descrito o procedimento de simulação de dados a ser utilizado.

Inicialmente, o vetor de médias $\boldsymbol{\mu}_1$, não será constante, o valor de cada coordenada μ_j será obtido aleatoriamente através de uma simulação de $\mathcal{N}(0, 1)$ univariada. Posteriormente, o vetor de médias $\boldsymbol{\mu}_2$ terá suas coordenadas definidas aleatoriamente através de uma simulação de $\mathcal{N}(\pm 2, 1)$, ou seja, fixado o desvio-padrão de cada coordenada do vetor aleatório, o segundo vetor de médias da distribuição mistura estará em torno da média μ_j desviada em 2 desvios-padrão para mais ou para menos. Para efeito da

construção metodológica, o valor 2, na simulação tem o efeito do valor τ enunciado na construção do método. O objetivo é criar um cenário similar ao deslocamento $\mu_j \pm 2\sigma_j$. A definição quanto a somar ou subtrair, também será dada por um sorteio aleatório com probabilidade 0,5 de somar e 0,5 de subtrair.

Por fim, a parte menos realista do mecanismo de geração de dados está na escolha das matrizes de correlações, usualmente constantes. A proposição de um método inovador para esse propósito de geração de matrizes de correlações foi apresentado por Duarte et al. (2021). A matriz $\mathbf{\Sigma}$ ilustrada anteriormente pode ser substituída por uma matriz de correlações mais realista, sem correlações idênticas par a par. A matriz de correlações pode, inclusive, ser customizada de acordo com os interesses de comparação para os métodos de detecção de *outliers*. A utilização de uma matriz de correlações realística permite um estudo comparativo sobre a eficiência dos métodos mais confiável através de dados simulados mais similares aos dados reais.

4.2 Análise dos efeitos da escolha da quantidade de agrupamentos k

Como mencionado anteriormente, a escolha do valor k (quantidade de agrupamentos) tem grande impacto na metodologia CAM. O estudo apresentado em Barbosa et al. (2020) buscou elucidar este efeito. Experimentos foram realizados com a metodologia CAM padrão, que adota $k = {}^{n}/_{10}$ e uma segunda metodologia, aqui nominada CAM otimizado. A metodologia CAM otimizado, escolhe o valor k que maximiza a acurácia média em uma sequência de procedimentos de detecção de *outliers* em dados simulados.

Vale ressaltar que a metodologia CAM otimizado não tem sentido prático para dados reais, uma vez que não há como escolher previamente o valor k que maximizará a acurácia no processo de detecção. Diante dessa investigação, o termo acurácia, no processo de detecção, precisa estar bem definido. Serão definidos também os conceitos de sensibilidade e especificidade no processo de detecção.

Dada uma simulação de n observações multivariadas, das quais algumas são candidatos potenciais à *outliers* multivariados. Defina então os seguintes conjuntos:

- Ω: conjunto composto por todas as observações que foram simuladas para o procedimento;
- $\mathbb{O}$: conjunto dos candidatos à *outlier*, ou seja, observações que foram geradas com o vetor de médias μ_2;
- $\mathbb{D}$: conjunto das observações *outliers* identificadas pelo método utilizado.

A sensibilidade é a probabilidade de, dado que uma observação pertence à $\mathbb{O}$, ela pertencer à $\mathbb{D}$, ou seja, $P(\mathbb{D}|\mathbb{O})$. A especificidade é a probabilidade de, dado que uma observação não pertence à $\mathbb{O}$, ela não pertencer à $\mathbb{D}$, ou seja, $P(\overline{\mathbb{D}}|\overline{\mathbb{O}})$, em que para um dado conjunto A, tem-se que $\overline{A}$ é o complemento do conjunto A. Por fim, a acurácia é a proporção total de acertos dentre positivos e negativos, ou seja, $P[(\mathbb{D} \cap \mathbb{O}) \cup (\overline{\mathbb{D}} \cap \overline{\mathbb{O}})]$.

A tabela 4.2.1 auxilia para um claro entendimento sobre as medidas de aferição da qualidade que serão utilizadas. A medida de sensibilidade é dada por $a/(a+c)$, a medida de especificidade por $d/(b+d)$ e a medida de acurácia por $(a+d)/n$.

Tabela 4.2.1: Medidas de aferição da qualidade

Método	*Outlier*		Total
	Sim	Não	
Positivo	a (verdadeiros positivos)	b (falsos positivos)	$a+b$ (positivos)
Negativo	c (falsos negativos)	d (verdadeiros negativos)	$c+d$ (negativos)
Total	$a+c$ (*outliers*)	$b+d$ (não *outliers*)	$a+b+c+d$ (n)

Experimentos foram executados com dados p-variados simulados, com $p \in \{5, 10, 20, 50, 100\}$, taxas de mistura $\delta \in \{0; 0{,}02; 0{,}05; 0{,}10\}$, tamanho amostral $n = 500$ e gerados com uma matriz de covariâncias sem correlações fixas, mas que garanta todas as correlações tais que $0{,}3 \leqslant \rho \leqslant 0{,}6$. Para cada um desses cenários de investigação foram geradas 30 réplicas.

O ambiente de execução para realização dos experimentos computacionais foi um Intel(R) Core(TM) i7-3720QM 2,60GHz, que executa Windows 7 Professional 64 bits, com 8 GB de memória RAM e a versão 4.0.2 do software estatístico R.

A tabela 4.2.2 apresenta a comparação entre o método CAM e o CAM otimizado para taxa de mistura $\delta = 0$. Ressalta-se que neste cenário não existe medida de sensibilidade, uma vez que esse formato considera que *outliers* não foram produzidos nos dados simulados. Além disso, a não existência de medida de sensibilidade faz com que as medidas de acurácia e especificidade sejam iguais.

Tabela 4.2.2: Comparação experimental com taxa de mistura $\delta = 0$.

p	acurácia ou especificidade			
	CAM		CAM otimizado	
	média (sd)	mín - máx	média (sd)	mín - máx
5	0,6367 (0,1081)	0,1920 - 0,8900	**0,9943** (0,0494)	0,3260 - 1,0000
10	0,6447 (0,1048)	0,2080 - 0,9220	**0,9998** (0,0061)	0,8060 - 1,0000
20	0,6962 (0,0824)	0,4200 - 0,9000	**1,0000** (0,0000)	1,0000 - 1,0000
50	0,7611 (0,0683)	0,4840 - 0,9520	**0,9998** (0,0042)	0,9020 - 1,0000
100	0,7974 (0,0592)	0,6180 - 0,9420	**0,9999** (0,0018)	0,9440 - 1,0000

Conforme previsto, uma vez que para o método CAM otimizado foi escolhido o valor k que maximiza a medida de acurácia, esse método apresentou valores de acurácia superior ao método CAM em todos os cenários avaliados. A tabela 4.2.3 apresenta a comparação entre os métodos com $\delta = 0{,}02$.

Tabela 4.2.3: Comparação experimental com taxa de mistura $\delta = 0{,}02$.

p	sensibilidade			
	CAM		CAM otimizado	
	média (sd)	mín - máx	média (sd)	mín - máx
5	**0,9652** (0,0930)	0,3000 - 1,0000	0,7889 (0,3886)	0,0000 - 1,0000
10	**0,9982** (0,0147)	0,8000 - 1,0000	0,9799 (0,1314)	0,0000 - 1,0000
20	**1,0000** (0,0000)	1,0000 - 1,0000	0,9984 (0,0325)	0,0000 - 1,0000
50	**1,0000** (0,0000)	1,0000 - 1,0000	**1,0000** (0,0000)	1,0000 - 1,0000
100	**1,0000** (0,0000)	1,0000 - 1,0000	**1,0000** (0,0000)	1,0000 - 1,0000
p	especificidade			
	CAM		CAM otimizado	
	média (sd)	mín - máx	média (sd)	mín - máx
5	0,8253 (0,1008)	0,4327 - 1,0000	**0,9801** (0,0776)	0,0000 - 1,0000
10	0,8387 (0,0828)	0,5388 - 0,9959	**0,9988** (0,0148)	0,6918 - 1,0000
20	0,8724 (0,0609)	0,6388 - 0,9939	**0,9998** (0,0046)	0,8735 - 1,0000
50	0,9026 (0,0485)	0,7714 - 0,9980	**1,0000** (0,0000)	1,0000 - 1,0000
100	0,9130 (0,0421)	0,7939 - 0,9980	**1,0000** (0,0000)	1,0000 - 1,0000
p	acurácia			
	CAM		CAM otimizado	
	média (sd)	mín - máx	média (sd)	mín - máx
5	0,8281 (0,0992)	0,4440 - 1,0000	**0,9762** (0,0771)	0,0200 - 1,0000
10	0,8419 (0,0812)	0,5480 - 0,9960	**0,9984** (0,0155)	0,6960 - 1,0000
20	0,8749 (0,0597)	0,6460 - 0,9940	**0,9998** (0,0046)	0,8740 - 1,0000
50	0,9046 (0,0475)	0,7760 - 0,9980	**1,0000** (0,0000)	1,0000 - 1,0000
100	0,9147 (0,0412)	0,7980 - 0,9980	**1,0000** (0,0000)	1,0000 - 1,0000

Para investigações com taxa de mistura $\delta = 0{,}02$, a metodologia CAM apresenta alguma superioridade para a medida de sensibilidade. Entretanto, à medida que o número de variáveis aumenta essa vantagem é reduzida. Já para as análises das medidas de especificidade e acurária, a metodologia CAM apresenta desvantagens para todos os cenários sob avaliação. Novamente, resultados de melhoria que são decorrentes de uma escolha mais adequada para a quantidade de agrupamentos k se tornam mais aparentes.

Simplesmente verificar uma melhoria na medida de acurácia não é uma constatação prática efetiva. A constatação de fato efetiva é verificar a significativa margem de avanço na medida de acurácia para o processo de detecção. Esse avanço ocorrerá se for possível determinar uma estratégia de escolha para o valor k melhor planejada e não puramente casual como no método CAM original. A tabela 4.2.4 apresenta a comparação entre os métodos com $\delta = 0{,}05$.

Tabela 4.2.4: Comparação experimental com taxa de mistura $\delta = 0{,}05$.

p	sensibilidade			
	CAM		CAM otimizado	
	média (sd)	mín - máx	média (sd)	mín - máx
	0,9691 (0,0742)	0,2000 - 1,0000	0,9306 (0,2090)	0,0000 - 1,0000
10	**0,9975** (0,0161)	0,7600 - 1,0000	0,9908 (0,0846)	0,0000 - 1,0000
20	**1,0000** (0,0000)	1,0000 - 1,0000	0,9989 (0,0317)	0,0000 - 1,0000
50	**1,0000** (0,0000)	1,0000 - 1,0000	**1,0000** (0,0000)	1,0000 - 1,0000
100	**1,0000** (0,0000)	1,0000 - 1,0000	**1,0000** (0,0000)	1,0000 - 1,0000
p	especificidade			
	CAM		CAM otimizado	
	média (sd)	mín - máx	média (sd)	mín - máx
5	0,8761 (0,0846)	0,5137 - 1,0000	**0,9842** (0,0744)	0,0000 - 1,0000
10	0,8879 (0,0788)	0,6021 - 1,0000	**0,9990** (0,0140)	0,6926 - 1,0000
20	0,9135 (0,0607)	0,6926 - 1,0000	**1,0000** (0,0015)	0,9516 - 1,0000
50	0,9336 (0,0490)	0,7411 - 1,0000	**1,0000** (0,0000)	1,0000 - 1,0000
100	0,9419 (0,0431)	0,7747 - 1,0000	**1,0000** (0,0000)	1,0000 - 1,0000
p	acurácia			
	CAM		CAM otimizado	
	média (sd)	mín - máx	média (sd)	mín - máx
5	0,8808 (0,0817)	0,5340 - 1,0000	**0,9815** (0,0722)	0,0500 - 1,0000
10	0,8934 (0,0749)	0,6220 - 1,0000	**0,9986** (0,0153)	0,6920 - 1,0000
20	0,9178 (0,0577)	0,7080 - 1,0000	**0,9999** (0,0030)	0,9040 - 1,0000
50	0,9369 (0,0465)	0,7540 - 1,0000	**1,0000** (0,0000)	1,0000 - 1,0000
100	0,9448 (0,0409)	0,7860 - 1,0000	**1,0000** (0,0000)	1,0000 - 1,0000

Para taxa de mistura $\delta = 0{,}05$, o efeito relatado anteriormente se repete, o método CAM apresenta uma superioridade bastante sutil para a medida sensibilidade. Entretanto, à medida que o número de variáveis aumenta essa vantagem desaparece. Já para as análises das medidas de especificidade e acurária, o método CAM mostra desvantagens em todos os cenários avaliados.

Novamente, a investigação destes resultados permite verificar capacidade de evolução na medida de acurácia para o processo de detecção se a escolha do valor k for mais adequada. A tabela 4.2.5 apresenta a comparação entre os métodos com $\delta = 0{,}10$.

Para $\delta = 0{,}10$, novamente, o método CAM apresenta uma superioridade marginal para a medida sensibilidade. À medida que o número de variáveis aumenta essa vantagem desaparece. Para as análises de especificidade e acurária, o método CAM é inferior, mostra desvantagens e novamente as margens de melhoria decorrente de uma escolha adequada para o valor k se tornam aparentes.

Em uma comparação dos resultados apresentados nas tabelas 4.2.3, 4.2.4 e 4.2.5 nota-se uma grande similaridade nos valores obtidos pela metodologia CAM otimizado. Isso mostra que o aumento da taxa de mistura δ de 0,02 até 0,10 parece ter efeito pequeno. Entretanto, para o método CAM, o

Tabela 4.2.5: Comparação experimental com taxa de mistura $\delta = 0{,}10$.

var	sensibilidade			
	CAM		CAM otimizado	
	média (sd)	min - max	média (sd)	min - max
5	**0,9574** (0,0829)	0,4800 - 1,0000	0,9402 (0,1727)	0,0000 - 1,0000
10	**0,9978** (0,0164)	0,7200 - 1,0000	0,9963 (0,0382)	0,1200 - 1,0000
20	**1,0000** (0,0000)	1,0000 - 1,0000	0,9999 (0,0025)	0,9200 - 1,0000
50	**1,0000** (0,0000)	1,0000 - 1,0000	**1,0000** (0,0000)	1,0000 - 1,0000
100	**1,0000** (0,0000)	1,0000 - 1,0000	**1,0000** (0,0000)	1,0000 - 1,0000
var	especificidade			
	CAM		CAM otimizado	
	média (sd)	min - max	média (sd)	min - max
5	0,9153 (0,0806)	0,4244 - 1,0000	**0,9883** (0,0446)	0,5200 - 1,0000
10	0,9394 (0,0596)	0,6667 - 1,0000	**0,9995** (0,0087)	0,7911 - 1,0000
20	0,9560 (0,0438)	0,6778 - 1,0000	**1,0000** (0,0000)	1,0000 - 1,0000
50	0,9707 (0,0292)	0,8111 - 1,0000	**1,0000** (0,0000)	1,0000 - 1,0000
100	0,9739 (0,0303)	0,7800 - 1,0000	**1,0000** (0,0000)	1,0000 - 1,0000
var	acurácia			
	CAM		CAM otimizado	
	média (sd)	min - max	média (sd)	min - max
5	0,9195 (0,0762)	0,470 - 1,0000	**0,9835** (0,0461)	0,5580 - 1,0000
10	0,9452 (0,0538)	0,698 - 1,0000	**0,9992** (0,0105)	0,7540 - 1,0000
20	0,9604 (0,0394)	0,710 - 1,0000	**1,0000** (0,0003)	0,9920 - 1,0000
50	0,9737 (0,0263)	0,830 - 1,0000	**1,0000** (0,0000)	1,0000 - 1,0000
100	0,9765 (0,0273)	0,802 - 1,0000	**1,0000** (0,0000)	1,0000 - 1,0000

aumento da taxa de mistura tende a mostrar uma melhora de desempenho, uma instabilidade que pode ser decorrente da maior presença de valores *outliers*.

4.3 Calibração do método DDCAM

Uma vez delimitado que o método CAM pode, de fato, passar por estratégias de melhorias para se tornar mais adaptativo aos conjuntos de dados sob investigação, uma avaliação experimental efevetiva dos possiveis efeitos deve ser discutida.

Todos os parâmetros inerentes ao funcionamento do método DDCAM serão avaliados a seguir. O intuito é encontrar uma calibração adequada, que seja eficaz e eficiênte para a maioria dos conjuntos de dados sob investigação.

4.3.1 Comparação entre os possíveis estimadores $\hat{\delta}$

O objetivo inicial foi comparar as duas propostas de estimativa para a taxa de mistura δ. As duas propostas de estimadores podem ser observadas nas equações 3.1, que utiliza uma média de estimativas univariadas, e 3.2,

que utiliza o máximo dentre estimativas univariadas. Experimentos foram executados com dados p-variados simulados, com $p \in \{10, 25, 50, 75, 100\}$, as taxas de mistura foram $\delta \in \{0; 0{,}01; 0{,}025; 0{,}04; 0{,}07\}$, os tamanhos amostrais foram $n \in \{50, 100, 200, 300, 400, 500\}$ e os dados foram gerados com uma matriz de covariâncias sem correlações fixas, mas de forma que garanta todas as correlações tais que $0{,}3 \leqslant \rho \leqslant 0{,}6$. Para todos os cenários foram geradas 1000 réplicas.

A figura 4.3.1 apresenta a comparação através do método DDCAM para os dois estimadores segundo a medida de acurácia. Vale ressaltar que, para as 1000 execuções, a medida de acurácia foi calculada e o gráfico apresenta a acurácia média verificada. Para este estudo, o método DDCAM utilizou o parâmetro η igual a 2, o parâmetro ϕ igual a 2,5 (como na versão original do CAM), já o valor k segue a proposição descrita nas subseções 3.1.2, 3.1.3 e 3.1.4.

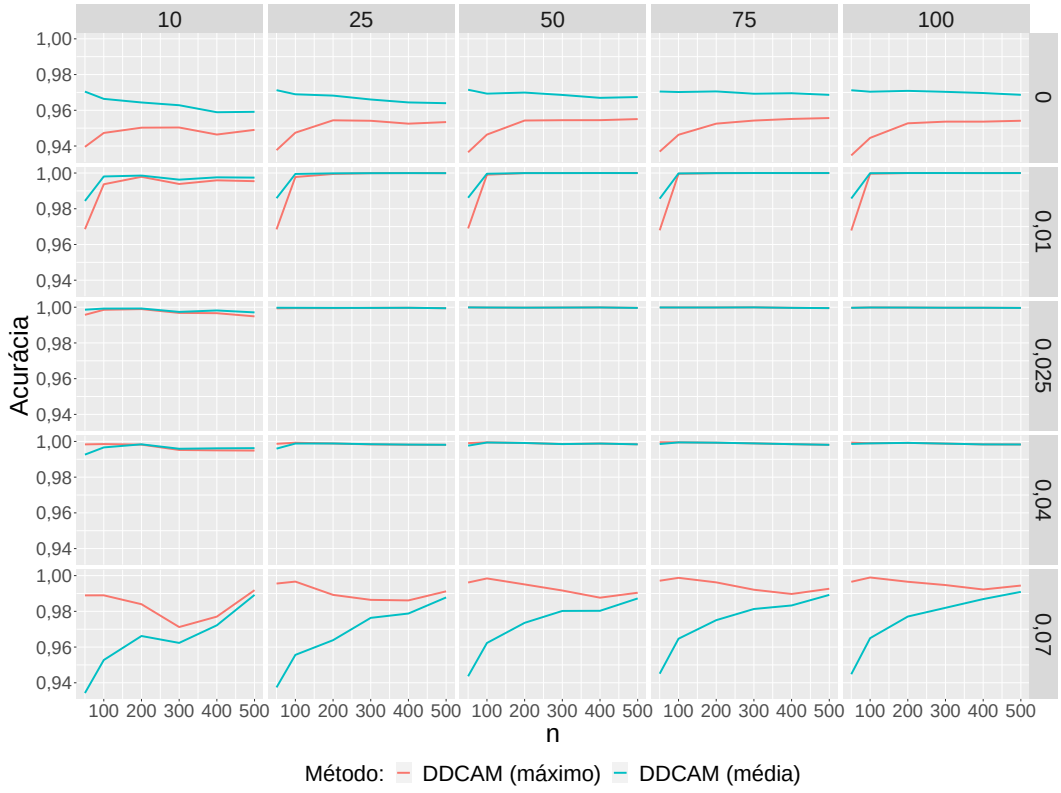

Figura 4.3.1: Análise de acurácia para a escolha do estimador $\hat{\delta}$ adequado.

A figura 4.3.1 mostra em suas colunas as quantidades de variáveis no banco de dados simulado, ou seja, o valor de p para dados p-variados, e ainda por linhas, as taxas de mistura utilizadas na simulação, ou seja, os valores de δ utilizados na simulação. Essa lógica será utilizada para as demais figuras na investigação de escolha do estimador adequado. Para taxa de mistura nula, o estimador através da média se comportou melhor. À medida que a taxa de mistura cresce, os desempenhos se tornam iguais. Entretanto, quando a taxa de mistura alcança o valor 0,07 o estimador através do máximo se torna mais efetivo.

A análise mostra algum equilibrio entre os dois estimadores, uma vez que mesmo nos piores casos as medidas verificadas para acurácia são elevadas. O aumento da quantidade de variáveis não revelou efeitos significativos nessa análise. A figura 4.3.2 mostra a comparação através do método DDCAM para os dois estimadores, agora segundo a medida de especificidade. Assim como na análise da acurácia, dentre as 1000 execuções, a especificidade foi calculada e o gráfico apresenta a especificidade média verificada.

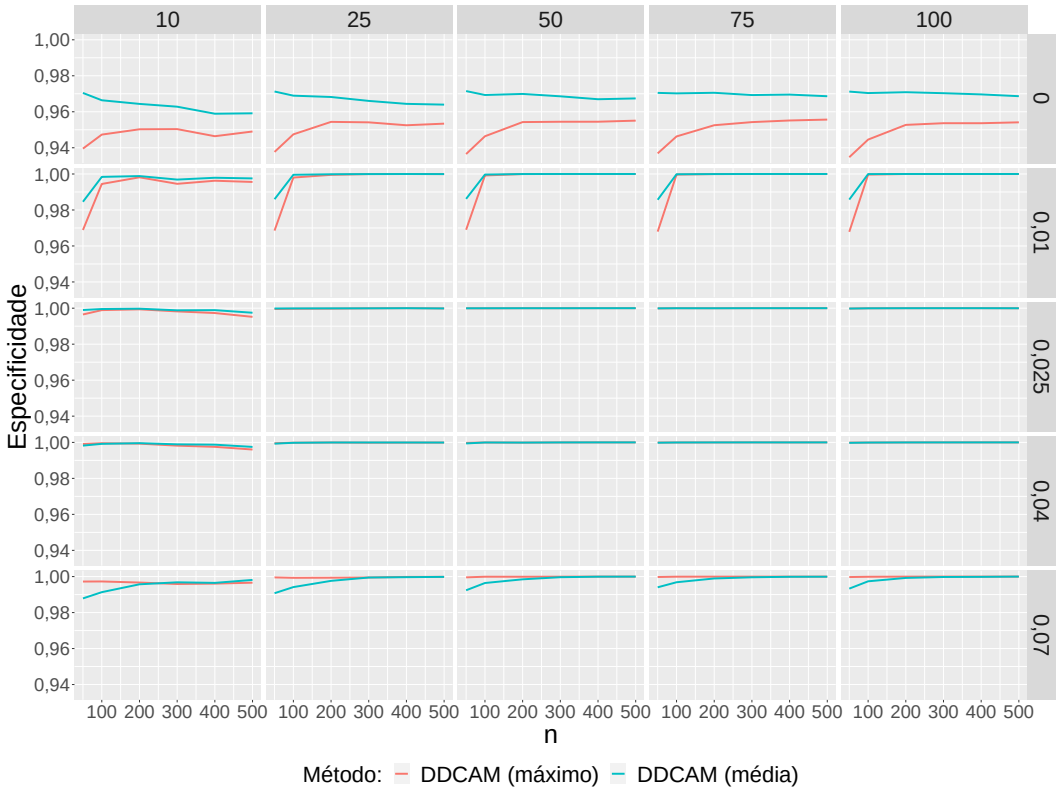

Figura 4.3.2: Análise de especificidade para a escolha do estimador $\hat{\delta}$ adequado.

Na análise de especificidade, os gráficos para $\delta = 0$, são coincidentes com os de acurácia para $\delta = 0$. Sempre idênticas. Para as situações em que a taxa de mistura é não nula, o desempenho é bastante similar entre os dois estimadores. Apenas para a taxa de mistura $\delta = 0{,}07$ uma superioridade nos resultados através do estimador baseado no máximo é verificada. A variação do número de variáveis não apresentou sinais efetivos nessa análise.

A figura 4.3.3 apresenta um comparativo entre os dois estimadores propostos, agora através da medida de sensibilidade. Os resultados através do método DDCAM com cada um dos estimadores propostos são utilizados. Novamente adota-se a medida de sensibilidade média entre as 1000 simulações.

Na avaliação de sensibilidade, não existem gráficos para $\delta = 0$. Com a não existência de *outliers* produzidos na amostra simulada, não há como medir sensibilidade no processo de detecção. Para as situações com $\delta \neq 0$,

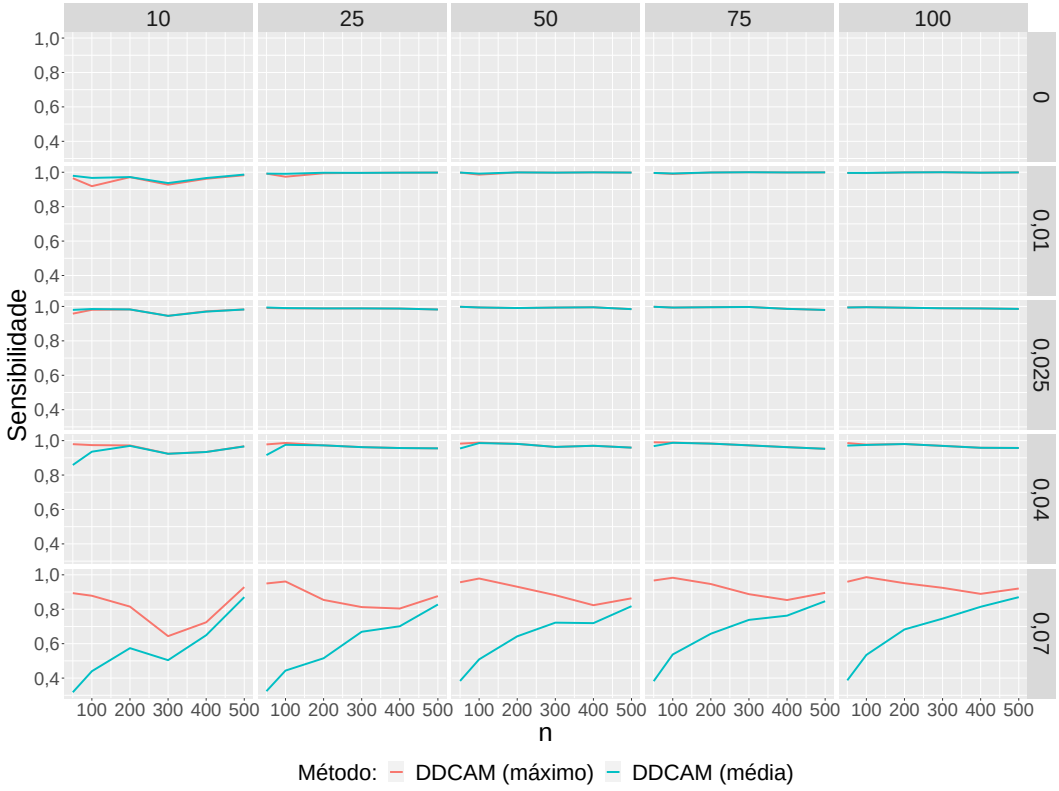

Figura 4.3.3: Análise de sensibilidade para a escolha do estimador $\hat{\delta}$ adequado.

os resultados são bastante próximos a menos do caso com $\delta = 0{,}07$. Para a taxa de mistura $\delta = 0{,}07$ uma superioridade marcante para o estimador através do máximo é verificada. Este efeito de superioridade aumenta à medida que o número de variáveis cresce. Para $p = 10$ o pior resultado para o estimador do máximo ocorreu quando $n = 300$, mas ainda com sensibilidade superior a 0,60. À medida que p cresce, os resultados com o estimador do máximo são superiores a 0,80. Em contrapartida, através do estimador da média, os resultados são bem inferiores.

Essa investigação não exclui por completo a possibilidade de utilização do estimador através da média. Por outro lado, o estimador através do máximo se mostrou mais estável e garantiu maior qualidade de resultados através das medidas comparativas utilizadas. Diante desse resultado, nos experimentos posteriormente executados, o método DDCAM sempre utilizará o estimador baseado no máximo apresentado na equação 3.2.

4.3.2 Comparação do efeito do parâmetro ϕ na utilização do método DDCAM

Aqui será apresentada uma análise para verificar o comportamento da técnica com a variação do valor de ϕ. Novamente os experimentos foram executados com dados p-variados simulados, com $p \in \{10, 25, 50, 75, 100\}$, $\delta \in \{0; 0{,}01; 0{,}025; 0{,}04; 0{,}10\}$, $n \in \{50, 100, 200, 300, 400, 500\}$ e os dados foram gerados com uma matriz de covariâncias sem correlações fixas, mas

de forma que garanta todas as correlaçoes tais que $0{,}3 \leqslant \rho \leqslant 0{,}6$. Para todos os cenários sob investigação, foram geradas 1000 réplicas através de simulações de Monte Carlo.

As figuras 4.3.4, 4.3.5 e 4.3.6 apresentam a comparação entre os valores de ϕ, tais que $\phi \in \{2{,}0; 2{,}5; 3{,}0\}$. Nas colunas as figuras mostram as quantidades de variáveis nos dados simulados e nas linhas os valores de δ (taxa de mistura) utilizados na simulação. Assim como na investigação para os estimadores, para as 1000 execuções, a acurácia, especificidade e sensibilidade foram calculadas e os gráficos apresentam valores médios verificados dentre estas três medidas.

O método DDCAM utilizou o parâmetro η igual a 2, o valor k já segue a proposição descrita nas subseções 3.1.2, 3.1.3 e 3.1.4 e o estimador através do máximo foi utilizado. A figura 4.3.4 apresenta análise para a medida de acurácia.

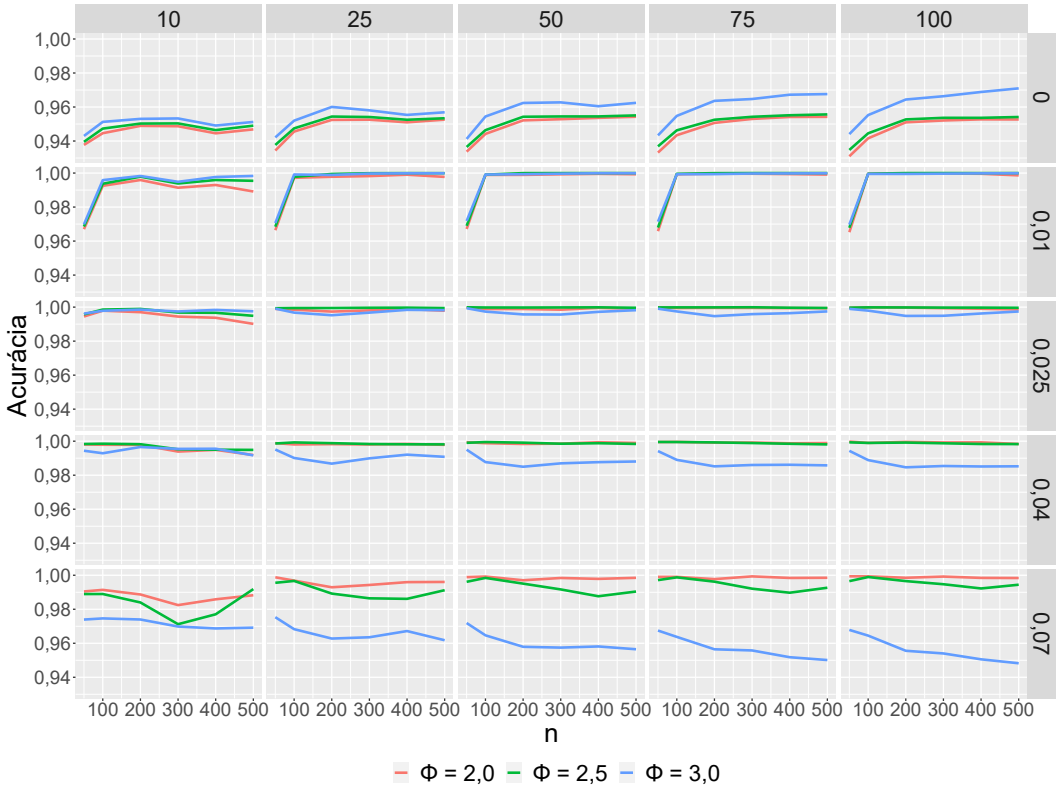

Figura 4.3.4: Avaliação comparativa de acurácia entre valores de ϕ.

Para taxa de mistura nula, os resultados com $\phi = 3{,}0$ foram ligeiramente superiores, mas à medida que a taxa de mistura cresce, essa superioridade desaparece. Por fim, para $\delta = 0{,}07$ os resultados com $\phi = 3{,}0$ são claramente piores. Os resultados para $\phi = 2{,}0$ e $\phi = 2{,}5$ são bastante semelhantes ao longo dos valores da taxa de mistura δ utilizados. O aumento do número de variáveis também não apresentou nenhum impacto semelhante nessa análise. Isso já havia sido verificado na comparação anterior entre os dois estimadores. A figura 4.3.5 apresenta resultados para a medida de especificidade.

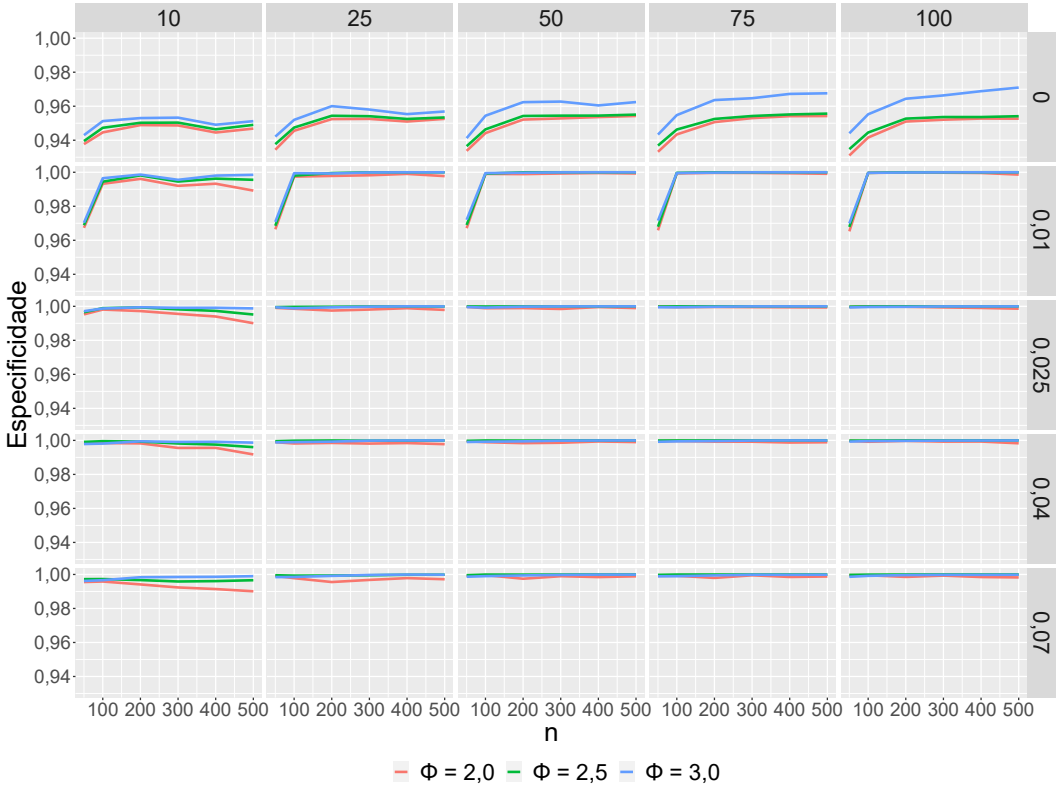

Figura 4.3.5: Avaliação comparativa de especificidade entre valores de ϕ.

Como verificado na avaliação prévia de acurácia, para a análise de especificidade, os resultados com $\phi = 3{,}0$ foram ligeiramente superiores. Novamente ocorre o mesmo efeito, à medida que a taxa de mistura cresce essa superioridade desaparece e um claro equilibrio entre os resultados pode ser verificado.

Para os casos em que a taxa de mistura é não nula, alguma diferença entre os três valores de ϕ são notadas de forma mais evidente somente para $p = 10$, ou seja, conjuntos de dados com 10 variáveis. A menos desse efeito, para $\delta > 0$, não são apresentadas evidências claras para destacar algum valor ϕ. Vale lembrar que o caso da especificidade para $\delta = 0$ é igual ao caso avaliado previamente para acurácia com $\delta = 0$. Dessa forma, as análises novas acerca do desempenho ser afetado por ϕ neste conjunto são somente os casos com $\delta > 0$. A figura 4.3.6 apresenta resultados para sensibilidade verificada no processo de detecção.

Para análise de sensibilidade, a medida não pode ser calculada quando $\delta = 0$. Os resultados aqui deixam claro que o aumento no valor de ϕ reduz a sensibilidade do processo de detecção. Essa redução da medida de sensibilidade com o crecimento de ϕ é ainda mais impactada com o crescimento do número de variáveis do banco de dados multivariado em estudo. Em particular, para $\delta = 0{,}07$ e $p = 100$, os resultado obtidos para $\phi = 3{,}0$ são extremamente baixos.

Os resultados verificados detectam um deficiência significativa ao utilizar $\phi = 3{,}0$. Os resultados comparativos entre $\phi = 2{,}0$ e $\phi = 2{,}5$ são bem mais equilibrados, mas alguma superioridade em sensibilidade pode

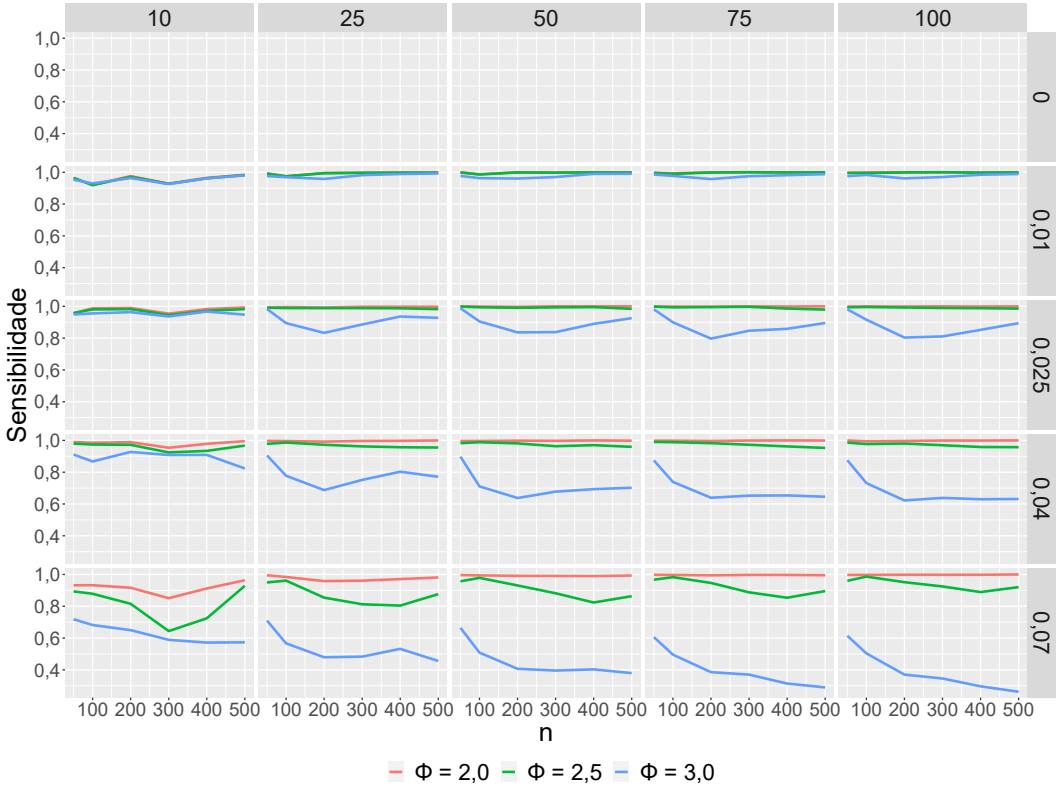

Figura 4.3.6: Avaliação comparativa de sensibilidade entre valores de ϕ.

ser observada para $\phi = 2{,}0$. Dessa forma, os experimentos posteriores adotaram este valor de ϕ para as análises seguintes.

4.3.3 Comparação do efeito do parâmetro η no utilização do método DDCAM

O valor η também tem impacto relevante nessa investigação. Como mencionado anteriormente, um formato simplista seria adotar η igual ao τ arbitrado pelo usuário do método. Para efeito das investigações simuladas, o valor τ foi fixado igual 2 e é condizente com a descrição dos dados de simulação apresentada na seção 4.1.

Diante disso, é de suma importância verificar se o valor η tem impacto significativo no procedimento de detecção de valores *outliers* através do método DDCAM. Para tanto, algumas escolhas de η que diferem de τ serão abordadas. Os experimentos foram executados com dados p-variados simulados, $p \in \{10, 25, 50, 75, 100\}$, $\delta \in \{0; 0{,}01; 0{,}025; 0{,}04; 0{,}10\}$, $n \in \{50, 100, 200, 300, 400, 500\}$ e os dados foram gerados com uma matriz de covariâncias sem correlações fixas, mas de forma que garanta todas as correlações tais que $0{,}3 \leqslant \rho \leqslant 0{,}6$. Para todos os cenários foram geradas 1000 réplicas.

As figuras 4.3.7, 4.3.8 e 4.3.9 apresentam a comparação entre os valores de η tais que $\eta \in \{1{,}5; 2{,}0; 2{,}5\}$, tais comparações são executadas para as medidas de acurácia, especificidade e sensibilidade. As figuras mostram por colunas as quantidades de variáveis nos dados simulados. Já

por linhas, os valores de δ (taxa de mistura) utilizados na simulação. Assim como na investigação para os estimadores, para as 1000 execuções, a acurácia, a especificidade e a sensibilidade foram calculadas e os gráficos apresentam valores médios verificados dentre essas três medidas. O método DDCAM utilizou o parâmetro τ em 2, portanto, foram abordados o caso $\tau = \eta = 2{,}0$; $\tau > \eta = 1{,}5$ e $\tau < \eta = 2{,}5$. A figura 4.3.7 apresenta análise em questão para a medida de acurácia.

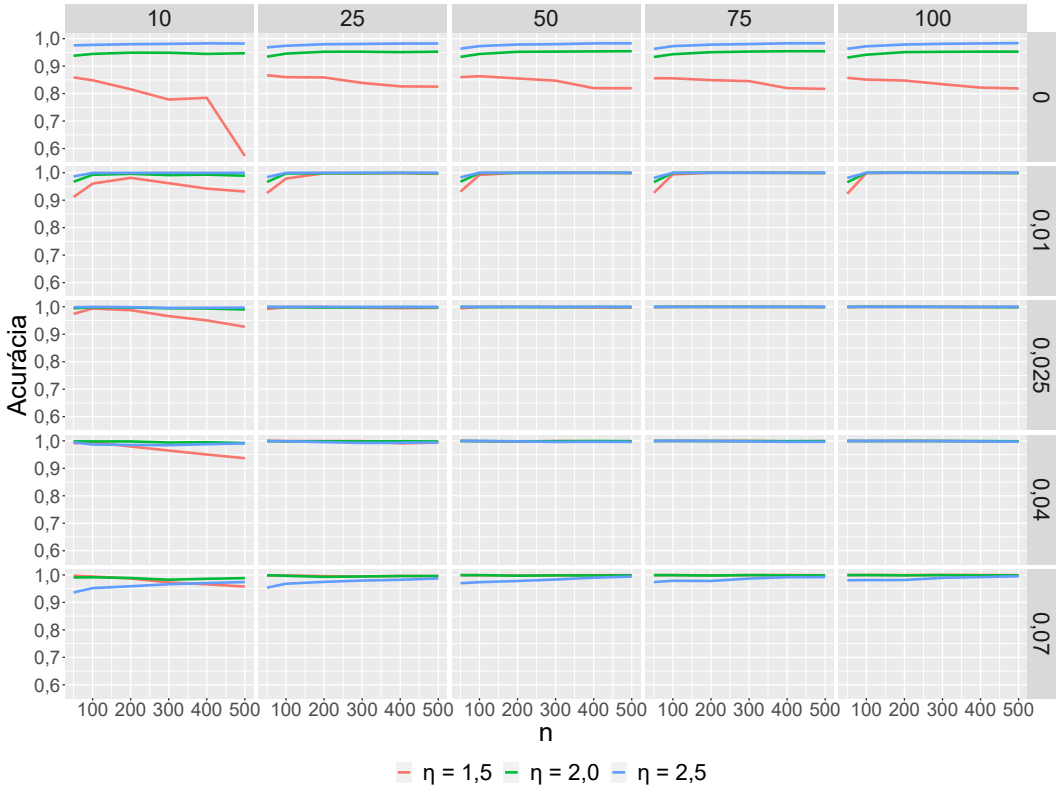

Figura 4.3.7: Estudo comparativo da acurácia para valores de η.

Para situação em que a taxa de mistura é nula, o caso $\eta < \tau$ apresenta um comportamento claramente inferior. À medida que a taxa de mistura cresce, as alterações no valor de η parecem não ter um efeito altamente impactante em termos da medida de acurácia no processo de detecção de valores *outliers*. Quando essa análise considera o número de variáveis nos dados em estudo, fica evidente que o aumento do número de variáveis também conduz para a constatação de que o efeito de η parece não ter impacto elevado em termos da medida de acurácia no processo de identificação de valores *outliers*.

Além disso, a figura 4.3.7 sugere que o comportamento com $\eta > \tau$ foi ligeiramente superior para taxa de mistura nula. Por outro lado, trata-se de uma superioridade marginal, os valores de acurácia média quando $\eta = \tau$ são também bastante elevados.

A figura 4.3.8 apresenta resultados para a medida de especificidade. Como verificado na avaliação da medida de acurácia, para o estudo baseado na medida de especificidade, os resultados com $\eta < \tau$ foram inferiores. Novamente, à medida que a taxa de mistura cresce e também que o

número de variáveis nos dados sob investigação aumentam η parece não revelar impacto decisivo.

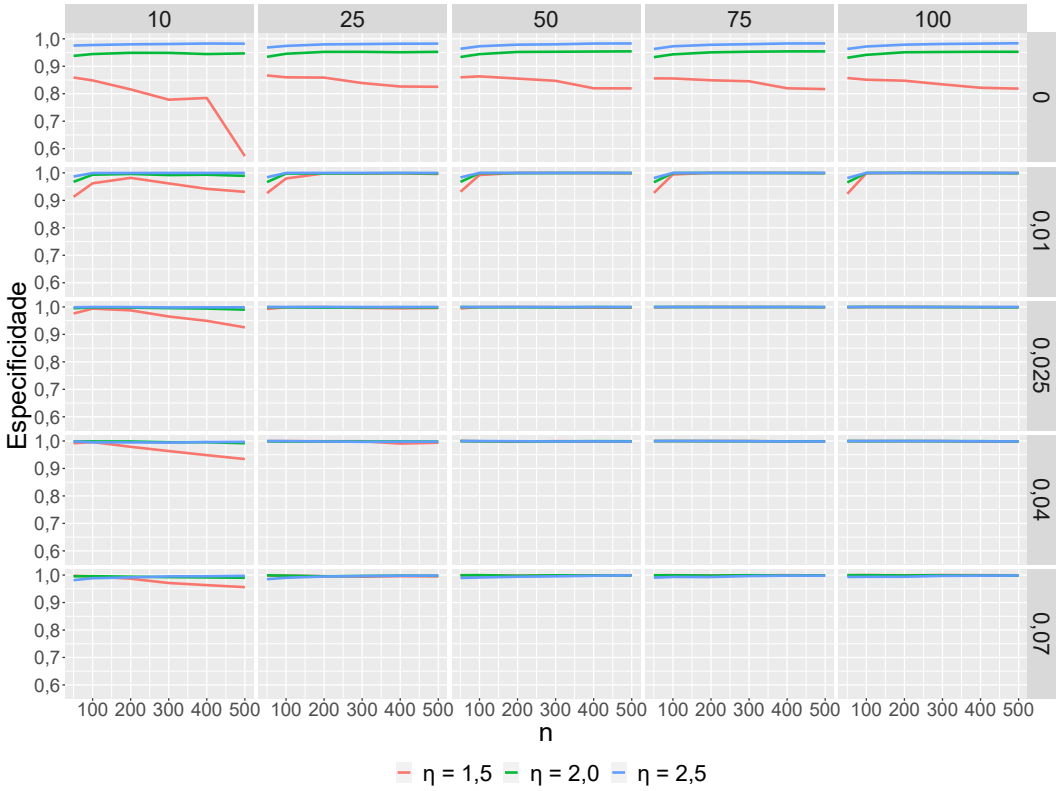

Figura 4.3.8: Estudo comparativo de especificidade para valores de η.

Como verificado anteriormente, a figura 4.3.8 parece sugerir alguma superioridade para o cenário em que $\eta > \tau$, como anteriormente verificado na figura 4.3.7. Entretanto, os valores de especificidade média obtidos para $\eta = 2{,}0$ são elevados e bastante competitivos aqui.

Por fim, são apresentados na figura 4.3.9 os valores para a sensibilidade média verificada no processo de detecção de valores *outliers*. Para análise de sensibilidade, a medida não pode ser calculada quando $\delta = 0$, em razão da não existência de valores *outliers* gerados no procedimento de simulação. Os resultados aqui apresentados ilustram bem a perda de performance para $\eta = 2{,}5$ à medida que a taxa de mistura aumenta. Esse efeito piora sensivelmente a sensibilidade do processo de detecção de valores *outliers* multivariados.

A configuração com $\eta = 2{,}0$ apresentou alguma flutuação para taxas de misturas elevadas combinadas com baixos números de variáveis, mas não foram reveladas medidas de sensibilidade baixas o suficiente para coibir a utilização de $\eta = 2{,}0$.

De uma forma geral, a investigação apresentada confirmou que a escolha $\eta = 2{,}0$ parece muito bem ajustada para o funcionamento do método. A elevação dos valores de η conduz para uma perdade de sensibilidade no processo de detecção de *outliers*. Ao passo que reduzir o valor de η leva para perdas nas medidas de acurácia e especificidade. Isso ocorre princi-

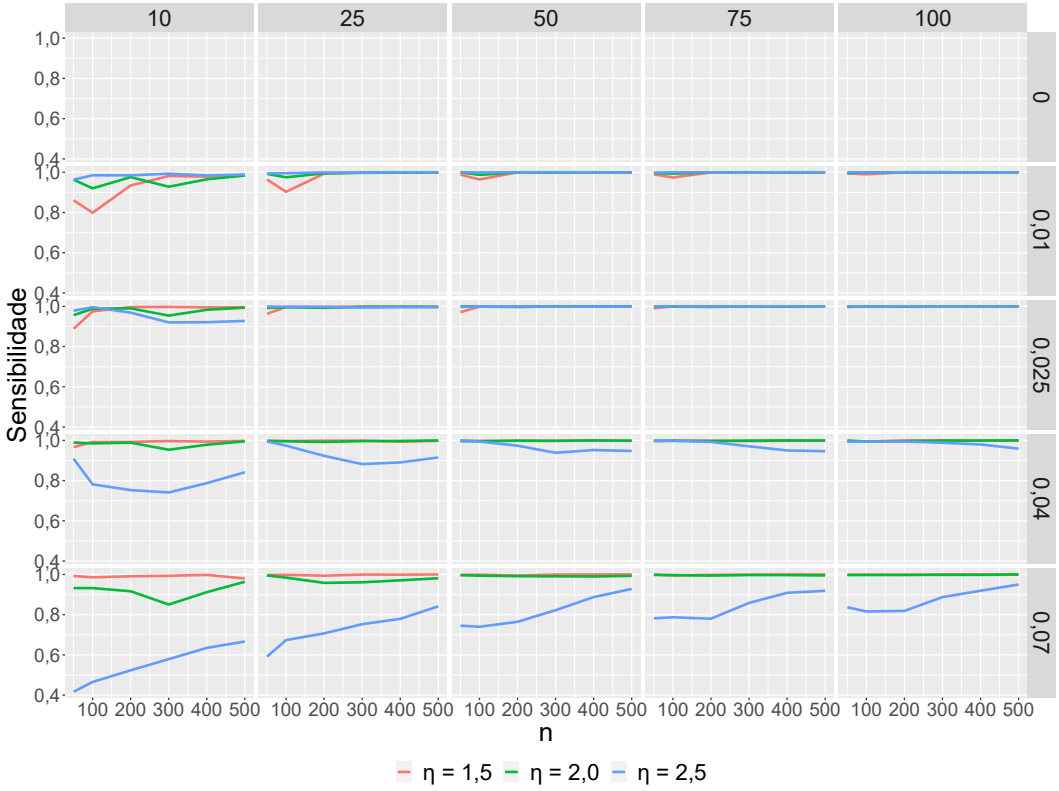

Figura 4.3.9: Estudo comparativo de sensibilidade para valores de η.

palmente em situações de não existência de valores *outliers*, ou seja, taxa de mistura nula.

Desse ponto em diante, entede-se que o método esteja adequadamente calibrado em termos de seus parâmetros operacionais. A sugestão é que o estimador utilizado seja o estimador baseado no máximo, que foi apresentado na equação 3.2. Além disso, que o parâmetro ϕ seja fixado em 2,0; que o parâmetro η seja fixado em 2,0 e ainda, que o parâmetro τ seja calibrado pelo usuário da metodologia de acordo com seu conhecimento sobre as variáveis envolvidas.

4.4 O efeito do parâmetro τ

Após a calibração adequada dos parâmetros discutidos anteriormente, que são inerentes ao funcionamento do método DDCAM, alguma indagação pode ser feita acerca do parâmetro τ. Desde a formulação da metodologia, foi estabelecido como adequado que τ seja especificado diretamente pelo usuário do método. Entretanto, uma pergunta natural é, se o usuário especifica muito mal o valor de τ, a qualidade do método é impactada expressivamente?

Será apresentado um estudo comparativo com respeito à variação da especificação de τ. O objetivo é garantir que obviamente τ tem efeito no procedimento de detecção, mas não um efeito altamente danoso para o método. Algumas constatações prévias são quase óbvias. Um valor τ mi-

norado, em comparação à verdade dos dados, tende a produzir algum confundimento entre os possíveis valores *outliers* e o restante dos dados. Este efeito tende a aumentar o volume de dados em cada um dos agrupamentos gerados e reduzir alguma capacidade do método, mas não de forma muito representativa.

Por outro lado, um valor τ majorado, quando comparado à população, pode dificultar a chance de identificar corretamente os valores *outliers* para qualquer método. Isso porque alguns possíveis *outliers* não seriam considerados suficientemente afastados, com respeito à distância para a média, por qualquer metodologia. Em particular, para a abordagem desse estudo, existiria alguma tendência de produzir agrupamentos muito pequenos (em volume de dados), isso dificultaria de alguma forma o método, e levaria à exclusão de muitos valores k que deveriam ser investigados no processo de refinamento.

Os experimentos foram executados novamente com dados p-variados simulados, com $p \in \{10, 25, 50, 75, 100\}$, $\delta \in \{0; 0{,}01; 0{,}025; 0{,}04; 0{,}10\}$, $n \in \{50, 100, 200, 300, 400, 500\}$ e os dados foram gerados com uma matriz de covariâncias sem correlações fixas, mas de forma que garanta todas as correlações tais que $0{,}3 \leqslant \rho \leqslant 0{,}6$. Para todos os cenários foram geradas 1000 réplicas. A configuração do método foi fixada na calibração previamente discutida. A figura 4.4.1 apresenta análise de desempenho através da medida de acurácia.

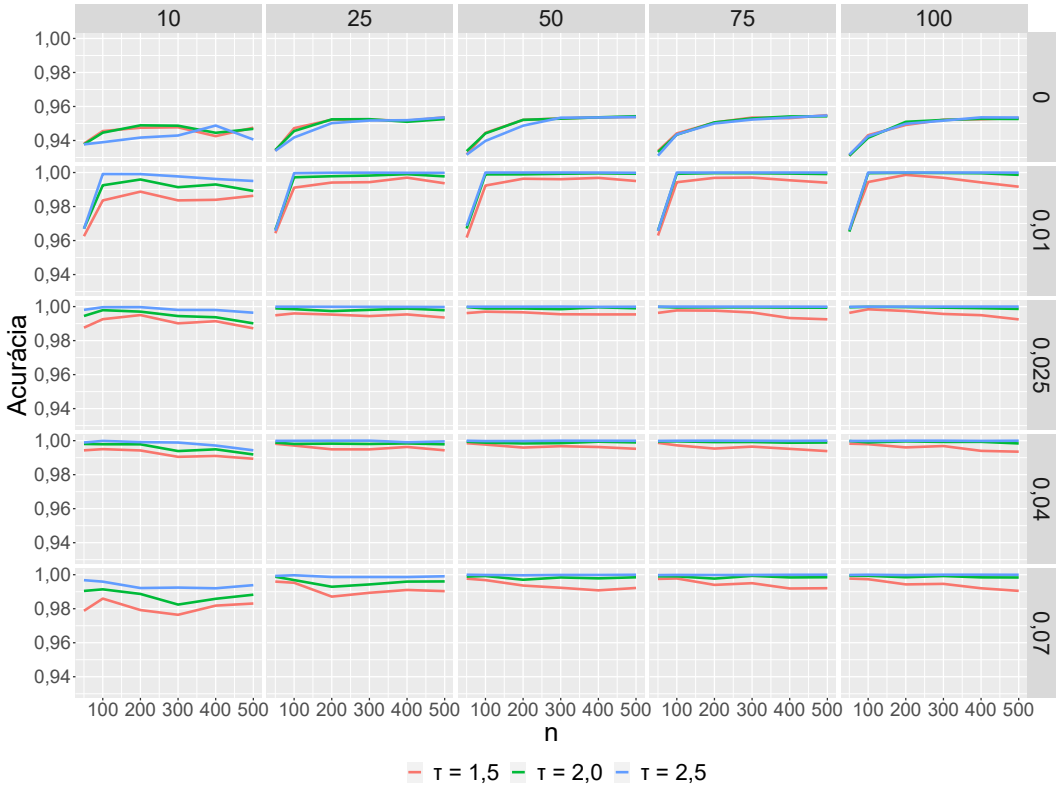

Figura 4.4.1: Gráfico de acurácia para a calibração de τ.

São apresentadas por colunas as quantidades de variáveis nos dados simulados e por linhas, os valores de δ utilizados na simulação. Como nas

avaliações de calibração da metodolgoia DDCAM, a acurácia, a especificidade e a sensibilidade foram calculadas e os gráficos apresentam valores médios verificados dentre estas três medidas. Como previamente previsto, não existem flutuações altamente significativas que levariam a acreditar que variações inadequadas na escolha de τ tenham um impacto altamente prejudicial. A figura 4.4.2 apresenta resultados avaliativos para a medida de especificidade.

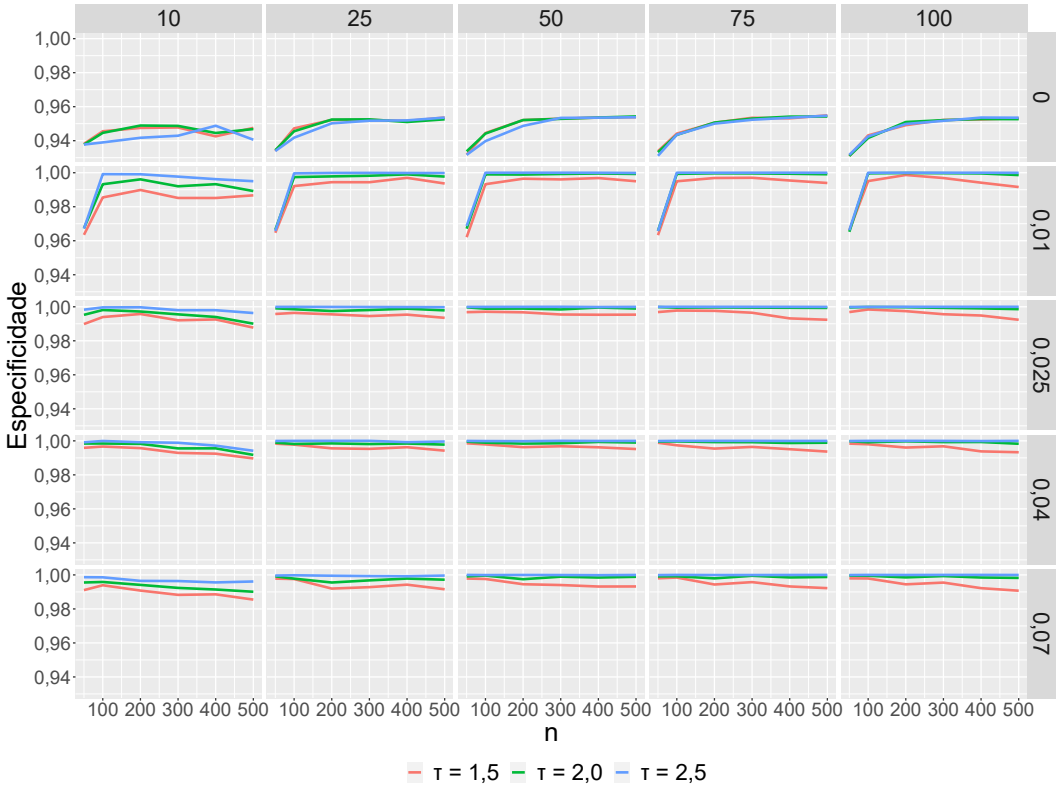

Figura 4.4.2: Gráfico de especificidade para a calibração de τ.

Os resultados de especificidade são bastante similares aos resultados de acurácia verificados na figura 4.4.1. Não existe nenhum efeito representativo de perda grande de performance em especificidade que seja decorrente de alguma flutuação na escolha do valor de τ. Novamente os resultados são menos estáveis para taxas de mistura menores. Não parece haver diferença efetiva entre as possíveis escolhas de valores para τ que foram investigadas. A figura 4.4.3 apresenta resultados para a medida de sensibilidade.

A comparação apresentada mostra alguma variabilidade na resposta de sensibilidade do processo de detecção de valores *outliers*. Isso ocorre quando o número de variáveis envolvidas no problema é menor. Porém, mesmo para os piores cenários, ou seja, quando $\tau = 1{,}5$ e o número de variáveis é 10, ainda assim, os piores valores de sensibilidade são superiores ao patamar de 0,80.

A investigação dos efeitos decorrentes da escolha do valor τ pelo usuário, mostram que de fato ela é relevante. O conhecimento prévio do usuário do método de detecção de *outliers* acerca das variáveis não pode ser desprezado. Porém, o método é suficientemente robusto quanto a isto para

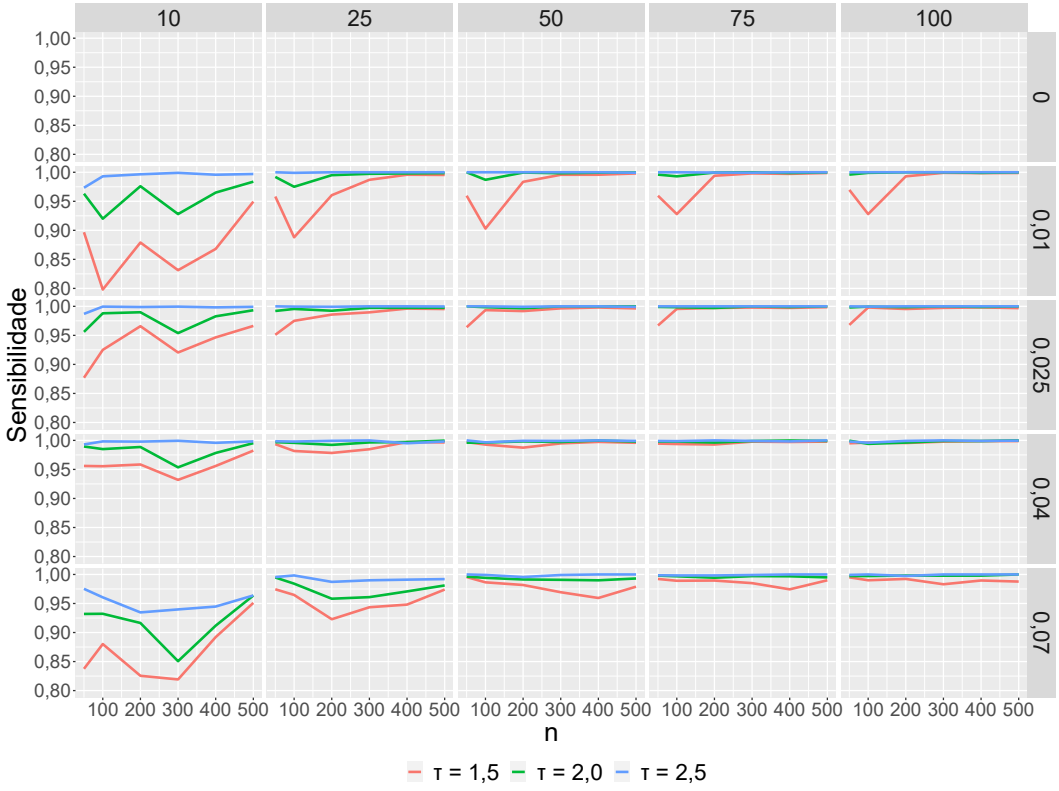

Figura 4.4.3: Gráfico de sensibilidade para a calibração de τ.

preservar sua estabilidade mesmo com escolhas que tragam algum grau de incorreção.

4.5 Análises comparativas entre o método CAM e o método DDCAM

A investigação de Barbosa et al. (2020) esclareceu que o método CAM deixava diversas margens para evolução. Em particular, captar informações inerentes ao banco de dados, que fosse capazes de gerar alguma natureza de aprendizado para o método estudar o respectivo banco de dados. Além disso, estabelecer um procedimento de escolha da quantidade de agrupamentos k que seja mais eficiente e não meramente casual como no método CAM.

A proposição da metodologia DDCAM, agora completamente definida e calibrada, carece de uma comparação direta com o método prévio CAM, para verificar se de fato ocorreu algum ganho efetivo com as alterações propostas neste estudo. Para essa comparação, experimentos foram executados com dados p-variados simulados, com $p \in \{10, 25, 50, 75, 100\}$, taxas de mistura $\delta \in \{0; 0{,}01; 0{,}025; 0{,}04; 0{,}10\}$, tamanhos amostrais $n \in \{50, 100, 200, 300, 400, 500\}$ e os dados foram gerados com uma matriz de covariâncias sem correlações fixas, mas de forma que garanta todas as correlações tais que $0{,}3 \leqslant \rho \leqslant 0{,}6$. Para todos os cenários foram geradas 1000

réplicas.

As figuras 4.5.1, 4.5.2 e 4.5.3 apresentam a comparação entre os dois métodos. As figuras apresentam nas colunas as quantidades de variáveis nos dados simulados e nas linhas os valores de δ (taxa de mistura) utilizados na simulação. Para as 1000 execuções, a acurácia, a especificidade e a sensibilidade foram calculadas. Os gráficos apresentam valores médios verificados dentre estas três medidas. A figura 4.5.1 mostra análise para a medida de acurácia.

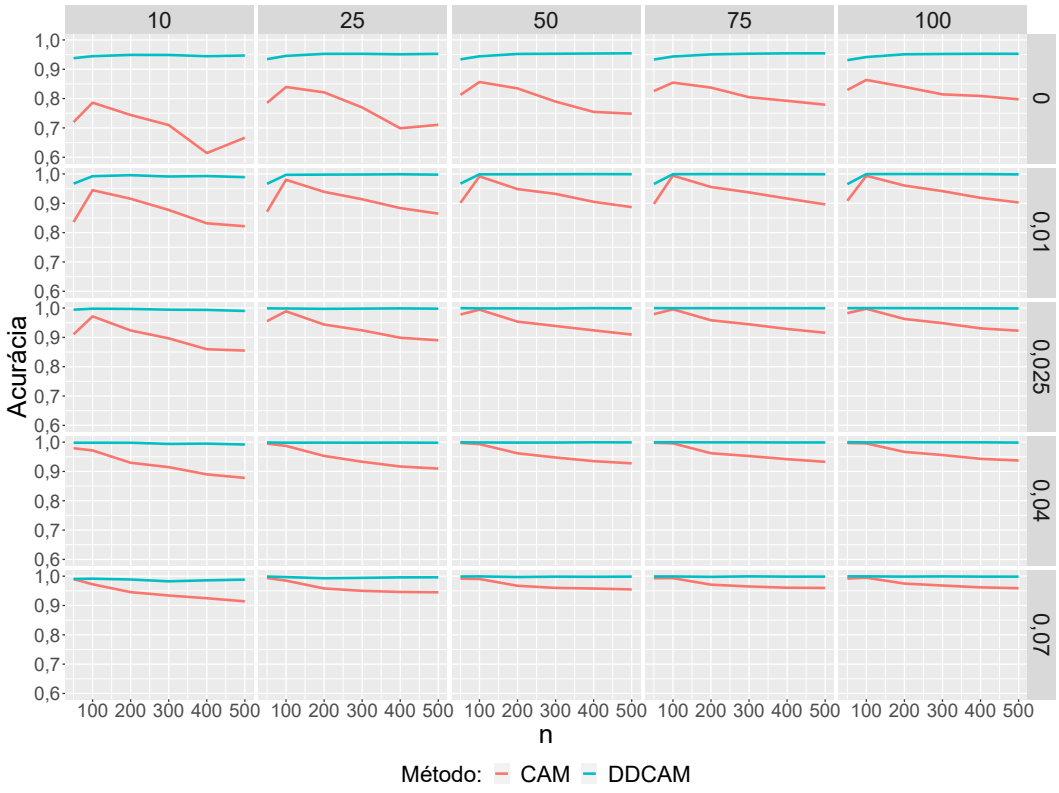

Figura 4.5.1: Comparação de acurácia entre CAM e DDCAM.

A análise da medida de acurácia mostra clara superioridade do método DDCAM em comparação ao método CAM. O referido efeito de superioridade diminui com o aumento da taxa de mistura e também com a aumento do número de variáveis em estudo. Entretanto, essa superioridade não diminui por alguma perda de performance do método DDCAM, o que ocorre na prática é que o método CAM consegue obter alguma melhora em seu desempenho. A melhora ocorre tanto com o aumento da taxa de mistura quanto com o aumento no número de variáveis envolvidas no problema em investigação.

O método DDCAM se mostra extremamente estável em termos da medida de acurácia durante as variações investigadas. Mesmo para taxa de mistura nula, situação mais difícil para métodos de detecção de valores *outliers* em geral, o método DDCAM apresenta valores elevados na medida de acurácia. Todo este ganho em performance é principamente decorrente do procedimento de escolha da quantidade de agrupamentos. O procedimento de refinamento e também a seleção do valor k baseada no critério

BIC garantem melhores desempenhos, além de maior estabilidade. A figura 4.5.2 apresenta análise para a medida de especificidade na comparação em estudo.

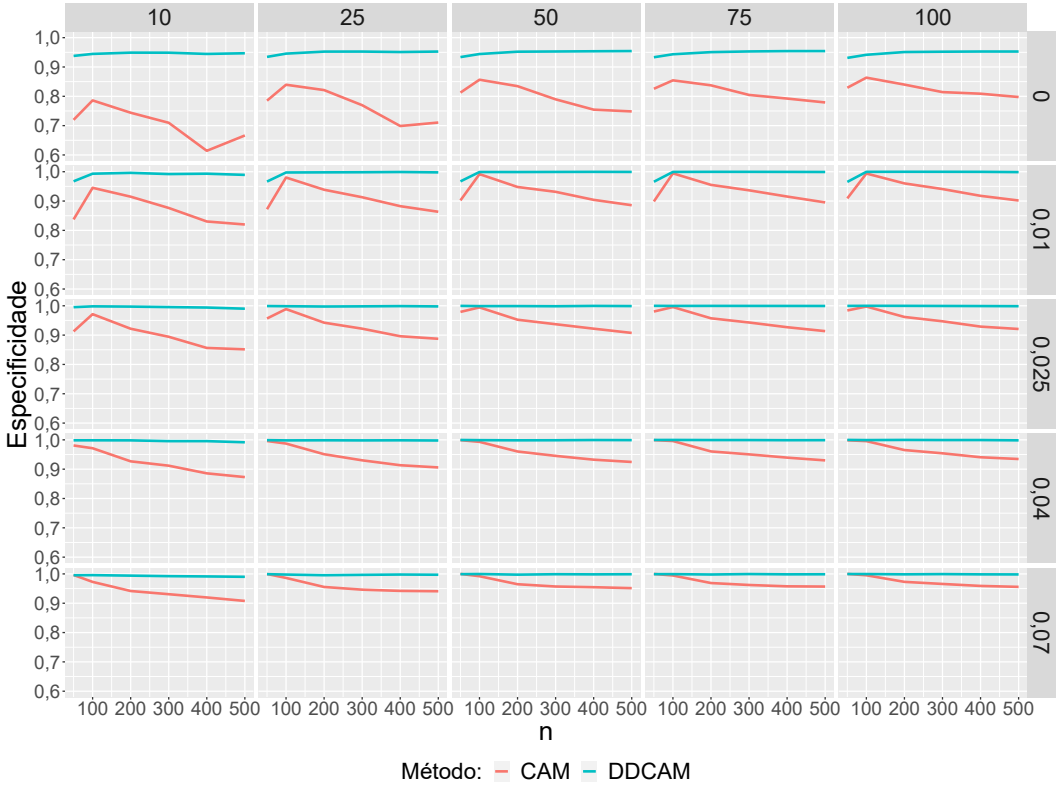

Figura 4.5.2: Comparação de especificidade entre CAM e DDCAM.

Os resultados apresentados para a medida de especificidade são bastante semelhantes aos verificados para acurácia. O desempenho do método DDCAM em comparação ao método CAM é notoriamente superior. Assim como verificado para a medida de acurácia, o aumento dos valores da taxa de mistura e também com a aumento do número de variáveis envolvidas no problema melhoram a performance do método CAM. Por outro lado, esse aumento de performance não é suficiente para fazer o método CAM alcançar um nível competitivo em relação ao método DDCAM. A figura 4.5.3 apresenta a análise com respeito à medida de sensibilidade na comparação em estudo.

Não existem resultados de sensibilidade para o caso de taxa de mistura nula, ou seja, $\delta = 0$, fato que já foi abordado anteriormente. Os resultados para medida de sensibilidade revelam um grande equilíbrio para quase todas as configurações de variações do número de variáveis p e taxas de mistura δ que foram investigadas. Para uma quantidade de variáveis menor ($p = 10$), o método CAM apresenta alguma vantagem na medida de sensibilidade, mas nada extremamente significativo uma vez que os valores de sensibilidade do método DDCAM neste cenário são também elevados.

Mesmo para o cenário de vantagem mais ampla para o método CAM, os valores da medida de sensibilidade, que foram obtidos através da metodologia DDCAM foram bastante elevados. O menor valor da medida

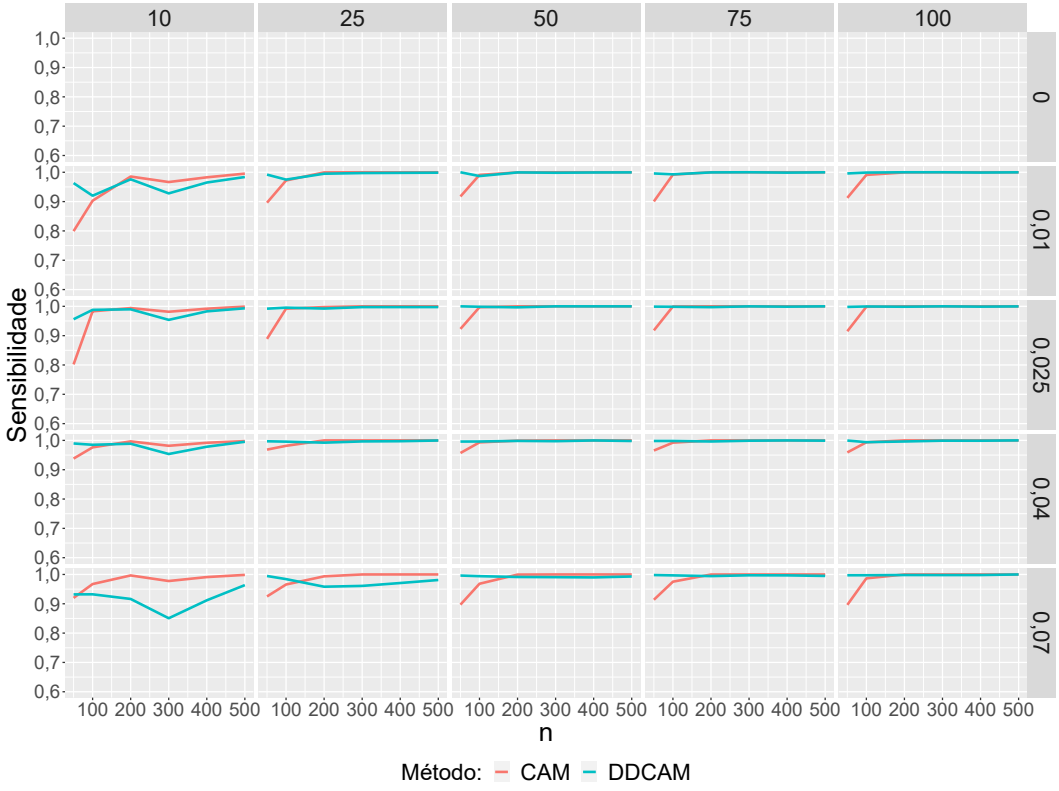

Figura 4.5.3: Comparação de sensibilidade entre CAM e DDCAM.

de sensibilidade que foi verificado, ocorreu para $p = 10$ e $\delta = 0{,}07$, mas ainda assim, a sensibilidade média fornecida pelo DDCAM esteve sempre superior ao patamar de 0,85.

De uma forma geral, os resultados obtidos deixam clara a superioridade do método DDCAM com respeito ao método CAM. Os resultados apresentados são superiores nas três medidas em estudo e para diversas configurações dentre as que foram investigadas. Além de superioridade, o método DDCAM se mostrou mais estável com respeito às variações de configurações testadas.

4.6 Análises comparativas entre dois métodos baseados em distância de Mahalanobis e o método DDCAM

É estabelecido também nessa investigação, um procedimento comparativo com outras metodologias já muito difundidas na literatura. A verificação comparativa será entre duas estratégias mencionadas, baseadas em distância de Mahalanobis e a metodologia proposta DDCAM (*Data-driven Cluster Analysis Method*). Os dois métodos baseados em distância de Mahalanobis dependem de estimadores robustos, aqui serão utilizados os estimadores MCD (*Minimum Covariance Determinant*) e MVE (*Minimum Volume Ellipsoid*). Por uma questão de facilidade para identifica-

ção, a partir desse ponto, os métodos baseados em distância de Mahalanobis serão identificados por MCD e MVE, respectivamente. Para essa comparação, experimentos foram executados com $p \in \{10, 25, 50, 75, 100\}$, $\delta \in \{0; 0{,}01; 0{,}025; 0{,}04; 0{,}10\}$, $n \in \{50, 100, 200, 300, 400, 500\}$ e os dados foram gerados com uma matriz de covariâncias sem correlações fixas, mas de forma que garanta todas as correlações tais que $0{,}3 \leqslant \rho \leqslant 0{,}6$. Para todos os cenários foram geradas 1000 réplicas.

As figuras 4.6.1, 4.6.2 e 4.6.3 apresentam a comparação entre os três métodos DDCAM, MCD e MVE. As figuras apresentam por colunas as quantidades de variáveis nos dados simulados e ainda, por linhas, os valores de δ (taxa de mistura) utilizados na simulação. Para as 1000 execuções, a acurácia, a especificidade e a sensibilidade foram calculadas. Os gráficos apresentam valores médios verificados dentre estas três medidas. Por uma questão de construção, os métodos MCD e MVE não são executados em cenários com $n < p$, ou seja, o tamanho da amostra de dados precisa ser superior à quantidade de variáveis envolvidas no problema multivariado em estudo. A figura 4.6.1 mostra análise comparativa entre as metodologias para a medida de acurácia.

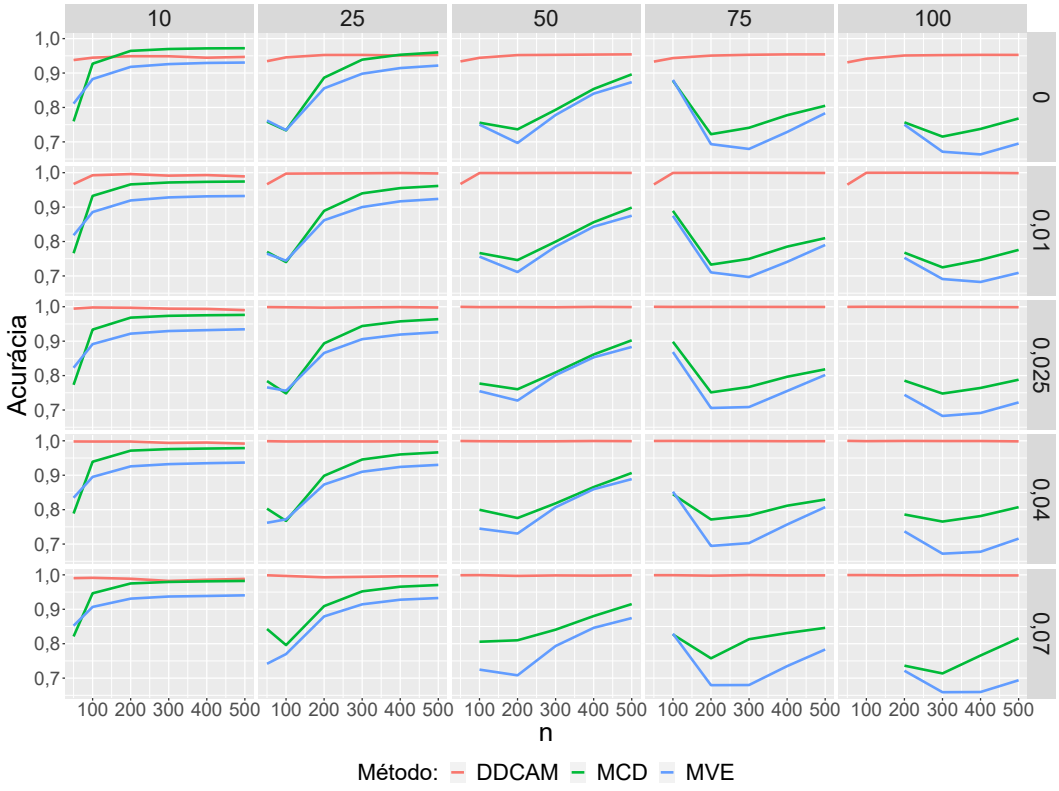

Figura 4.6.1: Comparação de acurácia entre DDCAM, MCD e MVE.

A avaliação para a medida de acurácia revela notória superioridade para o método DDCAM em comparação ao outros dois métodos em estudo. Além disso, os métodos MCD e MVE mostram clara perda de performance com o aumento do número de variáveis do problema. Por outro lado, o DDCAM se mostra absolutamente estável quanto ao número de variáveis. As variações na taxa de mistura não parecem afetar de forma

significativa nenhum dos métodos com respeito à medida de acurácia. A figura 4.6.2 mostra comparação para a medida de especificidade.

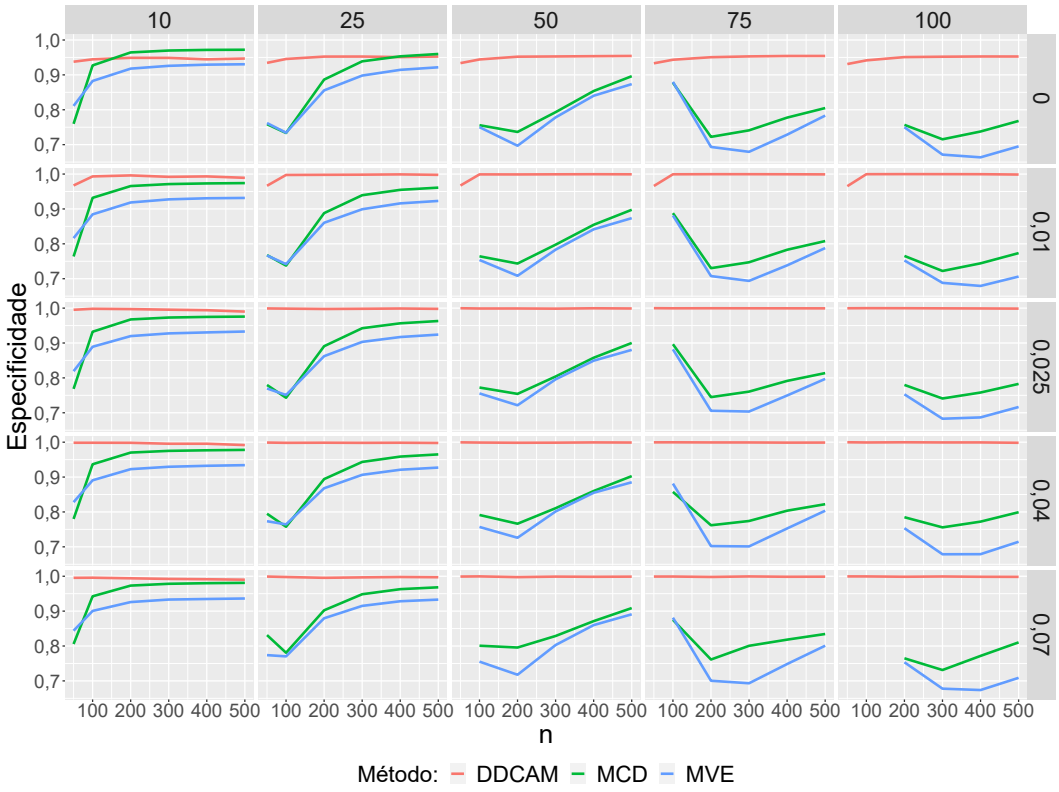

Figura 4.6.2: Comparação de especificidade entre DDCAM, MCD e MVE.

Como verificado anteriormente, na comparação entre os métodos CAM e DDCAM, nota-se que uma grande similaridade nos resultados da medida de acurácia e da medida de especificidade foram observados. Novamente, a avaliação apresenta com clareza o melhor desempenho do método DDCAM com respeito aos métodos MCD e MVE. Assim como na análise de acurácia, os métodos MCD e MVE perdem desempenho à medida que o número de variáveis p cresce. A estabilidade dos resultados através do método DDCAM se repete também quando é avaliada a medida de especificidade. Aqui também, as variações na taxa de mistura não parecem afetar de forma significativa nenhum dos métodos com respeito à medida de especificidade.

A figura 4.6.3 compara as metodologias em termos de sensibilidade. Aqui, vale relembrar que não existem resultados de sensibilidade para a situação em que a taxa de mistura δ é nula. Em particular, para a avaliação com número de variáveis baixo ($p = 10$), o desempenho dos métodos MCD e MVE foi superior, mas os resultados fornecidos pelo DDCAM são bastante competitivos. Para valores superiores de p essa superioridade desaparece.

O método MCD se mostra competitivo em relação ao método DDCAM em muitos cenários, entretanto, para situações com amostras pequenas (n pequeno) este efeito não é plenamente sustentável para o método MCD. O método MCD notoriamente depende de algum aumento no tamanho

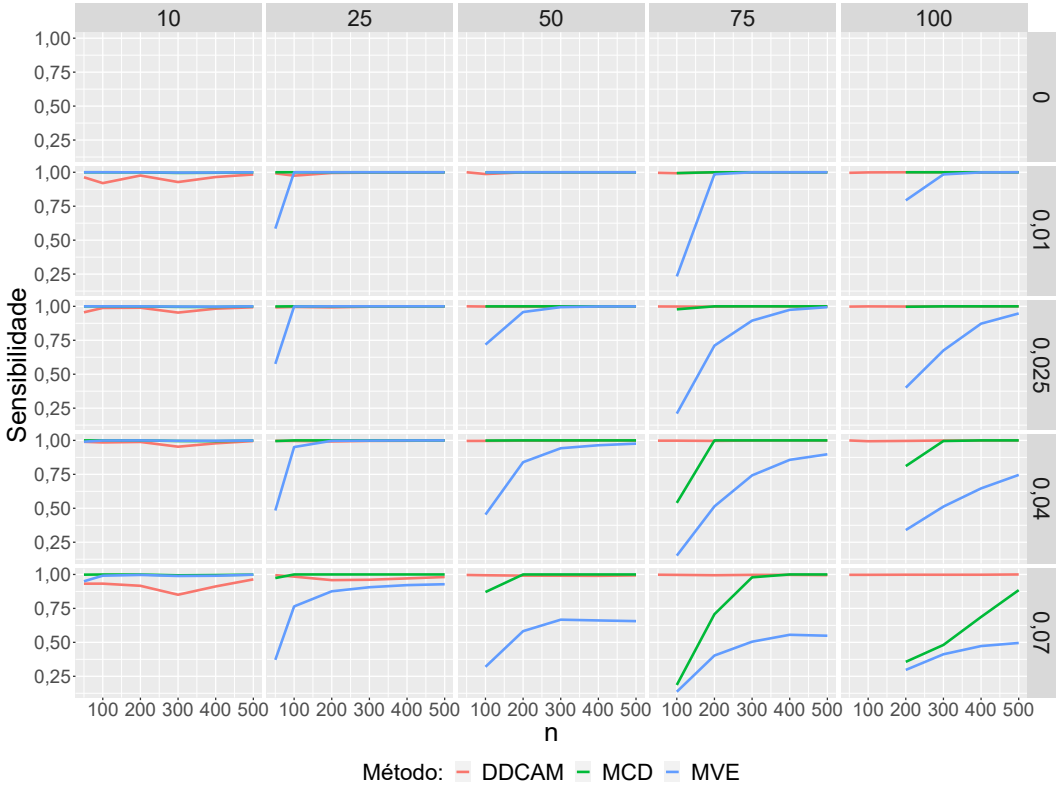

Figura 4.6.3: Comparação de sensibiliade entre DDCAM, MCD e MVE.

amostral para que sua performance seja melhorada. Efeito semelhante ocorre para o método MVE, entretanto o MVE tem muito mais dificuldade de alcançar os patamares alcançados pelo método DDCAM. O método DDCAM, mostrou grande estabilidade nos resultados, sofreu impacto diminuto do número de variáveis p, do tamanho amostral n e da taxa de mistura δ.

Depois de comparações quanto à eficiência do procedimento de detecção de valores *outlier*, será apresentada uma análise comparativa acerca do tempo computacional necessário para a utilização das três metodologias: DDCAM, MCD e MVE. A figura 4.6.4 apresenta os resultados em tempo de CPU consumido para as execuções.

Os resultados são expressos em tempo médio por execução, isso dentre as 1000 simulações executadas. Obviamente todos os métodos apresentam aumento em tempo computacional diretamente proporcional ao aumento do número de variáveis e também ao aumento do tamanho amostral. A taxa de mistura não apresenta qualquer impacto no tempo computacional. É bastante evidente que o desempenho do método DDCAM é superior aos demais métodos investigados. Este fato é claramente outra condição garantidora para a qualidade da metodologia proposta neste estudo.

Todos os códigos utilizados neste estudo, para aplicação das técnicas de detecção de *outliers* multivariados, para geração de amostars simuladas e também para a geração das matrizes de correlação utilizadas, foram implementados através do *software* estatístico R Core Team (2021). Todos estes códigos estão disponíveis, mediante solicitação junto aos autores, desde

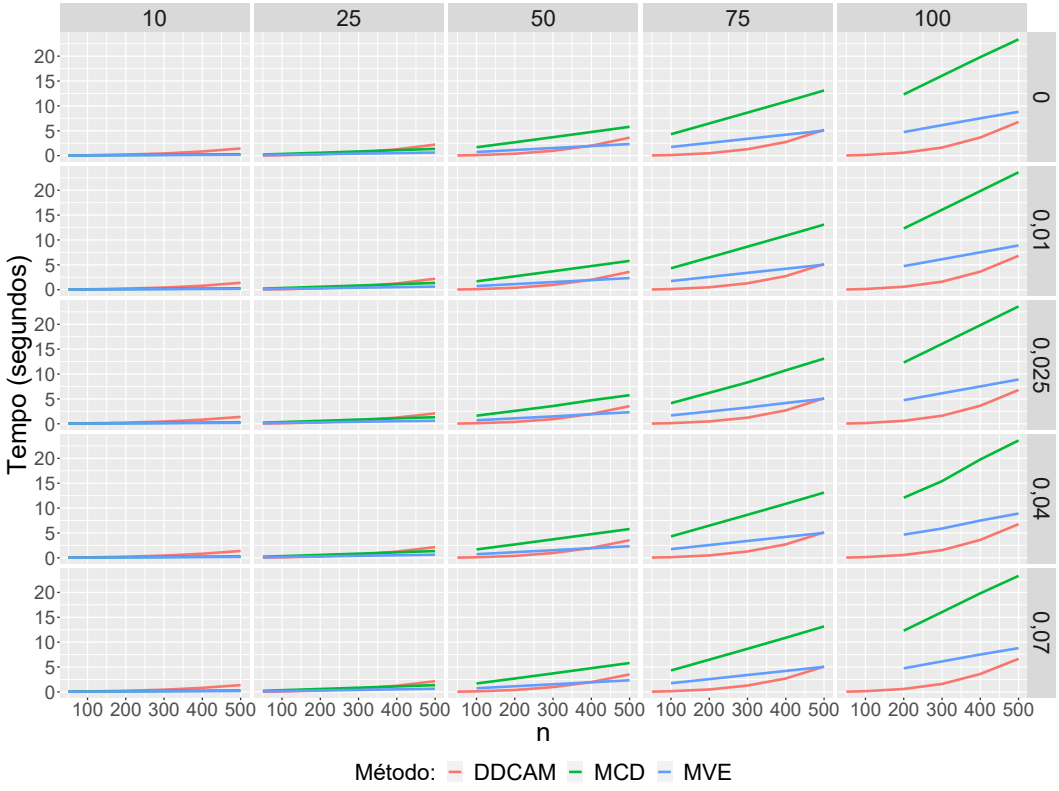

Figura 4.6.4: Gráfico comparativo de tempo computacional entre DDCAM, MCD e MVE.

que para serem utilizados apenas e tão somente para fins educacionais e de pesquisa.

4.7 Aplicação

Esta investigação apresenta também uma análise baseada em conjunto de dados reais. Um banco de dados com informações sobre avaliações de programas de pós-graduação *stricto sensu* realizadas pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) será analisado. O referido banco de dados já foi abordado em investigações estatísticas anteriormente por Martins e Barbosa (2019). Aqui neste estudo, serão aplicadas técnicas para identificação de possíveis programas de pós-graduação que apresentem comportamento *outlier* em uma base de dados da avaliação quadrienal realizada em 2017. O banco de dados sob análise, se refere apenas aos programas de pós-graduação da área de Ciências Agrárias I de acordo com os indicadores produzidos entre os anos de 2013 e 2016 e será descrito com mais detalhes.

Criada em 11 de julho de 1951, com o objetivo de assegurar a existência de pessoal especializado, em quantidade e qualidade suficientes para atender às necessidades dos empreendimentos públicos e privados que visam ao desenvolvimento do país, a então Campanha Nacional de Aperfeiçoamento de Pessoal de Nível Superior, hoje denominada CAPES, Coordena-

ção de Aperfeiçoamento de Pessoal de Nível Superior, tem como missão a busca pela expansão e consolidação da pós-graduação *stricto sensu* (mestrado e doutorado) em todos os estados da Federação.

A CAPES desenvolve ações voltadas para linhas de avaliação da pós-graduação *stricto-sensu*, acesso e divulgação da produção científica, investimentos na formação de recursos humanos de alto nível no Brasil e no exterior, promoção da cooperação científica internacional, além da indução e fomento da formação inicial e continuada de docentes da educação básica, em todas as modalidades. Dentre estas diversas linhas de ação, existe um grande interesse na avaliação da pós-graduação *stricto sensu*. O processo de avaliação tem por objetivo ratificar a capacidade de formação de recursos humanos e geração de conhecimento científico por parte dos referidos programas. Bem como certificar a qualidade da pós-graduação brasileira e identificar possíveis discrepâncias regionais com o intuito de orientar ações de melhoria e expansão de programas de pós graduação no território nacional.

O processo de avaliação é realizado de maneira escalonada. Para tanto, os programas de pós-graduação são agrupados em diferentes áreas de avaliação. Cada uma destas áreas, por meio de alguns instrumentos fundamentais, possui autonomia para elaborar estratégias de avaliação que se adequem às suas peculiaridades. Uma das mais tradicionais áreas de avaliação é a Ciências Agrárias, pois alguns dos primeiros programas de pós-graduação estabelecidos no Brasil encontram-se dentro desta área do conhecimento.

Para uma melhor compreensão dos dados deste estudo, faz-se necessária uma breve explanação acerca da metodologia de avaliação da CAPES. A avaliação da CAPES classifica os programas de pós-graduação em notas que variam na seguinte escala: 1 e 2, tem canceladas as suas autorizações de funcionamento e o reconhecimento dos cursos de mestrado e/ou doutorado por ele oferecidos; 3 significa desempenho regular, que atende ao padrão mínimo de qualidade; 4 é considerado um bom desempenho e 5 é a nota máxima para programas com apenas mestrado. Notas 6 e 7 indicam desempenho equivalente ao alto padrão internacional (Ferreira e Santiago, 2018).

O Ministério da Educação, por meio do Conselho Nacional de Educação, reconhece os resultados da avaliação dos cursos novos e da Avaliação Periódica da CAPES. Estes conceitos são atribuídos com base em uma série de atributos, quantitativos e qualitativos, que avaliam a proposta do programa, corpo docente, corpo discente, teses e dissertações, produção intelectual e inserção social. Todo o processo de avaliação é realizado por uma comissão de consultores *ad-hoc*, pesquisadores de destaque em cada área de avaliação. Com o método de avaliação devidamente esclarecido, serão apresentados os indicadores de produção intelectual, construídos com base no documento da área de Ciências Agrárias I.

Os dados em estudo foram obtidos via Plataforma Sucupira, sistema da CAPES que agrega as informações dos programas de pós-graduação brasileiros, por meio de técnicas de extração automatizada de dados, através de rotina computacional desenvolvida no *software* R Core Team (2021), com

base no pacote `RSelenium`. Foram coletadas informações de todos os 204 programas acadêmicos de pós-graduação pertencentes à área da CAPES de Ciências Agrárias I, referente à avaliação quadrienal realizada em 2017 e ao desempenho no período 2013-2016.

Existe um vasto leque de variáveis constantes da avaliação quadrienal da CAPES. Em particular, foi adotado um subconjunto de variáveis e indicadores que tem um caratér mais classificatório entre os programas. Desta forma, adotou-se como critério utilizar as variáveis cujo módulo da correlação com a nota obtida na avaliação fosse maior ou igual 0,5. A Tabela 4.7.1 apresenta a descrição das variáveis utilizadas.

Tabela 4.7.1: Descrição das variáveis utilizadas.

Variável	Descrição
A1	Número de artigos A1 na classificação QUALIS
A1D	Número de artigos A1 na classificação QUALIS com discente na autoria
A2	Número de artigos A2 na classificação QUALIS
A2D	Número de artigos A2 na classificação QUALIS com discente na autoria
BPDT	Número de bolsistas de produtividade em pesquisa ou em desenvolvimento tecnológico e extensão inovadora
ORI	Orientações $\left[(2{,}5 \times \text{Teses} + \text{Dissertações})/\text{docente permanente}\right]$
PAI	Produção média de alto impacto no quadriênio $\left[(A1 + A2 + B1^{\ddagger})/\text{docente permanente}\right]$
PAID	Produção média de alto impacto com discente no quadriênio $\left[(A1D + A2D + B1D^{\ddagger\ddagger})/\text{docente permanente}\right]$
MBP	Média de bolsistas de produtividade no quadriênio/docente permanente
THE	Número de teses/docente permanente

‡ Número de artigos B1 na classificação QUALIS.
‡‡ Número de artigos B1 na classificação QUALIS com discente na autoria.

De acordo com a base de dados sob investigação, dentre os 204 programas de pós-graduação, são: 11 programas classificados com o conceito 7, 19 programas classificados com o conceito 6, 51 programas classificados com o conceito 5, 78 programas classificados com o conceito 4, 44 programas classificados com o conceito 3 e 1 programa classificado com o conceito 2. É reconhecido que o padrão CAPES de avaliação de programas de pós-graduação é bastante rígido. O grau de exigência e competitividade é bastante elevado. Logo, é razoável admitir que um programa com conceito abaixo de 3, mas que em avaliações anteriores já possuiu conceito maior ou igual a 3, pode reunir condições bastante próximas dos programas de nível 3. Dada a competitividade, essa linha divisória pode ser considerada uma linha bastante tênue.

Por outro lado, os extratos dos conceitos superiores, são de fato programas de nível bastante elevados, que notoriamente se destacam dos demais.

Essa investigação objetiva detectar a possibilidade de que alguns dos programas de pós-graduação se destaquem como programas *outliers*. Isso com respeito aos demais programas de pós-graduação. Diante disso, foi estabelecido por este estudo para determinar programas de pós-graduação *outliers*, um modelo de padrão ouro. Os programas de pós-graduação com conceito 7 em pelo menos duas avaliações subsequentes, anteriores à avaliação do banco de dados em estudo, e que permaneceram com conceito 7 mesmo após a avaliação dos dados em estudo serão classificados como programas *outliers*. Essa definição parte da premissa que programas de excelência em tal nível diferem dos demais o bastante para serem classificados como suficientemente discrepantes, a ponto de serem classificados como programas de pós-graduação *outliers*.

De acordo com o padrão ouro descrito, seis programas de pós-graduação foram previamente selecionados como *outliers*. Dentre os programas selecionados, quatro obtiveram conceito 7 na avaliação trienal de 2007, referente ao desempenho no período 2004-2006, a saber: Entomologia da UFV, Agronomia (Genética e Melhoramento de Plantas) da USP/ESALQ, Agronomia (Solos e Nutrição de Plantas) da USP/ESALQ e Ciências (Energia Nuclear na Agricultura) da USP/CENA. Outros dois programas selecionados obtiveram conceito 7 na avaliação trienal de 2010, referente ao desempenho no período 2007-2009, a saber: Agronomia (Fitopatologia) e Ciências Agrárias (Fisiologia Vegetal), ambos da UFV.

Os mecanismos de detecção de *outliers* multivariados baseados em distância de Mahalanobis, via MCD e MVE, anteriormente descritos, serão aplicados à base de dados. Também será utilizada a nova metodologia proposta DDCAM. Uma vez que o método DDCAM é uma evolução da metodologia CAM, essa comparação baseada em dados reais se restringirá apenas aos três métodos mencionados.

O método MCD identificou, dentre os 204 programas de pós-graduação, que 87 programas deveriam ser classificados como programas de pós graduação *outliers*. Trata-se de um número bastante elevado, parece extremamente improvável que mais de 42% dos dados sejam valores *outliers*. Dentre os 87 identificados 11 são programas de pós-graduação classificados com o conceito 7, 17 são programas classificados com o conceito 6, 36 são programas classificados com o conceito 5, 21 são programas classificados com o conceito 4 e 2 são programas classificados com o conceito 3. Este resultado parece bastante contraditório com um processo de identificação de *outliers* razoável para os dados deste estudo.

Já o método MVE identificou, dentre os 204 programas de pós-graduação, que 62 programas deveriam ser classificados como programas de pós graduação *outliers*. Apesar de inferior em quantidade quando comparado ao resultado através do método MCD, ainda é um volume muito alto. Novamente, parece muito pouco provável que mais de 30% dos dados sejam valores *outliers*. Dentre os 62 identificados 11 são programas de pós-graduação classificados com o conceito 7, 13 são programas classificados com o conceito 6, 23 são programas classificados com o conceito 5, 13 são programas classificados com o conceito 4 e 2 são programas classificados com o conceito 3. Este resultado também parece um tanto controverso com

um processo de identificação de *outliers* aplicado neste banco de dados.

Por fim, o método DDCAM identificou, dentre os 204 programas de pós-graduação, que 5 programas deveriam ser classificados como programas de pós graduação *outliers*, todos com conceito 7. Dentre os programas classificados como *outliers*, 4 pertencem ao grupo do padrão ouro e 1 era conceito 6 na avaliação anterior (trienal 2010-2012), mas apresentou desempenho de grande destaque na avaliação quadrienal 2013-2016, por isso foi elevado ao nível 7. Este menor volume de *outliers* parece um resultado um pouco mais alinhado com os conceitos associados à valores *outliers*.

Ao considerar previamente por padrão ouro um conjunto de programas *outliers*, fica possivel calcular as medidas de desempenho de acurácia, especificidade e sensibilidade como nos experimentos anteriores. A tabela 4.7.2 apresenta estes resultados, é importante salientar que são medidas condicionadas à validade da premissa estabelecida para o padrão ouro mencionada anterioremente.

Tabela 4.7.2: Comparação entre MCD, MVE e DDCAM para dados reais.

Medidas	MCD	MVE	DDCAM
acurácia	0,6029	0,6961	**0,9902**
especificidade	0,5909	0,6869	**1,0000**
sensibilidade	**1,0000**	**1,0000**	0,6667

Os resultados demonstram clara superioridade nas medidas de acurácia e especificidade para o método DDCAM. A medida de especificidade pode remeter para alguma possível subestimação do conjunto de valores *outlier*, porém uma grande subestimação poderia deteriorar a medida de acurácia e principalmente fornecer valores muito pequenos em sensibilidade, o que não ocorreu. Por outro lado, valores unitários para sensibilidade obtidos pelos métodos MCD e MVE são bastante enganosos, os resultados deixaram clara uma enorme superestimação de resultados através dos dois métodos. Novamente é importante relembrar que toda esta análise está apoiada no conjunto padrão ouro previamente estabelecido antes da aplicação das metodologias. A Figura 4.7.1 apresenta a distribuição empírica das variáveis utilizadas nesta aplicação.

Nota-se que todas as variáveis possuem distribuição assimétrica à direita, logo, para estes dados, não são esperados, a priori, *outliers* na cauda à esquerda. Isso parece em clara consonância com o padrão ouro estabelecido, que não considera existência de possíveis programas *outliers* que possuirem conceitos baixos, apenas no nível dos conceitos elevados. Além disso, observa-se que os *outliers* detectados pelo método DDCAM apresentam destaque nas variáveis MBP, A1, A2, A1D, AD2 e BPDT, sendo possíveis *outliers* univariados. Por outro lado, nas demais variáveis esse comportamento não está completamente evidente.

Existem outros pontos de maior destaque que possivelmente são *outliers* univariados nas variáveis discutidas na figura 4.7.1. Isso mostra que de fato para ser um *outlier* multivariado não é necessário ser classificado com *outlier* univariado em todas as dimensões. Além disso, ser *outlier* univariado

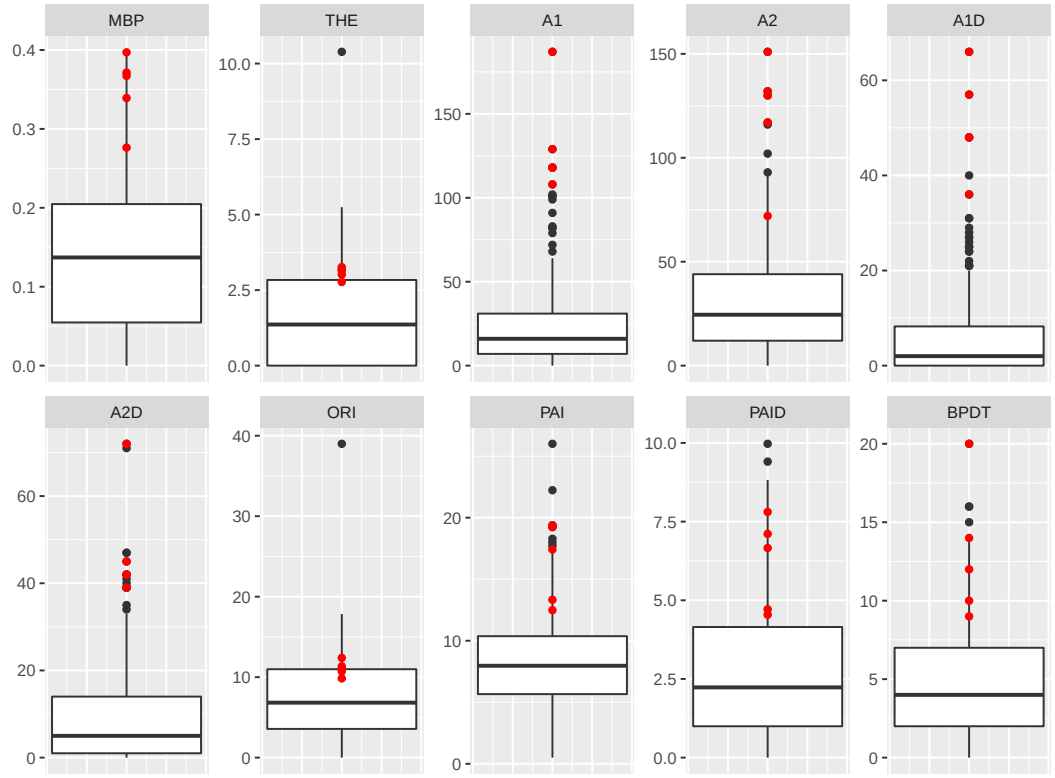

Figura 4.7.1: Representação gráfica da distribuição empírica das variáveis com *outliers* identificados pelo método DDCAM destacados em vermelho.

não é fato garantidor para ser classificado como *outlier* multivariado.

Referências Bibliográficas

Barbosa, J. J., Duarte, A. R. e Martins, H. S. R. (2020). A performance evaluation in multivariate outliers identification methods. Ciência & Natura, 42:1–14.

Duarte, A. R., Martins, H. S. R. e Oliveira, F. L. P. (2021). CM-generator: a methodology for generating customized correlation matrices. Communications in Statistics: Theory and Methods (submitted paper), páginas 1–22.

Ferreira, C. G. e Santiago, J. S. (2018). Considerações sobre o sistema CAPES de avaliação. Brazilian Journal of Development, 4(4):1274–1294.

Martins, H. S. R. e Barbosa, J. J. (2019). Impacto da produção científica na avaliação quadrienal da CAPES 2013-2016. Revista Brasileira de Biometria, 37(2):290–305.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Capítulo 5

Conclusão

O objetivo central desse estudo foi a apresentação de uma nova metodologia para o procedimento de detecção de *outliers* multivariados. O novo método considera informações inerentes ao próprio banco de dados e também informações de conhecimento prévio do pesquisador acerca das populações sob investigação. O método utiliza procedimentos de análise de agrupamentos para efetuar a detecção de *outliers* multivariados.

A motivação central para essa proposição foi subsidiada pelos resultados apresentados na seção 4.2. Essa análise deixa clara a margem de melhoria através de uma escolha mais adequada para a quantidade de grupos no procedimento de análise de agrupamentos inerente à aplicação da metodologia.

Além desse objetivo central, estudos sobre o efeito de parâmetros inerentes ao funcionamento da técnica foram discutidos. Uma investigação comparativa com outras metodologias eficazes para a detecção de *outliers* multivariados foi conduzida. O método CAM, utilizado como base dessa nova proposta, foi parâmetro comparativo assim como estratégias bem difundidas baseadas na distância de Mahalanobis, MCD (*Minimum Covariance Determinant*) e MVE (*Minimum Volume Ellipsoid*). O vasto conjunto de configurações de simulação fez com que os diversos aspectos comparativos pudessem ser melhor explorados. Isso possibilitou uma comparação justa e permitiu estabelecer que o novo método é mais adequado para este tipo de investigação.

O método DDCAM se mostrou mais efetivo nas diversas formas comparativas abordadas. A medida de acurácia, bastante indicada para tal comparação, garante a eficiência da metodologia. Os resultados são bons mesmo com o significativo aumento do número de variáveis envolvidas na investigação. Ainda assim, as medidas de sensibilidade e especificidade apresentam resultados bastante relevantes para o método DDCAM.

O conjunto de resultados apresentados ilustra a capacidade do método em diagnosticar, de forma adequada, os *outliers* multivariados e também os não *outliers*. O excelente comportamento do método DDCAM, mesmo com o aumento do número de variáveis, sem significativo comprometimento do tempo computacional necessário, é um indicativo de que a técnica se adapta bem, inclusive para grandes bancos de dados, assunto em voga na atualidade. Em comparação, os demais métodos apresentam tempo de

execução crescente de forma substancial com o aumento do número de variáveis.

A estratégia de refinamento do conjunto $\mathcal{K}$, de possíveis valores k, mostrou grande eficiência no processo de seleção do melhor valor k, bem como a utilização do critério BIC. Estes itens, em conjunto, transformaram a escolha k flutuante, dependente da informação contida nos dados em estudo.

Um estudo de aplicação, com dados reais, com informações sobre avaliações de programas de pós-graduação *stricto sensu* realizadas pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) foi introduzido. Foram utilizadas informações de 204 programas acadêmicos de pós-graduação pertencentes à área da CAPES de Ciências Agrárias I, referente à avaliação quadrienal realizada em 2017 e ao desempenho no período 2013-2016. Nesta análise, os resultados apresentados demonstram clara superioridade nas medidas de acurácia e especificidade para o método DDCAM. A medida de especificidade pode remeter para alguma possível subestimação do conjunto de valores *outlier*, porém uma grande subestimação poderia deteriorar a medida de acurácia e principalmente fornecer valores muito pequenos em sensibilidade, o que não ocorreu. Por outro lado, valores unitários para sensibilidade obtidos pelos métodos MCD e MVE são bastante enganosos, os resultados deixaram clara uma enorme superestimação de resultados através dos dois métodos. De um modo geral, o método DDCAM mostrou-se mais adequado na aplicação do conjunto de dados reais.

A continuidade desses estudos inclui, de forma quase natural, a inclusão de critérios para reduzir o valor k_{max}, sem trazer prejuízo para a escolha do valor sub-ótimo para k. Uma efetiva redução em k_{max}, mas que não acarrete em prejuízos ao método, tende a ampliar substancialmente a eficiência em termos de tempo computacional. Outra frente relevante, seria a avaliação de possibilidades de critérios de escolha dos valores k através de procedimentos do tipo *machine learning*. Mecanismos adaptativos para escolhas dos centróides no procedimento $k-$médias e também do critério de distância $\phi \times s_c$ também são assuntos já sob investigação, e que podem apresentar resultados futuros bastante interessantes e úteis para a comunidade estatística e demais pesquisadores. Adicionalmente, a apresentação de um pacote estatístico através do *software* R Core Team (2021) é mencionada como proposta de continuidade e já se encontra em pleno desenvolvimento pelos pesquisadores envolvidos neste estudo.

Referências Bibliográficas

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Referências Bibliográficas

Aggarwal, C. C. (2017). An Introduction to Outlier Analysis, páginas 1–34. Springer International Publishing.

Atkinson, A. C. e Riani, M. (2002). Forward search added-variable t-tests and the effect of masked outliers on model selection. Biometrika, 89(4):939–946.

Atkinson, A. C. e Riani, M. (2004). The forward search and data visualisation. Computational Statistics, 19(1):29–54.

Atkinson, A. C., Riani, M. e Cerioli, A. (2010). The forward search: Theory and data analysis. Journal of the Korean Statistical Society, 39(2):117–134.

Barbosa, J. J., Duarte, A. R. e Martins, H. S. R. (2020). A performance evaluation in multivariate outliers identification methods. Ciência & Natura, 42:1–14.

Barbosa, J. J., Pereira, T. M. e Oliveira, F. L. P. (2018). Uma proposta para identificação de outliers multivariados. Ciência & Natura, 40:1–8.

Barnett, V. e Lewis, T. (1994). Outliers in Statistical Data. John Wiley & Sons.

Berton, L., Huertas, J., Araújo, B. e Zhao, L. (2010). Identifying abnormal nodes in complex networks by using random walk measure. Em IEEE Congress on Evolutionary Computation, páginas 1–6. IEEE.

Bussab, W. O. (1990). Introdução à análise de agrupamentos. ABE.

Butler, R., Davies, P. e Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. The Annals of Statistics, páginas 1385–1400.

De, D., Song, W.-Z., Xu, M., Shi, L. e Tan, S. (2013). Advances in real-world sensor network system. Em Advances in Computers, volume 90, capítulo: 1, páginas 1–90. Elsevier.

De Moivre, A. (1738). The doctrine of chances: A method of calculating the probability of events in play. W. Pearson, second edition.

Duarte, A. R., Martins, H. S. R. e Oliveira, F. L. P. (2021). CM-generator: a methodology for generating customized correlation matrices. Communications in Statistics: Theory and Methods (submitted paper), páginas 1–22.

Everitt, B., Landau, S. e Leese, M. (2001). Cluster Analysis. Arnold, 4th edition.

Ferreira, C. G. e Santiago, J. S. (2018). Considerações sobre o sistema CAPES de avaliação. Brazilian Journal of Development, 4(4):1274–1294.

Ferreira, D. F. (2011). Estatística multivariada. Editora UFLA, 2 edition.

Filzmoser, P. (2005). Identification of multivariate outliers: a performance study. Austrian Journal of Statistics, 34(2):127–138.

Filzmoser, P., Garrett, R. e Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. Computers & Geosciences, 31(5):579–587.

Filzmoser, P., Hron, K. e Reimann, C. (2009). Principal component analysis for compositional data with outliers. Environmetrics: The Official Journal of the International Environmetrics Society, 20(6):621–632.

Filzmoser, P., Maronna, R. e Werner, M. (2008). Outlier identification in high dimensions. Computational Statistics & Data Analysis, 52(3):1694–1711.

Gauss, C. F. (1809). Theoria motus corporum coelestium in sectionibus conicis solem ambientium auctore Carolo Friderico Gauss. sumtibus Frid. Perthes et IH Besser.

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E. e Tatham, R. L. (2009). Análise multivariada de dados. Bookman Editora.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. e Stahel, W. A. (2011). Robust statistics: the approach based on influence functions, volume 196. John Wiley & Sons.

Hastie, T., Tibshirani, R. e Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Hawkins, D. M. (1980). Identification of Outliers, volume 11. Chapman and Hall.

Jolliffe, I. T. e Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065):20150202.

Kamalov, F. e Leung, H. H. (2020). Outlier detection in high dimensional data. Journal of Information & Knowledge Management, 19(01).

Kutsuna, T. e Yamamoto, A. (2017). Outlier detection using binary decision diagrams. Data Mining and Knowledge Discovery, 31(2):548–572.

Laplace, P. (1776). Mémoires de mathématique et de physique présentés à l'académie royale des sciences par divers savans et lûs dans ses assemblées. L'Académie Royale des Sciences par divers Savans et lûs dans ses Assemblées.

Lejeune, C., Mothe, J., Soubki, A. e Teste, O. (2020). Shape-based outlier detection in multivariate functional data. Knowledge-Based Systems.

Lu, G., Zhou, L., Lyu, S., Shi, C. e Su, K. (2020). Outlier node detection algorithm in wireless sensor networks based ongraph signal processing. Journal of Computer Applications, 40(3):783–787.

Luo, J., Frisken, S., Machado, I., Zhang, M., Pieper, S., Golland, P., Toews, M., Unadkat, P., Sedghi, A., Zhou, H., Mehrtash, A., Preiswerk, F., Cheng, C., Golby, A., Sugiyama, M. e Wells III, W. M. (2018). Using the variogram for vector outlier screening: application to feature-based image registration. International Journal of Computer Assisted Radiology and Surgery, 13(12):1871–1880.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Em Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, páginas 281–297. Oakland, CA, USA.

Malhotra, N. K. (2012). Pesquisa de marketing: uma orientação aplicada. Bookman Editora.

Martins, H. S. R. e Barbosa, J. J. (2019). Impacto da produção científica na avaliação quadrienal da CAPES 2013-2016. Revista Brasileira de Biometria, 37(2):290–305.

Mohamad, I. B. e Usman, D. (2013). Standardization and its effects on k-means clustering algorithm. Research Journal of Applied Sciences, Engineering and Technology, 6(17):3299–3303.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramsey, S. A., Klemm, S. L., Zak, D. E., Kennedy, K. A., Thorsson, V., Li, B., Gilchrist, M., Gold, E. S., Johnson, C. D., Litvak, V., Garnet Navarro, G., Roach, J. C., Rosenberger, C. M., Rust, A. G., Yudkovsky, N., Aderem, A. e Shmulevich, I. (2008). Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics. PLoS Computational Biology, 4(3):e1000021.

Resende, M., Brighenti, C. R. G. e Cirillo, M. Â. (2017). Procedure to identify outliers through cumulative distribution of extremes in a gamma response model. Communications in Statistics-Simulation and Computation, 46(9):6937–6946.

Rousseeuw, P. J. (1984). Least median of squares regression. Journal of the American statistical association, 79(388):871–880.

Rousseeuw, P. J. e Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41(3):212–223.

Rousseeuw, P. J. e Leroy, A. M. (1987). Robust regression and outlier detection. John Wiley & Sons.

Rousseeuw, P. J. e Zomeren, B. C. V. (1990). Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, 85(411):633–639.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2):461–464.

Sturges, H. A. (1926). The choice of a class interval. Journal of the American Statistical Association, 21(153):65–66.

Tabjula, J. L., Kanakambaran, S., Kalyani, S., Rajagopal, P. e Srinivasan, B. (2021). Outlier analysis for defect detection using sparse sampling in guided wave structural health monitoring. Structural Control and Health Monitoring.

Valadares, F. G., Aquino, A. L. L. e Rabelo, R. A. (2012). Detecção de outliers multivariados em redes de sensores sem fio. Em XLIV Simpósio Brasileiro de Pesquisa Operacional. SBPO.

Van Zoest, V., Stein, A. e Hoek, G. (2018). Outlier detection in urban air quality sensor networks. Water, Air, & Soil Pollution, 229(4):111.

Veloso, M. V. S. e Cirillo, M. A. (2016). Principal components in the discrimination of outliers: A study in simulation sample data corrected by pearson's and yates´ s chisquare distance. Acta Scientiarum. Technology, 38(2):193–200.

Wahid, A. e Rao, A. C. S. (2019). A distance-based outlier detection using particle swarm optimization technique. Em Information and Communication Technology for Competitive Strategies, páginas 633–643. Springer.

Wang, B. e Mao, Z. (2019). Outlier detection based on gaussian process with application to industrial processes. Applied Soft Computing, 76:505–516.

Wang, C., Liu, Z., Gao, H. e Fu, Y. (2019). Vos: A new outlier detection model using virtual graph. Knowledge-Based Systems, 185.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American statistical association, 58(301):236–244.

Zhu, J., Jiang, W., Liu, A., Liu, G. e Zhao, L. (2017). Effective and efficient trajectory outlier detection based on time-dependent popular route. World Wide Web, 20(1):111–134.