Modelo espaço-temporal generalizado misto: efeito aleatório via análise fatorial com interação não linear para detecção de conglomerados

Milton Pifano Soares Ferreira

Departamento de Estatística - ICEx - UFMG

Julho de 2020

Modelo espaço-temporal generalizado misto: efeito aleatório via análise fatorial com interação não linear para detecção de conglomerados

Milton Pifano Soares Ferreira

Orientador: Vinícius Diniz Mayrink

Coorientador: Antônio Luiz Pinho Ribeiro

Tese submetida ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais, como parte dos requisitos necessários à obtenção do grau de Doutor em Estatística.

Departamento de Estatística

Instituto de Ciências Exatas

Universidade Federal de Minas Gerais

Belo Horizonte, MG - Brasil

Julho de 2020

A minha esposa, Aglaia, e ao meu filho, Rodrigo.

"Our world, our life, our destiny, are dominated by Uncertainty; this is perhaps the only statement we may assert without uncertainty".

de Finetti

Agradecimentos

A realização deste trabalho não seria possível não fosse a coordenação efetiva de Vinícius D. Mayrink. Suas explicações, extremamente didádicas, foram fundamentais para o entendimento e a solução de problemas, e sua disciplina, na condução de todo o trabalho, me direcionou a manter o foco.

O segundo apoiador desta tese, mas não menos importante, foi Antônio Ribeiro, coordenador do grupo de pesquisa CODE (*Clinical Outcomes in Digital Electrocardiology*) do Centro de Telessaúde do Hospital das Clínicas da UFMG, responsável pela definição, orientação da questão a ser investigada e coordenação da equipe de pesquisadores multidisciplinar responsável, direta ou indiretamente, pela preparação e qualificação dos dados da aplicação real utilizada nesta tese.

Gostaria de agradecer à toda equipe do CODE, Antonio L. P. Ribeiro, Manoel Horta Ribeiro, Gabriela M. M. Paixão, Derick M. Oliveira, Paulo R. Gomes e Jéssica A. Canazarta, pelo trabalho desenvolvido e pelas análises efetuadas durante todo o processo de organização dos dados.

Finalmente, agradeço à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo apoio financeiro.

Resumo

Neste estudo, desenvolvemos um modelo fatorial para explorar dados de distribuições da família exponencial coletados no espaço e no tempo visando identificar a existência de *clusters* entre regiões similares e estudar a complexidade de relações entre essas regiões. *** O objetivo principal é incorporar ao modelo fatorial interações não lineares para lidar com um efeito aleatório espaço-temporal na estrutura de um modelo de regressão linear generalizado misto. A dependência espacial entre regiões é estabelecida através do modelo CAR especificado para cada coluna da matriz de cargas. A dependência temporal é modelada pela associação entre as colunas da matriz de escores. A presença de interações não lineares visa melhorar a detecção de conglomerados (*clusters*), uma vez que novos tipos de grupos podem surgir como uma combinação dos efeitos principais dos fatores e o efeito de interação. Nosso estudo é focado na regressão logística e Poisson, mas pode ser estendido a outros modelos lineares generalizados baseados em distribuições da família exponencial. Um estudo simulado extenso foi conduzido para investigar a performance do modelo proposto. Este trabalho foi motivado pela análise de um conjunto de dados de eletrocardiogramas (ECG's) de pacientes que sofreram infarto agudo do miocárdio (IAM). Os dados foram obtidos entre 2013-2016 por um sistema de telediagnóstico de ECG's que abrange todas as regiões do estado de Minas Gerais. O sistema foi desenvolvido e é mantido pelo Centro de Telessaúde do Hospital das Clínicas da Universidade Federal de Minas Gerais. A metodologia proposta em que se define uma interação não linear espaço-temporal e a análise de uma base de dados de ECG's inédita são as contribuições centrais desta tese.

Palavras-chave: Regressão logística, Regressão Poisson, Eletrocardiograma, MCMC, Processo Gaussiano, Modelo CAR.

Abstract

In this study, we develop factor analysis to explore areal data collected in space and time. The main goal is to incorporate the framework with nonlinear interactions to handle a spatio-temporal random effect in the structure of a mixed generalized linear regression. The spatial dependence between regions is established through the CAR model specified for each column of the loadings matrix. Temporal dependence is considered to associate the columns of the factor scores matrix. The presence of nonlinear interactions is intended to improve cluster detection, since new types of groups can emerge as a combination of the main factors effects and the interaction effect. Our study is focused on the logistic and Poisson cases, but it can be extended to other generalized linear models originated from distributions of the exponential family. A comprehensive simulation study is conducted to investigate the performance of the proposed approach. This work was motivated by the analysis of electrocardiogram (ECG) data related to patients affected by acute myocardial infarction (AMI). The data were collected between 2013–2016 through an ECG telediagnostic system covering the state of Minas Gerais in Brazil. The system is maintained by the Telehealth Center within the Hospital das Clínicas of the Federal University of Minas Gerais. The methodology proposed defining nonlinear interaction in the spatio-temporal setting and the analysis of the novel ECG data set are the central contributions of this thesis.

Keywords: Logistic Regression, Poisson Regression, Electrocardiogram, MCMC, Gaussian Process, CAR Model.

Sumário

Capítulo 1

Introdução

Da análise fatorial ao modelo proposto

A análise fatorial é uma técnica estatística muito utilizada para modelagem multivariada visando uma redução de dimensionalidade para o conjunto de dados com diversas variáveis. Ela surgiu com o principal objetivo de descrever a variabilidade original de um vetor aleatório X ($q \times 1$ observado), através de um número $k < q$ de variáveis (fatores) latentes. Modelos fatoriais são uma ferramenta flexível e poderosa para analisar a dependência multivariada e verificar padrões e relacionamentos entre os dados (Johnson e Wichern, 2007). A visão clássica da análise fatorial, conforme destacam Gamerman e Salazar (2013), foi introduzida por Thurstone (1931). Diversos métodos semelhantes podem ser encontrados na literatura com diferentes restrições e algoritmos computacionais. Dentre eles podemos citar Análise de Componentes Principais (Yeung e Ruzzo, 2001), Mínimos Quadrados Parciais (Nguyen e Rocke, 2002; Boulesteix e Strimmer, 2006) e *Single Value Decomposition* (Martin e Porter, 2012). O método Mínimos Quadrados Parciais (PLS) normalmente apresenta melhores resultados em análises preditivas do que o Análise de Componentes Principais (PCA), pois também leva em consideração a variável resposta da regressão (Mayrink e Lucas, 2013). Outra técnica usada para reduzir a dimensionalidade dos dados é a Fatoração de Matriz Não-negativa (NMF). A NMF pode ser útil no estudo de subsistemas biológicos, uma vez que é capaz de identificar padrões de similaridades locais e globais entre genes (Brunet et al., 2004; Kim e Tidor, 2003), ao contrário do PCA

que se concentra apenas nos padrões globais.

O modelo fatorial, em uma forma padrão e usual, é estruturado com uma matriz de cargas (ou *loadings*) que multiplica uma matriz contendo escores de fatores latentes. A este produto é adicionado uma matriz de erros para os quais adota-se uma distribuição Gaussiana e assume-se independência. A matriz de escores contém em suas linhas as novas variáveis latentes que sumarizam os principais padrões subjacentes existentes nas diversas variáveis do conjunto de dados.

O modelo fatorial dinâmico (MFD) diferencia-se do modelo fatorial padrão pela introdução de uma estrutura de correlação temporal flexível para os fatores latentes, inicialmente assumidos como independentes (Gamerman e Salazar, 2013). Isso confere ao MFD a capacidade de modelar a estrutura complexa dos dados de séries temporais. Esse modelo procura explicar a estrutura dinâmica comum de uma série temporal multivariada através de um conjunto de fatores comuns que variam no tempo. Trabalhos nesse sentido podem ser encontrados em Geweke (1977) e Sargent e Sims (1977), dentre outros autores.

A análise fatorial e a estatística espacial são áreas estatísticas que se desenvolveram bastante nas últimas décadas, motivadas pelos avanços computacionais e de algoritmos eficientes, em particular métodos de simulação do tipo Monte Carlo via Cadeias de Markov permitiram a utilização de inferência Bayesiana nesses contextos (Gamerman e Lopes, 2006; Lopes e West, 2004; Banerjee et al., 2004). Lopes et al. (2008) propoem o modelo fatorial espaço-temporal derivado do modelo fatorial dinâmico padrão. Nesse modelo, a dependência temporal é modelada por fatores latentes comuns que podem ser considerados para descrever semelhanças entre as séries temporais, tais como sazonalidade e tendência. A dependência espacial é incorporada nas colunas da matriz de cargas dos fatores, que são modeladas por Processo Gaussiano, determinando a importância dos fatores comuns na descrição das medidas em diferentes locais.

O modelo proposto por Lopes et al. (2008) foca, exclusivamente, em situações nas quais a variável resposta segue uma distribuição Gaussiana. Lopes et al. (2011) estendem esse modelo para variável resposta cuja distribuição pertence à família exponencial, dando origem ao modelo fatorial espacial dinâmico generalizado (GSDFM), em que uma função de ligação conecta a estrutura fatorial a um parâmetro natural na distribuição dos dados

observados. Da mesma forma que no modelo fatorial espacial dinâmico, a dependência espacial é modelada pela matriz de cargas dos fatores e a dependência temporal pelos fatores latentes comuns. Essa modelagem possibilita a detecção de regiões similares a partir da associação delas aos fatores, permitindo a identificação de *clusters* ou grupos de locais com comportamento temporal semelhante. Outros pontos comuns entre os dois estudos são a estimação do número de fatores através do algoritmo *Reversible Jump* MCMC (Green, 1995), a flexibilidade de se incorporar covariáveis variando no tempo e no espaço, e o fato de que o modelo fatorial é aplicado diretamente aos dados observados para detectar agrupamentos de regiões.

Na evolução dos modelos fatoriais, Mayrink e Lucas (2013) incorporam ao modelo padrão um termo para capturar interações não lineares entre os fatores latentes. Mayrink e Lucas (2013) apresentam a análise de modelos fatoriais com diferentes estruturas de interação entre os fatores latentes considerando duas abordagens : uma incluindo efeitos multiplicativos e outra, com formulação mais geral, para interações não lineares introduzidas através do Processo Gaussiano. Semelhante aos modelos propostos por Lopes et al. (2008) e Lopes et al. (2011) o modelo fatorial é aplicado diretamente aos valores observados, neste caso, para explorar problemas relacionados a dados de expressão genética de alta dimensionalidade.

No contexto de modelos lineares, Nelder e Wedderburn (1972) generalizaram a regressão linear ao especificar um modelo em que a variável dependente possui uma distribuição que pertence à família exponencial. Esse modelo foi intitulado como modelo linear generalizado (GLM). Uma extensão dos modelos GLMs são os modelos lineares generalizados mistos (GLMMs) em que a ideia principal é a incorporação de correlação através da modelagem contendo um ou mais termos de efeitos aleatórios. Efeitos que estão ligados a covariáveis observáveis são chamados de "fixos". A inclusão de componentes fixos e aleatórios determinam a nomenclatura de "modelo misto" (McCulloch e Neuhaus, 2005). Os modelos GLMMs são amplamente utilizados em vários setores da ciência (Brown e Prescott, 1999; Demidenko, 2004), apenas citando alguns. Para trabalhar com esse tipo de modelo sob a visão da inferência clássica é necessário integrar os efeitos aleatórios fora, o que, na maioria das situações, é algo intratável. Uma alternativa para estimação dos

parâmetros é a utilização de métodos aplicados sob a visão Bayesiana, como por exemplo os algoritmos do tipo Monte Carlo via Cadeias de Markov (Zhao et al., 2006; Browne e Draper, 2006).

Do modelo proposto

A proposta desta tese estende o modelo em Lopes et al. (2011), incorporando um termo para capturar interações não lineares entre os fatores conforme introduzido por Mayrink e Lucas (2013), mas com a diferença de que a estrutura fatorial é inserida como um termo aleatório dentro do modelo de regressão linear generalizada, conferindo a ele uma estrutura mista (GLMMs). A esta extensão intitulamos de Modelo Fatorial Espaço-Temporal Generalizado Misto com interações não lineares (*Generalized Spatio-Temporal Factor Mixed Model with Nonlinear Interactions*). O efeito de interação é definido por um Processo Gaussiano e configurado de tal forma que os locais a serem afetados são impactados por um mesmo tipo de interação. Um modelo hierárquico simples em que o efeito aleatório não possui uma estrutura fatorial não atenderia aos nossos objetivos de identificação de *clusters* entre as regiões e da existência de interação não linear entre elas. Uma análise comparativa quanto a capacidade preditiva entre esse modelo e o modelo proposto nesta tese poderia ser efetuado, mas como o objetivo principal da tese não é a análise preditiva, essa comparação não foi realizada. Neste estudo, desenvolvemos várias simulações contemplando diferentes cenários relativos ao número de locais, número de fatores, número de tempos e quantidade de locais afetados pelo efeito de interação, dentre outros. Os cenários artificiais contemplados foram configurados para dois modelos lineares generalizados : logístico e Poisson. No tocante aos dados reais, apresentamos uma aplicação do modelo fatorial espaço-temporal logístico para avaliação da probabilidade de morte de pacientes que sofreram infarto agudo do miocárdio (IAM) e identificação de conglomerados de regiões semelhantes em termos da relação com fatores principais e o efeito de interação. Os efeitos principais são definidos em relação ao IDH Renda (Índice de Desenvolvimento Humano) dos municípios de Minas Gerais. Os dados a serem ajustados na aplicação real são provenientes do sistema de telediagnóstico do Centro de Telessaúde do Hospital das Clínicas da UFMG que coleta e lauda eletrocardiogramas

de pacientes de municípios do estado. Cada observação da base de dados se refere a um indivíduo que sofreu IAM, contendo o município de residência, indicativo de morte ou não, dentre outras informações. Essa estruturação nos direciona para a especificação de um modelo amostral logístico, mas como destacado anteriormente, o componente espacial está aqui representado por um termo aleatório e não pelos dados observados, o que remete a um GLMMs. Cada "observação" desse termo contem informação sumária de um local e tempo que são somadas ao termo linear com os regressores. Com isso, a especificação do modelo espacial é relativo à configuração para dados de área. Como nem todos os municípios de Minas Gerais registraram indivíduos que sofreram IAM nos anos analisados, os municípios não representados foram unidos a municípios fronteiriços formando regiões, de tal forma que todas as áreas registrassem pelo menos um caso de IAM em cada ano. Diferentemente de Lopes et al. (2008) e Lopes et al. (2011), em que o número de fatores foi estimado por *Reversible Jump* MCMC, optamos por definir a quantidade de fatores a partir de uma análise de sensibilidade. Todo o trabalho foi desenvolvido sob a ótica da inferência Bayesiana.

O modelo espacial para dados de área considera um região fixa que é particionada em um número finito de unidades de área, com contornos bem definidos em formato regular ou irregular. Nesses modelos, os dados são normalmente estatísticas sumárias de variáveis coletadas ao longo das sub-regiões (estados, municípios, etc). A introdução de associação espacial é realizada a partir da definição de uma estrutura de vizinhança baseada na organização dos blocos dentro do mapa (Banerjee et al., 2004). Os dois modelos mais populares que incorporam esse tipo de informação de vizinhança são os modelos condicionalmente e simultaneamente autoregressivos, CAR e SAR, respectivamente (Besag, 1974). Nesta tese, as dependências espacial e temporal foram configuradas utilizando o modelo CAR.

Das contribuições

Dentre as contribuições desta tese para a área da estatítica, destacamos o ineditismo na incorporação de interações não lineares em um modelo fatorial construído para estruturar o efeito aleatório em uma aplicação do tipo espaço-temporal. Até o momento, o efeito de

interação já tinha sido proposto apenas para o modelo fatorial padrão em Mayrink e Lucas (2013). A presença de interações não lineares visa melhorar a explicação das complexas interrelações entre as regiões do espaço e, também, permitir a detecção de variados conglomerados (*clusters*). Novos tipos de *clusters* podem surgir como uma combinação dos efeitos principais dos fatores e do efeito de interação. Outra importante contribuição é a utilização de uma base de dados que nunca foi analisada em outros trabalhos na literatura com modelagem estatística. Essa base foi extraída de um banco de dados de milhões de pacientes com registro do laudo de exames de eletrocardiogramas (ECG's) indicando a ocorrência de anormalidades cardíacas, sendo a mais relevante o infarto agudo do miocárdio (IAM). Nosso estudo envolveu análises extensas, contemplando vários cenários simulados para os modelos Bernoulli e Poisson. Mostramos, com isso, que ele é extensivo a toda família exponencial que define os modelos lineares generalizados, disponibilizando, para a comunidade em geral, uma ferramenta que pode ser aplicada a contextos diferentes.

Da organização

Esta tese está organizada da seguinte maneira. No Capítulo 2 apresentamos a base de dados real, um dos principais motivadores para o desenvolvimento desta tese. Descrevemos a origem dos dados e os processos de limpeza e preparação para processamento da análise estatística. Os dados são referentes a indivíduos que sofreram infarto agudo do miocárdio em municípios do estado de Minas Gerais aliados a identificação de morte ou não. No Capítulo 3 apresentamos as formulações dos modelos logístico e Poisson, os dois modelos avaliados neste estudo. Detalhamos a estrutura hierárquica desses modelos destacando que toda análise foi realizada sob o ponto de vista da inferência Bayesina. Descrevemos os passos do algoritmo MCMC utilizado na estimação dos parâmetros dos modelos, bem como as configurações necessárias para execução do mesmo. Nos Capítulos 4 e 5 apresentamos as análises a partir de dados artificiais referentes aos modelos logístico e Poisson, respectivamente. Em ambos os casos, consideramos dados balanceados e desbalanceados em relação à variável resposta, os quais são explorados em diversos cenários. Esses cenários diferem-se pelo número de locais, número de tempos, número de vizinhos e quantidade de regiões afetadas pela interação não linear. Uma comparação

geral dos cenários é efetuada com destaque para análise de resíduos, curvas ROC (caso logístico) e finalizando com a análise a partir de 30 réplicas de Monte Carlo. No Capítulo 6 descrevemos os resultados obtidos a partir do ajuste do modelo logístico aos dados reais de ECG's. Discutimos como o modelo descreve o impacto do sistema de telediagnóstico no atendimento a pacientes que sofreram IAM e os conglomerados de regiões identificados. Finalmente, o Capítulo 7 contém as conclusões de todas as análises realizadas e a descrição de trabalhos que consideramos relevantes a serem desenvolvidos futuramente.

Capítulo 2

Dados de eletrocardiogramas

As doenças cardiovasculares (DC) são as principais causas de morte e, em 2015, causaram 18 milhões de óbitos no mundo (Roth et al., 2017). Dados do Departamento de Informática do Sistema Único de Saúde do Brasil (DATASUS), de 2013, revelam que o infarto agudo do miocárdio (IAM) foi a principal causa de morte por doença cardíaca no Brasil, tendo sido observado aumento de 48% entre 1996 e 2011 (Medeiros et al., 2018). Roth et al. (2017) apresentam estudo em que as doenças cardiovasculares são as principais causas de morte em todas as regiões do mundo. Eles destacam que no período de 1990 a 2015 registrou-se declínios dramáticos na ocorrênca de DC em regiões com elevado índice sociodemográfico, mas apenas uma diminuição gradual ou nenhuma mudança na maioria das regiões.

A introdução das radiografias de tórax em 1895 e do eletrocardiograma (ECG) em 1902 forneceram informações objetivas sobre a estrutura e a função do coração (Howell, 1991). Na primeira metade do século XX, vários indivíduos inovadores foram responsáveis por uma sequência de descobertas e invenções que levaram ao ECG de 12 derivações (registros da diferença de potencial elétrico entre dois pontos do corpo, a maioria no tórax), como se conhece hoje (Fye, 1994). Atualmente, o exame ECG é procedimento essencial na avaliação inicial de pacientes que reportam dores toráxicas. Especificamente, desempenha um papel importante como uma ferramenta não invasiva e de baixo custo para avaliar arritmias e cardiopatia isquêmica (Fye, 1994).

Da origem dos dados

A partir do Centro de Telessaúde do Hospital das Clínicas da Universidade Federal de Minas Gerais (UFMG), cuja coordenação está sob a responsabilidade do Professor Antônio Luiz Pinho Ribeiro do Departamento de Clínica Médica da Faculdade de Medicina da UFMG, um volume considerável de ECGs são gerados e armazenados em banco de dados através de um sistema de telediagnóstico de eletrocardiogramas (tele-ECGs) que abrange municípios da rede pública do estado de Minas Gerais. O Centro de Telessaúde do Hospital das Clínicas da UFMG faz parte da Rede de Telessaúde de Minas Gerais (TNMG), uma rede colaborativa de sete universidades públicas do estado, coordenada pelo Hospital Universitário da UFMG (Alkmim et al., 2012). O TNMG fornece serviço de telediagnóstico de ECGs, ou seja, ECGs registrados em lugares remotos e enviados para um servidor central, via internet, a partir do qual um sistema *Web* disponibiliza a vizualização do traçado e registro do laudo por um cardiologista (Alkmim et al., 2012). Os serviços, quando da preparação dos dados deste trabalho, cobriam 814 municípios de Minas Gerais, principalmente em centros de atenção primária à saúde (APS), mas também em departamentos de emergência e hospitais. Em 2017, como parte de um programa do Ministério da Saúde do Brasil, também começou a fornecer serviços de tele-ECG para outros estados brasileiros nas regiões amazônica e nordeste. Mais de 4 milhões de laudos de tele-ECGs já foram realizados (até abril de 2019), onde a qualidade é garantida por auditorias regulares (mensalmente um percentual dos ECGs laudados são enviados para auditores que reavaliam o relatório registrado pelos cardiologistas). Reuniões regulares ocorrem entre a equipe de auditores e os cadiologistas para avaliação dos resultados. Todas essas informações compõem um banco de dados grande e rotulado de ECGs digitais vinculados a hospitalizações e óbitos, constituindo uma amostra capaz de fornecer diversas análises úteis, seja prognósticos clínicos, seja padrões de atendimento regionais.

A motivação inicial para o desenvolvimento do presente trabalho surgiu da necessidade, apresentada pelo Professor Antônio Ribeiro, de se verificar o impacto na saúde dos pacientes após a implantação do sistema de Telessaúde em cada município. Segundo

Ribeiro, uma das maneiras de se avaliar esse impacto é a partir da análise do desfecho do paciente após a ocorrência de IAM, pois esse desfecho pode ser rastreado a partir da base de dados pública de mortalidade.

Os ECGs utilizados em nosso trabalho foram obtidos pelo TNMG, usando um aplicativo *Web* construído na linguagem de programação Java (Andrade et al., 2011). Os ECGs foram gravados utilizando um eletrocardiógrafo fabricado pela Tecnologia Eletrônica Brasileira (São Paulo, Brasil) - modelo TEB ECGPC - ou Micromed Biotecnologia (Brasília, Brasil) - modelo ErgoPC 13, de 2010 a 2017. Eles são enviados aos servidores centrais pela internet, usando um aplicativo desenvolvido pela equipe própria de programadores. Todos os ECGs são analisados por um grupo de cardiologistas treinados, utilizando critérios padronizados (Kligfield et al., 2007) para gerar o laudo em texto livre. O laudo de ECGs são auditados para reconhecer erros médicos e interpretações discordantes, a fim de garantir a qualidade e a uniformidade dos laudos cardiológicos (Ribeiro et al., 2019).

Da identificação de IAM e do desfecho do paciente

Com o objetivo de se classificar os ECGs quanto aos distúrbios cardiovasculares e, com isso, ser possível selecionar os indivíduos que sofreram IAM, bem como de se desenvolver pesquisas diversas, um algoritmo hierárquico de aprendizado de máquina de texto livre foi usado para reconhecer diagnósticos específicos. Foram criadas classes de diagnóstico (classes CODE), de acordo com diretrizes internacionais (Kligfield et al., 2007). Primeiramente, o texto foi pre-processado removendo preposições e conjuções, e gerando n-gramas (sequência contínua de n itens). Em seguida, foi utilizado o algoritmo Lazy Associative Classifier (LAC) (Veloso et al., 2006), alimentado por um dicionário de 2800 palavras-chave criado manualmente por especialistas, com base em textos de diagnósticos reais. O resultado da execução do algoritmo LAC é a classificação do texto livre dos laudos em uma das 79 classes CODE e a atribuição da classe identificada a cada exame de ECG (Ribeiro et al., 2019).

Além da classificação do texto livre, os traçados do ECG, sinais elétricos capturados por eletrodos instalados na superfície do corpo humano, foram analisados pelo programa

Glasgow (release 28.4.1, disponibilizado em 16 de junho de 2009), codificado pelas Declarações Diagnósticas de Glasgow (Macfarlane et al., 1990) e pelos códigos de Minnesota (Macfarlane e Latif, 1996). As correspondências entre as classes CODE, Glasgow e Minnesota foram mapeadas, de forma que ficaram registradas 3 classificações para cada ECG. A definição do diagnóstico final do ECG foi baseada na classe CODE. Na concordância da classe CODE com uma das restantes (Glasgow e Minnesota), a classificação CODE foi considerada. Não havendo essa concordância, especialistas treinados procederam a uma revisão manual e definiram a classe CODE do exame.

Para que nosso estudo pudesse ser desenvolvido era necessário que o resultado do laudo do ECG fosse vinculado aos dados pessoais do paciente (nome, sexo, data de nascimento, cidade de residência), bem como com os dados do sistema nacional de informações sobre mortalidade. Esse processo foi desenvolvido usando métodos de ligação probabilística padrão através do *software* FRIL (Fine-Grained Records Integration and Linkage tool, v. 2.1.5, Atlanta, GA. `http://fril.sourceforge.net/`) (Ribeiro et al., 2019).

Podemos resumir todo esse processo de organização dos dados nas seguintes etapas: (i) extração dos dados de pacientes, exames, e laudos do sistema de telediagnóstico, (ii) obtenção dos dados de mortalidade do sistema público, (iii) pareamento dos dados de pacientes com os dados de mortalidade, (iv) validação por especialistas do resultado do algoritmo de classificação do laudo textual, (v) comparação da classificação dos exames pelo diagnóstico dos 3 sistemas (Glasgow: `https://www.gla.ac.uk/researchinstitutes/healthwellbeing/research/robertsoncentreforbiostatistics/electrocardiology/glasgowecgprogram/`, Minnesota, CODE), (vi) análise por especialistas dos laudos discordantes entre os sistemas para definição da classe final.

Os dados são armazenados em um banco de dados PostgreSQL sendo que os mais importantes para análise são : paciente (id, sexo, idade, endereço, município), histórico clínico (comorbidades e medicamentos em uso), exame (id, data, centro de saúde), traçados do exame (número do registro, batimento cardíaco, velocidade, sensibilidade, sinais das 12 derivações, medidas dos sinais), laudo do cardiologista em formato texto, classes Minnesota, classes Glasgow, classes CODE e dados de mortalidade (data, município, causa). De uma base de dados de 2470424 ECGs, 1773689 pacientes foram identificados.

Após excluir os ECGs com problemas técnicos e pacientes com menos de 16 anos, um total de 1558415 pacientes foram considerados para análises. A média de idade foi de 51.6 anos, com 40.2% de homens e a taxa de mortalidade geral foi de 3.34% (Ribeiro et al., 2019).

Da modelagem espacial com interação

Visto que aos dados dos ECGs estão atrelados o local de realização do exame, e levando em conta o fato de Minas Gerais ser um estado grande com desigualdades marcantes entre suas regiões, a ideia de uma modelagem espacial faz sentido para este estudo. E ainda, como o desfecho, no caso de ocorrência de IAM, é a morte ou não do paciente, isso nos motivou a estudar essas situações a partir de um modelo logístico espacial, e mais, em identificar se há a existência de grupos de municípios que tenham características semelhantes, nos levando a propor um modelo logístico fatorial espacial que permita avaliar a existência de conglomerados. É razoável considerar que há uma complexa ligação ou estrutura de influências afetando o sistema TNMG. Veja que os dados envolvem municípios diferentes, com políticas públicas distintas e que trocam informações, principalmente, entre municípios vizinhos (Portaria emitida pelo Ministério da Saúde do Brasil No 1097 de 22 de maio de 2006 e Portaria emitida pelo Ministério da Saúde do Brasil No 2135, de 25 de setembro de 2013). A característica peculiar no diagnóstico e tratamento de doenças coronarianas aliado a estrutura da saúde pública no Brasil, onde há uma interação entre os orgãos de gestão pública dos municípios (Portaria emitida pelo Ministério da Saúde do Brasil 1097 de 22 de maio de 2006 e Portaria emitida pelo Ministério da Saúde do Brasil No 2135, de 25 de setembro de 2013), estabelece forte influência no atendimento dos pacientes. Essa estrutura, então, nos levou a pensar em um modelo que contemplasse a interação entre grupos de municípios. Essa interação é complexa envolvendo fluxo de pacientes e fluxo financeiro. Com isso, baseado no trabalho de Mayrink e Lucas (2013), pensamos em avaliar um modelo logístico fatorial espacial com interações não lineares.

Da base de dados final

Finalmente, os dados com os quais formulamos nosso modelo são de pacientes que sofreram IAM, em municípios de Minas Gerais, dos quais sabemos o desfecho (óbito ou não). Nos ajustes dos modelos realizados, consideramos apenas as variáveis sexo e idade, que são as principais características consideradas no estudo de mortalidade por IAM

realizado por Medeiros et al. (2018) e por van de Leur et al. (2020). Pela quantidade de variáveis disponíveis há a necessidade de desenvolvimento de um trabalho de seleção de variáveis. Zhang et al. (2015) e Zellner et al. (2004) avaliam vários métodos para seleção de variáveis em modelos logísticos, tais como : *stepwise* com *bagging*, *bootstrap*, *backward* e *forward*. Esta tese não tem o objetivo de obter um modelo explicativo das variáveis mais influentes em relação à variável resposta. O interesse é explorar nossa proposta de interação não linear no modelo fatorial espaço-temporal. A análise de seleção de variáveis está considerado para trabalho futuro desta tese.

A Tabela 2.1 apresenta o número de exames classificados como sendo IAM ao longo dos anos em municípios do estado de Minas Gerais, registrados pelo sistema de telediagnóstico. Devido a baixa quantidade de dados nos primeiros anos, concentramos nossa análise no período de 2013 a 2016. Após esta restrição, ainda optamos por manter apenas o primeiro exame do paciente, pois identifica o início do tratamento. Se considerássemos mais de um exame por paciente isso implicaria em um modelo de dependência. Com isso, a quantidade de contagens ficou alterada conforme Tabela 2.2, totalizando 15835 observações (indivíduos). Note que esses são os dados analisados nesta tese. Entretanto, um grande número de novos ECGs são laudados diariamente e incluídos na base de dados, o que possibilita a utilização de nosso modelo para análises futuras, com mais dados e anos.

Além da avaliação da análise descritiva apresentada acima, verificamos que alguns municípios não possuíam observações em todos os anos. Desta forma, fizemos um trabalho de união deles em grupos de forma que tivéssemos observações em todos os períodos selecionados. Os grupos foram formados analisando os dados ano-a-ano, ou seja, a cada ano aqueles que não tinham dados se uniam a outro grupo ou município de sua fronteira que possuia observaçõess. No final da análise de todos os anos, obtivemos uma única configuração de regiões válidas para todos os períodos, totalizando 441 áreas.

Uma análise importante no entendimento dos dados e na identificação da natureza espacial é a verificação de como estão distribuídos, geograficamente, os locais para os quais existem observações. A Figura 2.1 apresenta a distribuição geográfica das observações a partir da plotagem dos centróides dos municípios. Ressalta-se que apesar da Figura

2.1 mostrar a disposição espacial dos centróides, a modelagem espacial desta aplicação assumiu a configuração de dados de área com fronteiras e vizinhança bem definidas e fixas. Pela Figura 2.1 podemos ver que as observações estão distribuídas em todas as regiões de Minas Gerais havendo alguns pequenos espaçamentos no triângulo mineiro e na região noroeste, mas que não comprometem a análise. O espaçamento no triângulo mineiro é explicado pela não existência de observações para os municípios de Uberlândia e Uberaba, pois eles não fazem parte da rede de atendimento do Centro de Telessaúde do Hospital das Clínicas da UFMG. Após a execução do algoritmo, esses dois municípios se juntaram ao município vizinho Veríssimo.

2005	2006	2007	2009	2010	2011	2012	**2013**	**2014**	**2015**	**2016**
1	8	16	1	4	11	164	**4022**	**5420**	**6021**	**2302**

Tabela 2.1: Número de exames classificados como sendo infarto agudo do miorcárido (IAM) ao longo dos anos em municípios do estado de Minas Gerais registrados pelo sistema de telediagnóstico do Centro de Telessaúde do Hospital das Clínicas da UFMG.

2013	2014	2015	2016
3781	4913	5226	1915

Tabela 2.2: Número de indivíduos que sofreram IAM ao longo dos anos em municípios do estado de Minas Gerais registrados pelo sistema de telediagnóstico do Centro de Telessaúde do Hospital das Clínicas da UFMG.

A base de dados ficou, então, estruturada com os campos 'indicativo de morte', 'sexo', 'idade', 'ano do exame' e 'código do município'. A variável resposta binária é 'indicativo de morte' (1 = morte) e as covariáveis são 'idade/100' e 'sexo' (1 = masculino, 0 = feminino). A utilização da 'idade/100' foi por razões computacionais. Como o modelo utilizado foi o logístico, valores grandes do preditor linear dentro da função exponencial poderiam acarretar *overflow* numérico levando a erro de estimação dos parâmetros. Esse artifício computacional não gera problemas de interpretabilidade, pois basta multiplicar

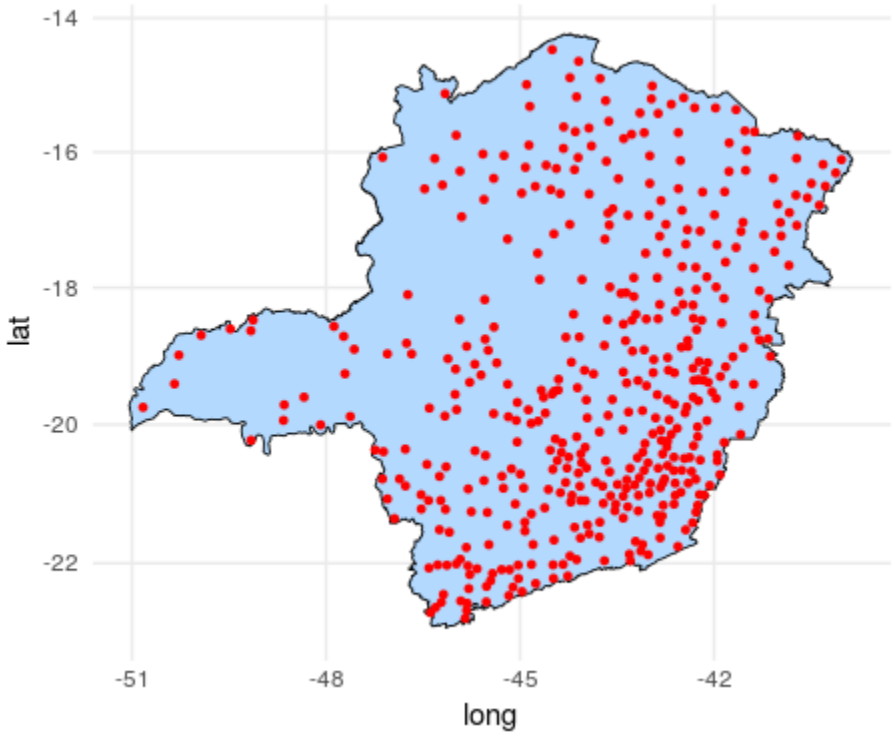


Figure 2.1: *Mapa com os centróides dos municípios/grupos que possuem dados.*

por 100 a estimativa do coeficiente.

Terminamos, aqui, a descrição da base de dados real, cuja a análise do ajuste do modelo se encontra no Capítulo 6. No próximo capítulo descrevemos a estrutura hieráquica dos modelos fatorial espaço-temporal logístico e Poisson com interação não linear na formulação proposta por esta tese.

Capítulo 3

Modelos lineares generalizados mistos

Nelder e Wedderburn (1972) generalizaram a regressão linear ao especificar um modelo onde a variável dependente, Y_i, possui uma função de distribuição pertencente à família exponencial que pode ser escrita como $f(y_i|\theta) = h(y_i)c(\theta_i)\ \exp\left\{\sum_{j=1}^{J} t_j(y_i)b_j(\theta_i)\right\}$, em que $h(y_i) \geq 0$ e $t_1(y_i), \cdots, t_J(y_i)$ são funções reais (não dependentes de θ_i), e $c(\theta_i) \geq 0$ e $b_1(\theta_i), \cdots, b_J(\theta_i)$ são funções reais do parâmetro θ_i, que pode ser um vetor, não dependente de y_i (Casella e Berger, 2002). Seja $X_{\bullet i}^{\top} = (1, X_{1i}, \cdots, X_{(q-1)i}), i = 1, 2, \cdots, n$, o vetor linha $(1 \times q)$ do $i-$ésimo elemento amostral da matriz de desenho X, $n \times q$, que possui 1's na primeira coluna e $q-1$ covariáveis nas demais colunas, e $\beta = (\beta_0, \beta_1, \cdots, \beta_{q-1})^{\top}$ vetor coluna de coeficientes. No modelo proposto por Nelder e Wedderburn (1972), intitulado modelo linear generalizado (GLM), além do preditor linear, $X_{\bullet i}^{\top}\beta$, ele é composto por uma função de ligação $g(.)$ que conecta o parâmetro θ_i ao preditor linear $X_{\bullet i}^{\top}\beta$ de tal forma que $\theta_i = g(X_{\bullet i}^{\top}\beta) = E[Y_i|X_{\bullet i}]$.

Os modelos lineares generalizados mistos (GLMMs) são uma extensão dos GLMs em que a ideia principal é a incorporação de correlação através da modelagem contendo um ou mais termos de efeitos aleatórios. Efeitos que não são aleatórios são chamados de "fixos" e, pelo fato do modelo incluir componentes fixos e aleatórios, é descrito como um modelo "misto" (McCulloch e Neuhaus, 2005). Ou seja, ao incorporarmos um termo aleatório δ_i, através da adição desse termo ao preditor linear $X_{\bullet i}^{\top}\beta$, transformamos um

modelo GLM em GLMM com $\theta_i = g(X_{\bullet i}^\top \beta + \delta_i) = E[Y_i | X_{\bullet i}, \delta_i]$. Devido à sua flexibilidade, os modelos GLMMs são amplamente utilizados em vários setores da ciência (Brown e Prescott, 1999; Demidenko, 2004), dentre outros. Por outro lado, sob a visão da inferência clássica, trabalhar com esse tipo de modelo exige integrar os efeitos aleatórios fora. Essa integração é, na maioria dos modelos, intratável. Com isso, os métodos do tipo Monte Carlo via Cadeias de Markov (MCMC), aplicados sob a visão Bayesiana, fornecem uma excelente alternativa para estimação dos parâmetros (Zhao et al., 2006; Browne e Draper, 2006).

Neste estudo, incorporamos à distribuição da resposta Y_i o efeito aleatório δ_i, cuja estrutura será modelada via análise fatorial estabelecendo uma correlação espaço-temporal (STFM) e permitindo a identificação de agrupamentos de regiões por meio de fatores latentes. Duas funções de distribuição membras da família exponencial serão o foco desta tese: Bernoulli(θ_i) e Poisson(θ_i). O modelo Bernoulli(θ_i) foi considerado por representar a distribuição dos dados reais que motivaram esta tese, e o modelo Poisson(θ_i) foi trabalhado por ser uma opção bastante popular.

3.1 STFM na Regressão Logística

O primeiro modelo sob investigação é construído a partir da regressão logística. Com isso temos:

$$Y_i | \theta_i \sim \text{Bernoulli}(\theta_i), \quad \theta_i = \frac{\exp\{X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}\}}{(1 + \exp\{X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}\})}, \tag{3.1}$$

em que $i = 1, 2, \cdots, n$ equivale ao $i-$ésimo elemento amostral. Seja l_i^* o local l da observação i com $l^* = (l_1^*, l_2^*, \cdots, l_n^*)^\top$ e $l_i^* = l$, se i pertence ao local/região $l \in \{1, 2, \cdots, L\}$. Considere t_i^* o tempo t da observação i com $t^* = (t_1^*, t_2^*, \cdots, t_n^*)^\top$ e $t_i^* = t$, se i ocorre no tempo $t \in \{1, 2, \cdots, T\}$. A função de verossimilhança para esse modelo assume a seguinte formulação :

$$
\begin{aligned}
\mathrm{p}(Y|X,\beta,\delta) &= \prod_{i=1}^{n}\left(\frac{\theta_i}{1-\theta_i}\right)^{y_i}(1-\theta_i) \\
&= \prod_{i=1}^{n}\exp\{y_i X_{\bullet i}^{\top}\beta + y_i\delta_{l_i^* t_i^*}\}\left[1+\exp\{X_{\bullet i}^{\top}\beta+\delta_{l_i^* t_i^*}\}\right]^{-1}
\end{aligned}
\tag{3.2}
$$

Diante da configuração estabelecida para l_i^* e t_i^* podemos organizar os efeitos aleatórios $\delta_{l_i^* t_i^*}$ em uma matriz δ de tamanho $L \times T$. Essa matriz será tratada por meio de uma modelagem fatorial como segue

$$
\delta = \alpha\lambda + \eta + \epsilon, \tag{3.3}
$$

em que α é uma matriz $(L \times K)$ contendo as cargas (*loadings*) dos fatores, λ é matriz $(K \times T)$ contendo os escores dos fatores $(K \ll L)$, η é uma matriz $(L \times T)$ em que suas linhas são nulas ou representam a interação não linear entre os fatores, ϵ é a matriz $(L \times T)$ de erros do modelo, e K é o número de fatores. A matriz η, em (3.3), foi acrescentada ao modelo fatorial tradicional ($\delta \sim \alpha\lambda + \epsilon$), estendendo-o para tratar interações não lineares. Essa extensão ao modelo fatorial foi inspirada no trabalho de Mayrink e Lucas (2013), o qual propõe algumas alternativas para tratar a interação não linear. Nesta tese, vamos utilizar apenas uma delas, que será explicada mais adiante.

O modelo indicado pela Equação (3.1), para a variável resposta Y_i, será trabalhado aqui sob o ponto de vista da inferência Bayesiana. Para isso devemos, naturalmente, especificar distribuições *a priori* relacionadas aos parâmetros dos modelos. Para o vetor de coeficientes $\beta = (\beta_0, \beta_1, \cdots, \beta_{q-1})^{\top}$ vamos adotar $\beta \sim N_q(M_\beta, S_\beta)$, com vetor de médias $M_\beta = (m_{\beta_0}, \cdots, m_{\beta_{q-1}})^{\top}$, fixados pelo pesquisador. O termo $S_\beta = v_\beta I_q$ é matriz de variância e covariância, em que v_β é variância comum do vetor β, também fixada pelo pesquisador, e I_q é matriz identidade $q \times q$. O próximo passo é tratar da estrutura dos δ's. O primeiro elemento na estrutura dos δ's é a distribuição *a priori* relacionada aos elementos de α. Como o objetivo deste trabalho é desenvolver uma análise espacial, a estrutura espacial será incorporada ao modelo (3.3) através das colunas da matriz de cargas α, inspirado nos trabalhos de Lopes et al. (2008) e Lopes et al. (2011). Assumimos para as colunas da matriz α, a configuração do modelo autoregressivo condicional (CAR)

(Besag, 1974). A distribuição *a priori* de $\alpha_{\bullet k}$ ficou definida como

$$\alpha_{\bullet k}|\tau_\alpha \sim N_L(\mathbf{0}, \tau_\alpha[D_\alpha - \rho_\alpha W_\alpha]^{-1}) \tag{3.4}$$

para $k \in \{1, 2, \cdots, K\}$. Admita que $\mathbf{0}$ é vetor $(L \times 1)$ de zeros representando a média; $D_\alpha = \text{diag}\{w_{1+}, \cdots, w_{L+}\}$, em que w_{l+} é o número de vizinhos do local l; ρ_α é parâmetro fixo, também usual no modelo CAR, que torna $[D_\alpha - \rho_\alpha W_\alpha]$ não-singular. Perceba que se $\rho_\alpha = 1$, a matriz $[D_\alpha - \rho_\alpha W_\alpha]$ será singular e a distribuição *a priori* adotada fornecerá o caso CAR impróprio para a distribuição conjunta de $\mathbf{Y}$. Banerjee et al. (2004) apresenta detalhes relativos às restrições ao valor de ρ, neste caso ρ_α. O intervalo em que ρ é escolhido contém o zero e seus limites são obtidos a partir dos autovalores da matriz $D_\alpha^{-1/2} W_\alpha D_\alpha^{-1/2}$. Em geral, por questão de interpretabilidade, ρ é escolhido positivo e mais próximo de 1. A opção por não trabalhar com o modelo CAR impróprio, a partir da inclusão do elemento ρ, é porque futuramente, na extensão deste trabalho, pode-se atribuir uma distribuição *a priori* para ρ e avaliar se a estimativa obtida é significativa ou não, e se está mais próxima de zero ou de 1. A inclusão do zero na estimativa de ρ permitirá a inversão da matriz $[D_\alpha - \rho_\alpha W_\alpha]$, mas o modelo resultante não conterá a parte espacial. Ou seja, um modelo com a estimação de ρ permite testar se a estrutura espacial existe ou não. Como neste trabalho queremos incluir a parte espacial, optamos por fixar ρ próximo de 1 (Banerjee et al., 2004). Outro ponto importante na inclusão de ρ é a obtenção da distribuição Normal multivariada, o que permite avaliar sua variabilidade e analisar a interpretabilidade *a priori*, o que não é possível no caso CAR impróprio. W_α é matriz de vizinhança $(L \times L)$ sendo $(W_\alpha)_{l_1 l_2} = 1$, se l_1 é vizinho de l_2 $(l_1 \sim l_2)$, e $(W_\alpha)_{l_1 l_2} = 0$, caso contrário. O elemento τ_α é o parâmetro de variância usualmente inserido na modelagem CAR. A distribuição *a priori* de τ_α é dada por $\tau_\alpha \sim GI(a_{\tau_\alpha}, b_{\tau_\alpha})$, em que GI sinaliza a distribuição Gama Inversa com hiperparâmetros $a_{\tau_\alpha} > 0$ e $b_{\tau_\alpha} > 0$, especificados pelo pesquisador. A escolha pelo modelo CAR se deve ao fato de que, conforme descrito no Capítulo 2, os dados reais com os quais trabalhamos envolvem registros − exames eletrocardiográficos e desfecho − de indivíduos em uma região. Além disso, nosso objetivo não é fazer previsão para locais não observados. Queremos apenas analisar a estrutura espacial e o modelo CAR permite suavização local de estimativas em

áreas vizinhas. Essa configuração da distribuição *a priori* para α induz a conglomerados próximos, mas permite ao modelo escolher *clusters* formado por regiões distantes uma das outras.

A dependência temporal é inserida também por meio do modelo CAR especificado na estrutura da matriz de escores λ. Seja $\lambda_{k\bullet}$ um vetor $(1 \times T)$ representando a $k-$ésima linha da matriz λ, temos a seguinte especificação

$$\lambda_{k\bullet}^{\top} \sim N_T(\mathbf{0}, \tau_\lambda[D_\lambda - \rho_\lambda W_\lambda]^{-1}); \tag{3.5}$$

em que $\mathbf{0}$ é vetor $(T \times 1)$ de zeros; τ_λ é o parâmetro de variância fixo. O motivo pelo qual fixamos o τ_λ será explicado mais adiante quando descrevemos o problema de identificabilidade do modelo devido ao termo multiplicativo $\alpha\lambda$. O termo $D_\lambda = \text{diag}\{1, 2, 2, \cdots, 2, 1\}$ é matriz diagonal $(T \times T)$ contendo o número de vizinhos de cada tempo t; ρ_λ tem o mesmo papel que o elemento ρ_α e deve ser escolhido seguindo as mesmas diretrizes para garantir a inversão da matriz de covariâncias $[D_\lambda - \rho_\lambda W_\lambda]$ (Banerjee et al., 2004). O elemento W_λ é matriz, banda diagonal, de vizinhança dos tempos com dimensão $T \times T$ (1 tempo tem 2 vizinhos: passado e futuro, exceto o primeiro e último tempos que possuem apenas 1 vizinho) com a seguinte estrutura:

$$W_\lambda = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}_{T \times T} .$$

O modelo CAR com essas configurações é semelhante ao modelo AR(1); veja os trabalhos Box et al. (2011) e Rue e Held (2005). Optamos por utilizar o modelo CAR para fazer uso das propriedades da normal multivariada.

Seguindo os passos de Mayrink e Lucas (2013), estabelecemos interações não lineares nas linhas de η por meio de uma mistura do tipo *"spike and slab"* (George e McCulloch, 1993) envolvendo um Processo Gaussiano. Para a $l-$ésima linha de η, representada por $\eta_{l\bullet} = (\eta_{l1}, \cdots, \eta_{lT})$, foi avaliada a seguinte configuração :

$$\eta_{l\bullet}|p_l, \eta^* \sim (1-p_l)\mathcal{D}_0 + p_l\mathcal{D}_{\eta^*}, \tag{3.6}$$

em que $\mathcal{D}_0$ significa $\eta_{l\bullet} = (0, \cdots, 0)$ com probabilidade 1, ou seja, distribuição degenerada em $\mathbf{0}$. O elemento $\mathcal{D}_{\eta^*}$ é uma distribuição degenerada no vetor η^* de dimensão $1 \times T$, sendo que para esse elemento adotamos um Processo Gaussiano dado por

$$\eta^*|\lambda \sim N_T(\mathbf{0}, \kappa(\lambda)). \tag{3.7}$$

Nesta notação, temos que $\mathbf{0}$ é vetor de zeros $(T \times 1)$, o termo $\kappa(\lambda)$ é matriz de variância e covariância $(T \times T)$ em que $\kappa(\lambda)_{t_1t_2} = \gamma \exp\{-\phi^2||\lambda_{\bullet t_1} - \lambda_{\bullet t_2}||^2\}$ é função de covariância exponencial quadrática (Banerjee et al., 2004) com γ e ϕ fixos, sendo $||\lambda||$ a norma Euclidiana do vetor λ. A função exponencial quadrática, bastante popular na literatura, apresenta as propriedades de ser estacionária, isotrópica e infinitamente diferenciável, o que confere a ela uma característica desejável de ser suave. Como exposto por Mayrink e Lucas (2013), note que se os pontos $\lambda_{\bullet t_1}$ e $\lambda_{\bullet t_2}$ estiverem muito próximos no espaço R^T, então as amostras de t_1 e t_2 são similares e $\kappa(\lambda)_{t_1t_2} \approx \gamma$. Por outro lado, quanto maior a distância entre esses pontos, maior é a dissimilaridade entre as amostras t_1 and t_2 , e mais próximo de 0 é $\kappa(\lambda)_{t_1t_2}$. O elemento ϕ é um parâmetro de escala ajustável que controla quão próximo os pontos λ_{t_1} e λ_{t_2} devem estar de forma a considerar que existe uma associação entre eles. Se γ for igual a 1, teremos uma função de correlação, pois $\kappa(\lambda)_{t_1t_2}$ iria variar entre 0 e 1. Já o parâmetro ϕ está associado à distância entre $\lambda_{\bullet t_1}$ e $\lambda_{\bullet t_2}$. Se ϕ for próximo de zero, essa distância tem pouca força, pois para qualquer magnitude dessa distânca o resultado da multiplicação $\phi^2||\lambda_{\bullet t_1} - \lambda_{\bullet t_2}||^2$ será próximo de zero. O núcleo $\kappa(\lambda)_{t_1t_2}$ é uma função não linear de elementos em λ. Se $p_l = 0\ \forall(l)$ na distribuição *a priori* dada pela Equaçao (3.6), o modelo é dito linear, pois ele pode ser expresso como uma combinação de $\lambda_{kt}\ \forall(k,t)$. Por outro lado, se $p_l \neq 0$ para algum l, $\delta_{l\bullet}$ dependerá de $\eta_{l\bullet}$ que tem uma relação não linear com λ através da função de covariância; por isso, definimos o modelo como não linear. O parâmetro p_l representa a probabilidade de $\eta_{l\bullet} = \eta^*$, ou seja, do $l-$ésimo local ser afetado por uma interação não nula.

Por questões computacionais, reestruturamos a Equação (3.6) considerando uma variável indicadora Z_l, no lugar da probabilidade p_l, conforme a seguir :

$$\eta_{l\bullet}|Z_l, \eta^* \sim (1 - Z_l)\mathcal{D}_0 + Z_l\mathcal{D}_{\eta^*}, \tag{3.8}$$

em que $Z_l \sim$ Bernoulli(p_l). Em termos de especificação *a priori* adotamos que $p_l \sim$ Beta(a_p, b_p), sendo $a_p > 0$ e $b_p > 0$ hiperparâmetros fixados pelo pesquisador.

Alternativamente, Mayrink e Lucas (2013) também consideraram uma outra versão da Equação (3.8) com a seguinte expressão :

$$\eta_{l\bullet}^{\top}|Z_l, \lambda \sim (1 - Z_l)\mathcal{D}_0 + Z_l N_T(\mathbf{0}, \kappa(\lambda)), \tag{3.9}$$

Nessa versão temos tipos de interações, entre fatores, diferentes para cada local l quando a indicadora $Z_l = 1$. Nessa configuração, muito mais parâmetros devem ser estimados (para cada linha de η). A modelagem escolhida para esta tese é mais simples e parcimoniosa, possuindo apenas dois tipos de interação, a saber : a nula e a não nula. Isso significa que se houver interação entre fatores, ela será a mesma para todos os locais afetados. A proposição da mistura em (3.6) é vantajosa, pois a partir da probabilidade p_l sabemos se a interação é significativa ou não. Caso a mistura não fosse utilizada, o η^* seria atribuído para todas as localidades. Como a modelagem considera que existem localidades que não são afetadas por interação, adotar η^* para todos os locais não é razoável. O mais natural é deixar que algumas regiões estejam associadas com interações e outras não (η^* não nulo e η^* nulo, respectivamente).

Finalizando a especificação da Equação (3.3), a $l-$ésima linha da matriz de erros é denotada por $\epsilon_{l\bullet} = (\epsilon_{l1}, \cdots, \epsilon_{lT})$. Assuma a seguinte distribuição :

$$\epsilon_{l\bullet}^{\top}|\sigma^2 \sim N_T(\mathbf{0}, \sigma^2 I_T); \tag{3.10}$$

em que σ^2 é a variância comum dos erros; I_T é matriz identidade $(T \times T)$. Em termos de distribuição *a priori* adotamos $\sigma^2 \sim GI(a_{\sigma^2}, b_{\sigma^2})$, sendo $a_{\sigma^2} > 0$ e $b_{\sigma^2} > 0$ hiperparâmetros fixos. Mais uma vez, optou-se por um modelo mais parcimonioso ao se considerar uma variância comum para os erros, pois no modelo fatorial tradicional assume-se σ_l^2, isto é, uma variância para cada local l.

Em um modelo fatorial é muito comum ocorrer a troca de sinais entre as colunas

da matriz de cargas, α, e as linhas da matriz de escores dos fatores, λ. Essa situação não é um problema, pois no caso simulado, quando isso ocorre, basta multiplicar por -1 a coluna e a linha correspondentes que foram afetadas. Para dados reais, não teremos os valores verdadeiros para verificação, o que terá impacto sobre a interpretabilidade do modelo. Na análise fatorial tradicional aplicada diretamente a dados observados, o recomendável é que o pesquisador avalie a interpretação da relação de uma das variáveis e o escore do fator. Ou seja, analisar se há sentido o escore decrescer ou crescer para locais que já se tem conhecimento prévio de seu comportamento. Essa análise se faz avaliando o sinal das cargas (*loadings*), pois são esses coeficientes que correlacionam as variávies aos fatores. Se o sinal for negativo, quando o fator decrescer, a variável também decresce.

Lopes et al. (2011) e Mayrink e Lucas (2013) destacam o problema da identificabilidade do modelo fatorial envolvendo α, λ e η. Lopes et al. (2011) comentam que para qualquer matriz ortogonal Q, temos $\alpha\lambda = \alpha Q^\top Q\lambda$. Conforme mencionado anteriormente, optamos por fixar τ_λ. A justificativa para isso é o fato de que assumir uma distribuição *a priori* permitindo grande variação de τ_λ junto com τ_α, acarretaria problema quanto à determinação das magnitudes de λ e α, ou seja, ora o α poderia crescer e o λ diminuir e, vice-versa, tornando o modelo não identificável. Fixando o τ_λ e atribuindo uma distribuição *a priori* para τ_α, damos liberdade para α variar, mas restringimos a variabilidade do λ. Mayrink e Lucas (2013) também destacam o problema da identificabilidade no modelo fatorial com interações não lineares (termo η). Eles informam que para a $l-$ésima linha $\alpha_{l\bullet}\lambda + \eta_{l\bullet} = C\alpha_{l\bullet}\lambda + \eta^\dagger_{l\bullet}$ em que $\eta^\dagger_{l\bullet} = (1-C)\alpha_{l\bullet}\lambda + \eta_{l\bullet}$ e C é qualquer número real. Neste caso, é fácil ver que existe a possibilidade de ajustar os dados de δ a partir de infinitos valores para o par $\alpha\lambda$ e η, além de poder ocorrer a troca de informação entre eles. Para evitar essa troca, η foi configurado conforme ilustrado pela Figura 3.1. Caso nenhuma restrição seja imposta, as colunas da matriz α e as linhas de λ podem alternar as posições. Lopes et al. (2011) adotam uma estratégia de distribuições *a priori* para o componente espacial, tanto para a média quanto para variância, centradas em valores prefixados para resolver o problema. Outras maneiras de tratar essa questão são descritas em Geweke e Zhou (1996), Aguilar e West (2000) e Lopes e West (2004). A estratégia

que adotamos neste trabalho é semelhante à utilizada por Mayrink e Lucas (2013) ao configurar a matriz de cargas α. Assumimos que o conjunto de regiões $\{1, 2, \cdots, L\}$ é particionado em $K + 1$ grupos disjuntos $G_1, G_2, \cdots, G_K$ e G_E (grupo extra). Cada grupo $G_{k \neq E}$ contém regiões associadas somente ao fator k. Os elementos de $G_{k \neq E}$ não são afetados por outros fatores e nem por interações. O grupo G_E representa os locais com associação desconhecida com qualquer fator k. O objetivo é medir a interação entre fatores e identificar os elementos em G_E afetados por essas interações. Considerando as restrições estabelecidas definimos : $\alpha_{lk} = 0$, se $l \notin G_k$ e $l \notin G_E$ e $\eta_{l\bullet} = 0$ para todo $l \in G_{k \neq E}$. A Figura 3.1 ilustra essa restrição em α e η no cenário $K = 3$.

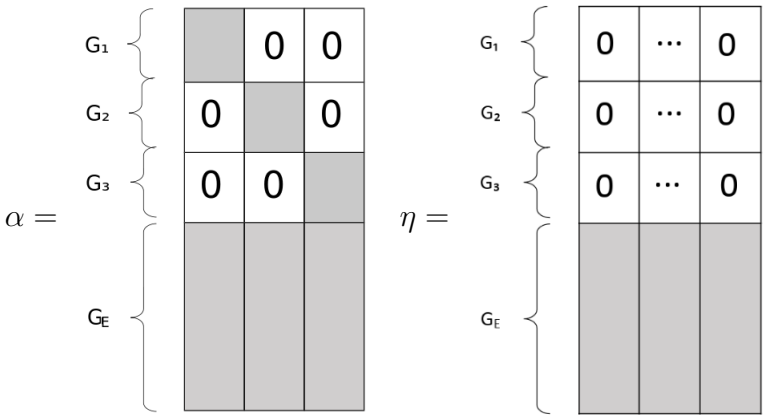


Figure 3.1: Exemplo de configuração das matrizes α e η para $K = 3$ fatores.

As configurações das matrizes α e η serão introduzidos na modelagem através da especificação das distribuições *a priori*. Em termos da matriz η isso é feito com base na Equação (3.6) que é uma mistura em que iremos valorizar uma probabilidade p_l baixa. Ao fazer isso, estamos assumindo que determinados locais não são afetados por interações, por isso é possível aplicar uma distribuição *a priori* favorecendo $\eta_{l\bullet} = \mathbf{0}$. Considerando essa suposição temos que $\alpha_{l\bullet}\lambda + \eta_{l\bullet} = \alpha_{l\bullet}\lambda + \mathbf{0}$ e, sendo assim, $\alpha\lambda$ e η podem ser identificados, pois suas linhas tem liberdade de se comunicarem e compartilhar valores. Já no caso de α considere a Equação (3.11) cuja a explicação é apresentada a seguir. Tendo como referência a matriz de α da Figura 3.1, os valores localizados nas posições de fundo cinza devem ser estimados, ou seja, $\alpha_{lk} \neq 0$, se $l \in G_k$ ou $l \in G_E$. Para

as próximas definições adotaremos a seguinte notação : α_{0k} é um vetor contendo os elementos do conjunto $\{\alpha_{lk} = 0; l \notin G_k \text{ e } l \notin G_E\}$, em que α_{0k} tem dimensão $(L_{0_k} \times 1)$. Denote, também, $\alpha_{\emptyset k}$ como um vetor contendo os elementos do conjunto $\{\alpha_{lk} \neq 0; l \in G_k$ ou $l \in G_E\}$, ou seja, temos um vetor não nulo de cargas dos fatores, de dimensão $(L_{\emptyset_k} \times 1)$, a ser estimado. A ilustração na Figura 3.1 está organizada apenas por questões didáticas, mas, obviamente, o posicionamento dos grupos $G_{k \neq E}$ e G_E não precisa estar em blocos bem definidos, ou seja, as linhas desses grupos podem estar dispersas na matriz.

Considerando as restrições descritas acima, a distribuição *a priori* de α, definida pela Equação (3.4), assume a seguinte formulação :

$$\alpha_{\bullet k} = (\alpha_{\emptyset k},\ \alpha_{0k})^\top | \tau_\alpha \sim N_L(\mathbf{0},\ \tau_\alpha B_k), \quad B_k = \begin{bmatrix} B_{k,11} & B_{k,12} \\ B_{k,21} & B_{k,22} \end{bmatrix}. \tag{3.11}$$

A matriz B_k é construída em blocos a partir da matriz de variância e covariância $B = [D_\alpha - \rho_\alpha W_\alpha]^{-1}$. Seja $\partial_{\emptyset_k}$ o conjunto de índices, contidos em $\{1, \cdots, L\}$, relacionados a $\alpha_{\emptyset k}$ e ∂_{0_k} os índices relacionados a α_{0k}. Usando essa notação, define-se as matrizes $B_{k,11} = B[\partial_{\emptyset_k}, \partial_{\emptyset_k}]_{(L_{\emptyset_k} \times L_{\emptyset_k})}$, $B_{k,22} = B[\partial_{0_k}, \partial_{0_k}]_{(L_{0_k} \times L_{0_k})}$, $B_{k,12} = B[\partial_{\emptyset_k}, \partial_{0_k}]_{(L_{\emptyset_k} \times L_{0_k})}$ e $B_{k,21} = B[\partial_{0_k}, \partial_{\emptyset_k}]_{(L_{0_k} \times L_{\emptyset_k})}$, em que o subscrito $(L_{\bullet_k} \times L_{\bullet_k})$ indica a dimensão da matriz B. Tomando como base propriedades da distribuição Normal Multivariada, obtemos a distribuição condicional

$$(\alpha_{\emptyset k} | \alpha_{0k}, \tau_\alpha) \sim N_{L_{\emptyset_k}}(\mu_{\emptyset_k|0_k}\ ,\ \tau_\alpha B_{\emptyset_k|0_k}). \tag{3.12}$$

em que $\mu_{\emptyset_k|0_k} = \mathbf{0} + B_{k,12}(B_{k,22})^{-1}(\alpha_{0k} - \mathbf{0}) = \mathbf{0}$ e $B_{\emptyset_k|0_k} = B_{k,11} - B_{k,12}B_{k,22}^{-1}B_{k,21}$. Lembrando ao leitor que α_{0k} é vetor nulo, conforme ilustrado na Figura 3.1.

3.1.1 Inferência Bayesiana

Conforme informado anteriormente, utilizamos a inferência Bayesiana para estimação dos parâmetros do modelo apresentado na última seção. Devido à complexidade da distribuição conjunta *a posteriori*, $\mathrm{p}(\beta, \delta, \alpha, \tau_\alpha, \lambda, \eta, p, z, \sigma^2 | Y, X)$, ela não pode ser avaliada analiticamente. Para viabilizar a amostragem indireta da distribuição *a posteriori* desconhecida, aplica-se neste problema o amostrador de Gibbs (Geman e Geman, 1984; Gelfand e

Smith, 1990). Para aqueles parâmetros cuja distribuição condicional completa não possui forma fechada aplicamos o algoritmo Metropolis-Hasting - MH (Metropolis et al., 1953; Hastings, 1970), com a proposta gerada a partir de um passeio aleatório envolvendo a distribuição Gaussiana. Os parâmetros atualizados em cada iteração do MCMC são : β, α, λ, η, δ, σ^2, τ_α, p and z. A seguir apresentamos as etapas do algoritmo. Denote $\mathrm{p}(\zeta|\bullet)$ para representar a distribuição condicional completa do parâmetro genérico ζ dado todos os demais.

1. Atribua valores iniciais para : o vetor β $(q \times 1)$, a matriz α $(L \times K)$, a matriz λ $(K \times T)$, a matriz η $(L \times T)$, a matriz δ $(L \times T)$, o escalar σ^2, o escalar τ_α, o vetor $p = (p_1, \cdots, p_L)^\top$ e o vetor $z = (z_1, \cdots, z_L)^\top$.

2. Inicie o algoritmo amostrando de $(\eta^*|\bullet) \sim N_T(M_{\eta^*},\ V_{\eta^*})$ em que $M_{\eta^*} = V_{\eta^*} \sum_{l=1}^{L} (z_l/\sigma^2)(\delta_{l\bullet}^\top - \lambda^\top \alpha_{l\bullet}^\top)$ e $V_{\eta^*} = [\sum_{l=1}^{L} (z_l/\sigma^2) I_T + \kappa(\lambda)^{-1}]^{-1}$.

3. Com o valor gerado de η^*, para cada $l \in \{1, \cdots, L\}$, calcule :
   $\mathrm{p}(z_l = 1|\bullet) \propto \exp\{(-1/2\sigma^2)\ \left[(\eta^*)^\top \eta^* - 2\eta^*(\delta_{l\bullet} - (\alpha_{l\bullet}\lambda))^\top\right]\ \mathrm{p}(\eta_{l\bullet} = \eta^*|\lambda, z_l = 1)\ p_l$ e
   $\mathrm{p}(z_l = 0|\bullet) \propto \exp\{(-1/2\sigma^2)\ \left[(\mathbf{0})^\top \mathbf{0} - 2\mathbf{0}(\delta_{l\bullet} - (\alpha_{l\bullet}\lambda))^\top\right]\ \mathrm{p}(\eta_{l\bullet} = \mathbf{0}|\lambda, z_l = 0)\ (1 - p_l) = (1 - p_l)$.
   Em seguida realize a normalização : $\mathrm{p}^*(z_l = 1|\bullet) = \frac{\mathrm{p}(z_l=1|\bullet)}{\mathrm{p}(z_l=1|\bullet)+\mathrm{p}(z_l=0|\bullet)}$.
   Se $u \sim U[0,1] < \mathrm{p}^*(z_l = 1|\bullet)$ então $z_l = 1$ e $\eta_{l\bullet} = \eta^*$, sendo η^* aquele amostrado no passo anterior. Naturalmente, se $u \sim U[0,1] \geq \mathrm{p}^*(z_l = 1|\bullet)$, faça $z_l = 0$ e $\eta_{l\bullet} = \mathbf{0}$.
   De posse de z_l amostre $p_l \sim \mathrm{Beta}(a_p + z_l,\ b_p - z_l + 1)$, lembrando que $a_p > 0$ e $b_p > 0$ são hiperparâmetros fixados pelo pesquisador.

4. Para amostrar de $\mathrm{p}(\beta|\bullet)$ considere :
   $\log \mathrm{p}(\beta_j|\bullet)\ \propto\ \beta_j \sum_{i=1}^{n} y_i X_{ij} - \sum_{i=1}^{n} \log\left[1 + \exp\{X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}\}\right] - \frac{1}{2v_\beta}\left[\beta_j^2 - 2\beta_j m_{\beta_j}\right]$, $j \in \{0, 1, 2, \cdots, q-1\}$. Essa expressão não possui forma fechada, então utilizamos o algoritmo MH para amostragem indireta. A geração de propostas é feita por meio de $(\beta_j^{(r)} \mid \beta_j^{(r-1)}) \sim N(\beta_j^{(r-1)},\ \omega_{\beta_j})$, em que $\beta_j^{(r)}$ é o valor gerado de β_j na interação r do MCMC. A tunagem da taxa de aceitação é realizada a partir da atribuição de valor adequado para ω_{β_j}, que garante uma taxa de aceitação entre 30% e 60%, conforme recomendado na literatura (Roberts e Sahu, 1997). Optamos por

amostrar da distribuição univariada β_j, e não do vetor β, devido ao fato de ser difícil a tunagem do MH visando uma taxa de aceitação dentro do intervalo recomendado (Roberts e Sahu, 1997). A escala log foi utilizada por questões computacionais.

5. Na sequência, amostre $(\delta_{l\bullet}|\bullet)$ que também, por não possuir forma fechada, necessita do algoritmo Metropolis-Hasting.

$$\begin{aligned}\log \pi(\delta_{lt}|\bullet) \propto & \sum_{i=1}^{n} y_i \delta_{l_i^* t_i^*} 1_{\{l_i^*=l\}} 1_{\{t_i^*=t\}} \\ & - \sum_{i=1}^{n} 1_{\{l_i^*=l\}} 1_{\{t_i^*=t\}} \log[1+\exp\{X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}\}] \\ & - \frac{1}{2\sigma^2}[\delta_{lt}^2 - 2\delta_{lt}(\alpha_{l\bullet}\lambda_{\bullet t} + \eta_{lt})]\end{aligned}$$

   A proposta MH utiliza um passeio aleatório Gaussiano dado por $(\delta_{lt}^{(r)} \mid \delta_{lt}^{(r-1)}) \sim N(\delta_{lt}^{(r-1)},\ \omega_{\delta_{lt}})$, em que $\delta_{lt}^{(r-1)}$ é o valor de δ_{lt} na iteração $r-1$ (anterior) do MCMC.

6. A amostragem de σ^2 é direta, pois a condicional completa é conhecida, ou seja, $\sigma^2|\bullet \sim GI(a^*_{\sigma^2},\ b^*_{\sigma^2})$, em que
   $a^*_{\sigma^2} = LT/2 + a_{\sigma^2}$ e $b^*_{\sigma^2} = b_{\sigma^2} + 1/2 \sum_{l=1}^{L} \left[\delta_{l\bullet}^\top - (\alpha_{l\bullet}\lambda + \eta_{l\bullet})^\top\right]^\top \left[\delta_{l\bullet}^\top - (\alpha_{l\bullet}\lambda + \eta_{l\bullet})^\top\right]$.

7. A geração de $\mathrm{p}(\alpha_{\emptyset k}|\alpha_{0k} = \mathbf{0})$ é direta, pois $(\alpha_{\emptyset k}|\alpha_{0k}) \sim N_{L_{\emptyset_k}}(M^*_{\emptyset_k|0_k},\ V^*_{\emptyset_k|0_k})$ tal que
   $M^*_{\emptyset_k|0_k} = 1/\sigma^2\ V^*_{\emptyset_k|0_k}\ \sum_{t=1}^{T}(\delta_{\emptyset_k t} - \eta_{\emptyset_k t} - \sum_{k'\neq k} \alpha_{\emptyset k'}\lambda_{k't})\ \lambda_{kt}$ e
   $V^*_{\emptyset_k|0_k} = (1/\tau_\alpha\ B^{-1}_{\emptyset_k|0_k} + 1/\sigma^2\ \sum_{t=1}^{T} \lambda_{kt}^2 I_{L_{\emptyset_k}})^{-1}$.

8. A amostragem de τ_α também é direta, pois $\tau_\alpha|\bullet \sim GI(a^*_{\tau_\alpha},\ b^*_{\tau_\alpha})$, em que
   $a^*_{\tau_\alpha} = a_{\tau_\alpha} + \sum_{k=1}^{K} L_{\emptyset_k}/2$ e $b^*_{\tau_\alpha} = b_{\tau_\alpha} + 1/2 \sum_{k=1}^{K} \alpha_{\emptyset_k}^\top B^{-1}_{\emptyset_k|0_k} \alpha_{\emptyset_k}$.

9. O último passo é a amostragem de $\lambda_{k\bullet}$ cuja condicional completa tem o núcleo
   $\mathrm{p}(\lambda_{k\bullet}|\bullet)\ \propto\ N_T[\lambda_{k\bullet}|M_\lambda, V_\lambda]\ |\kappa(\lambda)|^{-1/2} \exp\{-1/2(\eta^*)^\top \kappa(\lambda)^{-1}\eta^*\}$ sendo
   $N_T[\lambda_{k\bullet}|M_\lambda, V_\lambda]$ a densidade da $N_T(M_\lambda, V_\lambda)$ avaliada em $\lambda_{k\bullet}$. Considere
   $V_\lambda = \left[(1/\tau_\lambda)(D_\lambda - \rho_\lambda W_\lambda) + \sum_{l=1}^{L}\ (\alpha_{lk}^2/\sigma^2) I_T\right]^{-1}$ e
   $M_\lambda = V_\lambda\ (1/\sigma^2) \sum_{l=1}^{L} \alpha_{lk} \left(\delta_{l\bullet}^\top - \eta_{l\bullet}^\top - \sum_{k'\neq k} \alpha_{lk'}\lambda_{k'\bullet}^\top\right)$.
   Esse núcleo não pertence à nenhuma distribuição de probabilidade conhecida. Com isso, aplica-se aqui, novamente, o passo MH, cuja proposta é gerada de $(\lambda_{kt}^{(r)} \mid \lambda_{kt}^{(r-1)}) \sim$

$N(\lambda_{kt}^{(r-1)},\ \omega_{\lambda_{kt}})$. Neste caso $\omega_{\lambda_{kt}}$ é o parâmetro utilizado para tunagem da taxa de aceitação.

Complementando a especificação do amostrador de Gibbs, em todas as análises desenvolvidas nesta tese, consideramos 10000 amostras para o período de aquecimento (*burn-in* ou *warm-up*) e 10000 amostras *a posteriori* com *lag* de 1. A variância para geração das propostas de β_j, δ_{lt} e λ_{kt} foram tunadas de forma a se obter uma taxa de aceitação seguindo as recomendações apresentadas por Roberts e Sahu (1997). O vetor β foi iniciado com zeros, o parâmetro τ_α recebeu o número 1, a matriz α foi preenchida a partir da distribuição $U(-0.1, 0.1)$. Em todas as configurações os chutes iniciais para $\lambda_{1\bullet}$ foram gerados a partir da $U(-2, 2)$ e ordenado de forma decrescente. A inicialização de $\lambda_{2\bullet}$ também foi realizada a partir da $U(-2, 2)$ e os valores ordenados de forma crescente. Em configurações nas quais trabalhamos com mais de 2 fatores, os outros elementos da matriz λ foram fixados com 0 ou 1. O parâmetro σ^2 foi iniciado com 1, o vetor z com zeros, o vetor p de tal forma que $p_l = 0.5$ para $l \in \{\partial_{\emptyset_k}\}$ e $p_l = 0.01$ para $l \in \{\partial_{0_k}\}$.

Após tratar do modelo logístico, discutindo todos os detalhes relacionados à inferência Bayesiana para esse caso, apresentamos a seguir as especificações para o modelo STFM na regressão Poisson. Iremos destacar apenas as principais diferenças entre os dois GLMs.

3.2 STFM na Regressão Poisson

A especificação proposta para o efeito aleatório δ em (3.3) pode ser encaixada em outras distribuições da família exponencial. Para explorar esse aspecto, o trabalho também incluiu o estudo do modelo de regressão Poisson, que é bastante popular, assumindo :

$$
\begin{aligned}
&Y_i|\theta_i \sim \text{Poisson}(\theta_i)\\
&\theta_i = \exp\{X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}\}, \quad i \in \{1, \cdots, n\}\\
&\delta = \alpha\lambda + \eta + \epsilon\\
&\alpha_{\bullet k}|\tau_\alpha \sim N_L(\mathbf{0}, \tau_\alpha[D_\alpha - \rho_\alpha W_\alpha]^{-1}), \quad k \in \{1, \cdots, K\}, \text{ todos independentes}\\
&\lambda_{k\bullet}^\top \sim N_T(\mathbf{0}, \tau_\lambda[D_\lambda - \rho_\lambda W_\lambda]^{-1}), \text{ todos independentes}\\
&\eta_{l\bullet}|p_l, \eta^* \sim (1-p_l)\mathcal{D}_0 + p_l\mathcal{D}_{\eta^*}, \quad l \in \{1, \cdots, L\}, \text{ todos independentes}\\
&\eta^*|\lambda \sim N_T(\mathbf{0}, \kappa(\lambda)), \text{ todos independentes}\\
&\epsilon_{l\bullet}^\top|\sigma^2 \sim N_T(\mathbf{0}, \sigma^2 I_T), \text{ todos independentes}\\
&\beta \sim N_q(M_\beta, v_\beta I_q), \quad M_\beta = (m_{\beta_0}, \cdots, m_{\beta_{q-1}})^\top \quad \text{fixo}\\
&\tau_\alpha \sim GI(a_{\tau_\alpha}, b_{\tau_\alpha}), \quad a_{\tau_\alpha} > 0, \quad \text{e} \quad b_{\tau_\alpha} > 0 \quad \text{fixos}\\
&p_l \sim \text{Beta}(a_p, b_p), \quad a_p > 0, \quad \text{e} \quad b_p > 0 \quad \text{fixos, } p_\bullet \text{ todos independentes}\\
&\sigma^2 \sim GI(a_{\sigma^2}, b_{\sigma^2}), \quad a_{\sigma^2} > 0, \quad \text{e} \quad b_{\sigma^2} > 0 \quad \text{fixos.}
\end{aligned}
\tag{3.13}
$$

Considere, ainda, as mesmas especificações da seção anterior para os termos a seguir : $X_{\bullet i}$, D_α, ρ_α, W_α, D_λ, W_λ, τ_λ, ρ_λ, $\mathcal{D}_0$, $\mathcal{D}_{\eta^*}$ e $\kappa(\lambda)$. As distribuições para $(\alpha_{\emptyset k},\ \alpha_{0k})^\top$ e $(\alpha_{\emptyset k}|\alpha_{0k}, \tau_\alpha)$ são as mesmas definidas pelas Equações (3.11) e (3.12). No entanto, nesta modelagem o termo $\theta_i > 0$, pois θ_i é a taxa da Poisson. A função de verossimilhança em nível amostral assume a seguinte formulação :

$$
\text{p}(Y|X, \beta, \delta) = \prod_{i=1}^{n} \frac{\theta_i^{y_i}}{y_i!} e^{-\theta_i} = \prod_{i=1}^{n} \frac{1}{y_i!} \exp\{y_i X_{\bullet i}^\top \beta + y_i \delta_{l_i^* t_i^*}\} \exp\{-\exp\{X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}\}\}
\tag{3.14}
$$

As condicionais completas que tem formulação diferentes, em relação à Seção 3.1, são aquelas que dependem de Y, ou seja, β e δ. As demais condicionais permanecem inalteradas. Com isso, os únicos passos a serem modificados no algoritmo apresentado na Seção 3.1.1 são os Passos 4 e 5, que reescrevemos a seguir :

4. Para obter uma amostra de β utilizamos a condicional completa
   $\log \text{p}(\beta_j|\bullet) \ \propto \ \beta_j \sum_{i=1}^n y_i X_{ij} - \sum_{i=1}^n \exp\{X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}\} - \frac{1}{2v_\beta}\left[\beta_j^2 - 2\beta_j m_{\beta_j}\right]$,

$j \in \{0, 1, 2, \cdots, q-1\}$. Como esse núcleo não possui forma fechada, aplica-se o algoritmo MH. A geração da proposta e a tunagem foram efetuadas da mesma forma definida no Passo 4 da Seção 3.1.1.

5. A constante normalizadora de $\mathrm{p}(\delta_{l\bullet}|\bullet)$ também é desconhecida, portanto o algoritmo MH é requisitado nesta etapa. Aqui foi aplicado o mesmo procedimento para geração da proposta e a tunagem definida no Passo 5 da Seção 3.1.1. Temos o núcleo dado por

$$\begin{aligned}\log \mathrm{p}(\delta_{lt}|\bullet) \propto & \sum_{i=1}^{n} y_i \delta_{l_i^* t_i^*} 1_{\{l_i^*=l\}} 1_{\{t_i^*=t\}} \\ & - \sum_{i=1}^{n} 1_{\{l_i^*=l\}} 1_{\{t_i^*=t\}} \exp\{X_{\bullet i}^{\top}\beta + \delta_{l_i^* t_i^*}\} \\ & - \frac{1}{2\sigma^2}[\delta_{lt}^2 - 2\delta_{lt}(\alpha_{l\bullet}\lambda_{\bullet t} + \eta_{lt})].\end{aligned}$$

As configurações do algoritmo MCMC, que é um Gibbs *sampling* com passos Metropolis, foi totalmente detalhada para os dois GLMs. No caso do modelo Poisson assumimos as mesmas configurações especificadas para o modelo logístico, ou seja, tamanho da amostra de 10000, *burn-in* de 10000 e *lag* igual 1. Os valores iniciais de todos os parâmetros, exceto β, também foram os mesmos. O chute inicial de β_0, para o modelo Poisson, teve que ser diferente para o caso de $\approx 3\%$ de contagens iguais zero. Essa situação está descrita na Seção 5.1.

3.3 Comentários finais do capítulo

Concluimos, aqui, o capítulo sobre modelos lineares generalizados mistos com estrutura do STFM relacionada às regressões logística e Poisson. Primeiramente foi apresentado o modelo logístico que é o principal motivador desta tese porque ele está relacionado ao banco de dados que iremos descrever no Capítulo 6. O outro modelo discutido foi o Poisson que é uma aplicação extra envolvendo outro membro da família exponencial.

Nos próximos dois capítulos, 4 e 5, iremos desenvolver um estudo simulado bastante amplo considerando vários aspectos desses modelos para explorar o comportamento deles

diante de vários cenários de configurações, como por exemplo : mesma quantidade de fatores entre dados simulados e estimados, diferentes quantidade de fatores, variação na quantidade de locais e de vizinhos, número de locais que tiveram interações não lineares.

Capítulo 4

Estudo simulado logístico

O modelo logístico definido na Seção 3.1 foi especificado para analisar um vetor Y de tamanho n, preenchido com 0s e 1s, indicando ausência e presença de determinado evento, respectivamente. Temos também uma matriz X com dados de q variáveis preditoras para as n amostras. A aplicação de interesse considera que as amostras foram observadas durante T tempos em L localidades e que existem K grupos (fatores) de localidades com características semelhantes. Considere G_k um grupo de localidades que se sabe estarem associadas ao fator k, em que, por exemplo, $k \in \{1, 2\}$. Não se assume qualquer associação entre as amostras de G_1 com o Fator 2, ou de G_2 com o Fator 1. Conforme descrito no Capítulo 3, considere G_E um conjunto de amostras extras para o qual se desconhece o grupo a que pertencem, G_1 ou G_2. Os conjuntos G_1, G_2 e G_E são disjuntos, e apenas G_E deve ser afetado por interações. Essas premissas são necessárias para abordar os problemas de identificação no modelo fatorial. A partir do número de locais e de vizinhos por região é gerada a matriz de vizinhança W_α e a matriz diagonal D_α com o número de vizinhos. Levando em conta que T é a quantidade de tempos, gera-se a matriz de vizinhança W_λ que, juntamente com D_λ, definem a estrutura temporal do modelo.

A Tabela 4.1 enumera todas as configurações utilizadas na geração das bases de dados. O número de locais igual a 400 foi escolhido baseado no número de municípios existentes na base de dados real. Os demais valores para L foram selecionados para avaliar o comportamento do modelo em situações nas quais temos um menor número de regiões. Da mesma forma, o número de tempos igual a 4 também se baseou na quantidade de

Configurações utilizadas na geração dos dados						
Modelo	Locais (L)	Tempos (T)	Fatores (K)	Vizinhos	% Interação	% $Y = 1$
$M^{K_2I_{30\%}}_{L_{100}T_4V_v}$	100					
$M^{K_2I_{30\%}}_{L_{200}T_4V_v}$	200	4	2	$v \in \{4, 6\}$	$\approx 30\%$	$\approx 50\%$
$M^{K_2I_{30\%}}_{L_{400}T_4V_v}$	400					
$M^{K_2I_{50\%}}_{L_{100}T_4V_v}$	100					
$M^{K_2I_{50\%}}_{L_{200}T_4V_v}$	200	4	2	$v \in \{4, 6\}$	$\approx 50\%$	$\approx 50\%$
$M^{K_2I_{50\%}}_{L_{400}T_4V_v}$	400					
$M^{K_3I_{50\%}}_{L_{400}T_4V_4}$		4	3			
$M^{K_4I_{50\%}}_{L_{400}T_4V_4}$		4	4			
$M^{K_5I_{50\%}}_{L_{400}T_4V_4}$	400	4	5	4	$\approx 50\%$	$\approx 50\%$
$M^{K_2I_{50\%}}_{L_{400}T_{10}V_4}$		10	2			
$M^{K_3I_{50\%}}_{L_{400}T_{10}V_4}$		10	3			
$M^{K_2I_{50\%}}_{L_{400}T_{10}V_6}$	400	10	2	6	$\approx 50\%$	$\approx 50\%$
$M^{K_3I_{50\%}}_{L_{400}T_4V_6}$	400	4	3	6	$\approx 50\%$	$\approx 50\%$
$M^{K_2I_{50\%}}_{L400T_4V_4}$	400	4	2	4	$\approx 50\%$	$\approx 20\%$

Tabela 4.1: Configurações para geração de dados com variável resposta binária (modelo logístico) variando o número de locais, número de tempos, número de fatores, número de vizinhos por local, percentual de locais de G_E que possuem interação não linear e porcentagem de respostas iguais a 1.

anos disponíveis na base de dados real. Como os dados reais fazem parte de um processo onde novas observações são incluídas diariamente (Sistema de Telessaúde), a opção por $T = 10$ teve como objetivo avaliar a situação na qual se terá, futuramente, um volume maior de dados históricos. O número de fatores $K = 2$ está relacionado à aplicação real, em que apenas 2 fatores são definidos para representar o comportamento global de cidades apresentando baixo e alto índices de desenvolvimento. Os grupos G_1 e G_2 incluem locais nos quais supõe-se que há tendência de diminuição (G_1) ou aumento (G_2) na probabilidade de sucesso (mortalidade por IAM) em determinado período de tempo.

Configurações **comuns** entre as bases de dados geradas		
Covariável	Dimensão	Valor gerado
X	$(35L) \times 3$	
$X_{\bullet 1}$	$(35L) \times 1$	$\mathbf{1}$
$X_{\bullet 2}$	$(35L) \times 1$	Bernoulli(0.5)
$X_{\bullet 3}$	$(35L) \times 1$	$U(-1, 1)$
Parâmetro	Dimensão	Valor real
τ_α	escalar	2
ρ_α	escalar	0.9
σ^2	escalar	0.8
β_{C_1}	(1×3)	$(0.5, -1.0, 1.0)$
β_{C_2}	(1×3)	$(-1.5, -1.5, 1.0)$
ϵ	$(L \times T)$	$N(0, \sigma^2)$

Tabela 4.2: Geração das covariáveis (matriz X), valores reais dos coeficientes β e de parâmetros comuns para todas as bases de dados utilizadas na geração dos diferentes cenários apresentados na Tabela 4.1. A configuração β_{C_1} foi escolhida para garantir $\approx 50\%$ de sucessos em Y, enquanto que a configuraçao β_{C_2} estabelece $\approx 20\%$.

O cenário com $K = 3$ foi utilizado para avaliação do modelo em contextos de escolha diferentes para K.

Optamos por trabalhar, nesta análise, com estrutura de vizinhança contendo 4 ou 6 vizinhos para a maioria das regiões. Por serem dados simulados, a definição de vizinhança foi configurada por uma matriz banda diagonal dupla (diagonal principal com zeros e 2 diagonais acima e 2 abaixo da diagonal principal com 1's.). Note que 2 vizinhos por região remete a uma série temporal. Visto que desejamos uma vizinhança com mais de 2 conexões, escolheu-se o valor mínimo de 4. A matriz de vizinhança é banda diagonal (Mayrink e Gamerman, 2009) estabelecendo a mesma quantidade de vizinhos para a maioria dos locais. Visando maior semelhança com os dados reais da Telessaúde, estudamos a versão de 6 vizinhos por região, pois tal valor equivale à mediana do número

Configurações de α para geração das bases de dados nos contextos de K = 2 e K = 3.			
$K = 2$			
Parâmetro	Índice	Dimensão	Valor real
α_{lk}	$l \in \{1, \cdots, 10\},\ k = 1$ $l \in \{1, \cdots, 10\},\ k = 2$	10×1	$U(1, 2)$ $\mathbf{0}$
α_{lk}	$l \in \{11, \cdots, 20\}, k = 1$ $l \in \{11, \cdots, 20\},\ k = 2$	10×1	$\mathbf{0}$ $U(1, 2)$
α_{G_E}	$l \in \{21, \cdots, L\},\ k \in \{1, 2\}$	$(L - 20) \times 2$	$\mathrm{N}_{(L-20)}(\mathbf{0}, [D_\alpha - \rho_\alpha W_\alpha]^{-1})$
$K = 3$			
α_{lk}	$l \in \{1, \cdots, 10\},\ k = 1$ $l \in \{1, \cdots, 10\},\ k = 2$ $l \in \{1, \cdots, 10\},\ k = 3$	10×1	$U(1, 2)$ $\mathbf{0}$ $\mathbf{0}$
α_{lk}	$l \in \{11, \cdots, 20\}, k = 1$ $l \in \{11, \cdots, 20\},\ k = 2$ $l \in \{11, \cdots, 20\},\ k = 3$	10×1	$\mathbf{0}$ $U(1, 2)$ $\mathbf{0}$
α_{lk}	$l \in \{21, \cdots, 30\}, k = 1$ $l \in \{21, \cdots, 30\},\ k = 2$ $l \in \{21, \cdots, 30\},\ k = 3$	10×1	$\mathbf{0}$ $\mathbf{0}$ $U(1, 2)$
α_{G_E}	$l \in \{31, \cdots, L\},\ k \in \{1, 2, 3\}$	$(L - 30) \times 3$	$\mathrm{N}_{(L-30)}(\mathbf{0}, [D_\alpha - \rho_\alpha W_\alpha]^{-1})$

Tabela 4.3: Valores da matriz de cargas (*loadings*), utilizados na geração dos diferentes conjutos de dados baseados nos cenários apresentados na Tabela 4.1 e considerando os números de fatores $K = 2$ e $K = 3$.

de vizinhos dos municípios do estado de Minas Gerais.

É razoável pensar que quanto maior a porcentagem de locais afetados pela interação, melhor deve ser a estimação de η^*. Esta ideia surge do fato de que informações de mais regiões seriam usadas para atualizar a incerteza *a priori* sobre η^*. Seguindo esse pensamento, propomos a avaliação de duas porcentagens, 30% e 50%, que indicam o número de municípios no grupo G_E que são afetados por η^*. Percentuais acima de 50%

Configurações de λ para geração das bases de dados.		
$K = 2$ e $T = 4$		
Parâmetros	Dimensão	Valor real
$\lambda_{1\bullet}$	$1 \times T$	$(2.0, 1.5, 0.5, -0.5)$
$\lambda_{2\bullet}$		$(-1.0, 1.0, 1.5, 2.0)$
$K = 2$ e $T = 10$		
$\lambda_{1\bullet}$	$1 \times T$	$(2.0, 1.7, 1.5, 1.2, 1.0, 0.7, 0.5, -0.1, -0.3, -0.5)$
$\lambda_{2\bullet}$		$(-1.0, -0.7, -0.5, 0.5, 1.0, 1.2, 1.5, 1.6, 1.8, 2.0)$
$K = 3$ e $T = 4$		
$\lambda_{1\bullet}$	$1 \times T$	$(2.0, 1.5, 0.5, -0.5)$
$\lambda_{2\bullet}$		$(-1.0, 1.0, 1.5, 2.0)$
$\lambda_{3\bullet}$		$(1.0, -1.0, 1.0, -1.0)$
$K = 3$ e $T = 10$		
$\lambda_{1\bullet}$	$1 \times T$	$(2.0, 1.7, 1.5, 1.2, 1.0, 0.7, 0.5, -0.1, -0.3, -0.5)$
$\lambda_{2\bullet}$		$(-1.0, -0.7, -0.5, 0.5, 1.0, 1.2, 1.5, 1.6, 1.8, 2.0)$
$\lambda_{3\bullet}$		$(1.0, -1.0, 1.0, -1.0, 1.0, -1.0, 1.0, -1.0, 1.0, -1.0)$
$K = 4$ e $T = 4$		
$\lambda_{1\bullet}$	$1 \times T$	$(2.0, 1.5, 0.5, -0.5)$
$\lambda_{2\bullet}$		$(-1.0, 1.0, 1.5, 2.0)$
$\lambda_{3\bullet}$		$(1.0, -1.0, 1.0, -1.0)$
$\lambda_{4\bullet}$		$(1.0, -1.0, -1.0, 1.0)$
$K = 5$ e $T = 4$		
$\lambda_{1\bullet}$	$1 \times T$	$(2.0, 1.5, 0.5, -0.5)$
$\lambda_{2\bullet}$		$(-1.0, 1.0, 1.5, 2.0)$
$\lambda_{3\bullet}$		$(1.0, -1.0, 1.0, -1.0)$
$\lambda_{4\bullet}$		$(1.0, -1.0, -1.0, 1.0)$
$\lambda_{5\bullet}$		$(-1.0, -1.0, 1.0, 1.0)$

Tabela 4.4: Valores da matriz de fatores, λ, utilizados na geração dos diferentes conjuntos de dados baseados nos cenários apresentados na Tabela 4.1 e considerando diversas variações de números de fatores e tempos.

Dados das bases geradas para simulação de η para $T = 4$.			
$K = 2$ e $T = 4$			
Parâmetros	Dimensão	Valor	Índices
η^*	$1 \times T$	$(-2.0, 1.5, 0.75, -1.0)$	conjunto aleatório C_{η^*} de índices $l \in \{21, \cdots, (L - 20)\}$ de tamanho $\approx 50\%$ ou 30% de G_E.
$\eta_{l\bullet}$	$1 \times T$	$\mathbf{0}$	$\forall\, l \notin C_{\eta^*}$
$K = 3$ e $T = 4$			
η^*	$1 \times T$	$(-2.0, -1.5, 0.75, 1.0)$	conjunto aleatório C_{η^*} de índices $l \in \{31, \cdots, (L - 30)\}$ de tamanho $\approx 50\%$ ou 30% de G_E.
$\eta_{l\bullet}$	$1 \times T$	$\mathbf{0}$	$\forall\, l \notin C_{\eta^*}$

Tabela 4.5: Valores da matriz de interações não lineares, η, utilizados na geração dos diferentes conjuntos de dados para os cenários apresentados na Tabela 4.1 e considerando $T = 4$ tempos e $K \in (2, 3)$ fatores.

geraria um número muito grande de regiões com interações e isso pode não ser uma suposição razoável. Para percentuais abaixo de 30% (ex.: 10%) iremos estimar mal o valor de η^*, devido ao pequeno número de locais. Neste caso, estaríamos trabalhando com um modelo mais complexo que talvez tenha o mesmo desempenho que aquele mais parcimonioso, ou seja, sem interações.

Complementando a análise da Tabela 4.1, as opções para o percentual de sucesso, $\approx$ 50% e $\approx$ 20%, levou em consideração situações de bases de dados balanceadas e desbalanceadas em relação à quantidade de 0s e 1s, respectivamente.

A Tabela 4.2 detalha os valores gerados e reais atribuídos a termos comuns (covariáveis e parâmetros) a todas as bases de dados das simulações. O valor 35, que aparece na dimensão da matriz X, foi escolhido tendo como referência a proporção de número de

Dados das bases geradas para simulação de η para $T = 4$ - continuação.			
$K = 4$ e $T = 4$			
η^*	$1 \times T$	$(-2.0, 1.5, -0.75, 1.0)$	conjunto aleatório C_{η^*} de índices $l \in \{41, \cdots, (L - 40)\}$ de tamanho $\approx 50\%$ ou 30% de G_E.
$\eta_{l\bullet}$	$1 \times T$	$\mathbf{0}$	$\forall\, l \notin C_{\eta^*}$
$K = 5$ e $T = 4$			
η^*	$1 \times T$	$(2.0, -1.5, -0.75, 1.0)$	conjunto aleatório C_{η^*} de índices $l \in \{51, \cdots, (L - 50)\}$ de tamanho $\approx 50\%$ ou 30% de G_E.
$\eta_{l\bullet}$	$1 \times T$	$\mathbf{0}$	$\forall\, l \notin C_{\eta^*}$

Tabela 4.6: Valores da matriz de interações não lineares, η, utilizados na geração dos diferentes conjuntos de dados para os cenários apresentados na Tabela 4.1 e considerando $T = 4$ tempos e $K \in (4, 5)$ fatores.

indivíduos por local da base de dados real. A escolha das distribuições, Bernoulli(0.5) para X_{i2} e $U(-1, 1)$ para X_{i3}, imitam situações com uma covariável binária (categórica) e outra contínua, semelhante ao que temos nos dados de eletrocardiogramas do Sistema de Telessaúde. A atribuição do valor de $\rho_\alpha = 0.9$ objetiva tornar a matriz de covariância $[D_\alpha - \rho_\alpha W_\alpha]$ não singular (Banerjee et al., 2004). Assuncao e Krainski (2009) fazem um estudo sobre a interpretação dos valores de ρ em que mostram que a dependência espacial não cresce linearmente. Como nosso objetivo não é avaliar o comportamento desse elemento, fixamos em 0.9 seguindo o estudo de Banerjee et al. (2004). Dentre os trabalhos futuros está a estimação desse parâmetro para avaliar se existe uma dependência espacial entre os elementos de α. Os parâmetros τ_λ e ρ_λ não precisam ser especificados porque, como pode ser visto na Tabela 4.4, os valores para λ foram fixados. Os elementos τ_λ e ρ_λ serão definidos nos ajustes dos modelos. Para o parâmetro β foram utilizadas duas

Dados das bases geradas para simulação de η para $T = 10$.			
$K = 2$ e $T = 10$			
η^*	$1 \times T$	$(-2.00, -1.19, -0.75, 0.60, 1.00,$ $0.84, 0.75, -0.16, -0.54, -1.00)$	conjunto aleatório C_{η^*} de índices $l \in \{21, \cdots, (L - 20)\}$ de tamanho $\approx 50\%$ ou 30% de G_E.
$\eta_{l\bullet}$	$1 \times T$	$\mathbf{0}$	$\forall\, l \notin C_{\eta^*}$
$K = 3$ e $T = 10$			
η^*	$1 \times T$	$(-2.00, 1.19, -0.75, -0.60, 1.00,$ $-0.84, 0.75, 0.16, -0.54, 1.00)$	conjunto aleatório C_{η^*} de índices $l \in \{31, \cdots, (L - 30)\}$ de tamanho $\approx 50\%$ ou 30% de G_E.
$\eta_{l\bullet}$	$1 \times T$	$\mathbf{0}$	$\forall\, l \notin C_{\eta^*}$

Tabela 4.7: Valores da matriz de interações não lineares, η, utilizados na geração dos diferentes conjuntos de dados para os cenários apresentados na Tabela 4.1 e considerando $T = 10$ tempos e $K \in (2, 3)$ fatores.

configurações, uma direcionada para a obtenção de $\approx 50\%$ de sucessos ($Y_i's = 1$) e outra para estabelecer $\approx 20\%$. Em uma análise de sensibilidade, verificamos que β_0 é o parâmetro que mais influencia na propabilidade de sucesso, justificando uma maior variação de seu valor nas duas configurações. Os erros, ϵ, foram estabelecidos seguindo o padrão de média 0 e, para se ter um modelo mais parcimonioso, consideramos uma variância única, σ^2, para todas as localidades. Note que em uma análise fatorial usual admite-se uma variância para cada local.

A Tabela 4.3 mapea a configuração da matriz de cargas (*loadings*) nos contextos $K = 2$ e $K = 3$. Para cada fator k é definido o grupo de locais, G_k, associados a ele. Isso é feito atribuindo valores da distribuição $U(1, 2)$ para as posições da matriz

α que referenciam as regiões ligadas ao fator k, e zeros para as demais posições que definem a inexistência de relação desses locais com os demais fatores. A distribuição $U(1,2)$ foi escolhida porque sinais de cargas iguais (neste caso, positivos) simulam uma mesma direção para a associação de todos os municípios do grupo com o fator. Forçamos cargas elevadas (maiores do que zero), caso contrário a estimação não seria significativa, perdendo, com isso, a interpretação do fator. O grupo extra, G_E, define a estrutura espacial do modelo. Desta forma, as posições da matriz são preenchidas a partir de uma distribuição Normal Multivariada de acordo com as especificações do modelo CAR. A configuração da matriz de cargas para os casos em que $K = 4$ e $K = 5$ encontra-se no Apêndice A.

A Tabela 4.4 especifica os valores atribuídos para a matriz de fatores λ. Note que, em todas as configurações, $\lambda_{1\bullet}$ apresenta um padrão de decrescimento e $\lambda_{2\bullet}$ de crescimento. O objetivo dessas atribuições está no fato de querermos simular uma situação que faça sentido na análise dos dados reais, em que o objetivo é avaliar a existência de municípios cuja probabilidade de morte diminui ou aumenta com o tempo. Os valores de λ não afetam diretamente essa probabilidade, mas mantendo todos os demais termos fixos, a variação de λ acarretará uma variação de δ no mesmo sentido, consequentemente atuando no valor da razão de probabilidades e, indiretamente, na probabilidade de morte. As atribuições para os demais fatores foram realizadas de forma a se ter padrões de comportamento no tempo diferentes entre si.

As Tabelas 4.5, 4.6 e 4.7 apresentam os valores atribuídos para as linhas de η, elemento do modelo configurado para captar a interação não linear entre os fatores. A interação real escolhida para avaliação foi o produto entre os escores de cada fator para cada tempo. A 3º coluna mostra o resultado desse produto para diferentes valores de K. Verifique a Tabela 4.4 para identificar os produtos quando $K = 2$, $K = 3$, $K = 4$ e $K = 5$.

Finalmente, para a geração dos valores em Y (0 ou 1), é necessário definir a relação observação-local-tempo. Essa tarefa foi realizada de forma aleatória sorteando, com reposição, os índices com igual probabilidade para formar os vetores l^* e t^* de tamanho n. Uma vez definido esses vetores e as atribuições descritas pelas Tabelas 4.2, 4.3, 4.4, 4.5, 4.6 e 4.7, o preenchimento do vetor Y é imediato. Primeiramente calcula-se $\delta = \alpha\lambda + \eta + \epsilon$.

Em seguida, $\theta_i = \frac{\exp\{X_{i\bullet}\beta+\delta_{l_i^* t_i^*}\}}{1+\exp\{X_{i\bullet}\beta+\delta_{l_i^* t_i^*}\}}$, em que $i \in (1, \cdots, n)$, e, completando a geração dos dados temos, $Y_i = \text{Bernoulli}(\theta_i)$.

Conforme apresentado na Tabela 4.1, vários cenários foram avaliados para o modelo descrito na Seção 3.1. Em aplicações reais de classificação é muito comum encontrarmos um desbalanceamento entre o número de observações definidas nas classe 1 e 0 da resposta binária. Nos dados da Telessaúde, considerados na aplicação real desta tese, tal situação ocorre. No estudo a ser desenvolvido agora, descreveremos os resultados completos para os cenários $M^{K_2 I_{50\%}}_{L_{400}T_4V_4}$, $M^{K_2 I_{30\%}}_{L_{400}T_4V_4}$ ambos com $\approx 50\%$ e $\approx 20\%$ de $Y_i^{'}s = 1$, ou seja, dados balanceados e desbalanceados em relação ao número de $Y's = 1$. Na sequência, teremos a comparação das estimativas obtidas para η^* e λ para todas as configurações de números de locais (100, 200 e 400) e variando o percentual de regiões de G_E afetadas por interação (30% e 50%). Em seguida ilustraremos os resultados quando variamos o número de fatores e de tempos. Depois teremos a análise dos resultados quando ocorre sobreparametrização do modelo ao assumir uma quantidade de fatores acima do verdadeiro. Finalizamos a seção com a avaliação de resíduos, das curvas ROC e do vício relativo após o ajuste para 30 réplicas de Monte Carlo. Os resultados para os demais cenários estão descritos nos Apêndices B e C.

4.1 Análise para dados balanceados

As análises desenvolvidas aqui são dedicadas às configurações com 400 regiões, 2 fatores, 4 tempos, 4 vizinhos, número de sucessos de aproximadamente 50%, ou seja, $\approx 50\%$ de $Y_i^{'}s = 1$, mas variando a quantidade de locais de G_E afetados por interação, 50% ou 30%. O cenário com 400 locais e 4 anos foi selecionado por ser o mais próximo do encontrado na base de dados do sistema da Telessaúde. A escolha de 4 vizinhos foi para simplificar, uma vez que a mediana de vizinhos nas regiões de Minas Gerais é igual a 6 e os resultados para esse caso foram muito semelhantes aos cenários com 4 vizinhos. O nosso estudo não se restringiu apenas a esses cenários, conforme mostrado na Tabela 4.1. Alguns resultados adicionais serão apresentados no Apêndice.

Cenário com $\approx 50\%$ de locais em G_E afetados por interação

A Tabela 4.8 contém as estimativas *a posteriori* dos parâmetros β, σ^2, τ_α e η^* para o cenário $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$. Podemos verificar que o resultado de todos eles, com exceção de η^*_3 e η^*_4, está dentro do intervalo HPD de 95%. A estimativa para β_2 foi, especialmente, muito boa e obteve o menor erro padrão. Como pode ser visto pela Figura 4.1, η^* conseguiu captar a tendência inicial de crescimento e, depois, decrescimento, apesar das estimativas para η^*_3 e η^*_4.

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.50	0.35	0.35	0.10	0.17	0.56
β_1	-1.00	-0.99	-0.99	0.05	-1.09	-0.90
β_2	1.00	0.99	0.99	0.04	0.90	1.08
σ^2	0.80	0.93	0.93	0.13	0.68	1.17
τ_α	2.00	1.51	1.35	0.58	0.65	2.79
η^*_1	-2.00	-2.10	-2.10	0.33	-2.75	-1.42
η^*_2	1.50	1.81	1.80	0.28	1.28	2.35
η^*_3	0.75	1.27	1.28	0.22	0.84	1.72
η^*_4	-1.00	-0.33	-0.32	0.26	-0.87	0.18

Tabela 4.8: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Cenário: $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i' s = 1$.

A Figura 4.2 ilustra mapas de calor comparando valores verdadeiros e estimados para α, λ e δ. Vemos que o padrão real das matrizes foi bem capturado no ajuste do modelo Bayesiano. Visualmente, podemos destacar um exemplo de diferença maior nas estimativas para regiões próximas do número 100, em que o valor verdadeiro parece menor que o estimado. No entanto, os resultados para λ foram bastante similares ao real, tendo como destaque, a captura do padrão de decrescimento para o Fator 1 e de crescimento para o Fator 2. Analisando o mapa para δ percebemos que a diferença ocorrida na estimação de α, para locais próximos do número 100, não se repetiu, sendo compensada,

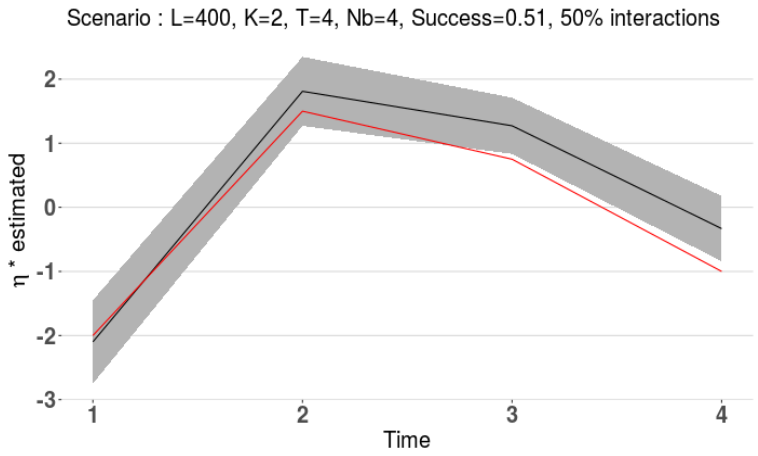


Figure 4.1: Gráfico da média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha) para o cenário $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i' s = 1$.

provavelmente, pela interação. Ainda podemos identificar diferenças para alguns locais, como por exemplo, próximo ao número 162 no Tempo 2, e próximo aos locais de número 281 e 321, no Tempo 4. Entretanto, mais uma vez, o padrão global também foi seguido em todos os períodos para a grande maioria das regiões.

Na Figura 4.3, as estimativas de α (a) e δ (c) foram consolidadas e ordenadas para facilitar a análise comparativa. No Painel (a), o intervalo em que o valor verdadeiro e estimado de α são iguais a 0, refere-se aos locais de G_1 e G_2 para os quais a distribuição *a priori* foi aquela definida em (3.12), lebrando que essa configuração é necessária para identificabilidade do modelo. Nos dois gráficos, (a) e (c), podemos ver que os valores estimados seguem a mesma tendência dos valores verdadeiros. Percebe-se que para as cargas há um desalinhamento (sobre-estimação ou subestimação) nos valores extremos, mas para o parâmetro δ, o qual concentra toda a variabilidade estocástica espaço-temporal e da interação ($\delta = \alpha\lambda + \eta + \epsilon$), o ajuste está, na maior parte, bem alinhado com o valor verdadeiro. O Painel (b) ilustra os resultados para os escores dos fatores em λ. Todas as estimativas estão muito próximas ou, praticamente, iguais ao valor verdadeiro, condizente com os resultados apresentados nos Paineis (c) e (d) da Figura

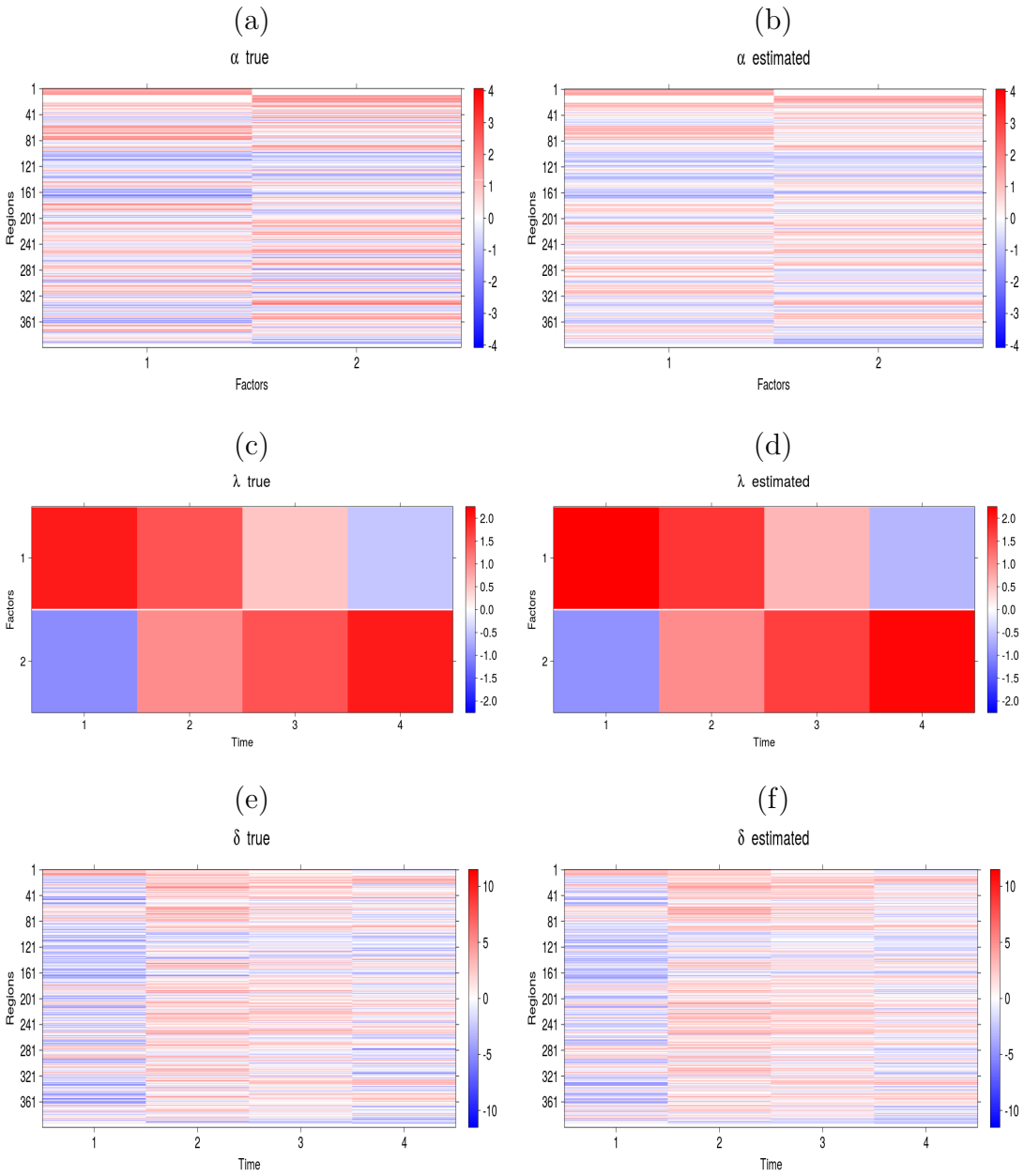


Figure 4.2: Mapas de calor comparando valores verdadeiros e estimados para o cenário $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$. Paineis: (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

4.2. Finalmente, o Painel (d) mostra as probabilidades estimadas de haver interações para cada local (ver passo 3 da Seção 3.1.1). As regiões são destacadas de acordo com a geração dos dados. A cor azul representa os locais configurados para não ter interação em G_1 ou G_2. A cor vermelha ilustra aquelas regiões, do grupo G_E, que tiveram interação. A cor preta representa as regiões de G_E que não tiveram interação. Analisando os locais e as probalilidades estimadas, constata-se que a maioria das probalilidades estimadas concorda com o valor verdadeiro. A maior parte (81.58%) das regiões de cor vermelha aparece com probabilidades acima de 0.5 e as de cor preta (80.95%) com probabilidades abaixo de 0.5.

A Figura 4.4 mostra mapas fictícios com a estrutura espacial dos dados artificiais com 4 vizinhos por região, sendo cada ponto representante de um local. A partir deles pode-se avaliar, visualmente, a formação de agrupamentos (*clusters*) relacionados aos fatores. Nos Paineis (c) e (d) temos os locais associados ao Fator 1 cujas cargas assumem valor negativo (cor azul) e positivo (cor vermelha). De forma equivalente, os Paineis (e) e (f) ilustram os locais associados ao Fator 2. Os Paines (a) e (b) destacam as regiões associadas aos Fatores 1 e 2, concomitantemente. O Painel (g) destaca locais que não estão associados a nenhum fator, mas que foram afetados por interação, e o Painel (h) identifica os locais que não foram afetados por qualquer tipo de efeito. O critério para definir a região afetada pela interação é verificar se a probabilidade $\mathrm{p}^*(z_l = 1|\bullet) > 0.5$. Essa probabilidade está definida no Passo 3 do algoritmo MH descrito na Seção 3.1.1. Para verificar se o local foi afetado por algum efeito principal (Fator 1 e/ou Fator 2), o critério utilizado é examinar se o intervalo HPD de 95% das cargas não inclui o 0.

A análise desenvolvida a partir da Figura 4.4 ilustra um tópico presente no título desta tese que é a avaliação da existência de *clusters* que são capturados pelo modelo fatorial. Conforme pode ser visto pelos grafos, os pontos (locais) formando conglomerados não precisam estar próximos uns dos outros, isto é, não precisam ser vizinhos. A formação dos *clusters* não é por localização, mas sim pela característica de ser influenciado ou não por algum efeito principal ou interação.

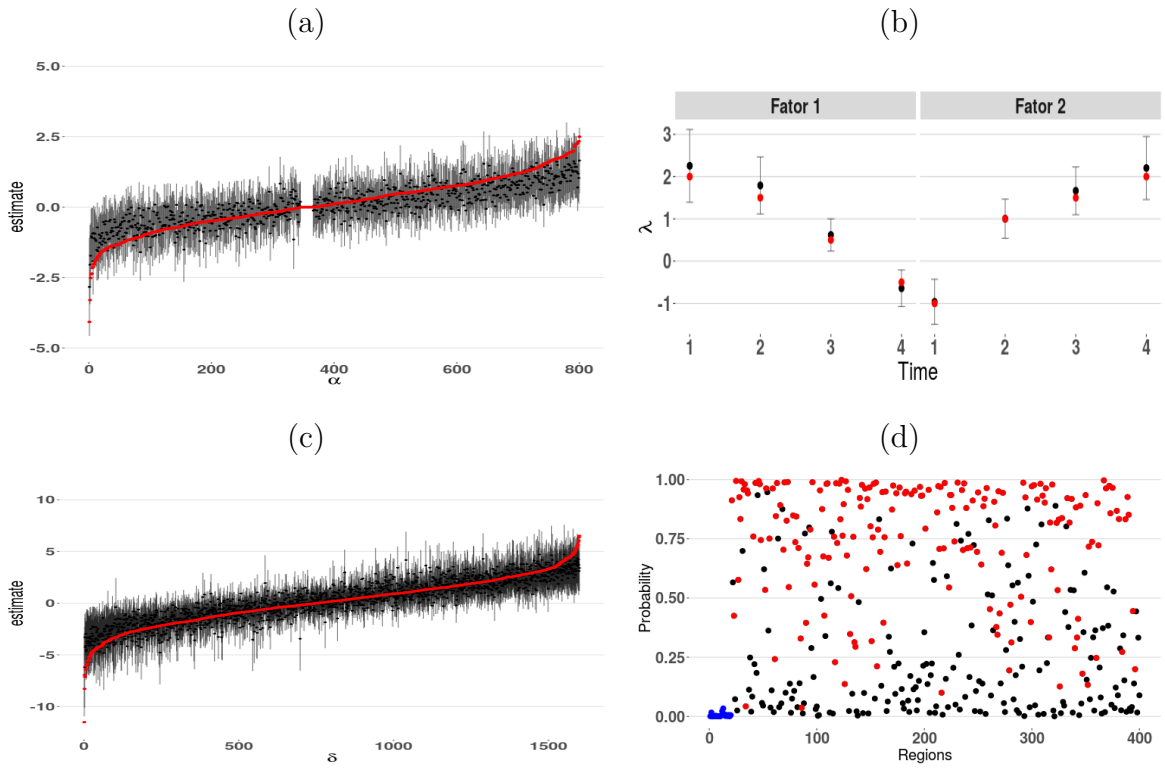


Figure 4.3: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais do grupo G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o cenário: $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

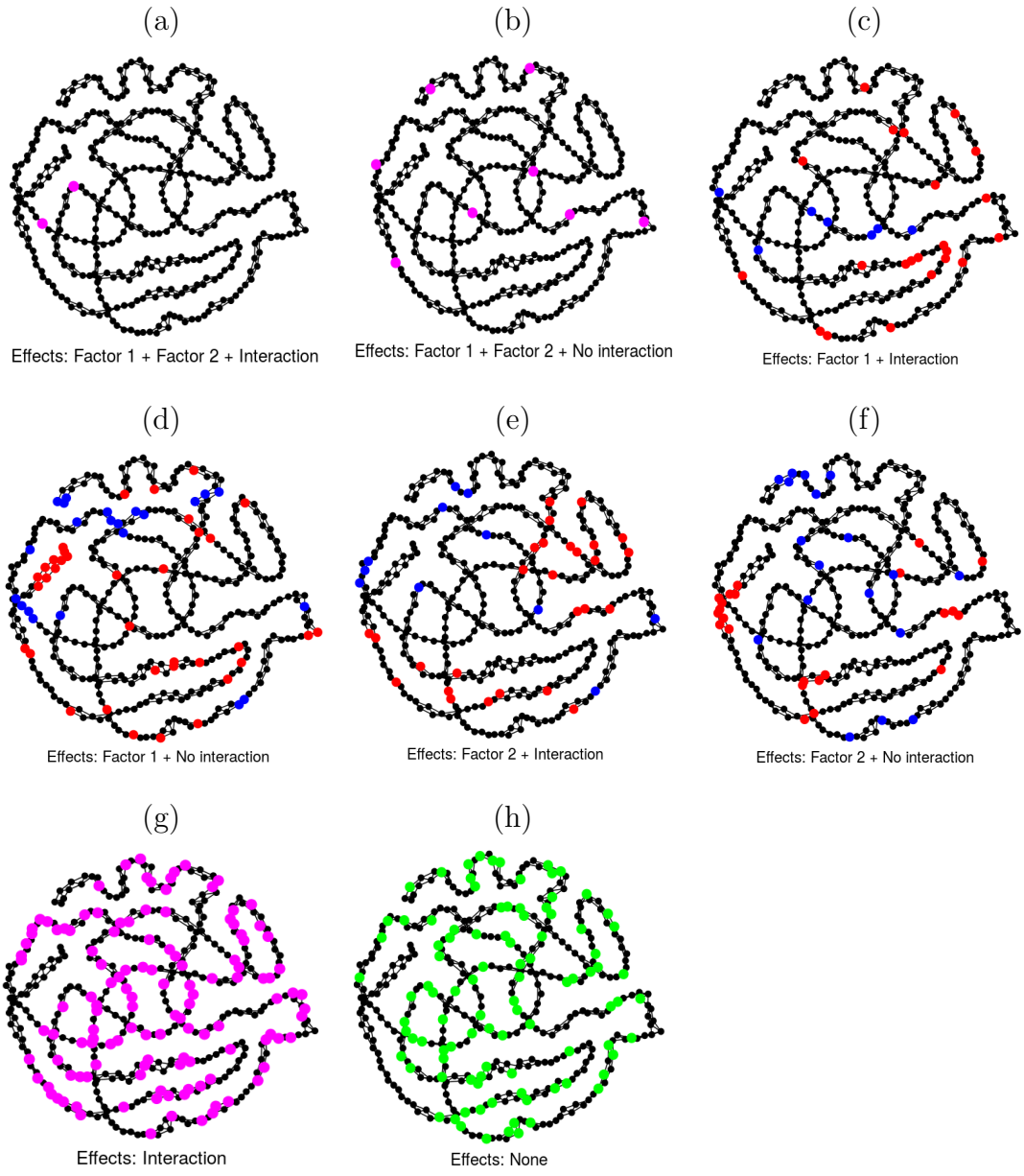


Figure 4.4: Grafos com 4 vizinhos por região imitando a estrutura espacial dos dados artificiais. Cada ponto representa um local. A cor vermelha (carga positiva) ou azul (carga negativa) identifica os locais associados a algum efeito principal e/ou interação (Paineis c, d, e, f). A cor verde (Painel h) denota as regiões não afetadas por qualquer efeito. Nos Paineis (a, b, g), a cor magenta indica os locais afetados por mais de um efeito principal ou somente por interação. Significado da legenda: Effects: Factor k + interaction, indica os locais que sofreram o efeito do fator k e do efeito de interação não-linear. Considere o cenário: $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

Cenário com $\approx 30\%$ de locais em G_E afetados por interação

Conforme descrito anteriormente, também consideramos o cenário no qual temos apenas 30% de locais de G_E afetados por interação. As análises a seguir se referem a esse caso. O leitor deve-se lembrar que menos locais afetados por interação pode levar a uma pior estimação da interação.

Comparando a Tabela 4.9 com a 4.8 vemos que as estimativas ficaram com qualidade semelhante para todos os parâmetros. Particularmente para η^*, note que, aqui, apenas η_4^* ficou um pouco fora do intervalo HPD, entretanto o leitor deve ter em mente que esta é uma análise baseada em um único ajuste (1 amostra). Uma análise com réplicas Monte Carlo, provavelmente, indicará maior semelhança na estimação. Esse estudo está dentre os itens considerados para avaliação futura. Pela Figura 4.5 vemos que a estimativa do parâmetro de interação η^*, no cenário $M_{L_{400}T_4V_4}^{K_2I_{30\%}}$, também acompanha a tendência inicial de crescimento e depois decrescimento.

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.50	0.53	0.53	0.13	0.25	0.78
β_1	-1.00	-1.01	-1.01	0.05	-1.11	-0.91
β_2	1.00	1.03	1.03	0.04	0.94	1.12
σ^2	0.80	0.94	0.94	0.13	0.70	1.21
τ_α	2.00	1.65	1.59	0.58	0.68	2.87
η_1^*	-2.00	-1.84	-1.84	0.31	-2.46	-1.22
η_2^*	1.50	1.35	1.36	0.27	0.83	1.88
η_3^*	0.75	1.17	1.18	0.25	0.68	1.64
η_4^*	-1.00	-0.16	-0.14	0.29	-0.75	0.42

Tabela 4.9: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Cenário: $M_{L_{400}T_4V_4}^{K_2I_{30\%}}$ com $\approx 50\%$ de $Y_i's = 1$.

Como pode ser visto pelas Figuras 4.6 e 4.7, o modelo se comporta muito bem para a situação na qual o número de locais afetados por interação é menor, mas ainda grande

o suficiente (30% de 380, pois 20 locais foram configurados para ter um efeito principal e não ter interação) para obter uma boa estimativa de η^*. Os resultados são tão próximos que, nos mapas de calor (4.2) e (4.6), as diferenças são quase imperceptíveis. Veja como as estimativas para λ são equivalentes através do Painel (b) das Figuras 4.3 e 4.7. No Painel (d) da Figura 4.7 vemos, mais uma vez, que a grande maioria dos locais configurados para ter interação obteve as estimativas de probabilidade acima de 0.5, e o inverso ocorreu para a maioria daqueles configurados para não ter interação. Ao colocar as Figuras 4.4 e 4.8 lado-a-lado, verifica-se que um grande número de regiões foi afetado pelos mesmos tipos de efeitos. Como o número de parâmetros é o mesmo para cenários em que se varia apenas o número de locais afetados por interação, lembrando que o modelo considera um único tipo de interação para todos os locais, era de se esperar que o cenário com mais interação (50%) teria um ajuste melhor, pois temos mais observações (locais) para estimar η^*. A avaliação do impacto das configurações 30% e 50% de locais de G_E afetados por interações deve levar em conta o tamanho de G_E, que por sua vez, depende de L. Na Seção 4.3 faremos o estudo comparativo considerando $L = 100, 200$ e 400, no qual pode-se ver que para o cenário com $L = 100$ o percentual de 30% de locais de G_E afetados por interação levou a uma pior estimação de η^*, o que era de se esperar, pois temos poucos locais (30% de 80, pelo mesmo motivo descrito anteriormente para $L = 400$).

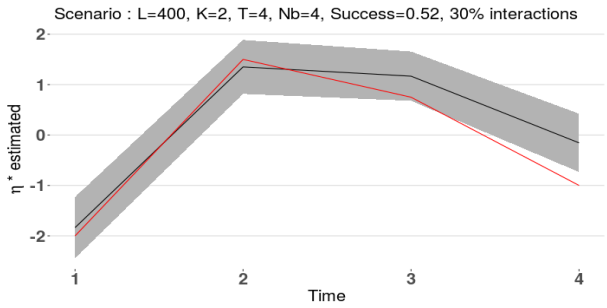


Figure 4.5: Gráfico da média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha) para o cenário $M^{K_2 I_{30\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

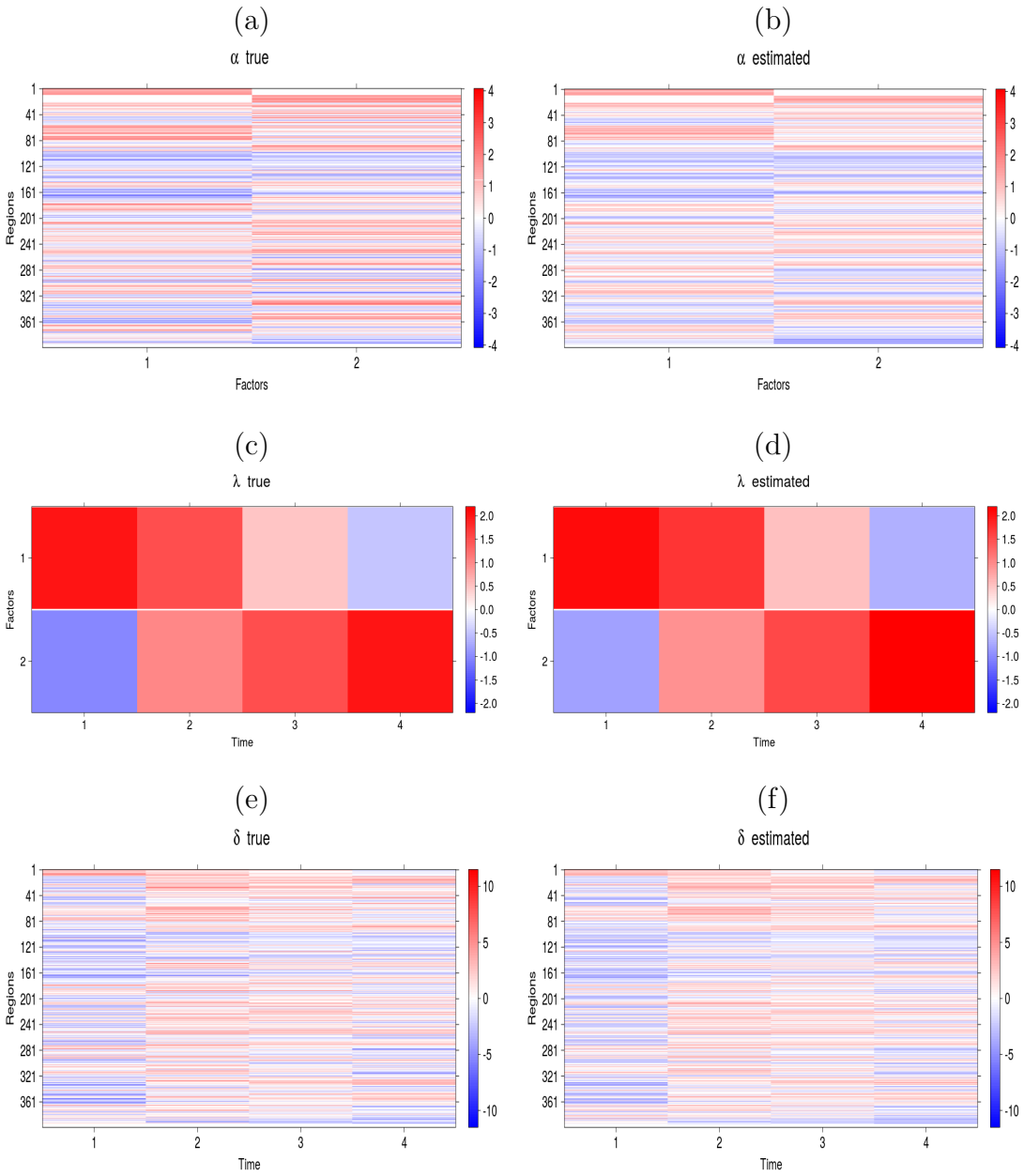


Figure 4.6: Mapas de calor comparando valores verdadeiros e estimados para o cenário $M^{K_2 I_{30\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i's = 1$. Paineis (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

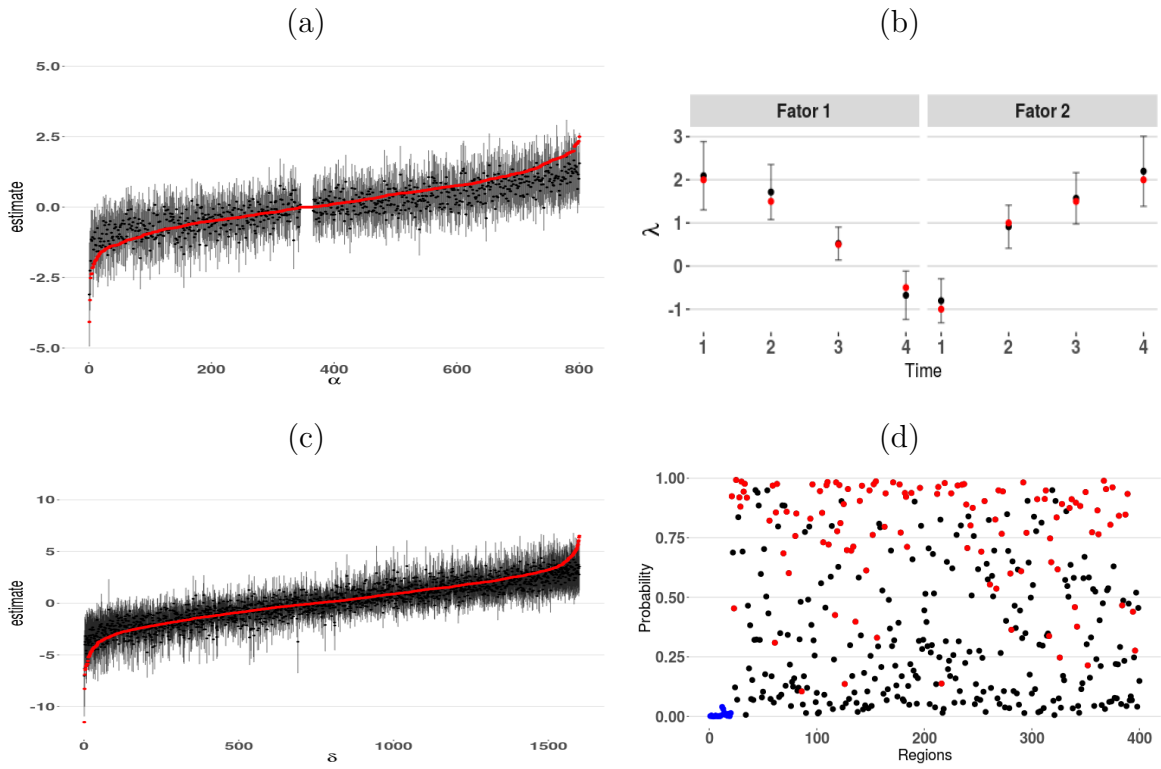


Figure 4.7: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas pela interação; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais de G_E com interação na geração dos dados. A cor preta aos locais de G_E que não tiveram interação na geração. Considere o cenário: $M^{K_2 I_{30\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

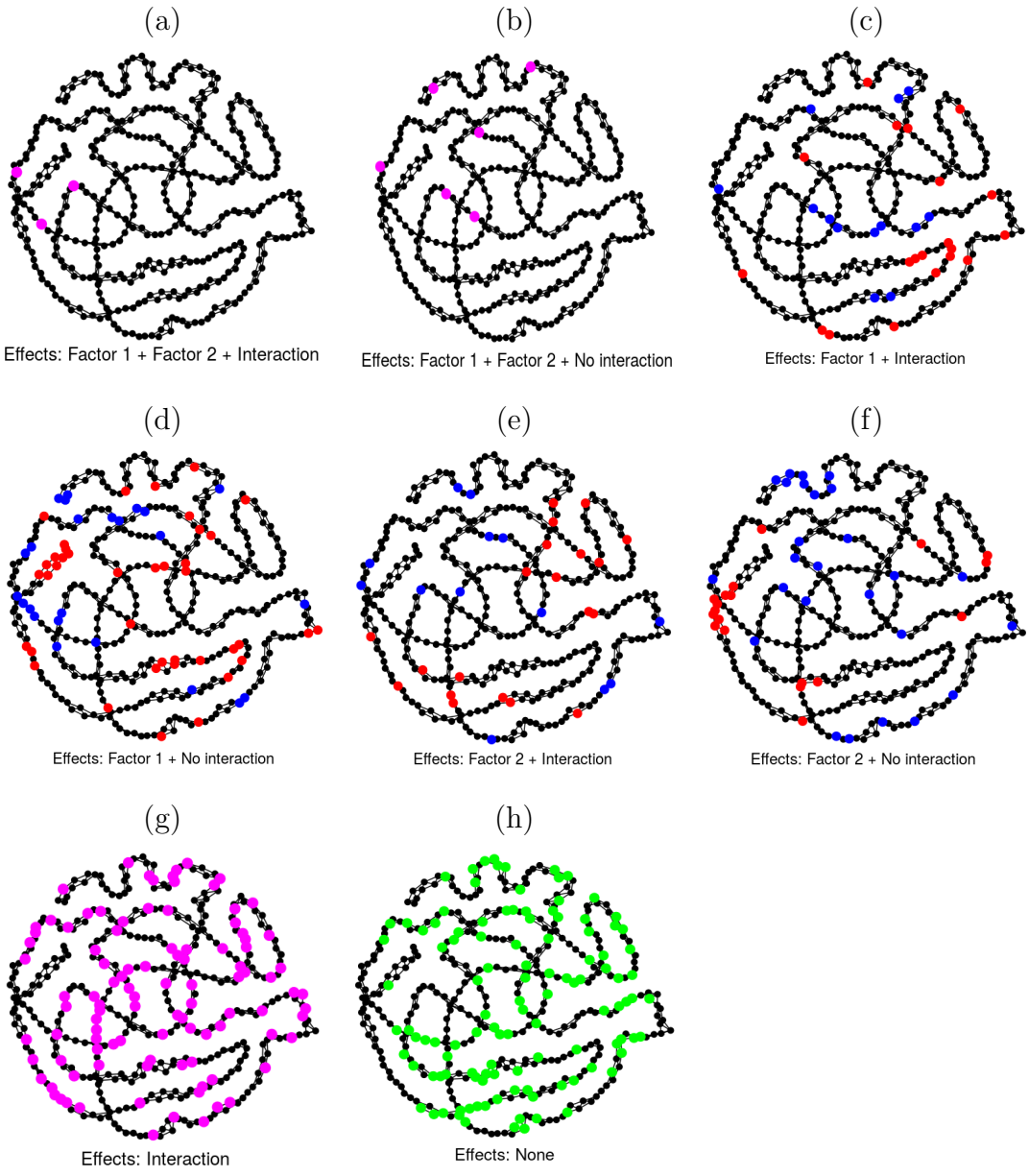


Figure 4.8: Grafos com 4 vizinhos por região imitando a estrutura espacial dos dados artificiais. Cada ponto representa um local. A cor vermelha (carga positiva) ou azul (carga negativa) identifica os locais associados a algum efeito principal e/ou interação (Paineis c, d, e, f). A cor verde (Painel h) denota as regiões não afetadas por qualquer efeito. Nos Paineis (a, b, g), a cor magenta indica os locais afetados por mais de um efeito principal ou somente por interação. Considere o cenário: $M^{K_2 I_{30\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

Encerramos aqui a análise do modelo logístico considerando dados balanceados para quantidade de $1's$ e $0's$ na variável resposta. Nosso próximo passo é fazer uma análise comparativa com as mesmas estratégias utilizadas nesta seção, mas considerando dados desbalanceados. O principal objetivo é verificar se o desbalanceamento dos dados traz alguma dificuldade de estimação para o modelo que estamos propondo.

4.2 Análise para dados desbalanceados

Conforme comentado no início da seção anterior, em aplicações reais de classificação é muito comum encontrarmos um desbalanceamento entre o número de observações definidas nas classes 1 e 0 da resposta binária. Nesta seção, as análises se referem ao caso no qual temos $\approx 20\%$ de $Y_i^{'}s = 1$. Lembrando ao leitor que as configurações consideradas são para 400 regiões, 2 fatores, 4 tempos, 4 vizinhos, e variando a quantidade de locais de G_E afetados por interação, 50% ou 30%.

Cenário com $\approx 50\%$ de locais em G_E afetados por interação

Equivalente às análise anteriores, iniciamos nossa discussão pelas estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^* destacadas na Tabela 4.10. O leitor deve se lembrar que na Tabela 4.2 mostramos que os valores para β, na geração dos dados, foram alterados para que pudéssemos obter a proporção aproximada de 20% de $Y_i^{'} = 1s$. Note que todas as estimativas ficaram dentro do intervalo HPD de 95%. Percebe-se grande proximidade entre a média e a mediana de todos os parâmetros, indicando simetria *a posteriori*. O modelo mostrou-se capaz de recuperar bem os valores verdadeiros. O parâmetro que obteve maior desvio padrão foi τ_α. Os desvios padrão para os $\beta^{'}$s foram bem próximos de zero indicando que a incerteza *a posteriori* é baixa para os coeficientes da regressão. Podemos ver pela Figura 4.9 que as estimativas para η^* (linha preta) acompanham a tendência dos valores verdadeiros (linha vermelha). Veja, também, que o padrão médio foi bem capturado para α, λ e η^* conforme ilustrado pela Figura 4.10. Entretanto, destaca-se que avaliar apenas a estimação pontual levando em conta as médias *a posteriori* expressas nos mapas de calor não informa sobre a incerteza da estimação. Pelo Painel (b) da Figura 4.11 vemos que a estimativa para λ foi muito bem capturada pelos intervalos HPD de 95% e, pelo Painel (d), que a maioria dos municípios que foram marcados com interação na geração dos dados (cor vermelha), obtiveram estimativas da probabilidade de interação acima de 0.5. De forma equivalente, a maioria daqueles que não tiveram a marcação de interação (cor preta), obtiveram probabilidade abaixo de 0.5.

A diferença maior entre o cenário desbalanceado ($\approx$ 20% de $Y_i^{'s} = 1$) e o cenário balanceado ($\approx$ 50% de $Y_i^{'s} = 1$) está nas estimativas para α e δ. Os Paineis (a) e (c) da Figura 4.11 mostram que a tendência para esses parâmetros é bem capturada pelas estimativas *a posteriori*, porém para valores negativos extremos (entre -10 e -5 para δ), as estimativas são piores e os intervalos HPD apresentam faixas maiores, mas ainda incluindo o valor verdadeiro. Isso pode ser explicado analisando a Equação (3.1). Considere os valores extremos de $\delta \in [-10, -5]$. Além disso, $\beta = (-1.5, -1.5, 1.0)$, $X_{2i} = 1$, pois $X_{2i} \in \{0, 1\}$, $X_{3i} = 0$ uma vez que $X_{3i} \sim U(-1, 1)$ estamos tratando de probabilidades (θ_i) que variam, aproximadamente, de $e^{-15}/(1 + e^{-15})$ a $e^{-8}/(1 + e^{-8})$, o que equivale ao intervalo [0.0000003, 0.00034]. Ou seja, probabilidades muito próximas de zero, sendo que a mudança de uma unidade em δ faz pouca diferença nesse cálculo, o que dificulta a obtenção mais precisa da estimativa.

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	-1.50	-1.53	-1.53	0.15	-1.81	-1.23
β_1	-1.50	-1.50	-1.51	0.07	-1.63	-1.37
β_2	1.00	1.04	1.04	0.05	0.93	1.15
σ^2	0.80	0.79	0.78	0.13	0.53	1.04
τ_α	2.00	2.22	1.89	1.30	0.60	4.90
η_1^*	-2.00	-1.76	-1.79	0.45	-2.60	-0.87
η_2^*	1.50	0.78	0.78	0.37	0.05	1.51
η_3^*	0.75	0.28	0.29	0.28	-0.29	0.83
η_4^*	-1.00	-1.55	-1.54	0.35	-2.27	-0.87

Tabela 4.10: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o cenário: $M_{L_{400}T_4V_4}^{K_2 I_{50\%}}$ com $\approx$ 20% de $Y_i^{'} s = 1$.

Semelhante ao que foi apresentado nos cenários anteriores, o gráfico da Figura 4.12 ilustra uma estrutura em rede com 4 vizinhos por nó (local), que imita a estrutura espacial de um mapa, em que podemos identificar os diversos conglomerados que representam

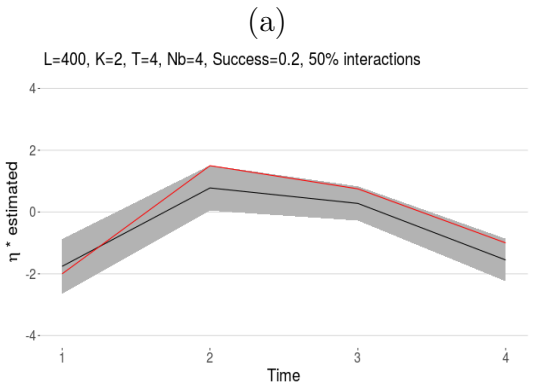


Figure 4.9: Gráfico da média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha) para o cenário $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 20\%$ de $Y_i^{'} s = 1$.

conjuntos de locais impactados por algum tipo de efeito comum. O Painel (h) mostra que muitas regiões (109) não sofrem efeito significativo de qualquer fator ou interação. Veja no Painel (g) que vários locais (nós) foram identificados sob efeito exclusivo da interação. O leitor deve se lembrar que estamos estudando, aqui, uma situação com 50% de regiões de G_E contendo interação. Na ocorrência de efeito de apenas um dos fatores principais (Fator 1 ou Fator 2), com ou sem interação, o *cluster* formado pode ser subdividido em dois casos: regiões com cargas positivas (cor vermelha) e com cargas negativas (cor azul). Esse aspecto está em destaque nos Paineis (c), (d), (e) e (f) da Figura 4.12. Reforçamos que a região é dita ser afetada pela interação se a probabilidade $\mathrm{p}^*(z_l = 1|\bullet) > 0.5$ (veja Passo 3 da Seção 3.1.1). O local é identificado como sendo afetado por algum efeito principal (Fator 1 e/ou Fator 2), quando o intervalo HPD de 95% das cargas não inclui o 0.

Finalizamos, aqui, a análise para dados desbalanceados no que se refere à quantidade de $1's$ e $0's$ na variável resposta. Concluimos que o ajuste para este caso é, em geral, satisfatório e, com isso, a estrutura hierárquica proposta neste capítulo parece atender bem as aplicações com resposta desbalanceada no modelo logístico. Uma análise comparativa das estimativas para α, λ e δ entre o cenário desbalanceado e os balanceados para $K = 2$ fatores, $T = 4$ tempos e 4 vizinhos por região será mostrada adiante na Seção 4.9, em

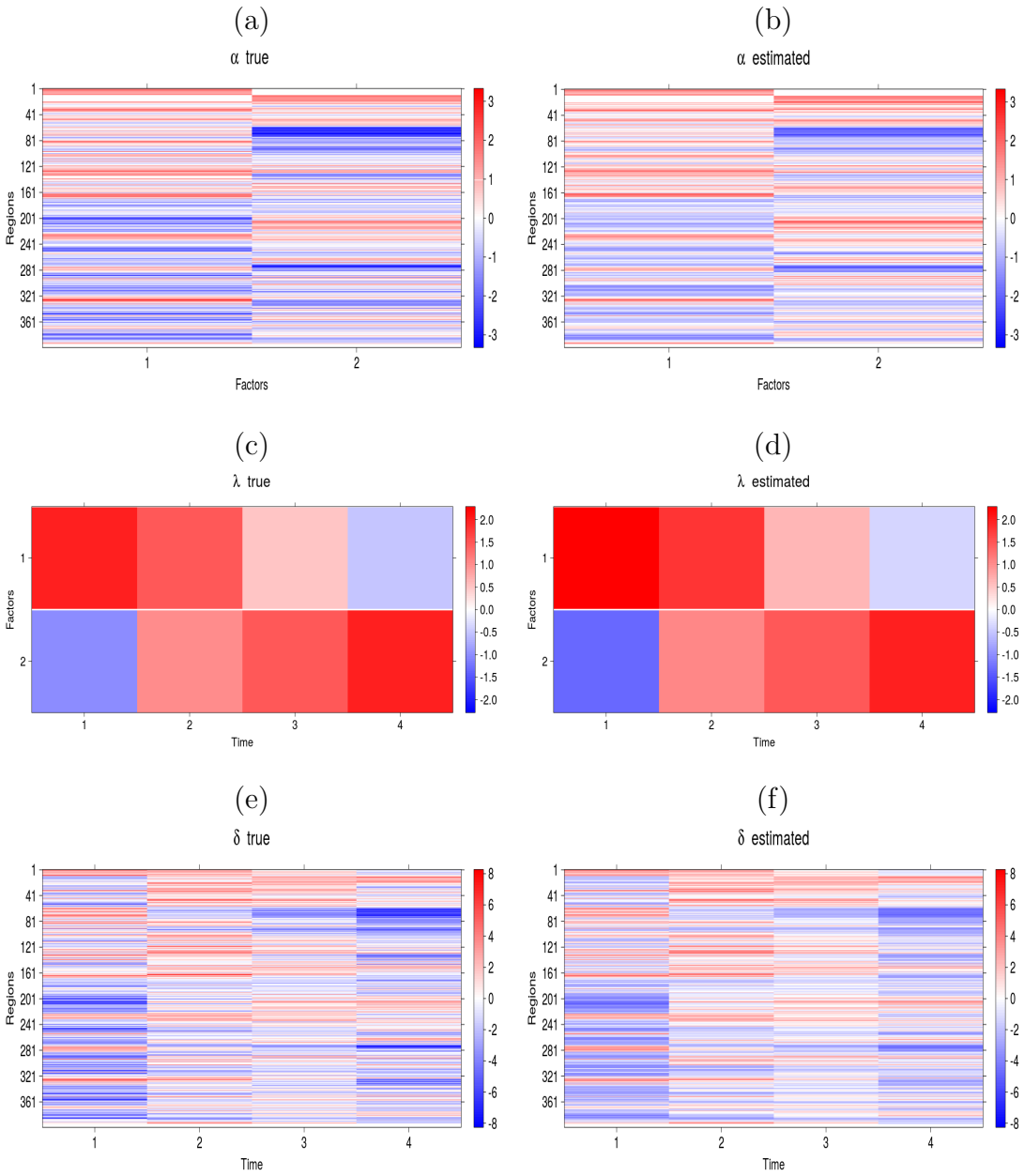


Figure 4.10: Mapas de calor comparando valores verdadeiros e estimados para o cenário $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 20\%$ de $Y_i^{'} s = 1$. Paineis (a) e (b) são referentes à α, (c) e (d) referem-se à λ, (e) e (f) representam δ.

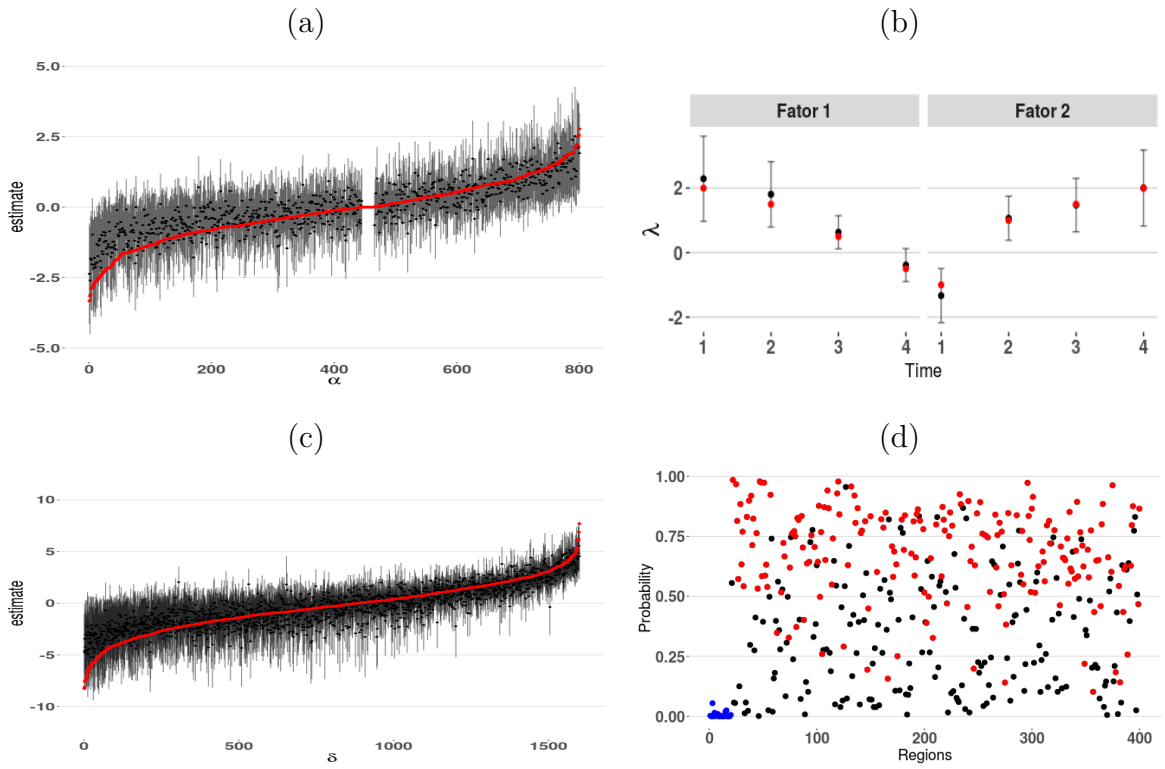


Figure 4.11: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas pela interação; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais de G_E que tiveram interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o cenário: $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 20\%$ de $Y_i^{'} s = 1$.

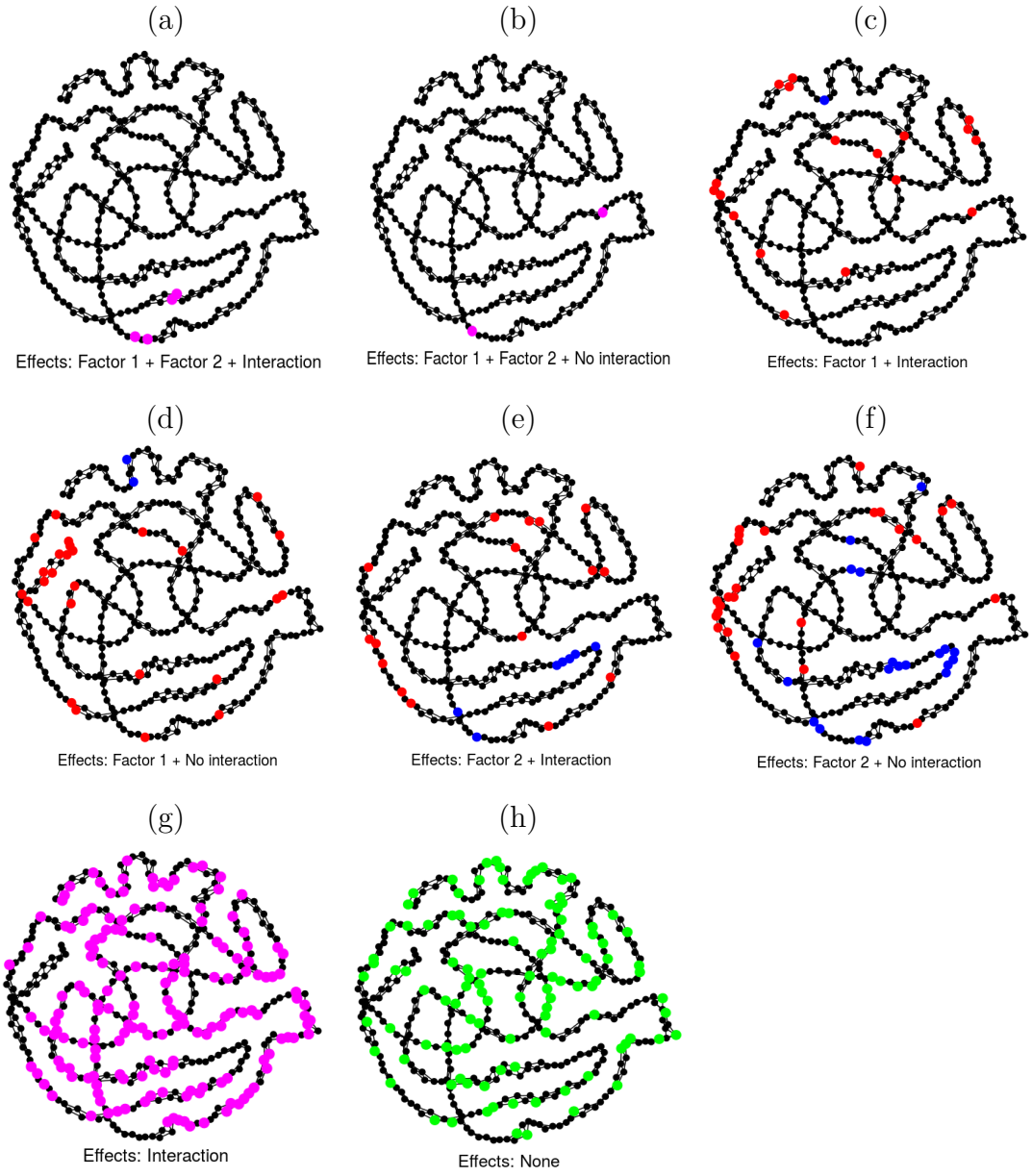


Figure 4.12: Grafos com 4 vizinhos por região imitando a estrutura espacial dos dados artificiais. Cada ponto representa um local. A cor vermelha (carga positiva) ou a cor azul (carga negativa) identifica os locais associados a algum efeito principal e/ou interação (Paineis c, d, e, f). A cor verde (Painel h) denota as regiões não afetadas por qualquer efeito. Nos Paineis (a, b, g), a cor magenta identica os locais afetados por mais de um efeito principal ou somente interação. Considere o cenário: $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 20\%$ de $Y_i^{'} s = 1$

.

que apresentamos resultados após a execução de 30 réplicas de Monte Carlo. Na próxima seção, vamos desenvolver uma análise comparativa do ajuste do modelo para η^* e λ em alguns cenários em que variamos a quantidade de locais, mas considerando apenas uma amostra *a posteriori*.

4.3 Comparação geral dos cenários

Iniciamos, agora, um estudo variando o número de locais em que consideramos $L = 100, 200$ e 400, lembrando que $L = 400$ é o cenário mais parecido com os dados reais de Minas Gerais do sistema de Telessaúde. Nosso objetivo é comparar o quanto o número de regiões afeta o ajuste do modelo. Ressaltamos ao leitor que, nesta seção, consideramos apenas o cenário no qual os dados são balanceados, ou seja, temos $\approx 50\%$ de $Y_i^{'} s = 1$.

Gostaríamos de destacar que vamos avaliar apenas as estimativas para η^* e λ, pois eles são os únicos parâmetros cujos valores são os mesmos para todas as gerações dos dados de todos os cenários com $K = 2$ fatores e $T = 4$ tempos. A Figura 4.13 mostra a média *a posteriori* para η^*, nas configurações com 4 vizinhos por região e com número de locais afetados por interação sendo 30% (Paineis da esquerda) e 50% (Paineis da direita).

Como era de se esperar, os cenários que possuem mais locais afetados por interação (50%) apresentam variância menor, o que pode ser identificado pelo intervalo HPD de 95% mais estreito (área sombreada). Para $L = 200$ e 400, não identificamos, visualmente, muita diferença em termos do formato e tamanho dos envelopes HPD. A quantidade de locais utilizados para estimar η^* nos 2 casos (30% e 50% de G_E) é grande o suficiente para determinar uma boa estimação. Quanto mais locais forem afetados pela interação em η^*, mais informação será usada pelo modelo para estimar este elemento. No caso $L = 100$, perceba uma maior diferença no tamanho dos envelopes HPD para as situações de 30% e 50% de locais de G_E com efeito η^*. Essa diferença pode ser explicada pela baixa quantidade de regiões que contribuem para estimar η^*, ou seja, temos apenas 30% de $80 = 24$ locais contra 50% de $80 = 40$ locais.

As Figuras 4.14, 4.15 e 4.16 ilustram as estimativas para λ nos casos $L = 100, 200$ e 400, respectivamente. Os paineis da esquerda se referem ao primeiro fator e os da direita,

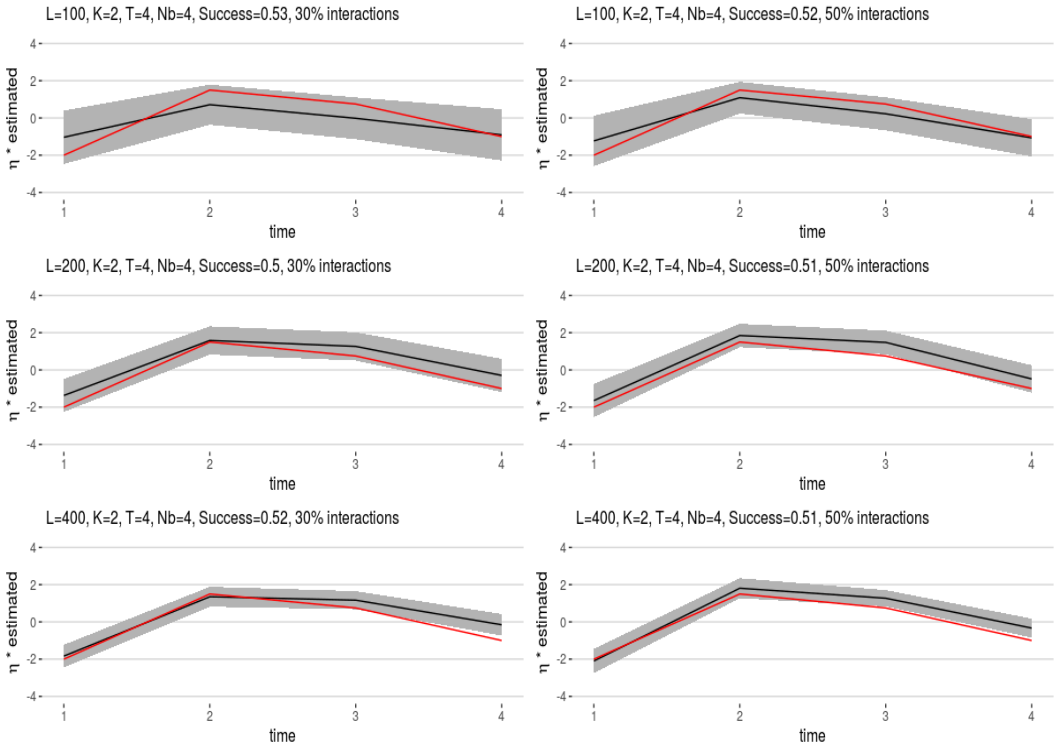


Figure 4.13: Média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha) em todas as configurações de números de regiões (L = 100, 200 e 400). Considere os cenários: $M^{K_2I_{30\%}}_{L_{100}T_4V_4}$ e $M^{K_2I_{50\%}}_{L_{100}T_4V_4}$, $M^{K_2I_{30\%}}_{L_{200}T_4V_4}$ e $M^{K_2I_{50\%}}_{L_{200}T_4V_4}$, $M^{K_2I_{30\%}}_{L_{400}T_4V_4}$ e $M^{K_2I_{50\%}}_{L_{400}T_4V_4}$ com $\approx 50\%$ de $Y_i^{'}s = 1$, ou seja, $\beta = (0.5, -1.0, 1.0)$

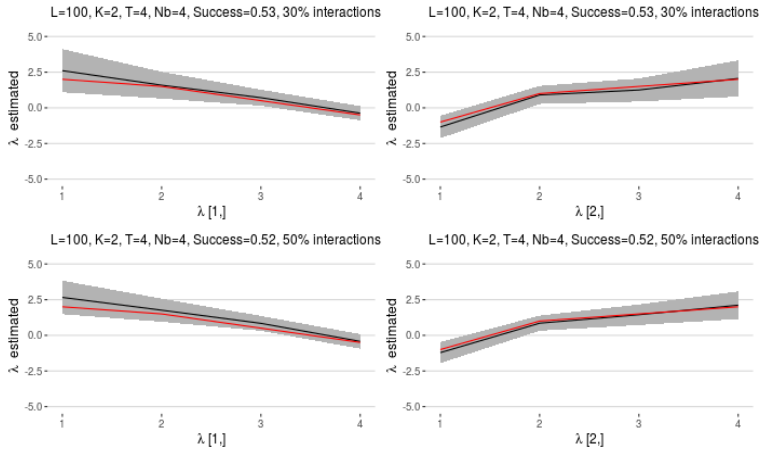


Figure 4.14: Média *a posteriori* (linha preta), intervalo HPD de 95% para λ (área sombreada) e valor verdadeiro (linha vermelha). Considere os cenários: $M^{K_2 I_{30\%}}_{L_{100} T_4 V_4}$ e $M^{K_2 I_{50\%}}_{L_{100} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$, ou seja, $\beta = (0.5, -1.0, 1.0)$

ao segundo. Vemos que as estimativas capturaram bem o valor verdadeiro em todos os cenários, com destaque para a identificação dos padrões de decrescimento e crescimento, no tempo, dos Fatores 1 e 2, respectivamente. Exceção ocorre apenas para o Fator 1 quando $L = 100$ e $T = 1$, em que as estimativas não ficaram tão próximas do valor verdadeiro quanto as demais, mas bem dentro do intervalo HPD. Outros gráficos relativos aos cenários avaliados podem ser analisados nos Apêndices B e C. Comparando as três figuras, observa-se o efeito claro do tamanho amostral (valor de L) sobre a incerteza *a posteriori* de λ. Os envelopes HPD são bem menores no caso $L = 400$.

O próximo passo de nossa análise é avaliar como o modelo logístico, proposto nesta tese, se comporta perante outras quantidades de fatores e de tempos. Na próxima seção, primeiramente apresentamos a análise consideranto $T = 10$ tempos e $K = 2$ fatores. Na sequência, temos a análise com a configuração de $K = 3$ fatores, mas mantendo $T = 4$. Lembramos ao leitor que o modelo sendo ajustado assume o conhecimento, pelo analista, do valor verdadeiro de K. A análise com erro de especificação de K será tratada na Seção 4.5.

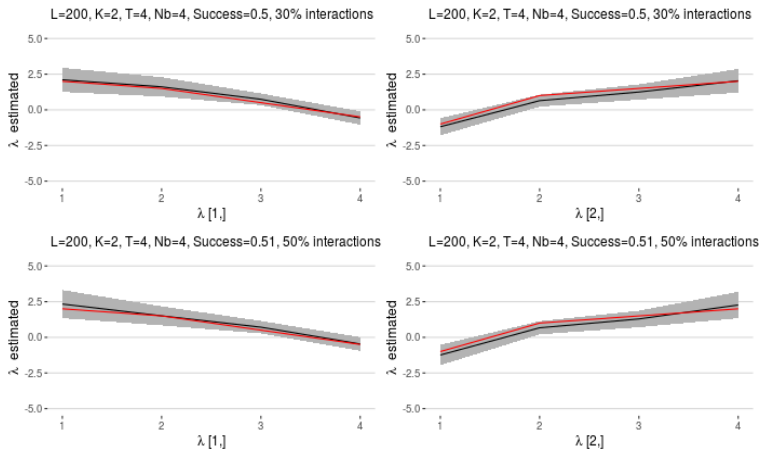


Figure 4.15: Média *a posteriori* (linha preta), intervalo HPD de 95% para λ (área sombreada) e valor verdadeiro (linha vermelha). Considere os cenários: $M^{K_2 I_{30\%}}_{L_{200} T_4 V_4}$ e $M^{K_2 I_{50\%}}_{L_{200} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$, ou seja, $\beta = (0.5, -1.0, 1.0)$

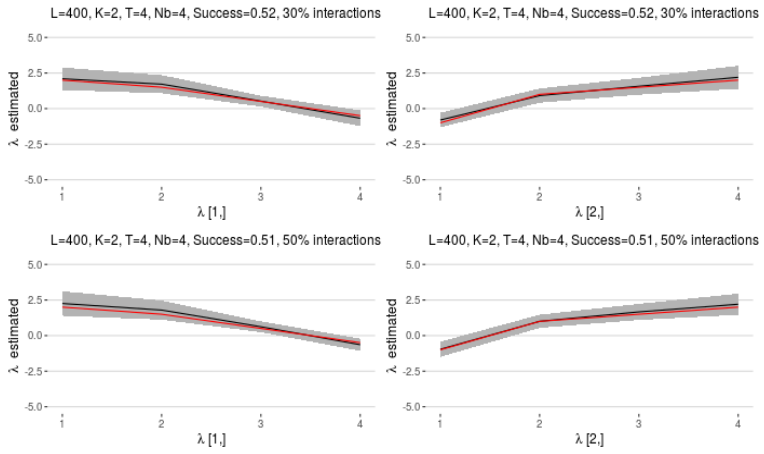


Figure 4.16: Média *a posteriori* (linha preta), intervalo HPD de 95% para λ (área sombreada) e valor verdadeiro (linha vermelha). Considere os cenários: $M^{K_2 I_{30\%}}_{L_{400} T_4 V_4}$ e $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$, ou seja, $\beta = (0.5, -1.0, 1.0)$

4.4 Análise com nova especificação de K e T

A análise de cenários nos quais temos outros números de fatores e de tempos é muito relevante, pois em um conjunto de 441 municípios (dados reais) é razoável se pensar que possamos ter, por exemplo, locais afetados por 3 tipos diferentes de efeito ($K = 3$). Tomando como referência o sistema de Telessaúde, o qual acumula, diariamente, milhares de exames de ECG, podemos considerar um horizonte maior de anos, por exempo, $T = 10$, que estará disponível no futuro com a coleta de dados sendo feita continuamente. A presente seção avalia os impactos na estimação em dois cenários. No primeiro temos $K = 2$ e $T = 10$, e no segundo, $K = 3$ e $T = 4$. Lembrando ao leitor que estamos trabalhando com $\approx 50\%$ de $Y_i^{'}s = 1$, $L = 400$, 4 vizinhos por região e com 50% de locais de G_E afetados por interação.

Análise para T igual a 10 tempos

O cenário com $T = 10$ simula a situação em que o banco de dados continua a ser atualizado com a obtenção de mais informações ao longo do tempo, equivalente ao que ocorre no sistema de Telessaúde que, atualmente, recebe cerca de 2500 ECG's por dia.

A Tabela 4.11 apresenta as estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2 e do parâmetro de variância τ_α. Note que os valores para β são aqueles relativos à configuração 1 da Tabela 4.2, utilizados na geração de dados balanceados. Os valores para η^* são os apresentados na Tabela 4.7, resultado do produto dos $\lambda^{'}s$ descritos na Tabela 4.4. Veja que, apenas a estimativa de τ_α ficou fora do intervalo HPD e dentre as demais, apenas η_8^*, η_9^* e η_{10}^* obtiveram estimativas distantes do valor verdadeiro. Veja que para o η_8^* houve troca de sinal, mas em uma intervalo bem próximo de 0. Média e mediana são estimativas muito próximas, em todos os casos, indicando simetria. Os desvios padrão para os coeficientes da regressão são bem pequenos e os maiores são os obtidos para os $\eta^{*'}s$ e para o τ_α. Veja pela Figura 4.17 como o ajuste para η^* segue a tendência do valor verdadeiro e totalmente dentro do intervalo HPD.

Analisando os mapas de calor da Figura 4.18 vemos que o padrão global foi capturado. Perceba pelas tonalidades de cores, na comparação dos mapas verdadeiros versus estimados,

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.50	0.48	0.49	0.11	0.27	0.70
β_1	-1.00	-1.01	-1.01	0.05	-1.11	-0.92
β_2	1.00	0.96	0.96	0.04	0.87	1.04
σ^2	0.80	0.95	0.94	0.10	0.76	1.14
τ_α	2.00	0.91	0.85	0.27	0.50	1.41
η_1^*	-2.00	-2.27	-2.26	0.34	-2.92	-1.61
η_2^*	-1.19	-1.53	-1.53	0.35	-2.18	-0.83
η_3^*	-0.75	-0.72	-0.72	0.26	-1.22	-0.20
η_4^*	0.60	0.54	0.54	0.25	0.05	1.02
η_5^*	1.00	1.20	1.19	0.28	0.67	1.73
η_6^*	0.84	0.77	0.77	0.28	0.21	1.32
η_7^*	0.75	0.77	0.76	0.32	0.15	1.38
η_8^*	-0.16	0.16	0.16	0.34	-0.51	0.85
η_9^*	-0.54	-0.09	-0.10	0.32	-0.69	0.57
η_{10}^*	-1.00	-0.55	-0.55	0.39	-1.29	0.21

Tabela 4.11: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o cenário: $M^{K_2 I_{50\%}}_{L_{400} T_{10} V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

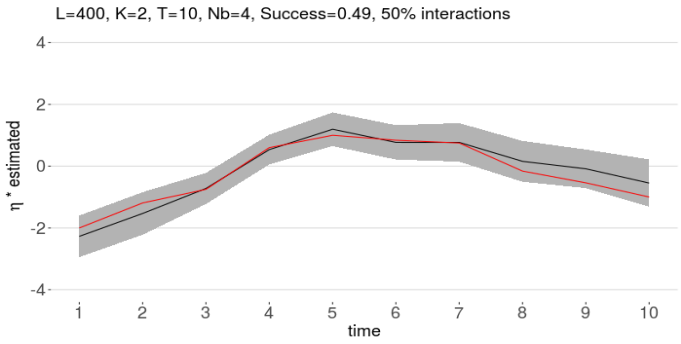


Figure 4.17: Média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha). Considere o cenário: $M^{K_2 I_{50\%}}_{L_{400} T_{10} V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

que há uma diferença de estimação para α e λ. Cores mais vibrantes aparecem nos Paineis (a) e (d). Parece que houve uma troca ou compensação na qualidade da estimação entre eles. Esta compensação entre α e λ reflete em uma boa estimativa para δ, lembrando ao leitor de que isso é razoável considerando que $\delta = \alpha\lambda + \eta + \epsilon$.

A Figura 4.19 mostra os intervalos HPD de 95% para α, λ e δ. Perceba nos Paineis (a) e (b) o erro de estimação. No Painel (a) temos sobrestimação para valores de α menores que ≈ -0.5 e subestimação para valores maiores que ≈ 0.5. Nessa mesma linha vemos que no Painel (b) também existe um pequeno erro de estimação para λ. Vê-se, claramente, que a maioria dos valores verdadeiros se encontram nas extremidades dos intervalos HPD ou até mesmo fora dele. Verifica-se, também, que os valores maiores são sobrestimados e os valores menores e negativos, subestimados. No entando, o Painel (c) indica que, dado a compensação existente entre α e λ (apresentada anteriormente), a estimação de δ é satisfatória. Veja que o valor verdadeiro (linha vermelha) praticamente divide ao meio a núvem de intervalos. Em outras palavras, os desvios identificados no Painel (a) não se repetem no Painel (c).

Finalmente, vemos novamente pelo Painel (d) que, a maioria das probalilidades estimadas concorda com o valor verdadeiro, ou seja, pode-se ver que a maior parcela de regiões de cor vermelha (regiões verdadeiramente afetadas pela interação) aparece

com probabilidades acima de 0.5 e a maior parcela de cor preta (regiões não afetadas por η^*), com probabilidades abaixo de 0.5.

Complementando a análise para $T = 10$, a Figura 4.20 ilustra mapas, através de grafos, em que pode-se analisar os conglomerados de locais sob algum tipo de efeito comum. Novamente, os Paineis (c, d, e, f) mostram os *clusters* subdivididos em cargas positivas (cor vermelha) e negativas (cor azul). Note que, a quantidade de locais sob efeito de interação, Painel (g), e não afetada por qualquer efeito, Painel (h), é grande. Mais uma vez é possível identificar, pelos Paineis (a) e (b), que os locais afetados pelos Fatores 1 e 2 (cor magenta), concomitantemente, são poucos. Lembrando que os locais sob efeito de interação são aqueles com $\mathrm{p}^*(z_l = 1|\bullet) > 0.5$ e as regiões sob algum efeito principal (Fator 1 ou Fator 2) são aquelas para as quais o intervalo HPD das cargas não inclui o 0.

Terminamos, aqui, a análise para $T = 10$ e vimos que, novamente, o modelo ficou bem ajustado, com exceção de poucos parâmetros, situação que pode ser atribuída ao fato das análises apresentadas se referirem a apenas uma amostra *a posteriori*. Importante destacar que, neste caso, o aumento no número de observações é muito maior do que o aumento no número de parâmetros. A cada novo tempo inserido temos o acréscimo de milhares de indivíduos contra um aumento duas dezendas de parâmetros ($\sum_{l=1}^{L} n_l$ indivíduos contra $18 = 2 \times 6 + 6$ novos parâmetros, em que 2×6 significa o produto de 2 fatores e 6 tempos, e o segundo 6 somado equivale ao aumento de elementos em η^*). Resumindo, temos muito mais informação disponível para estimar os parâmetros. No próximo tópico do estudo iremos mostrar avaliações das estimativas quando os dados foram gerados e estimados com $K = 3$ fatores, mas mantendo $L = 400$ locais, $T = 4$ tempos, 4 vizinhos por região, 50% de locais de G_E afetados por interação e $\approx 50\%$ de $Y_i^{'} s = 1$.

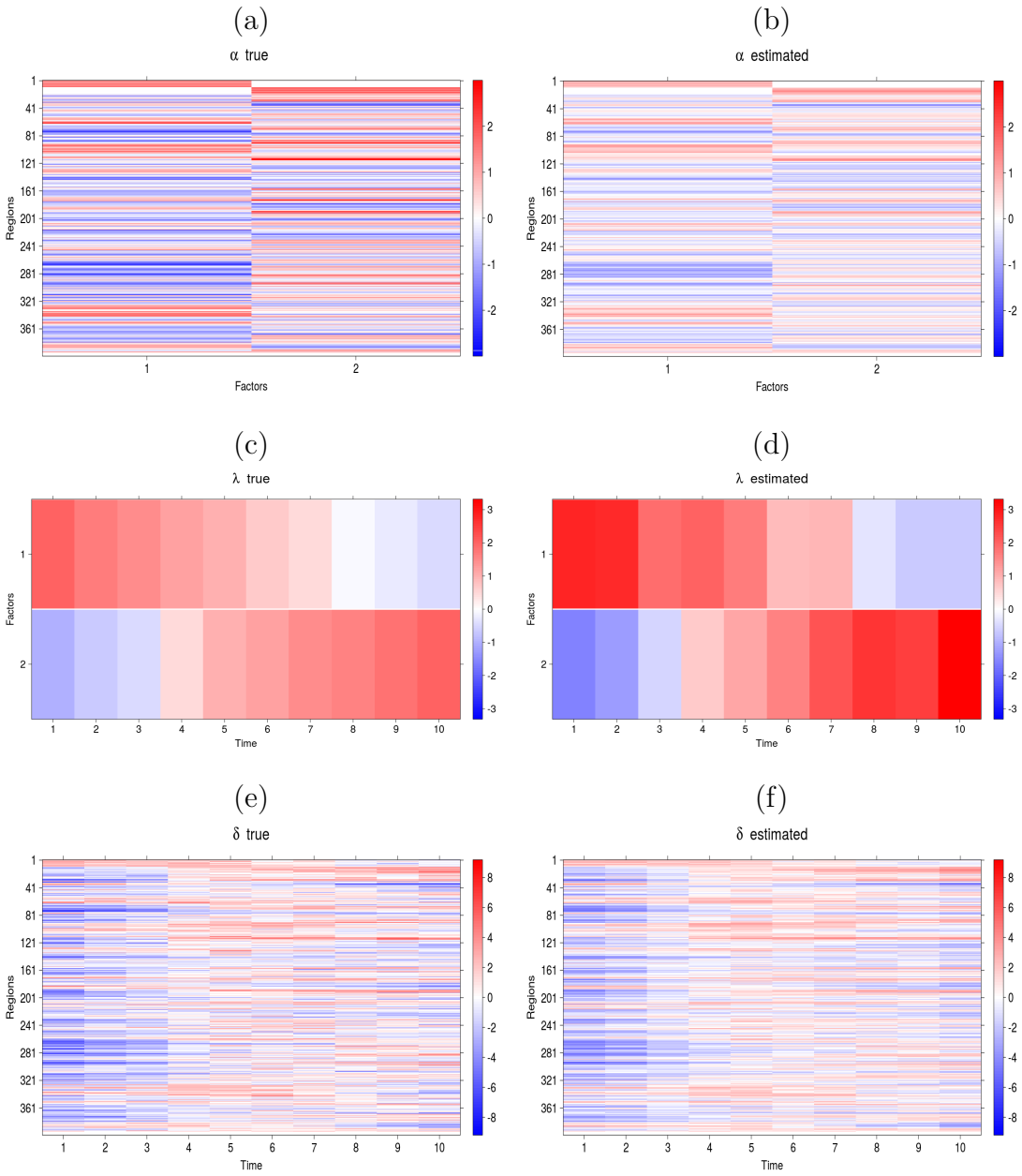


Figure 4.18: Mapas de calor comparando valores verdadeiros e estimados para o cenário $M^{K_2 I_{50\%}}_{L_{400} T_{10} V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$. Paineis (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

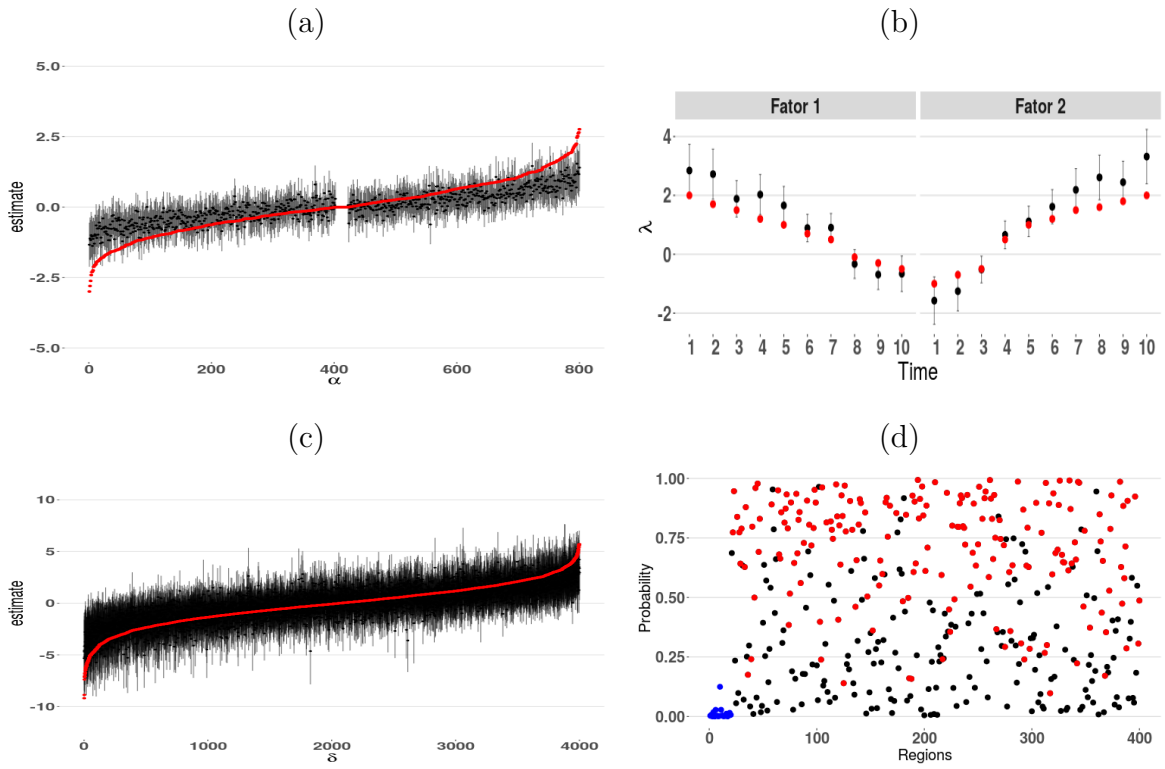


Figure 4.19: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais de G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o cenário: $M^{K_2 I_{50\%}}_{L_{400} T_{10} V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

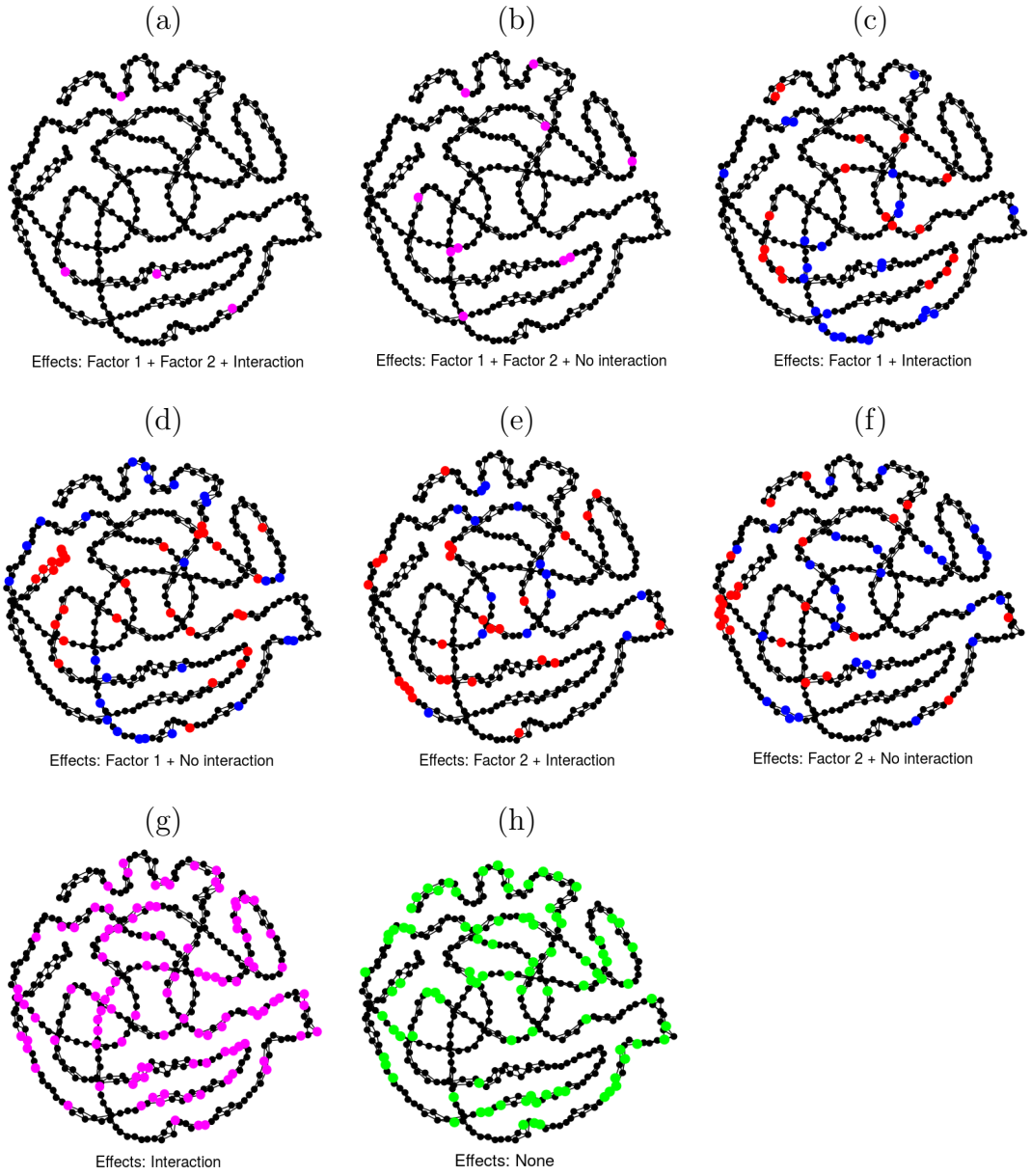


Figure 4.20: Grafos com 4 vizinhos por região imitando a estrutura espacial dos dados artificiais. Cada ponto representa um local. A cor vermelha (carga positiva) ou azul (carga negativa) identifica os locais associados a algum efeito principal e/ou interação (Paineis c, d, e, f). A cor verde (Painel h) denota as regiões não afetadas por qualquer efeito. Nos Paineis (a, b, g), a cor magenta identifica os locais afetados por mais de um efeito principal ou somente interação. Considere o cenário: $M^{K_2 I_{50\%}}_{L_{400} T_{10} V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$

.

Ajuste e geração dos dados com K igual a 3 fatores

Neste tópico, vamos apresentar a análise do ajuste do modelo logístico, proposto nesta tese, para o cenário com $K = 3$ na geração dos dados e na estimação. Os valores para o número de locais, de tempos, de vizinhos por local, percentuais de locais afetados por interação e de $Y_i^{'}s = 1$, se mantém os mesmos do tópico anterior.

Na Tabela 4.12 vemos que todos os parâmetros foram bem estimados, ficando dentro do intervalo HPD de 95%. Novamente observa-se uma similaridade entre a média e a mediana indicando simetria da distribuição *a posteriori* para todos os parâmetros. Os desvios padrão para os coeficientes em β são os menores e bem pequenos. Em geral, a Tabela 4.12 está indicando que, em termos de estatística descritiva para esta amostra simulada, o modelo foi capaz de fornecer uma boa estimação.

A Figura 4.21 mostra a incerteza *a posteriori* sobre a estimação de η^* neste caso envolvendo três fatores latentes para construir η^*. Apesar de existir um erro de estimação para os tempos 1 e 2, a captura do valor verdadeiro pelo HPD é visível.

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.50	0.52	0.53	0.14	0.23	0.78
β_1	-1.00	-1.11	-1.11	0.05	-1.21	-1.01
β_2	1.00	0.99	0.99	0.04	0.90	1.08
σ^2	0.80	0.70	0.70	0.13	0.45	0.95
τ_α	2.00	3.08	2.95	1.01	1.29	5.20
η_1^*	-2.00	-1.41	-1.38	0.43	-2.29	-0.61
η_2^*	-1.50	-0.94	-0.94	0.51	-1.94	0.06
η_3^*	0.75	0.93	0.95	0.44	0.02	1.74
η_4^*	1.00	1.29	1.30	0.43	0.45	2.13

Tabela 4.12: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Cenário: $M_{L_{400}T_4V_4}^{K_3I_{50\%}}$ com $\approx 50\%$ de $Y_i^{'}s = 1$.

Os Paines (a) e (b), da Figura 4.22, registram que o padrão global também foi

capturado neste caso. Analisando as estimativas de α para o Fator 2, coluna 2 dos Paineis (a) e (b), vemos que para os locais entre os números 210 e 240, os valores foram sobrestimados. Mas de alguma forma, essa sobrestimação foi compensada pelas estimativas de λ, uma vez que não se identifica grandes desvios em δ; veja Paineis (e) e (f). Pelos Paineis (c) e (d), pode-se avaliar que o padrão global para λ também foi bem capturado. Esse fato pode ser melhor verificado pelo Painel (b) da Figura 4.23, em que, com exceção do Fator 2 no Tempo 4, todas as estimativas foram estimadas dentro do intervalo HPD de 95% e seguindo a tendência do valor verdadeiro. O leitor deve retornar à Tabela 4.4 para ver os valores verdadeiros de λ quando $K = 3$ e $T = 4$. O Fator 1 tem padrão decrescente, o Fator 2 crescente e o Fator 3 desce-sobe-desce $(1.0, -1.0, 1.0, -1.0)$. Esse comportamente é claramente identificado na análise.

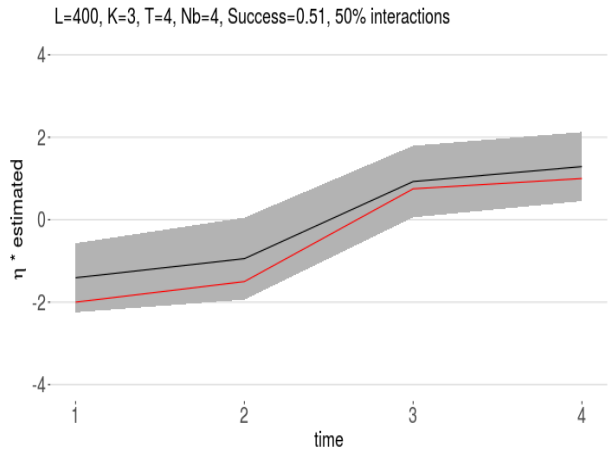


Figure 4.21: Média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha). Considere o cenário: $M^{K_3 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i's = 1$.

Seguindo os resultados das análises anteriores, os Paineis (a) e (c), da Figura 4.23, mostram que a maioria das médias *a posteriori* de α e de δ estão dentro do intervalo HPD de 95%, e apresentam a mesma tendência dos valores verdadeiros (linha de cor vermelha). Da mesma forma que os gráficos de interações dos cenários anteriores, o Painel (d) mostra as probalilidades estimadas por local. Vemos, novamente, que a maioria das regiões de cor

vermelha aparece com probabilidades acima de 0.5, e as de cor preta com probabilidades abaixo de 0.5. Lembrando que, na geração dos dados artificiais, as regiões de cor vermelha foram afetadas por interações e as de cor preta, não.

Na Figura 4.24, por simplicidade e devido à semelhança dos resultados anteriores, optamos em mostrar, neste cenário, apenas os grafos que imitam a estrutura espacial dos dados artificiais para os casos em que os locais são afetados por um único fator principal. Além disso, mostramos as situações em que as regiões são afetadas apenas por interação, Painel (g), e as regiões que não foram afetadas por qualquer efeito, Painel (h). Não houveram locais afetados pelos três fatores concomitantemente, por isso tal gráfico não foi inserido. Os locais afetados por pares de fatores (1 e 2, 1 e 3, 2 e 3) não são mostrados. Lembramos ao leitor que os gráficos da Figura 4.24 são uma ilustração do propósito da tese de detectar *clusters*. Enfatizamos que o critério é encontrar cargas significativas (pela análise do HPD) e obter uma probabilidade *a posteriori* acima de 0.5 para a existêcia de efeito de interação.

Finalizamos, aqui, as análises alterando o número de fatores verdadeiro e ajustado para $K = 3$. Na próxima seção vamos tratar de cenários nos quais avaliamos o impacto na estimação quando o analista comete o erro de especificar 1 fator acima ou abaixo da quantidade verdadeira para K, mas mantendo o mesmo número de tempos. Importante ressaltar que quando aumentamos o número de fatores e mantemos fixo o número de tempos estamos acrescentando mais parâmetros ao modelo. Com isso, vícios maiores são esperados por termos um modelo menos parcimonioso. Duas análises são conduzidas: a primeira considera que existe 1 fator extra para estimar, como por exemplo, $K_V = 2$ e $K_A = 3$, em que K_V equivale ao número verdadeiro de fatores e K_A o número ajustado, e a segunda efetua a análise contrária com $K_V = 3$ e $K_A = 2$.

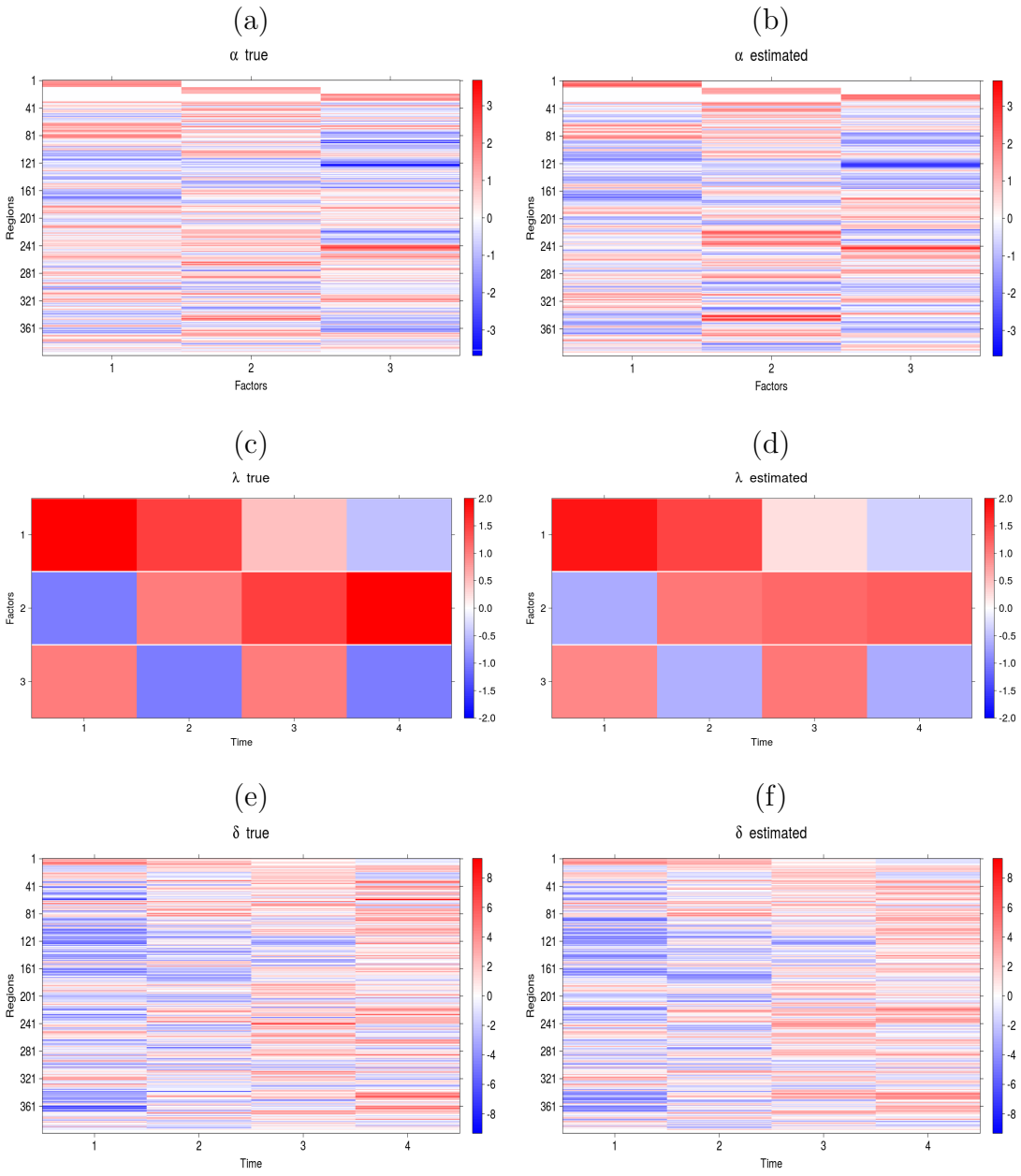


Figure 4.22: Mapas de calor comparando valores verdadeiros e estimados para o cenário $M^{K_3 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i's = 1$. Paineis (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

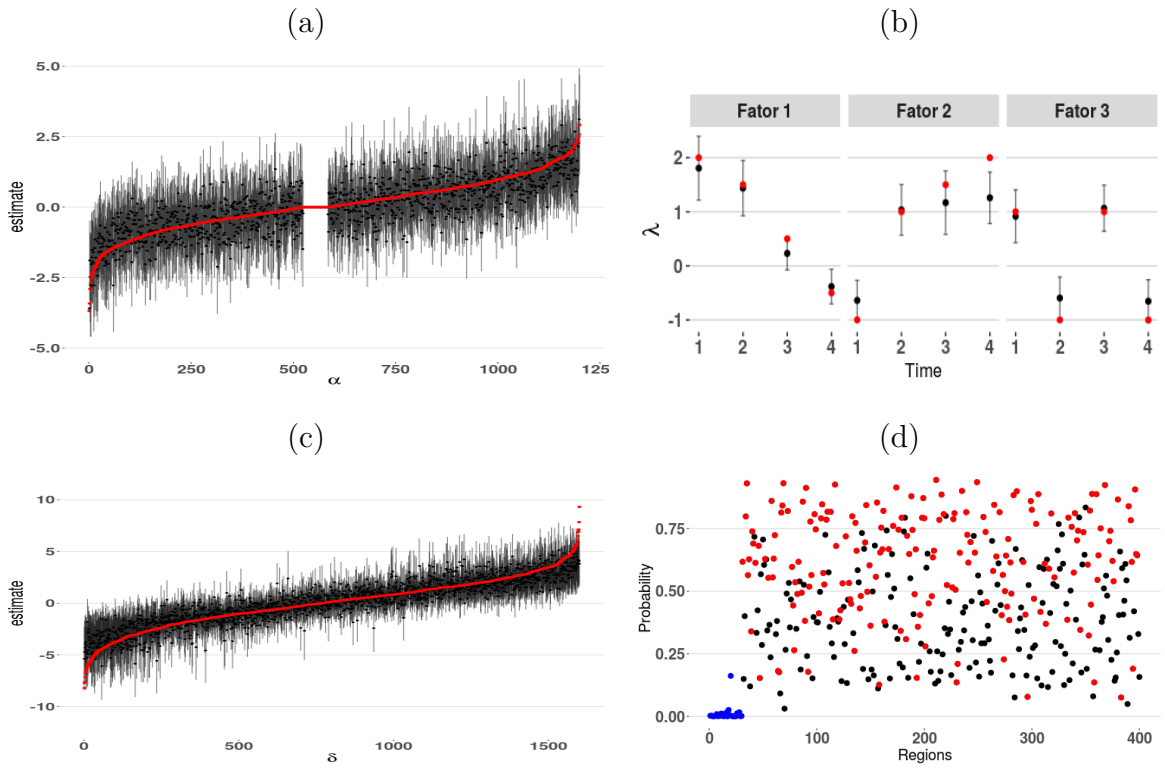


Figure 4.23: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1, G_2 e G_3, a cor vermelha representa locais de G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o cenário: $M^{K_3 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

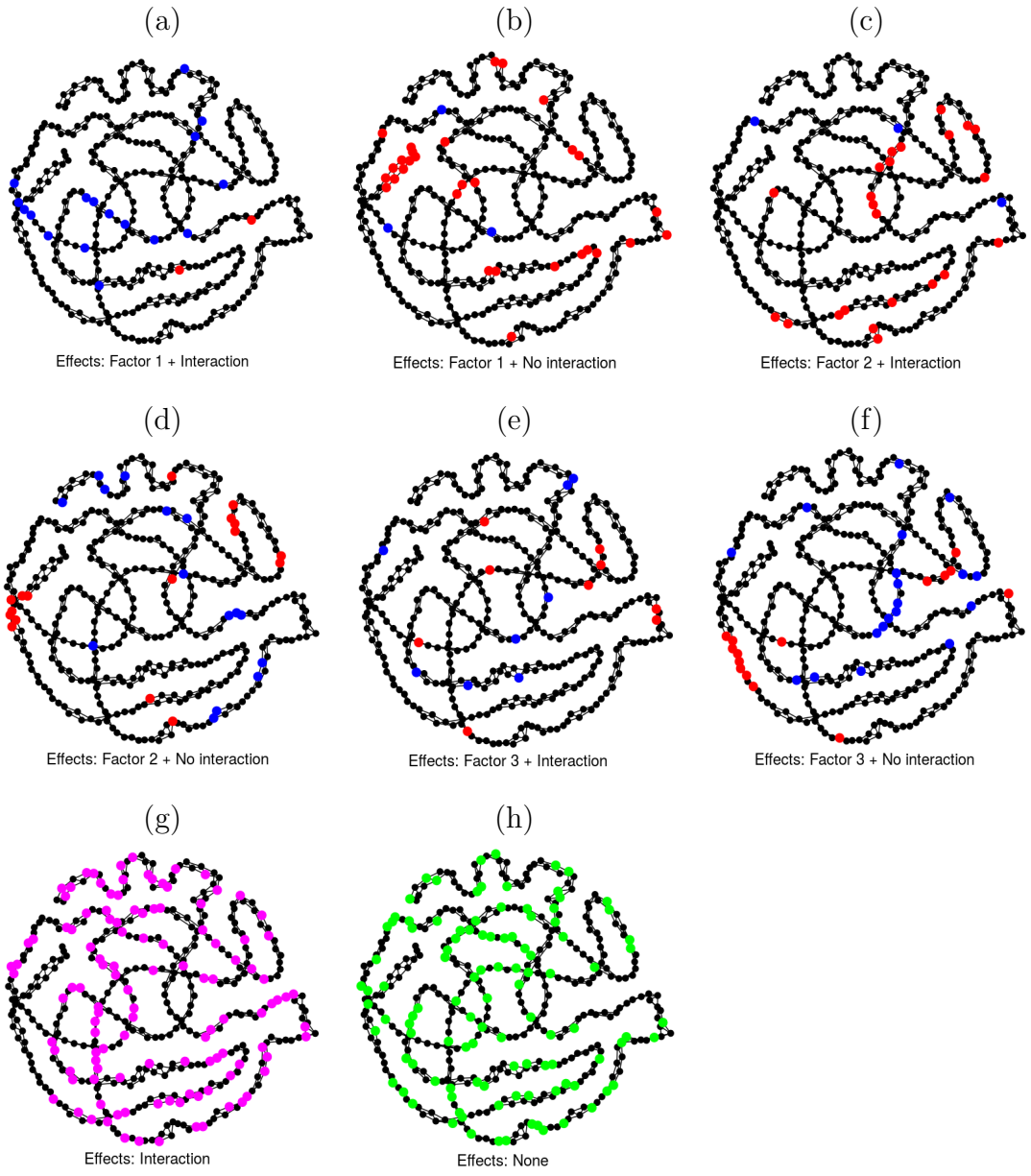


Figure 4.24: Grafos com 4 vizinhos por região imitando a estrutura espacial dos dados artificiais. Cada ponto representa um local. A cor vermelha (carga positiva) ou azul (carga negativa) identifica os locais associados a algum efeito principal e/ou interação (Paineis a, b, c, d, e, f). No Painel (g), a cor magenta indica os locais afetados somente por interação. A cor verde (Painel h) denota as regiões não afetadas por qualquer efeito. Considere o cenário: $M^{K_3 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

4.5 Análise com erro de especificação de K

Ao analisarmos dados reais, ao contrário do que acontece com dados artificiais, não sabemos, *a priori*, o número de fatores existentes para procedermos à estimação correta. Nesta seção analisamos como o modelo se comporta em dois cenários nos quais o número de fatores escolhido para ajuste do modelo difere do valor verdadeiro. Esclarecemos ao leitor que os gráficos com intervalo HPD de 95% não são apresentados devido ao fato do número de parâmetros verdadeiros ser diferente do número ajustado, tornando tal análise intervalar sem sentido. Primeiramente consideramos o caso em que $K_V = 2$ e $K_A = 3$, lembrando que K_V equivale ao valor verdadeiro para K e K_A, o valor ajustado. Em seguida analisamos a situação na qual temos $K_V = 3$ e $K_A = 2$.

Erro de especificação em que $K_V = 2$ e $K_A = 3$

A Tabela 4.13 apresenta as estimativas dos coeficientes em β, da variância dos erros σ^2 e do parâmetro de variância τ_α. A estimação para esses parâmetros foi bem capturada pelo modelo e todos eles estão dentro do intervalo HPD de 95%. Os desvios padrão dos coeficientes em β são os menores na tabela indicando que há pouca incerteza quanto à estimação dos parâmetros da regressão. A média e a mediana de todos os elementos são bem próximas indicando, mais uma vez, que as distribuições *a posteriori* são simétricas. Os maiores desvios padrão continuam sendo dos elementos em η^*. A Figura 4.25 mostra que o evolução da estimativa de η^*, no tempo, seguiu o valor verdadeiro com pequena sobrestimação para o tempo 4.

Analisando os Paineis (a, b, c, d) da Figura 4.26, as estimativas para α e λ, do Fator 1, seguiram o padrão verdadeiro. Comparando o Fator 2 verdadeiro com os Fatores 2 e 3 estimados, parece que houve uma diluição do efeito original (linha 2, Painel (c)) nas linhas 2 e 3 do Painel (d), mas com o Fator 2 capturando a maior parte do efeito verdadeiro, ou seja, o padrão de crescimento. Considerando o resultado para δ, Paineis (e) e (f), vemos que o padrão global foi bem capturado, especialmente para o Tempo 1. Concluimos que o erro na especificação de um fator a mais do que o verdadeiro não traz prejuízo para a estimação de δ, e consequentemente, os coeficientes da regressão são

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.50	0.47	0.47	0.13	0.21	0.74
β_1	-1.00	-0.99	-0.99	0.05	-1.10	-0.89
β_2	1.00	0.97	0.97	0.05	0.88	1.06
σ^2	0.80	1.01	1.00	0.14	0.72	1.27
τ_α	2.00	1.21	1.06	0.50	0.43	2.19
η_1^*	-2.00	-2.07	-2.07	0.30	-2.67	-1.47
η_2^*	1.50	1.55	1.56	0.30	0.95	2.14
η_3^*	0.75	1.06	1.07	0.24	0.56	1.54
η_4^*	-1.00	-0.50	-0.50	0.28	-1.05	0.05

Tabela 4.13: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o cenário: $M_{L_{400}T_4V_4}^{K_2K_3I_{50\%}}$ com $\approx 50\%$ de $Y_i^{'}s = 1$, em que o sobrescrito "K_2K_3" significa que o número de fatores verdadeiro é $K = 2$ e o número de fatores ajustados é $K = 3$.

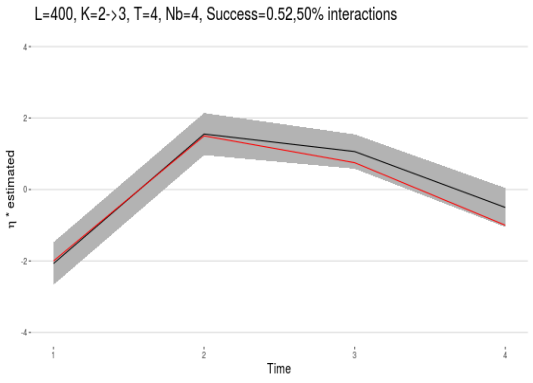


Figure 4.25: Média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha). Considere o cenário: $M_{L_{400}T_4V_4}^{K_2I_{50\%}}$ com $\approx 50\%$ de $Y_i^{'}s = 1$, mas com a estimação de $K = 3$.

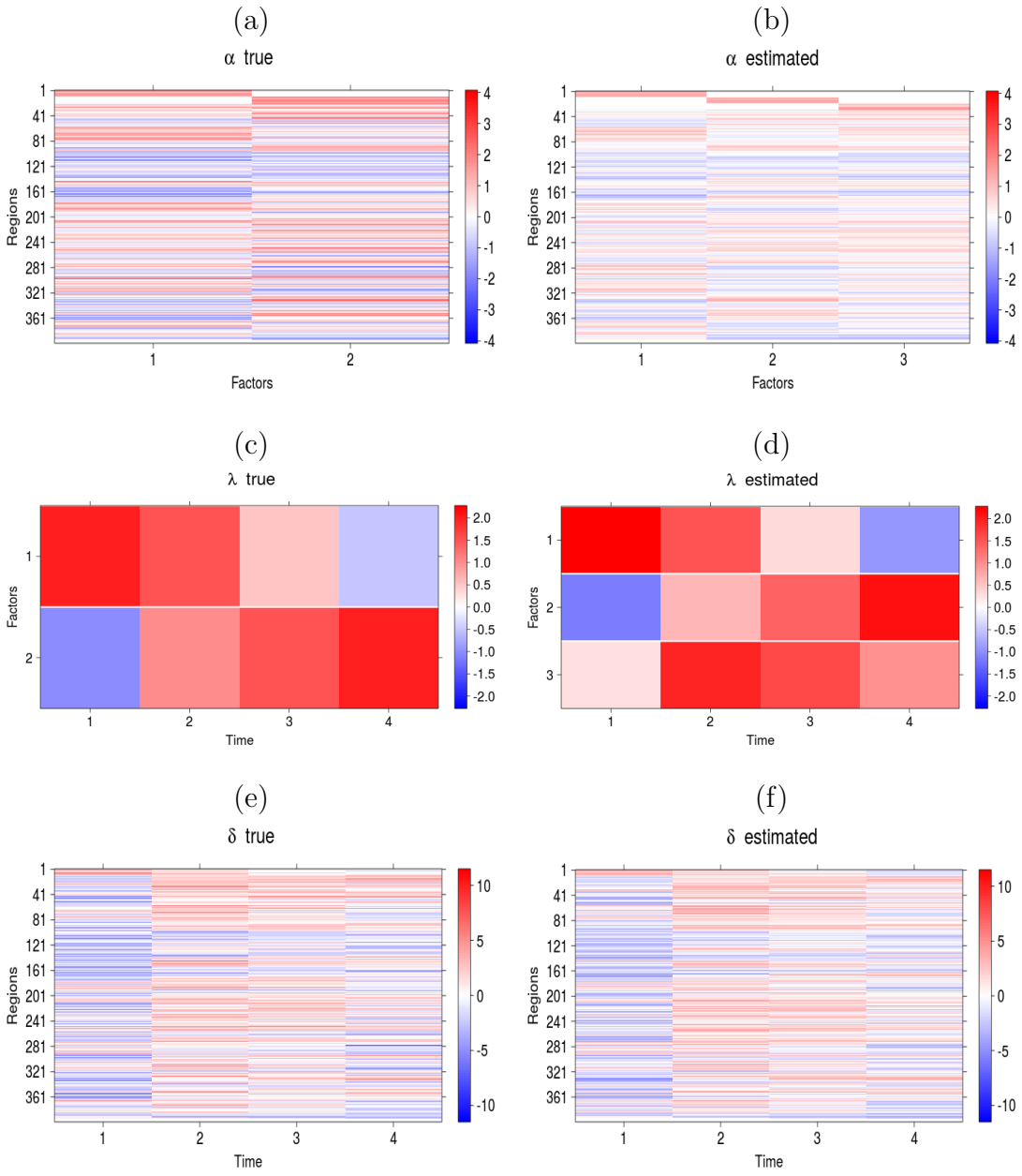


Figure 4.26: Mapas de calor comparando valores verdadeiros e estimados para o cenário $M_{L400T4V4}^{K_2K_3I_{50\%}}$ com $\approx 50\%$ de $Y_i's = 1$, em que o sobrescrito "K_2K_3" significa que o número de fatores verdadeiro é $K = 2$ e que o número de fatores ajustados é $K = 3$. Paineis à esquerda se referem aos valores verdadeiros e os à direita aos estimados.

bem estimados (veja Tabela 4.13). A seguir vamos analisar se a ocorrênca da situação inversa, número de fatores verdadeiro maior que o número ajustado, compromete ou não a estimativa final.

Erro de especificação em que $K_V = 3$ e $K_A = 2$

Neste tópico vamos analisar o comportamento do modelo para o caso em que temos $K_V = 3$ e $K_A = 2$. Pela Tabela 4.14, vemos que as estimativas não foram tão boas quanto as do cenário anterior. Note que os valores para σ^2 e τ_α ficaram fora do intervalo HPD, mas lembrando ao leitor que os resultados se referem a execução de apenas uma réplica de Monte Carlo. No entando, os desvios padrão para β continuaram sendo os menores e próximos de zero, indicando que existe pouca incerteza *a posteriori* para os coeficientes da regressão. Os desvios padrão de η^*, juntamente com o desvio padrão de τ_α, apontam uma maior incerteza *a posteriori* de suas estimativas. Entretanto, a Figura 4.27 mostra que o padrão de crescimento de η^* foi totalmente capturado, ficando os valores verdadeiros completamente dentro do intervalo HPD de 95% (área sombreada).

Pela Figura 4.28 podemos verificar que o Fator 2 estimado, Painel (d), capturou o padrão desce-sobe-desce do Fator 3 verdadeiro, Painel (c), mas com alguma subestimação. Diferentemente do cenário do tópico anterior, as cargas associadas ao Fator 1 não foram tão bem estimadas, Painel (b), e nem o Fator 1, Painel (d). Não conseguimos identificar uma relação entre as cargas associadas aos Fatores 2 e 3 verdadeiros e as cargas estimadas, segunda coluna do Painel (b). Veja que a maioria das cargas da coluna 2 do Painel (b) são negativas, ao contrário do que verificamos nas colunas 2 e 3 do Painel (a). Veja, por exemplo, os locais entre os números 240 e 250 cujos valores verdadeiros são positivos e os valores estimados ficaram muito próximos de zero. E ainda, as cargas dos locais entre os números 10 e 20 possuem valores verdadeiros positivos e os valores estimados negativos. No entanto, analisando os Paineis (e) e (f), o padrão global de δ foi bem capturado.

Terminamos aqui, a análise do erro de especificação de K, concluindo que com variações de 1 unidade, para mais ou para menos, do número de fatores, a estimação de δ ainda é bem satisfatória, não compremetendo o resultado final da parte logística contendo os efeitos das covariáveis. Na próxima seção analisaremos cenários nos quais ocorre a

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.50	0.79	0.80	0.16	0.48	1.08
β_1	-1.00	-1.11	-1.11	0.06	-1.22	-1.00
β_2	1.00	0.97	0.97	0.05	0.88	1.07
σ^2	0.80	1.22	1.21	0.15	0.93	1.51
τ_α	2.00	1.29	1.25	0.34	0.72	1.95
η_1^*	-2.00	-2.23	-2.26	0.35	-2.90	-1.52
η_2^*	-1.50	-1.94	-1.96	0.31	-2.52	-1.31
η_3^*	0.75	1.04	1.04	0.33	0.39	1.69
η_4^*	1.00	0.95	0.95	0.34	0.28	1.62

Tabela 4.14: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o cenário: $M_{L_{400}T_4V_4}^{K_3K_2I_{50\%}}$ com $\approx 50\%$ de $Y_i^{'}s = 1$, em que o sobrescrito "K_3K_2" significa que o número de fatores verdadeiro é $K = 3$ e que o número de fatores ajustados é $K = 2$.

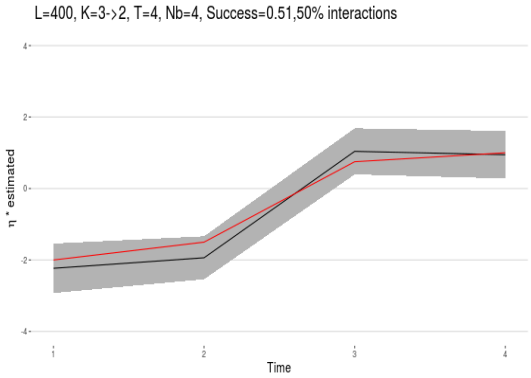


Figure 4.27: Média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha). Considere o cenário: $M^{K_3 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$, mas com a estimação de $K = 2$.

sobreparametrização do modelo fatorial. Dois cenários serão analisados: o primeiro se refere ao caso em que o número de fatores é igual ao número de tempos ($K = T$) e o segundo quando o número de fatores é maior do que a quantidade de tempos ($K > T$). Especificamente, iremos analisar os casos: "$K = 4$, $T = 4$" e "$K = 5$, $T = 4$".

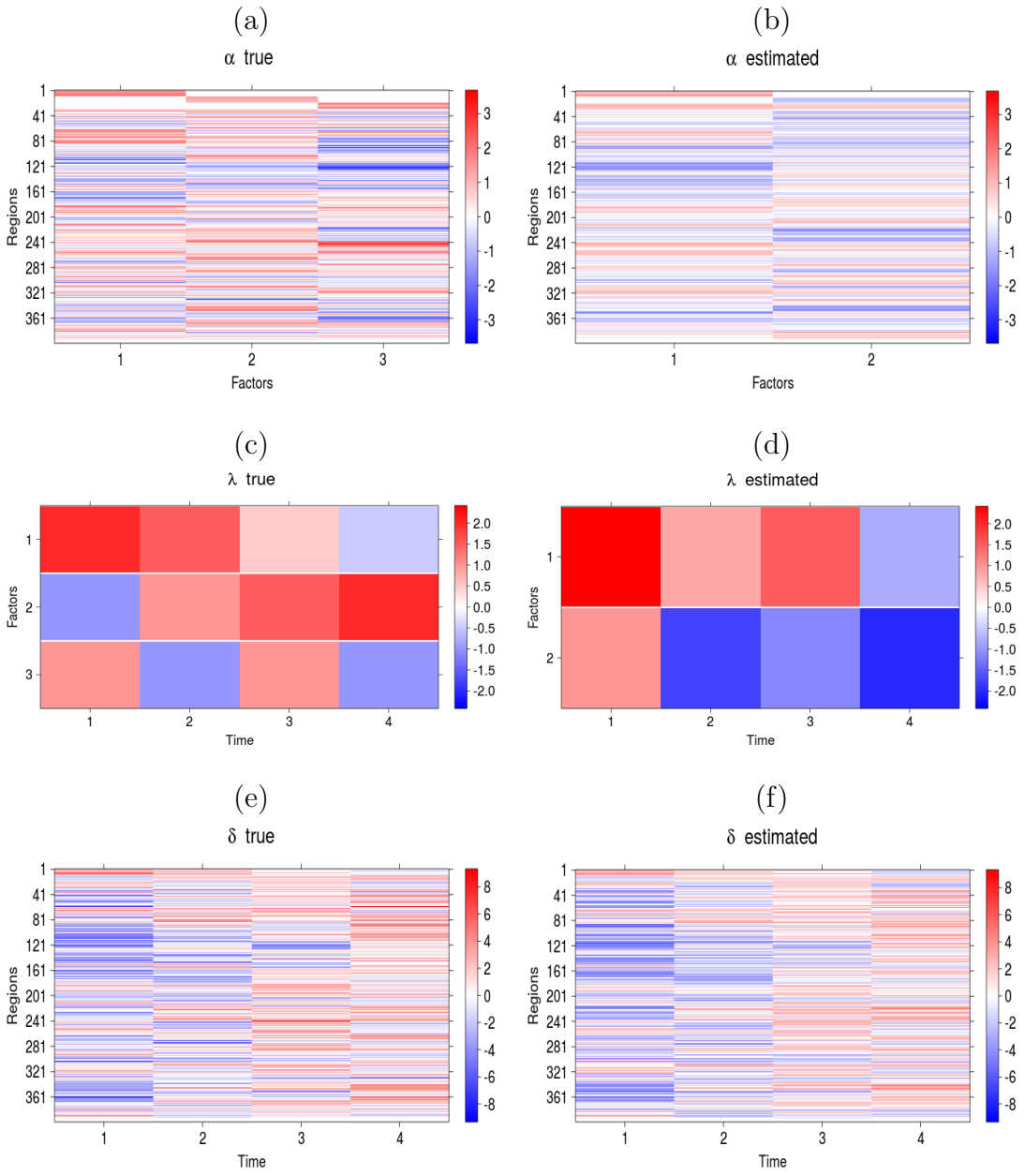


Figure 4.28: Mapas de calor comparando valores verdadeiros e estimados para o cenário $M_{L400T4V4}^{K_3K_2I_{50\%}}$ com $\approx 50\%$ de $Y_i's = 1$, em que o sobrescrito "K_3K_2" significa que o número de fatores verdadeiro é $K = 3$ e que o número de fatores ajustados é $K = 2$. Paineis à esquerda se referem aos valores verdadeiros e os à direita aos estimados.

4.6 Sobreparametrização do modelo fatorial

Uma questão importante a ser avaliada ao ajustarmos o modelo proposto nesta tese, em que a dependência temporal é modelada pelos fatores comuns, é analisar o comportamento do modelo quando o número de fatores latentes a serem estimados é igual ou superior ao número de tempos disponíveis. Nesta seção vamos analisar duas situações em que temos $T = 4$ tempos, o qual é um valor baseado na aplicação real. Na primeira configuramos $K = 4$ fatores e na segunda $K = 5$. Em ambos os casos o número de fatores ajustados é igual ao número verdadeiro. O cenário avaliado considera $L = 400$ locais, 4 vizinhos por região, $\approx 50\%$ de locais de G_E afetados pela interação não linear e $\approx 50\%$ de $Y_i^{'}s = 1$.

Sobreparametrização em que $K = 4$ e $T = 4$

A Tabela 4.15 apresenta as estimativas dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α e da interação não linear η^*. Analisando o valor verdadeiro e o intervalo HPD de 95% vemos que os valores verdadeiros dos parâmetros estão todos dentro do envelope. O desvio padrão para os coeficientes em β são os menores com os casos β_1 e β_2 bem próximos de zero, indicando pouca incerteza *a posteriori* para essas estimativas. Veja que os desvios padrão para os quatro parâmetros de η^* são os maiores, mas que as estimativas, com exceção de η_4^*, são bem próximas do valor verdadeiro e seguindo o padrão sobe-desce-sobe. Essa situação pode ser melhor verificada pela Figura 4.29 em que a linha vermelha representa o valor verdadeiro e a linha preta, o estimado.

Pela Figura 4.30, paineis da esquerda versus paineis da direita, podemos ver, pelas tonalidades de cores das colunas, que o padrão das estimativas foi bem próximo do padrão dos valores verdadeiros. Essa constatação pode ser melhor verificada analisando a Figura 4.31, Paineis (a),(b) e (c), em que temos os intervalos HPD de 95% de α, λ e δ, respectivamente. Perceba como a linha vermelha (valores verdadeiros) corta, praticamente ao meio, todos os intervalos de α e δ. No caso de δ, ocorre alguma sobrestimação e subestimação para valores extremos negativos e positivos, respectivamente. No Painel (b) vemos com mais distinção como λ foi bem estimado, com exceção dos escores do Fator 1 no Tempo 2 e do Fator 3 no Tempo 3, em que o valor verdadeiro (ponto vermelho)

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.50	0.52	0.51	0.15	0.20	0.81
β_1	-1.00	-0.97	-0.97	0.05	-1.08	-0.87
β_2	1.00	0.98	0.98	0.05	0.89	1.07
σ^2	0.80	0.81	0.79	0.19	0.47	1.20
τ_α	2.00	2.93	2.87	1.01	1.32	4.83
η_1^*	-2.00	-1.78	-1.79	0.46	-2.79	-0.91
η_2^*	1.50	1.91	1.93	0.36	1.20	2.60
η_3^*	-0.75	-0.99	-1.00	0.38	-1.73	-0.17
η_4^*	1.00	0.52	0.52	0.42	-0.30	1.31

Tabela 4.15: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o cenário: $M^{K_4 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

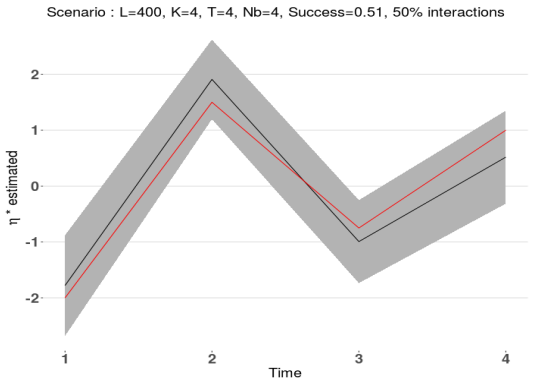


Figure 4.29: Média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha). Considere o cenário: $M^{K_4 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

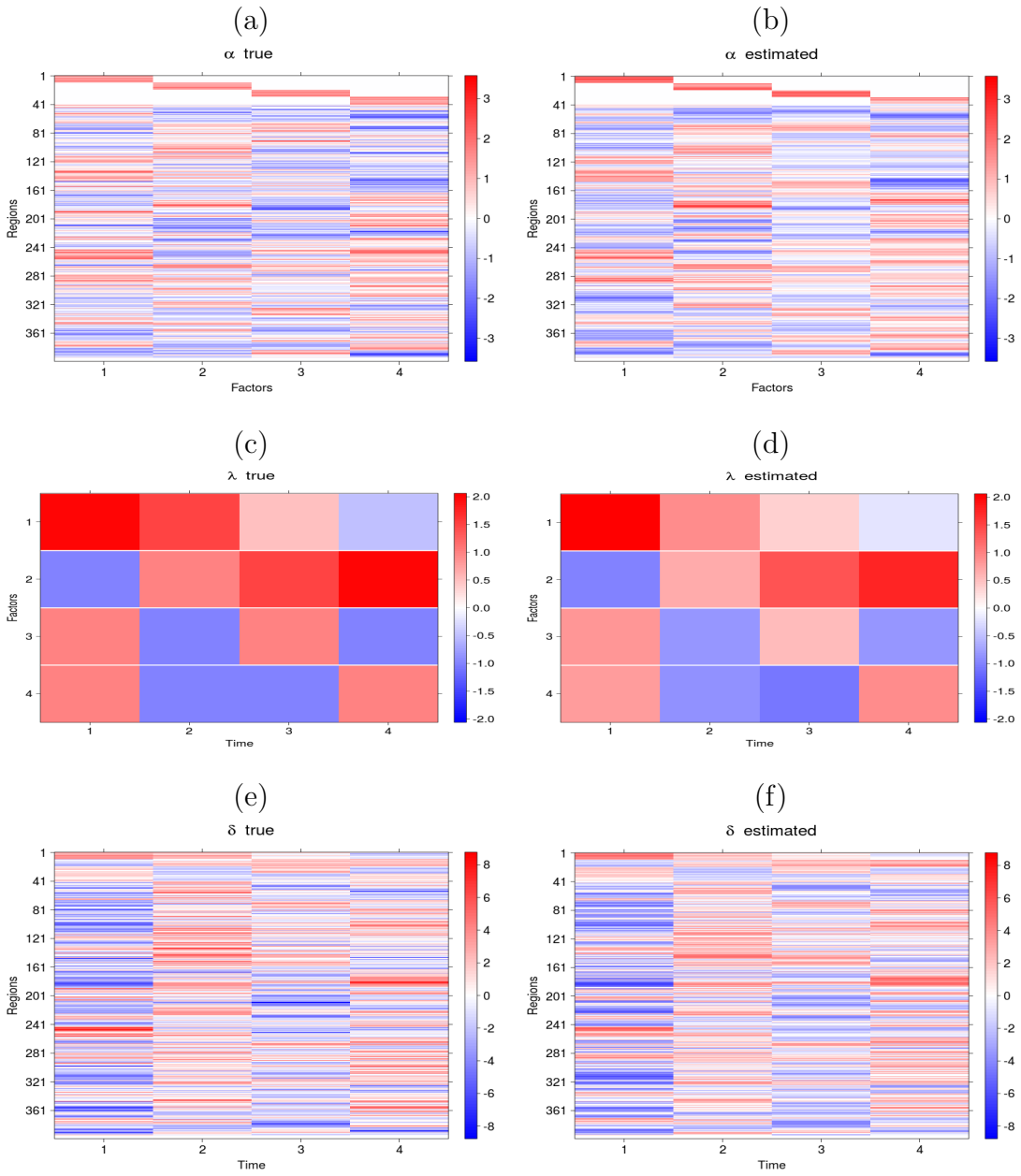


Figure 4.30: Mapas de calor comparando valores verdadeiros e estimados para o cenário $M^{K_4 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$. Paineis (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

ficou um pouco fora do intervalo HPD de 95%. Finalmente, pelo Painel (d) podemos ver as regiões afetadas por interação na geração dos dados (pontos vermelhos). Comparando com os casos em que temos $K = 2$ fatores, Painel (d) das Figuras 4.3, 4.7, 4.11 e 4.19, vemos que o número de locais cuja probabilidade de interação é < 0.5 é maior do que nesses casos citados, mas o número de locais com probabilidade > 0.5 ainda é a maioria.

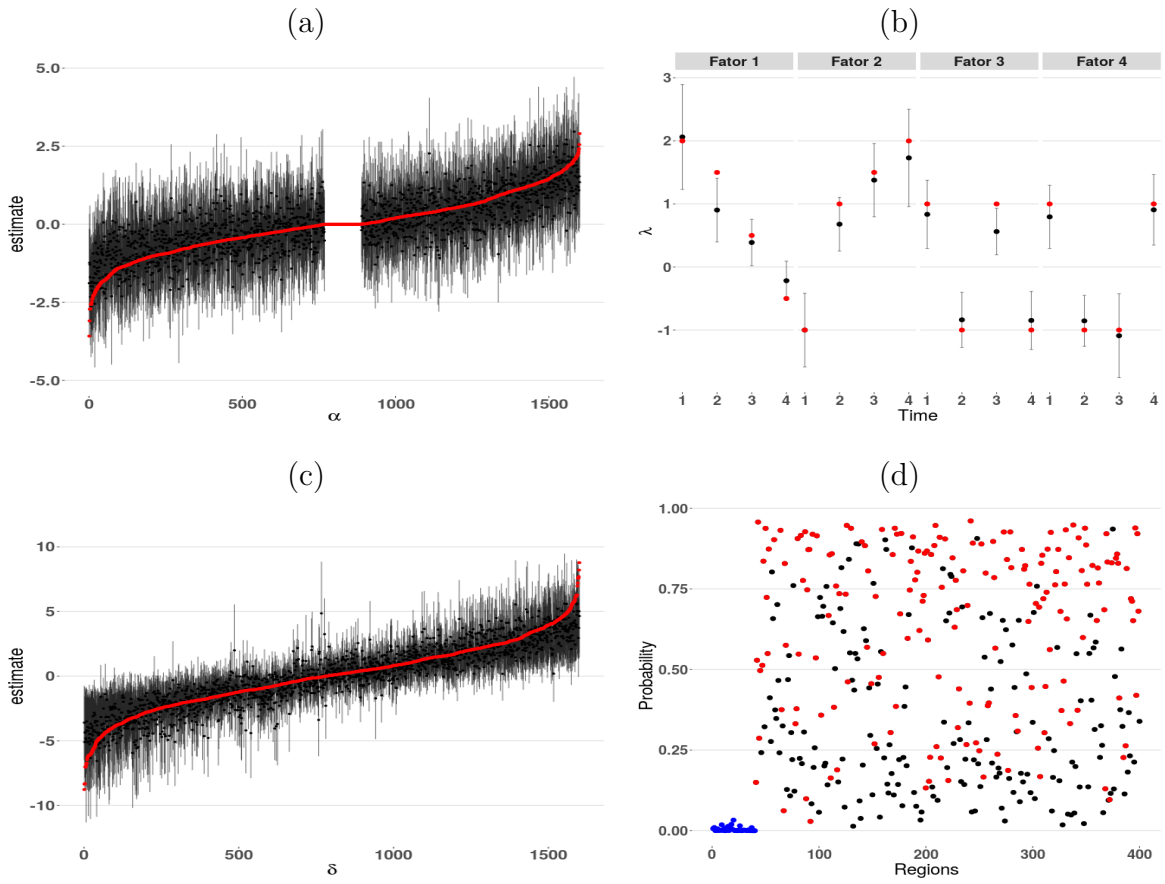


Figure 4.31: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1, G_2, G_3 e G_4, a cor vermelha representa locais de G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o cenário: $M^{K_4 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

Finalizamos esse tópico concluindo que o modelo apresentou estimativas satisfatórias

mesmo na ocorrência de sobreparametrização quando o número de fatores é igual ao número de tempos. Na análise visual nota-se, em relação aos casos com $K = 2$ referenciados anteriormente, um maior número de locais os quais foram gerados com interação, mas que apresentaram, *a posteriori*, probabilidade baixa de serem afetados por esse efeito. De fato, seja $\text{prop}_{<0.5}$ a proporção de locais, dentre todos os locais gerados com interação, que obtiveram a probabilidade *a posteriori* de interação menor que 0.5. A proporção $\text{prop}_{<0.5}$ aumenta de $\approx 7.44\%$ em média, nos cenários com $K = 2$, para $\approx 13.5\%$, indicando que a sobreparametrização afeta a estimação de η^*. Importante informar que $\text{prop}_{<0.5} = 13\%$ na situação em que $K = 3$ e $T = 4$, reforçando o fato de que quanto maior o número de fatores, mantendo o tempo fixo, maior a proporção $\text{prop}_{<0.5}$, gerando um erro maior na estimação de η^*. No próximo tópico vamos apresentar como o modelo se comporta quando o número de fatores é maior do que o de tempos, especificamente para $K = 5$ fatores e $T = 4$ tempos.

Sobreparametrização em que $K = 5$ e $T = 4$

Neste tópico vamos analisar outra configuração de sobreparametrização do modelo fatorial dinâmico considerando o número de fatores maior que o número de tempos. Essas análises são importantes para esclarecer como o modelo proposto se comporta em diversas situações de forma a orientar o analista no uso adequado do mesmo. Mais uma vez temos as estimativas dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α e da interação não linear η^* apresentadas na Tabela 4.16. Neste caso, observamos que apenas o valor verdadeiro de σ^2 ficou fora do envelope HPD de 95%. As mesmas observações, discutidas no tópico anterior, valem para os demais parâmentros.

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.50	0.60	0.60	0.10	0.40	0.81
β_1	-1.00	-0.93	-0.93	0.05	-1.03	-0.82
β_2	1.00	1.05	1.05	0.05	0.96	1.15
σ^2	0.80	0.48	0.46	0.15	0.23	0.78
τ_α	2.00	3.24	3.19	0.75	1.72	4.65
η_1^*	2.00	1.69	1.70	0.47	0.79	2.53
η_2^*	-1.50	-1.58	-1.60	0.41	-2.35	-0.73
η_3^*	-0.75	-0.92	-0.93	0.43	-1.85	-0.12
η_4^*	1.00	1.13	1.15	0.48	0.21	2.02

Tabela 4.16: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o cenário: $M_{L_{400}T_4V_4}^{K_5 I_{50\%}}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

Destaca-se os elementos de η^* para os quais o desvio padrão foi superior aos do caso $K = 4$ e $T = 4$, indicando uma maior incerteza *a posteriori* dessas estimativas em relação à situação quando temos $K = T$. Por outro lado, as estimativas pontuais (média e mediana) ficaram mais próximas do valor verdadeiro, veja a Figura 4.32.

Diferentemente do tópico anterior, percebemos na Figura 4.33 que o padrão das

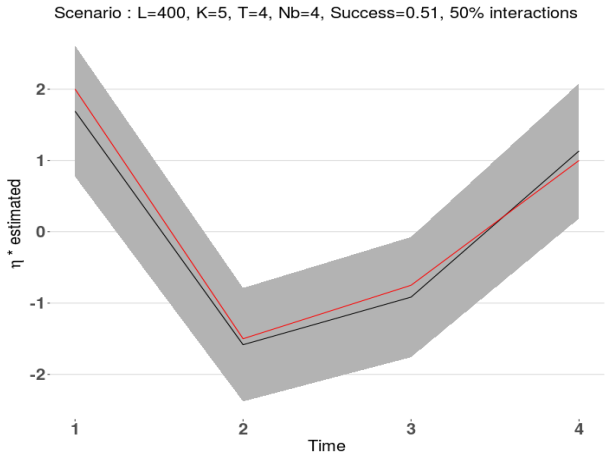


Figure 4.32: Média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha). Considere o cenário: $M^{K_4 I_{50\%}}_{L_{400} T_5 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

estimativas de α, Paineis (a), (b), e de λ, Paineis (c) e (d), diverge em algumas partes do padrão verdadeiro. No caso de α podemos destacar essa divergência no Fator 1 para os locais de ≈ 50 a 80, de ≈ 160 a 190 e de ≈ 360 a 400; no Fator 2 para os locais de ≈ 240 a 270 e de ≈ 330 a 400; e para alguns locais nos Fatores 3, 4 e 5. Para λ podemos salientar as diferenças de tonalidades do Fator 1 nos Tempo 2 e 3; do Fator 2 nos Tempos 1 e 2; do Fator 3 no Tempo 3; do Fator 4 nos Tempos 1, 2, e 3; e do Fator 5 nos Tempos 1, 3, e 4. Essas situações são melhor verificadas pela Figura 4.34, Painel (b), lembrando que pontos vermelhos são os valores verdadeiros e os pretos, a média *a posteriori* das estimativas. Interessante ressaltar que apesar de haver essas divergências em α e λ, as estimativas de δ, Paineis (e) e (f) da Figura 4.33 se parecem com as do tópico anterior em que temos $K = 4$ fatores e $T = 4$ tempos, ocorrendo, também, sobrestimação e subestimação para valores nos extremos negativos e positivos, respectivamente; ver Figura 4.34 (c).

Avaliando, visualmente, o Painel (d) da Figura 4.34 percebemos o mesmo efeito descrito no caso do tópico anterior, em que, em relação ao cenários com $K = 2$ (veja o Painel (d) das Figuras 4.3, 4.7, 4.11 e 4.19), temos que a probabilidade *a posteriori* de existência do efeito ficou menor que 0.5 para uma maior quantidade de locais de

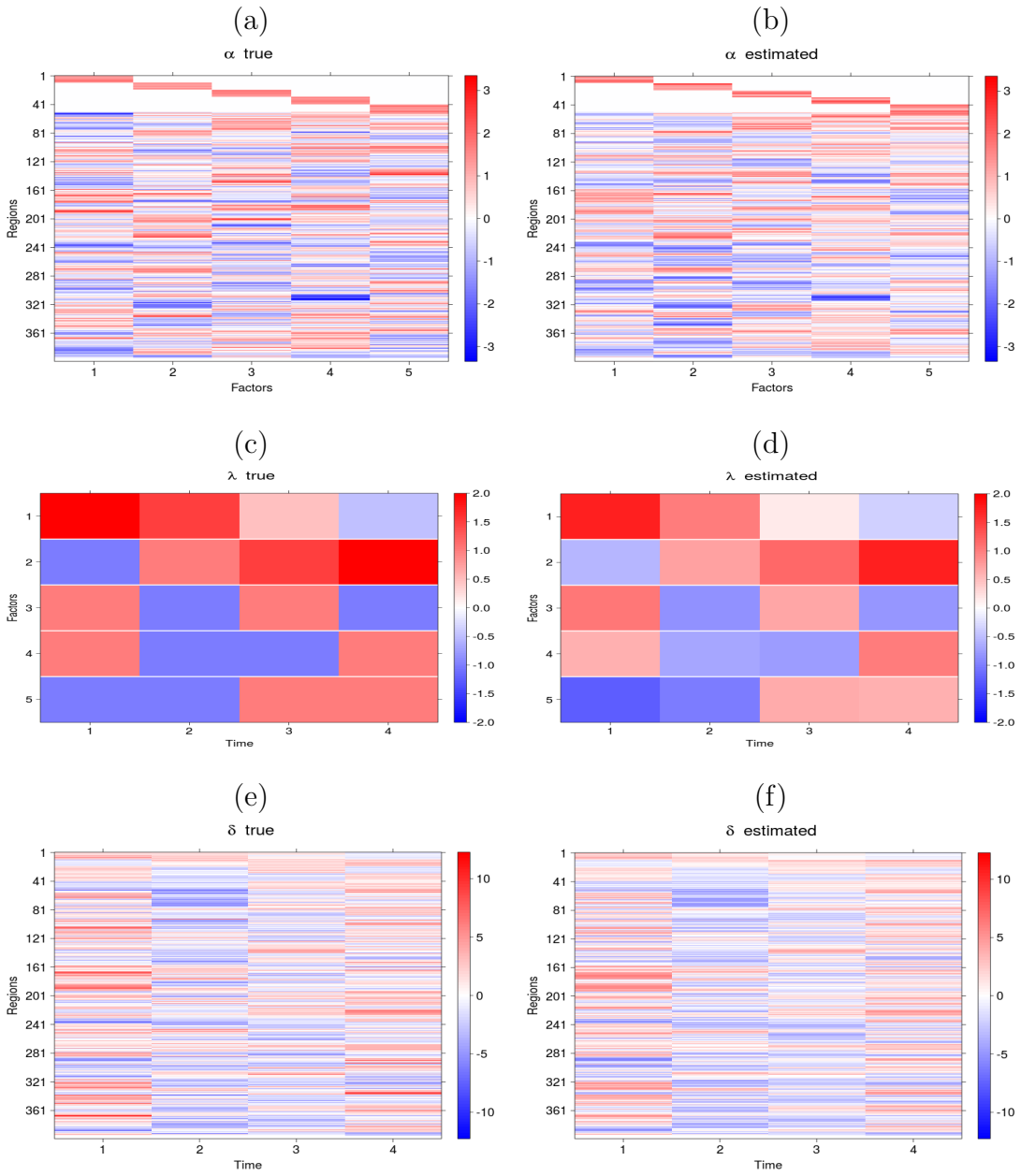


Figure 4.33: Mapas de calor comparando valores verdadeiros e estimados para o cenário $M^{K_4 I_{50\%}}_{L_{400} T_5 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$. Paineis (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

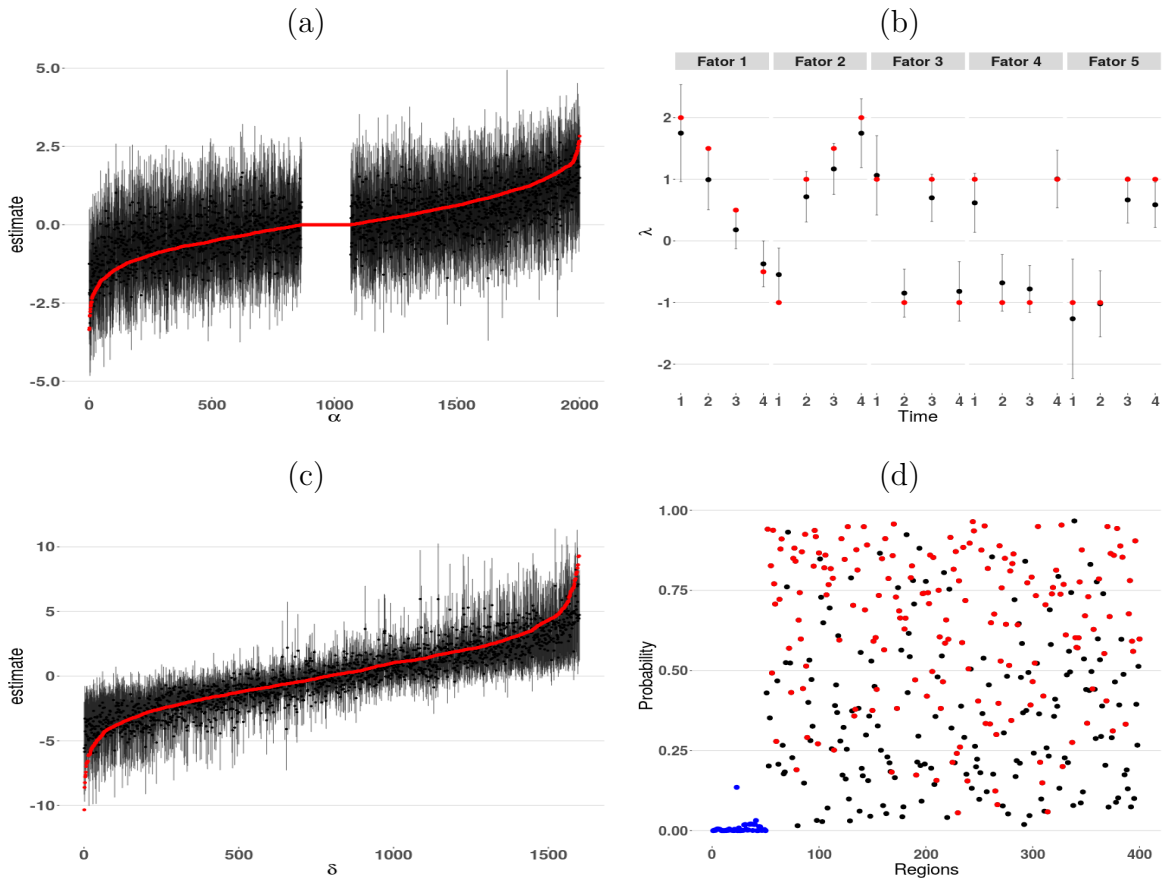


Figure 4.34: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1, G_2 e G_3, a cor vermelha representa locais de G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o cenário: $M^{K_4 I_{50\%}}_{L_{400} T_5 V_4}$ com $\approx 50\%$ de $Y_i's = 1$.

G_E verdadeiramente afetados por interação (pontos vermelhos). De fato, para este caso temos que $\text{prop}_{<0.5} = 11.75\%$ (veja o último parágrafo do tópico anterior para a definição de $\text{prop}_{<0.5}$). Esse resultado é um pouco inferior ao caso $K = T = 4$, em que o resultado foi de 13.5%, mas é importante lembrar que essas estimativas são relativas a apenas uma execução do algoritmo MCMC. Mesmo assim, a proporção $\text{prop}_{<0.5}$, aqui, ainda é muito superior do que as proporções ocorridas para os casos com $K = 2$, em que o valor médio foi de 7.44%.

Encerramos a análise para a sobreparametrização do modelo fatorial constatando que a situação com $K > T$, em relação à $K = T$, gera estimativas um pouco piores para α e λ, podendo comprometer a identificação de agrupamentos de locais, que é um dos objetivos do modelo proposto nesta tese. Em ambos os casos também identificamos que a proporção $\text{prop}_{<0.5}$ (veja o último parágrafo do tópico anterior) é bem menor do que nos casos que $K = 2$ e $T = 4$, ou seja, uma situação em que o número de fatores é bem menor (metade) do que o número de tempos, indicando que a sobreparametrização causa imprecisão na detecção dos locais com interação, afetando a interpretabilidade geral do modelo, pois interfere na probabilidade de $Y_i^{'} s = 1$ e na identificação de conglomerados formados a partir de efeitos principais com o efeito de interação.

Concluimos esta seção reportando que quanto maior o número de fatores em relação ao número de tempos, pior será o ajuste do modelo para a identificação de locais associados a fatores latentes, e da verificação de conglomerados formados por locais afetados por efeitos principais e interação não linear. Desta forma, o pesquisador deve ficar atendo ao definir o número de fatores a serem considerados na modelagem. Na próxima seção apresentaremos a análise de resíduos de Pearson para comparação do comportamento do modelo em diversos cenários.

4.7 Análise de resíduos

Os resíduos trazem informações importantes sobre a adequação dos modelos estatísticos. Eles são usados para identificar discrepâncias em relação aos dados, desempenhando um papel fundamental na verificação da qualidade do ajuste do modelo (Cordeiro e

Simas, 2009). *** Nesta seção, analisamos os resíduos de Pearson para os cenário com $L = 100,\ 200$ e 400 locais, $K = 2$ fatores, $T = 4$ tempos, 4 e 6 vizinhos por região e $\approx$ 50% de $Y_i^{'}s = 1$ em que temos 30% e 50% de locais afetados por interação. Os resíduos de Pearson são obtidos pela formulação $R_i = \frac{(Y_i - \hat{\theta}_i)}{\sqrt{\hat{\theta}_i(1-\hat{\theta}_i)}}$. Para utilizarmos um único valor para análise comparativa entre os cenários, calculamos a média quadrática dos resíduos, ou seja, $\sum_{i=1}^{n} R_i^2/n$.

Além desses cenários, analisamos, também, os casos $M^{K_2 I_{50\%}}_{L_{400}T_4V_4}$ com $\approx 20\%$ de $Y_i^{'}s = 1$ e $M^{K_3 I_{50\%}}_{L_{400}T_4V_4}$ com $\approx 50\%$ de $Y_i^{'}s = 1$. Em todas as situações o número de fatores verdadeiros é igual à quantidade de fatores ajustados. Conforme comentado anteriormente sobre a similaridade entre os cenários com 4 e 6 vizinhos por região, podemos ver pela Tabela 4.17 que os resíduos para esses casos são muito próximos.

Cenário	Num.Parâmetros	% Int	% $Y_i^{'}s = 1$	Resíduo (4 viz.)	Resíduo (6 viz.)
$M^{K_2}_{L_{100}T_4}$	317	$\approx 30\%$	$\approx 50\%$	0.7316	0.7456
		$\approx 50\%$		0.7422	0.7407
$M^{K_2}_{L_{200}T_4}$	617	$\approx 30\%$	$\approx 50\%$	0.7351	0.7572
		$\approx 50\%$		0.7334	0.7533
$M^{K_2}_{L_{400}T_4}$	1217	$\approx 30\%$	$\approx 50\%$	0.7495	0.7646
		$\approx 50\%$		0.7326	0.7517
$M^{K_3}_{L_{400}T_4}$	1621	$\approx 50\%$	$\approx 50\%$	0.7171	
$M^{K_2}_{L_{400}T_4}$	1217	$\approx 50\%$	$\approx 20\%$	0.6926	

Tabela 4.17: Média quadrática dos resíduos para os cenários $M^{K_2}_{L_{100}T_4}$, $M^{K_2}_{L_{200}T_4}$, $M^{K_2}_{L_{400}T_4}$, $M^{K_3}_{L_{400}T_4}$ em diversas configurações: 4 e 6 vizinhos por região, $\approx 20\%$ ou 50% de $Y_i^{'}s = 1$, 30% e 50% de locais de G_E afetados por interação não linear. Em todas as situações o número de fatores verdadeiros é igual à quantidade de fatores ajustados.

Importante notar que os resíduos para os cenários com 50% de locais afetados por interação, comparados com o caso 30%, são um pouco inferiores (exceção para o cenário com $L = 100$ e 4 vizinhos por região) indicando que ter mais locais afetados pelo efeito η^* determina em uma melhor estimação para o efeito de interação, que influencia na estimação final da variável resposta. Esse resultado, apesar de ser referente a apenas

uma amostra Monte Carlo, vai de encontro com as análises para 30 réplicas apresentadas na Seção 4.9 descrita mais adiante. Veja que, dentre os cenários balanceados ($\approx 50\%$ de $Y_i^{'}s = 1$) com $K = 2$ e 4 vizinhos por região, coluna "Resíduo (4 viz.)", o resíduo para o cenário $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ só não é menor do que o do cenário $M^{K_2 I_{30\%}}_{L_{100} T_4 V_4}$, mas o número de parâmetros do primeiro é praticamente 4 vezes maior do que o modelo ajustado para o caso com $L = 100$. Da mesma forma, considerando os casos com 6 vizinhos por região, coluna "Resíduo (6 viz.)", o resíduo para o cenário $M^{K_2 I_{50\%}}_{L_{400} T_4}$ só perde para os cenários com 100 locais. Esse fato indica que o aumento no número de observações trouxe maior ganho na estimação apesar do grande aumento no número de parâmetros resultando em um melhor ajuste do modelo.

A Figura 4.35 ilustra a evolução da média quadrática do resíduo de Pearson (eixo vertical) com o aumento no número de parâmetros (eixo horizontal). Para fazer o mapeamento entre o número de parâmetros e os modelos, o leitor deve avaliar o gráfico em conjunto com a Tabela 4.17. Com isso, é fácil ver que, com exceção do cenário $M^{K_2}_{L_{100} T_4}$ com 4 vizinhos por região e com menos parâmetros (317), todos os casos com 50% de locais afetados por interação possuem o resíduo menor do que quando se tem 30% de locais afetados, indicando que quanto mais locais estiverem contribuindo para a estimação de η^*, melhor será o ajuste do modelo em termos de média quadrática dos resíduos. Lembrando ao leitor que nos casos com $L = 100$ a diferença em relação ao número de locais afetados por interação nas situação com 30% e 50% é pequena, justificando a similaridade dos resutados para esses casos.

As duas seções finais deste capítulo são destinadas à análise de métricas de ajuste para comparar os diversos modelos estudados nesta tese. Na Seção 4.8 vamos tratar da análise das curvas ROC para todos os cenários com 400 locais, e finalizando o capítulo, a Seção 4.9 compara o vício relativo do cenário desbalanceado com os vícios de diversos casos balanceados com 400 locais e 50% de regiões afetadas por interação, após a execução de 30 réplicas Monte Carlo.

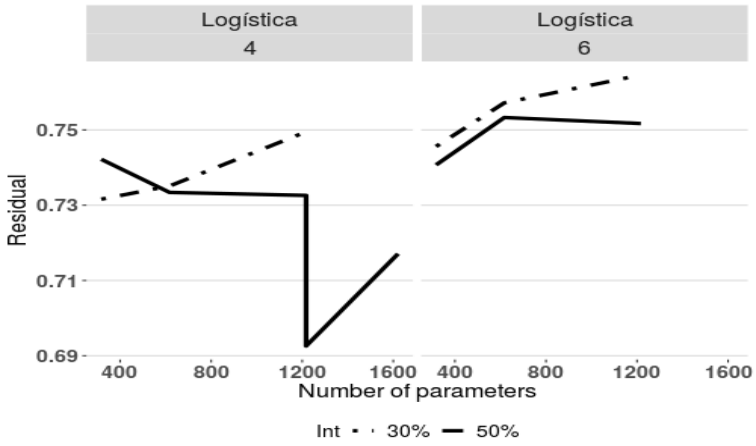


Figure 4.35: Média quadrática dos resíduos de Pearson por número de parâmetros considerando os cenários $M^{K_2}_{L_{100}T_4}$, $M^{K_2}_{L_{200}T_4}$, $M^{K_2}_{L_{400}T_4}$, $M^{K_3}_{L_{400}T_4}$ em diversas configurações: 4 e 6 vizinhos por região, $\approx$ 20% ou 50% de $Y_i^{'}s = 1$, 30% e 50% de locais de G_E afetados por interação não linear, ou seja, os mesmos cenários listados na Tabela 4.17.

4.8 Análise das curvas ROC

O gráfico da curva ROC é uma técnica de visualização do desempenho de um algoritmo de classificação binária. A curva ROC há muito tem sido utilizada na teoria de detecção de sinais para representar o *tradeoff* entre taxas de acerto (*hit rates*) e taxas de alarme falso (*false alarm rates*) de classificadores (Egan, 1975; Swets et al., 2000). A análise da curva ROC foi estendida para uso na visualização e análise do comportamento de sistemas óticos (Swets, 1988). A tomada de decisão pela comunidade médica tem uma extensa literatura sobre o uso de curvas ROC para testes de diagnóstico (Zou, 2002). Swets et al. (2000) trouxeram as curvas ROC à atenção geral com seu artigo da *Scientific American*.

Nesta seção mostramos os gráficos das curvas ROC para diversos cenários com 400 locais. Lembramos ao leitor que a base de dados real é desbalanceado em relação à variável resposta e foi simulada com $\approx 20\%$ de $Y_i^{'}s = 1$. Os demais cenários são balanceados, ou seja, contém $\approx 50\%$ de $Y_i^{'}s = 1$. Cada base de dados artificial foi dividida em 2 conjuntos: treino e teste. O conjunto de treino foi utilizado para o ajuste do modelo e o de teste para averiguação das predições. O conjunto de teste ficou composto de $\approx$ 10% das observações e as mesmas regras dos dados originalmente gerados foi aplicada. Isso quer dizer que a base de teste possue indivíduos em todos os tempos e um indivíduo só aparece em apenas um tempo. Com os parâmetros estimados com a base de treino e com as covariáveis existentes na base de teste, calculamos as estimativas para θ_i. A partir dos $\theta_i^{'}s$ estimados ($\hat{\theta}_i$), calculamos a variável resposta, $Y_i = 0$ ou $Y_i = 1$, aplicando limiares de probabilidades variando de 0.01 a 0.99. Para $\hat{\theta}_i$ maior que o limiar fazemos $Y_i = 1$, caso contrário, $Y_i = 0$. A partir desses cálculos criamos uma matriz de confusão para cada limiar e, consequetemente, as taxas de verdadeiros positivos e falsos positivos, que compoem o eixo vertical e horizontal, respectivamente, da curva ROC. Os gráficos foram gerados através do pacote ROCR (Sing et al., 2005), da linguagem de programação R (R Core Team, 2020).

A Figura 4.36,Paineis (b), (c), (d), (e), ilustra as curvas ROC para todos os cenários com $L = 400$ e $\approx 50\%$ de $Y_i^{'}s = 1$. O Painel (a) ilustra a curva ROC para o caso

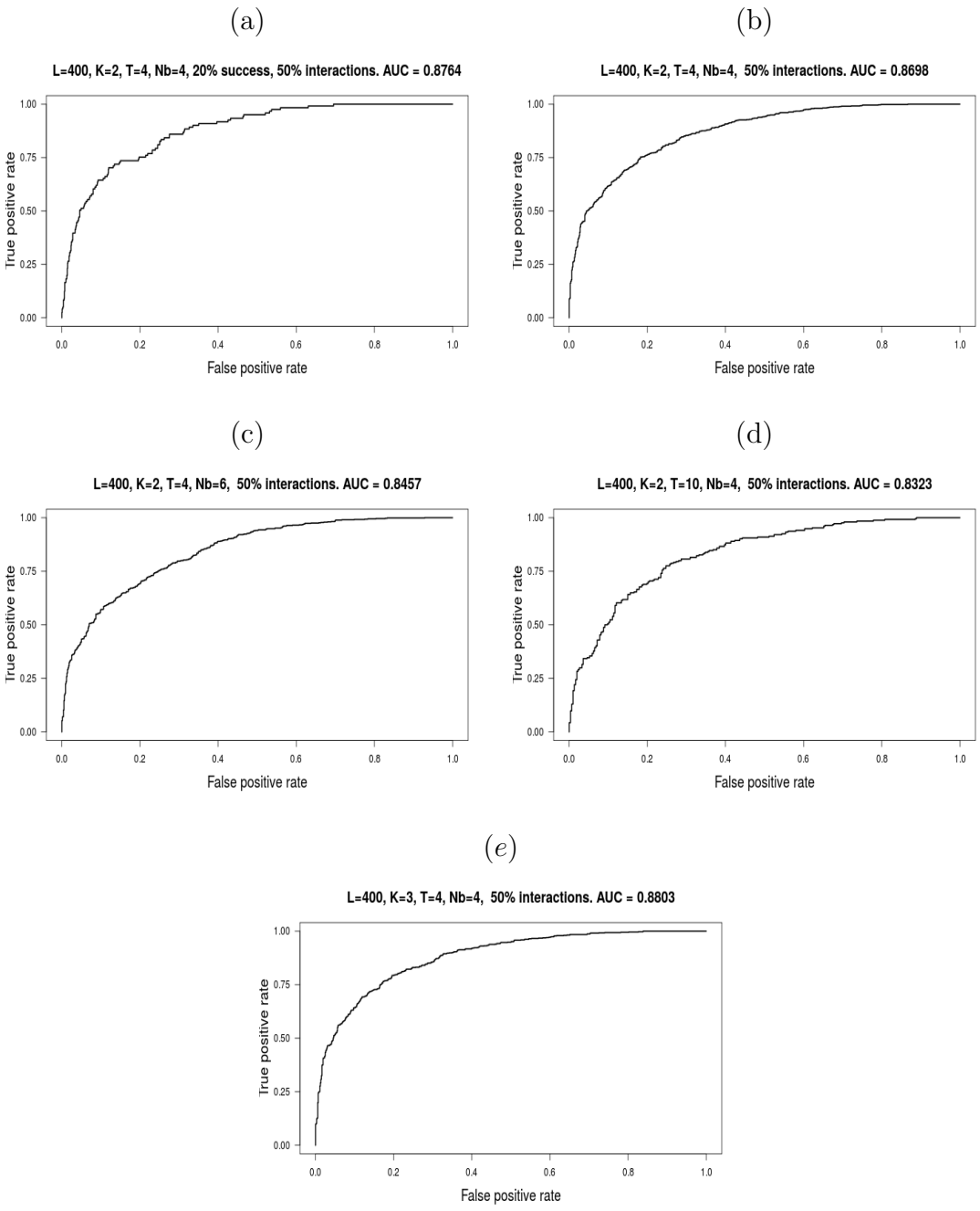


Figure 4.36: Curvas ROC e AUC para os cenários com $L = 400$ e 50% de regiões de G_E afetados por η^*. O Painel (a) apresenta o caso desbalanceado e os demais o balanceado. Nos Paineis (a, b) temos $K = 2$, $T = 4$, e 4 vizinhos por região. Os Paineis (b, c) se diferenciam pelo número de vizinhos por região, 4 e 6, respectivamente; os Paineis (b, d) pelo no número de tempos, 4 e 10; e os Paineis (b, e) pelo número de fatores, 2 e 3.

desbalanceado ($\approx 20\%$ de $Y_i^{'}s = 1$). Veja que todas as curvas são bem parecidas e, em todos os paineis, ela se distanciou da diagonal em direção ao canto superior esquerdo do gráfico, que é o ponto ótimo, ou seja, ponto no qual se tem a maior taxa de verdadeiros positivos e a menor de falsos positivos.

Uma curva ROC é uma representação bidimensional do desempenho de um modelo de classificação binária. Para comparar modelos logísticos podemos reduzir a representação para um único valor escalar. Um método comum é calcular a área sob a curva ROC, ou seja, AUC (*Area Under the Curve*). Como a AUC (Bradley, 1997; Hanley e McNeil, 1982) é uma parte da área do quadrado unitário, seu valor sempre estará entre 0 e 1. Podemos ver pela Figura 4.36 que todas as AUC's foram superiores a 0.83 indicando que as predições dos modelos foram satisfatórias.

Conforme informado reiteradamente, todas as análises apresentadas, até então, foram baseadas em uma única amostra *a posteriori*. Neste capítulo várias análises foram conduzidas, a saber: comparamos cenários balanceados versus desbalanceados e casos com diferentes números de locais; mostramos o comportamento do modelo quando temos mais fatores e mais tempos; avaliamos como o modelo reaje ao se errar na especificação do número de fatores, 1 unidade acima ou abaixo; apresentamos casos de sobreparametrização do modelo fatorial (mais ou o mesmo número de fatores do que tempos); checamos o ajuste dos modelos ao averiguar os resíduos de Pearson e ao avaliar as curvas ROC e as AUC's. Finalizamos este capítulo mostrando, na próxima seção, diversas análises comparativas de vários cenários, mas considerando estatísticas calculadas após a execução de 30 réplicas de Monte Carlo.

4.9 Análise Monte Carlo

Esta seção é dedicada à análise de estimativas calculadas a partir 30 réplicas Monte Carlo. Todas as análises, aqui descritas, são relativas ao vício relativo dos termos α, λ e δ por serem os elementos que contemplam mais parâmetros a serem estimados. O objetivo desta análise é estudar o comportamento do modelo (logístico, neste caso) em situações onde temos diferentes tamanhos de amostras dado que alguns elementos

são fixos. Especificamente, comparamos cenários com diferentes números de locais ($L = 100, 200$ e 400), com 30% e 50% de regiões afetadas por interação para os casos balanceados ($\approx 50\%$ de $Y_i^{'}s = 1$) e desbalanceados ($\approx 20\%$ de $Y_i^{'}s = 1$). Os elementos mantidos fixos são: $K = 2$ fatores, $T = 4$ tempos e 4 vizinhos por região. Não analisamos o caso com 6 vizinhos por região, conforme já dissemos anteriormente, por serem muito semelhantes ao de 4 vizinhos; os resultados podem ser vistos no Apêndice C.

O vício relativo foi calculado a partir da seguinte formulação $\frac{(\hat{\zeta}-\zeta)}{|\zeta|}$, em que $\hat{\zeta}$ representa, genericamente, o valor estimado e ζ, o verdadeiro. O termo $|\zeta|$ simboliza o valor verdadeiro absoluto. A divisão pelo módulo ($|\zeta|$) é para evitar sinais negativos no denominador e permitir averiguar se ocorreu subestimação ou sobrestimação. Para α e δ, em cada réplica, foi selecionada aleatoriamente e, sem reposição, uma amostra de tamanho 100, sendo que, para α consideramos apenas estimativas relacionadas a locais de G_E. O conjunto total de observações de α, que foi considerado na seleção da amostra de tamanho 100 em cada réplica, envolveu todos os fatores ($K = 2$), e de δ, todos os tempos ($T = 4$). Ou seja, a seleção em cada réplica de Monte Carlo pode conter $\alpha^{'}s$ dos 2 fatores e $\delta^{'}s$ dos 4 tempos.

A Figura 4.37 mostra o resultado do cálculo do vício relativo para α, Paineis (a) e (b); λ, Paineis (c) e (d), e δ, Paineis (e) e (f). Os paineis da esquerda apresentam uma visão da dispersão e variabilidade das estimativas. Os paineis da direita mostram com maior precisão o valor da mediana nos diversos cenários. Veja que, nos Paineis (a, c, e), não foi apresentado o *boxplot* para o cenário $M_{L_{400}}^{I_{30\%}20\%_{y=1}}$. Por simplicidade, não consideramos necessário analisar esse caso, porque os cenários balanceados foram suficientes para avaliarmos como quantidades diferentes de locais afetados por interação influencia na estimação de α, λ e δ. Analisando o Painel (c), vício relativo de λ, vemos que o maior número de observações e de locais afetados pela interação não linear reduziu a variabilidade das estimativas refletindo em menor vício relativo. Perceba como o intervalo interquartil do cenário $M_{L_{400}}^{I_{50\%}}$ é menor do que os demais. Lembrando ao leitor de que $\delta = \alpha\lambda + \eta + \epsilon$, esse fato mostra que a matriz de interação não linear, η, parece contribuir para uma melhor estimação de λ. Para α, Painel (a), observe que a redução do intervalo interquartil ocorre quando temos 50% de locais afetados por interação (*boxplot* sombreado), com exceção do caso $L = 100$. Essa exceção pode estar relacionada à

pequena diferença da quantidade de locais de G_E entre os cenários com 30% e 50% de regiões afetadas por interação quando se tem $L = 100$ no ajuste do modelo.

Ainda na Figura 4.37, o cenário desbalanceado ($M_{L_{400}}^{I_{50\%}20\%y=1}$), ilustrado nos Paineis (a), (c) e (e), obteve o intervalo interquartil maior do que os demais (balanceados) em α e λ. Mas essa diferença não se repetiu em δ, indicando que apesar de se ter mais dificuldade na estimação de α e λ quando se tem um desbalanceamento na quantidade de $1's$ e $0's$ na variável resposta, o termo $\alpha\lambda + \eta$ estabelece uma compensação que reflete em *boxplot* de vício relativo parecido com os demais no Painel (e).

Os Paineis (b), (d) e (f) da Figura 4.37 ilustram com mais precisão o valor da mediana que foi calculada para α e δ considerando todas as amostras de tamanho 100 e para λ, a amostra completa *a posteriori* de todas as réplicas. Perceba como o vício relativo para os cenários com mais locais afetados pela interação não linear ($\approx 50\%$) é mais próximo de zero na maioria das variações de quantidade de locais existente nas bases de dados. Esse é um fato esperado, pois temos mais locais para contribuirem na estimação de η^* que afeta, principalmente, a estimação de δ e, consequentemente, a estimação dos coeficientes da regressão (β).

A Figura 4.38 ilustra o vício relativo de λ para cada fator em cada tempo. Os Paineis (a), (b), (c) e (d) se referem ao Fator 1 para os tempos 1, 2, 3 e 4, respectivamente. Similarmente, os Paineis (e), (f), (g) e (h) são relativos ao Fator 2. Considerando o Fator 1 podemos ver que ter mais locais contribuindo para a estimação de η^* gera uma melhora (maior proximidade do zero) no vício relativo de λ para $L = 200$ e $L = 400$ (Paineis b, c, d), mas, como analisado anteriormente, quando temos poucos locais ($L = 100$), não se observa grande diferença no resultado do ajuste. Avaliando o Fator 2 não podemos identificar com tanta clareza a tendência de melhoria nos vícios relativos de λ quando se tem mais locais contribuindo para a estimação de η^*, pois há uma melhora no ajuste nos Paineis (f), (g) e (h) quando $L = 200$, mas para $L = 400$ o ganho é maior nos Paineis (e) e (f), e nos Paineis (g) e (h) eles são similares.

No entanto, direcionando o leitor para a Figura 4.39, pode-se identificar que os vícios para δ são, consistentemente, melhores quando se tem mais locais afetados por η^*. Esses resultados reforçam a influência da interação não linear para a boa estimação de δ; η^*

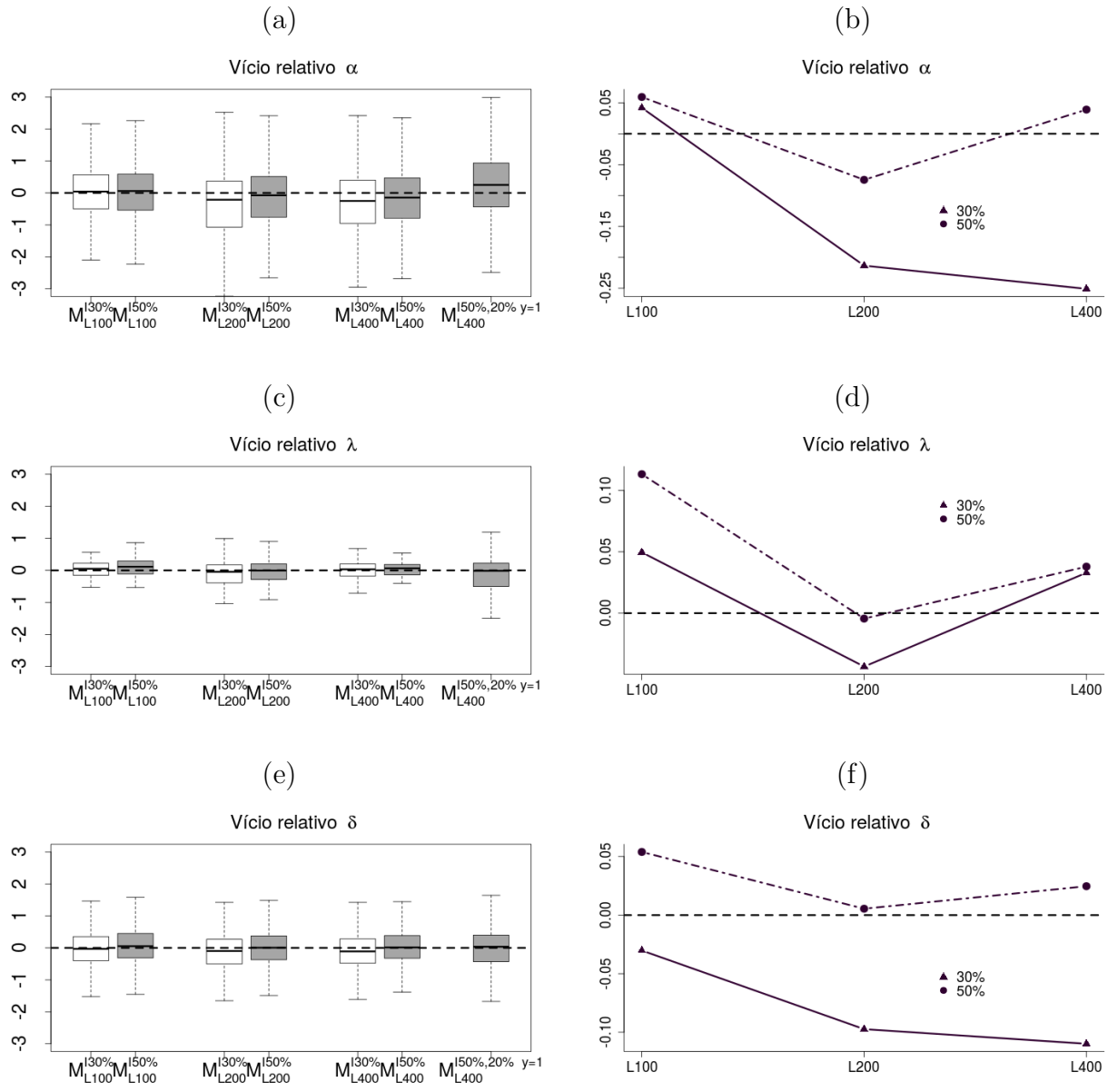


Figure 4.37: Mediana do vício relativo de α, λ e δ calculada a partir de amostras de tamanho 100 (α e δ) de cada uma das 30 réplicas de Monte Carlo. A expressão do vício é dada por $\frac{(\hat{\zeta}-\zeta)}{|\zeta|}$, em que $\hat{\zeta}$ representa, genericamente, o valor estimado e ζ, o verdadeiro, e $|\zeta|$ simboliza o valor verdadeiro absoluto. Considere os cenários com $K = 2$ fatores, $T = 4$ tempos, 4 vizinhos por região, número de locais $L \in \{100, 200, 400\}$, 30% e 50% de locais de G_E afetados pela interação não linear.

tende a ser melhor estimado quando 50% de locais de G_E está sob o efeito da interação. Percebe-se uma tendência decrescente do vício para o valor zero a medida que aumenta o número de locais. Mais uma vez, verificamos que para $L = 100$ não se identifica grande diferença do número de locais afetados pela interação na influência do ajuste de δ.

Finalizamos, aqui, a seção referente à análise do vício relativo calculado com base em amostras de 30 réplicas de Monte Carlo. Concluimos que o maior número de locais afetados pela interação não linear contribuem para se obter um melhor ajuste de α, λ e, principalmente, δ. Esse fato influencia, diretamente, na boa estimação dos coeficientes da regressão e, consequentemente, de θ_i. Encerra-se neste ponto, o capítulo referente ao estudo simulado logístico. No próximo capítulo faremos o estudo simulado Poisson seguindo a mesma estrutura utilizada para o modelo logístico.

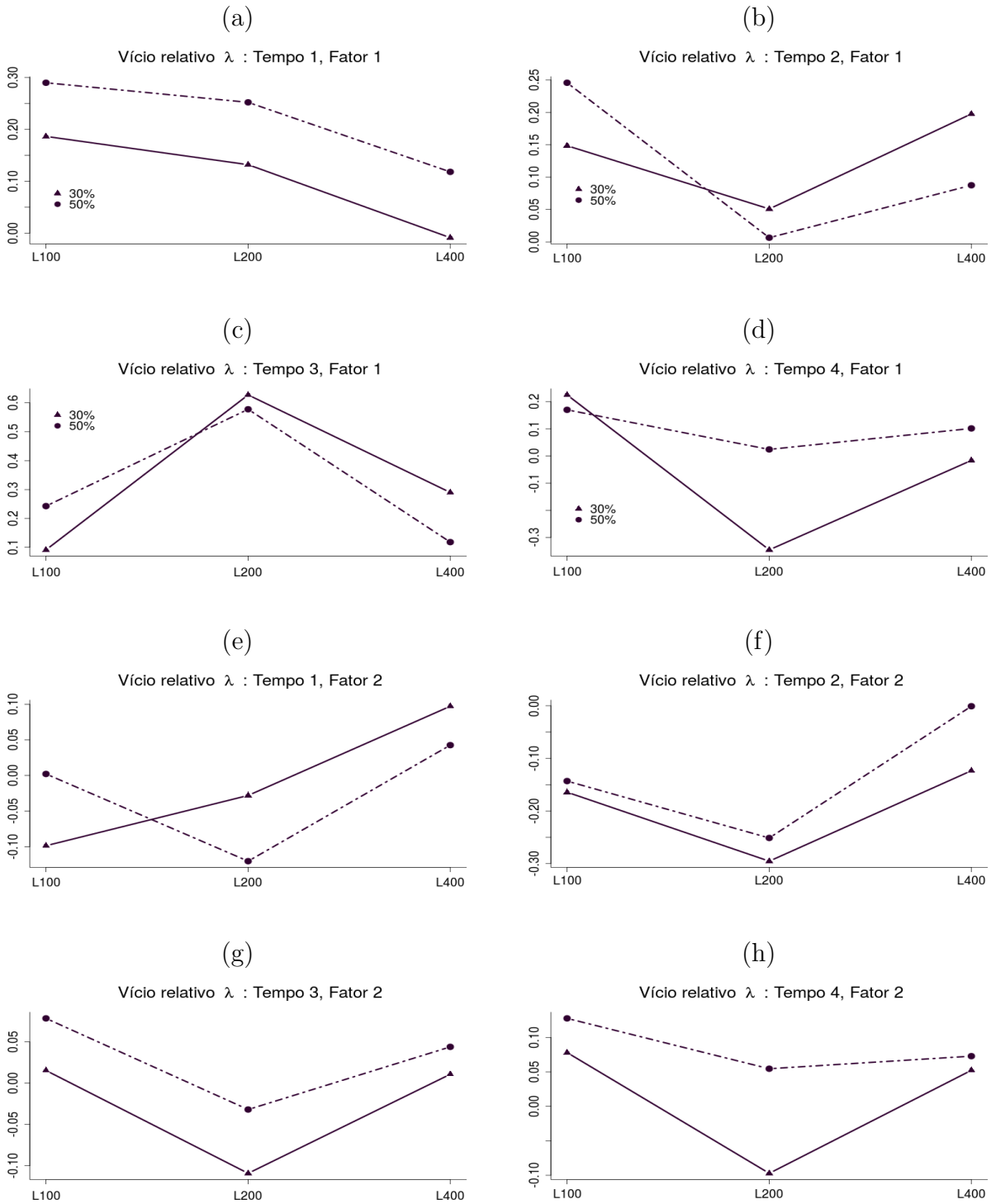


Figure 4.38: Mediana do vício relativo de λ $\left(\frac{(\hat{\lambda}-\lambda)}{|\lambda|}\right)$ calculada a partir de amostras de tamanho 100 de cada uma das 30 réplicas de Monte Carlo. Considere os cenários com $K = 2$ fatores, $T = 4$ tempos, 4 vizinhos por região, número de locais $L \in \{100, 200, 400\}$, 30% e 50% de locais de G_E afetados por η^*.

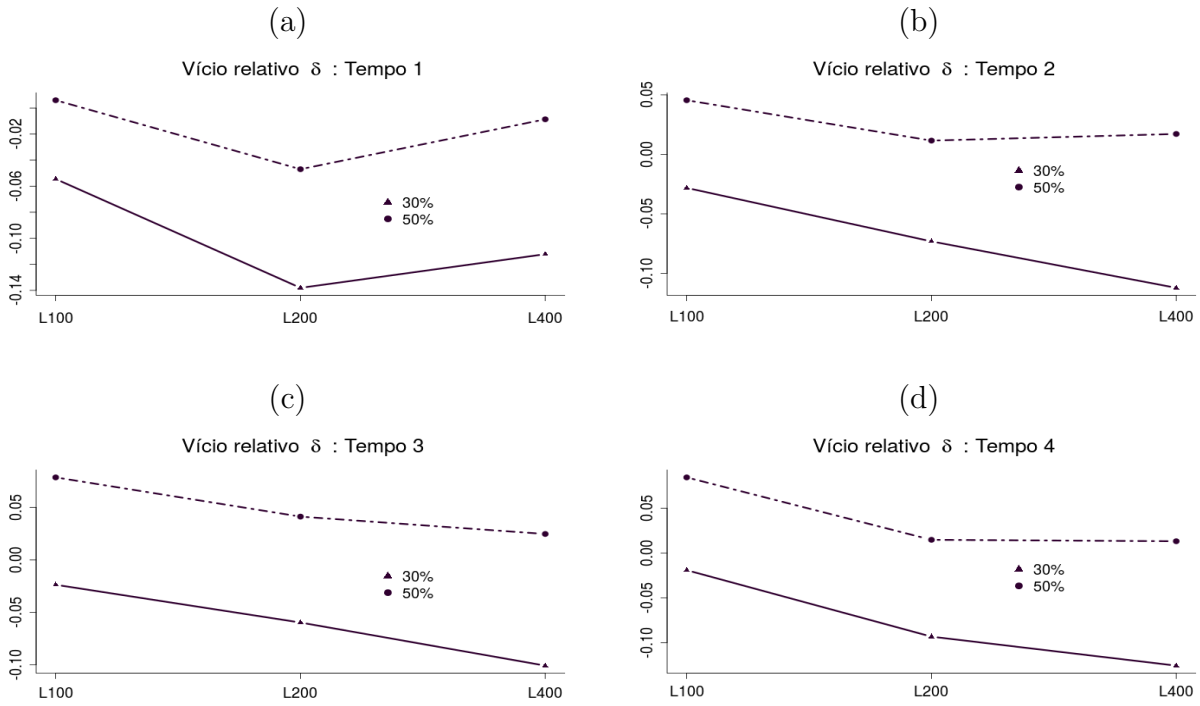


Figure 4.39: Mediana do vício relativo de δ para cada tempo calculado a partir de amostras de tamanho 100 de cada uma das 30 réplicas de Monte Carlo. O vício é calculado pela expressão $\frac{(\hat{\delta}-\delta)}{|\delta|}$. Considere os cenários com $K = 2$ fatores, $T = 4$ tempos, 4 vizinhos por região, número de locais $L \in \{100, 200, 400\}$, 30% e 50% de locais de G_E afetados por η^*.

Capítulo 5

Estudo simulado Poisson

O modelo Poisson, definido na Seção 3.2, foi especificado para analisar um vetor Y de tamanho n, preenchido com o número de ocorrências de um determinado evento de interesse. Da mesma forma que no modelo logístico, temos aqui uma matriz X com dados de q variáveis preditoras para as n amostras. A aplicação de interesse, também, considera que as amostras foram observadas durante T tempos em L localidades e que existem K grupos (fatores) de regiões com características semelhantes. Assim como no modelo logístico, os locais foram organizados em grupos e especificou-se distribuições *a priori* informativas para resolver o problema de identificabilidade do modelo fatorial (veja detalhes no Capítulo 3). Os cenários considerados neste contexto de contagens foram os mesmos avaliados para o caso logístico (veja Tabela 5.1), com exceção de $M^{K_2 I_{50\%}}_{L_{400} T_{10} V_6}$ e $M^{K_3 I_{50\%}}_{L_{400} T_4 V_6}$, pois, como apresentado no caso logístico, as situações com 6 vizinhos por região não diferem muito dos casos com 4 vizinhos. Outra peculiaridade deste estudo é a definição que estipulamos para os casos balanceado e desbalanceado. Lembramos ao leitor que, na modelagem logística, essa definição está relacionada à quantidade de 1's e $0's$ na variável resposta, mas no caso Poisson isso tem relação com o número de contagens distantes ou próximas de 0. Ou seja, no modelo Poisson, quando nos referimos ao caso balanceado estamos tratando da situação na qual existem poucas contagens próximas de 0 ($\approx 3\%$). O caso desbalanceado é caracterizado pelo existência de muitas contagens próximas de 0 ($\approx 40\%$).

Igualmente ao caso logístico, a partir do número de locais e de vizinhos por região é

gerada a matriz de vizinhança W_α e a matriz diagonal D_α. Levando em conta que T é a quantidade de tempos, gera-se a matriz de vizinhança W_λ que, juntamente com D_λ, definem a estrutura temporal. Conforme descrito acima, as especificações do modelo Poisson são muito semelhantes ao caso logístico. Convidamos o leitor a rever essas definições, descritas em detalhes, no Capítulo 4.

Modelo	Locais	Tempos	Fatores	Vizinhos	% Interação	Contagens 0
$M^{K_2I_{30\%}}_{L_{100}T_4V_v}$	100					
$M^{K_2I_{30\%}}_{L_{200}T_4V_v}$	200	4	2	$v \in \{4, 6\}$	$\approx 30\%$	$\approx 40\%$
$M^{K_2I_{30\%}}_{L_{400}T_4V_v}$	400					
$M^{K_2I_{50\%}}_{L_{100}T_4V_v}$	100					
$M^{K_2I_{50\%}}_{L_{200}T_4V_v}$	200	4	2	$v \in \{4, 6\}$	$\approx 50\%$	$\approx 40\%$
$M^{K_2I_{50\%}}_{L_{400}T_4V_v}$	400					
$M^{K_3I_{50\%}}_{L_{400}T_4V_4}$		4	3			
$M^{K_4I_{50\%}}_{L_{400}T_4V_4}$		4	4			
$M^{K_5I_{50\%}}_{L_{400}T_4V_4}$	400	4	5	4	$\approx 50\%$	$\approx 40\%$
$M^{K_2I_{50\%}}_{L_{400}T_{10}V_4}$		10	2			
$M^{K_3I_{50\%}}_{L_{400}T_{10}V_4}$		10	3			
$M^{K_2I_{50\%}}_{L400T_4V_4}$	400	4	2	4	$\approx 50\%$	$\approx 3\%$

Tabela 5.1: Configurações para geração de dados de contagem (modelo Poisson) variando o número de locais, número de tempos, número de fatores, número de vizinhos por local, percentual de locais de G_E que possuem interação não linear e o número de contagens do evento de interesse. Muitas contagens próximas de 0 ($\approx 40\%$) representa aqui o caso desbalanceado Poisson e poucas contagens próximas de 0 ($\approx 3\%$) representa o caso balanceado.

Todas as especificações definidas na geração de dados para o modelo logístico, descritas na introdução do Capítulo 4 (veja Tabelas 4.2, 4.3, 4.4, 4.5, 4.6 e 4.7), foram as mesmas utilizadas para os dados artificiais do modelo Poisson, com exceção da configuração β_{C_2} (Tabela 4.2) para o qual foi assumido os valores $(5.0, -1.0, 1.0)$. O motivo desta diferenciação em relação à β está relacionado ao controle para geração de muitas ou

poucas contagens próximas de 0. Mais detalhes serão apresentados na Seção 5.1.

Para obter as contagens em Y_i, a diferença em relação ao modelo logístico está no cálculo de θ_i e na distribuição utilizada para geração das observações. Assuma $\theta_i = \exp\{X_{i\bullet}\beta + \delta_{l_i^* t_i^*}\}$, em que $i \in 1, ..., n$, e $Y_i = \text{Poisson}(\theta_i)$. Perceba que θ_i mudou de interpretação no contexto Poisson. No caso logístico ele representa a probabilidade de sucesso e aqui, θ_i é a taxa média de eventos a serem observados para o indivíduo i.

Conforme apresentado na Tabela 5.1, vários cenários foram avaliados para o modelo Poisson. Na primeira seção, apresentamos as análises considerando $\approx 3\%$ de contagens zero (caso balanceado) e decidimos avaliar, por simplicidade e semelhança dos resultados, apenas o cenário $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$. Na seção seguinte tratamos do caso $\approx 40\%$ de contagens zero (caso desbalanceado) para os cenários $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ e $M^{K_2 I_{30\%}}_{L_{400} T_4 V_4}$. Em seguida, apresentamos a comparação geral das estimativas obtidas para η^* e λ admitindo as configurações de números de locais 100, 200 e 400 dos modelos com $K = 2$, $T = 4$ e $V = 4$. Para esses cenários variamos o percentual de regiões de G_E afetadas por interação ($\approx 30\%$ e $\approx 50\%$). De forma equivalente ao modelo logístico, ilustramos os resultados quando ocorre erro na escolha do número de fatores a serem ajustados. Na sequência, também espelhando no caso logístico, apresentamos as inferências quando trabalhamos com outras configurações de número de fatores e de tempos ($K = 3$ fatores e $T = 10$ tempos). Outra análise importante, desenvolvida e apresentada aqui, é quando ocorre sobreparametrização do modelo ao assumir uma quantidade de fatores igual ou acima do número de tempos. Ao final do capítulo analisamos os resíduos de Pearson e comparamos o vício relativo dos cenários após o ajuste para 30 réplicas de Monte Carlo. Os resultados detalhados para os demais cenários listados na Tabela 5.1 e 4 vizinhos por região são mostrados no Apêndices D.

5.1 Ajustes para poucas contagens zero

Conforme destacado no último parágrafo da Seção 3.2, o chute inicial da cadeia MCMC de β_0, $\beta_0^{(0)}$, teve que ser revisto para se obter um melhor ajuste do modelo. Nesta seção apresentamos as análises completas referentes às duas configurações iniciais da

cadeia utilizadas para o vetor β. A primeira com $\beta^{(0)} = \mathbf{0}$, valor utilizado na maioria dos ajustes dos modelos logísticos e Poisson, e a segunda com $\beta^{(0)} = (5, 0, 0)^\top$, cujo motivo está descrito adiante. Optamos por não apresentar o estudo detalhado para a configuração $\beta^{(0)} = (10, 0, 0)^\top$, mas ao final desta seção, traçamos um estudo comparativo para 3 cenários de valores iniciais da cadeia de β_0 (0, 5 e 10), mantendo $\beta_1^{(0)} = 0$ e $\beta_2^{(0)} = 0$. O leitor deve lembrar que esta análise considera apenas o cenário $M_{L400T_4V_4}^{K_2I_{50\%}}$.

Ajuste com valor inicial da cadeia de β_0 menor que o valor verdadeiro

O chute inicial da cadeia $\beta = \mathbf{0}$ foi utilizado, inicialmente, nos ajustes de todos os cenários logístico e Poisson. A necessidade de alterar esse valor para se obter estimativas melhores surge quando a variável resposta possui $\approx 3\%$ de contagens zero. Nesta seção, apresentamos as análises quando o chute inicial da cadeia de β_0 é menor do que o valor verdadeiro (β_0^V), especificamente quando $\beta_0^{(0)} = 0$, em que explicitamos como esse valor influencia na estimação de α, λ e δ.

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	5.00	0.20	0.20	0.00	0.20	0.21
β_1	-1.00	-1.00	-1.00	0.00	-1.00	-1.00
β_2	1.00	1.00	1.00	0.00	1.00	1.00
σ^2	0.80	1.13	1.13	0.07	1.00	1.27
τ_α	2.00	2.35	2.34	0.75	0.99	3.76
η_1^*	-2.00	-0.75	-0.74	0.30	-1.34	-0.18
η_2^*	1.50	2.64	2.63	0.26	2.13	3.14
η_3^*	0.75	2.17	2.17	0.25	1.69	2.66
η_4^*	-1.00	0.61	0.61	0.28	0.07	1.17

Tabela 5.2: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o caso Poisson: $M_{L_{400}T_4V_4}^{K_2I_{50\%}}$ com $\approx 3\%$ de contagens zero em que o valor verdadeiro de $\beta_0 = 5$ e o chute inicial da cadeia de $\beta_0 = 0$, ou seja, $\beta_0^V = 5$ e $\beta_0^{(0)} = 0$.

A Tabela 5.2 apresenta as estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α e da interação não linear η^*. Veja como a diferença entre o valor verdadeiro e a média/mediana *a posteriori* de β_0 tem magnitude praticamente igual ao valor verdadeiro. Verifique, também, que o valor verdadeiro de η^* ficou muito fora do intervalo HPD de 95% para todos os tempos, o mesmo acontecendo para σ^2. Apenas β_1, β_2 e τ_α foram bem estimados, com β_1 e β_2 sendo iguais ao valor verdadeiro. Apesar de η^* ter estimativas distantes do valor verdadeiro, pela Figura 5.1 pode-se ver que o padrão de crescimento e decrescimento foi bem capturado.

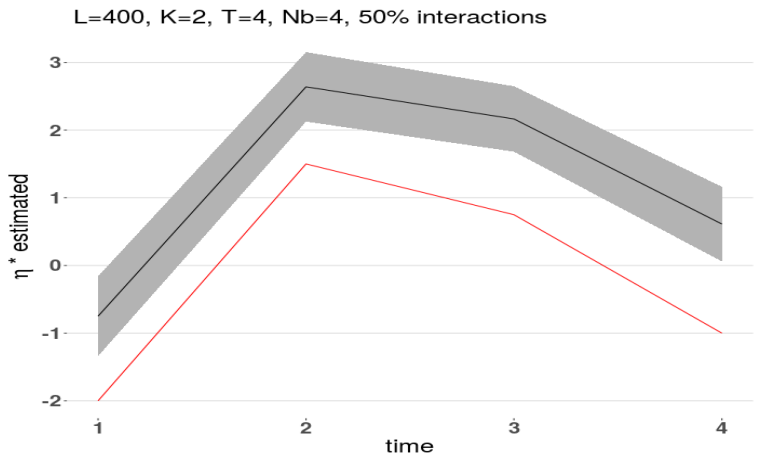


Figure 5.1: Gráfico da média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha) para o caso Poisson $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$, com poucas contagens zero em que o valor verdadeiro de $\beta_0 = 5$ e o chute inicial da cadeia de $\beta_0 = 0$, ou seja, $\beta_0^V = 5$ e $\beta_0^{(0)} = 0$.

Na Figura 5.2 podemos identificar que houve sobrestimação para α, β e δ, pois em todos os Paineis (b, d, f) referentes às suas estimativas, respectivamente, temos apenas tons vermelhos. Ou seja, valores verdadeiros positivos (tons vermelhos), Painel (a), tiveram estimativas maiores (tons de vermelho mais vivos), Painel (b); e para valores verdadeiros negativos (tons azuis), as estimativas ficaram positivas (tons vermelhos). Visualmente, a menos das diferenças em tonalidades, parece que o padrão geral foi

capturado. Isso é mais facilmente observado nos Paineis (c) e (d) referentes à λ. Mas é na Figura 5.3 que fica explícito essa sobrestimação de α, λ e δ. Observe como a diferença entre os valores verdadeiros e estimados é marcante e persiste ao longo do eixo horizontal, especialmente para δ, Painel (c), mas também é bem nítido para λ, Painel (b). Perceba que para os 3 casos, o padrão global é bem capturado pelas estimativas. Avalie pelo Painel (d) que, apesar das diferenças na estimação desses parâmetros, as probabilidades estimadas dos locais serem afetados, ou não, por interação foram muito bem capturadas, com os locais gerados com interação obtendo estimativas bem próximas de 1 e os sem interação, com probabilidades muito próximas de 0.

Finalmente a Figura 5.4, que imita a estrutura espacial dos dados artificiais quando existem 4 vizinhos por local, ilustra como os locais foram afetados por efeitos principais e/ou interação. Diferentemente dos efeitos de todos os cenários analisados até então, os locais foram afetados, no que se refere aos efeitos principais analisados individualmente (Paineis c, d, e, f), apenas por cargas positivas (cor vermelha). Também observamos mais locais afetados pelos 2 fatores concomitantemente, isto é, Fator 1 e Fator 2, Paineis (a) e (b). Enfatizamos que nesta avaliação não detectou-se locais sem qualquer tipo de efeito. Além disso, poucos foram os locais afetados apenas pela interação, veja o Painel (g).

Concluimos a análise para contagens com poucos zeros e com valor do chute inicial da cadeia de β_0 menor do que o valor verdadeiro, especificamente quando $\beta_0^V = 5$ e $\beta_0^{(0)} = 0$, reforçando o fato de que a diferença entre valores verdadeiros e estimados de δ, Painel (c) da Figura 5.3, é praticamente a mesma ao longo do eixo horizontal. Nesse mesmo painel, veja que a linha tracejada vertical corta a curva vermelha (valor verdadeiro) quando $\hat{\delta} \approx 0$ (valor estimado de δ) e a linha tracejada horizontal corta a curva preta (valor estimado) em $\hat{\delta} \approx 5$. Isso indica que a diferença entre os valores estimados (curva preta) e os verdadeiros (curva vermelha) é ≈ 5. Essa conclusão nos leva à análise seguinte, em que avaliamos as estimativas quando o valor do chute inicial do MCMC de β_0 é igual ao valor verdadeiro.

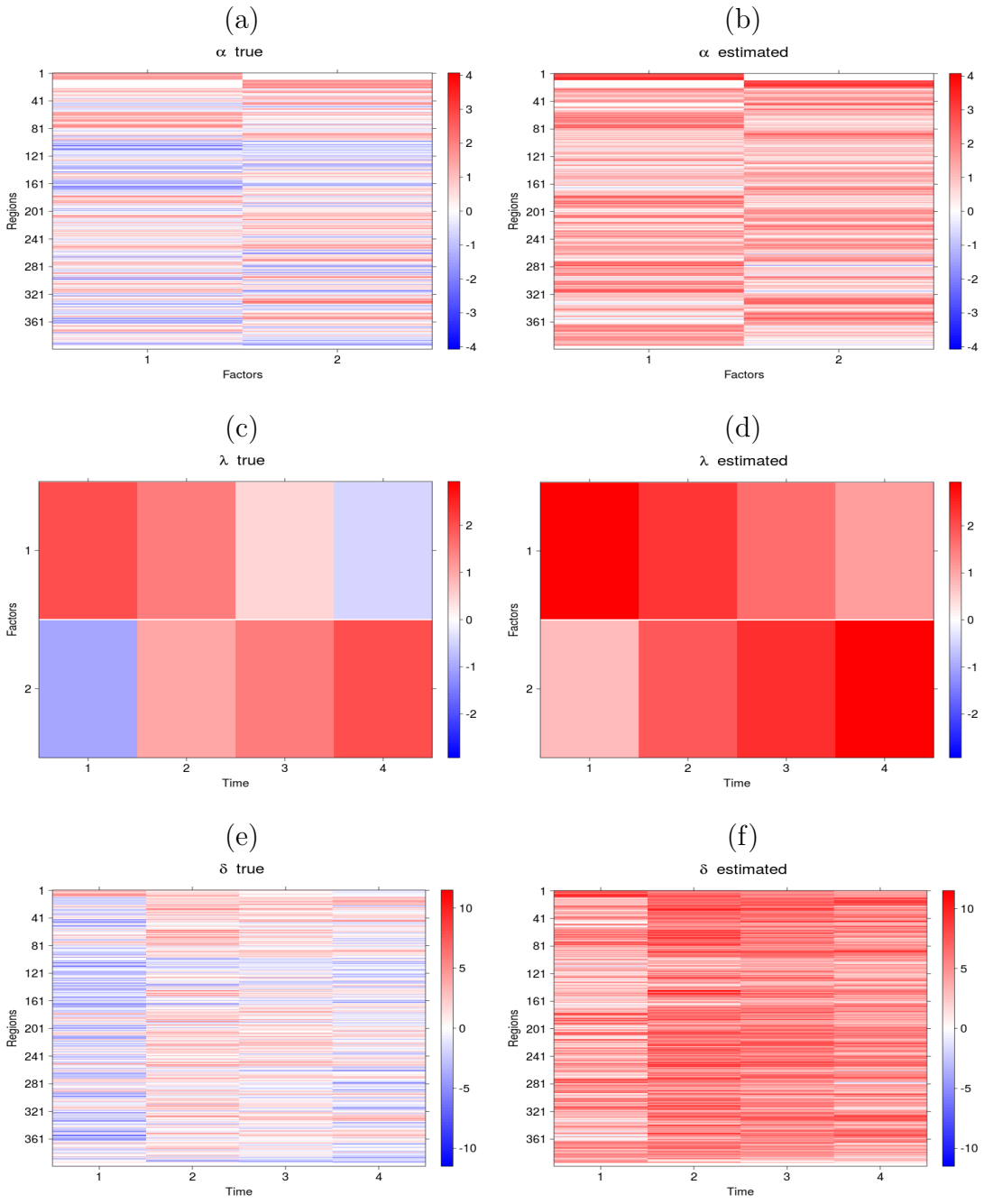


Figure 5.2: Mapas de calor comparando valores verdadeiros e estimados para o caso Poisson $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$, com $\approx 3\%$ de contagens zero em que o valor verdadeiro de $\beta_0 = 5$ e o chute inicial da cadeia de $\beta_0 = 0$, ou seja, $\beta_0^V = 5$ e $\beta_0^{(0)} = 0$. Paineis: (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

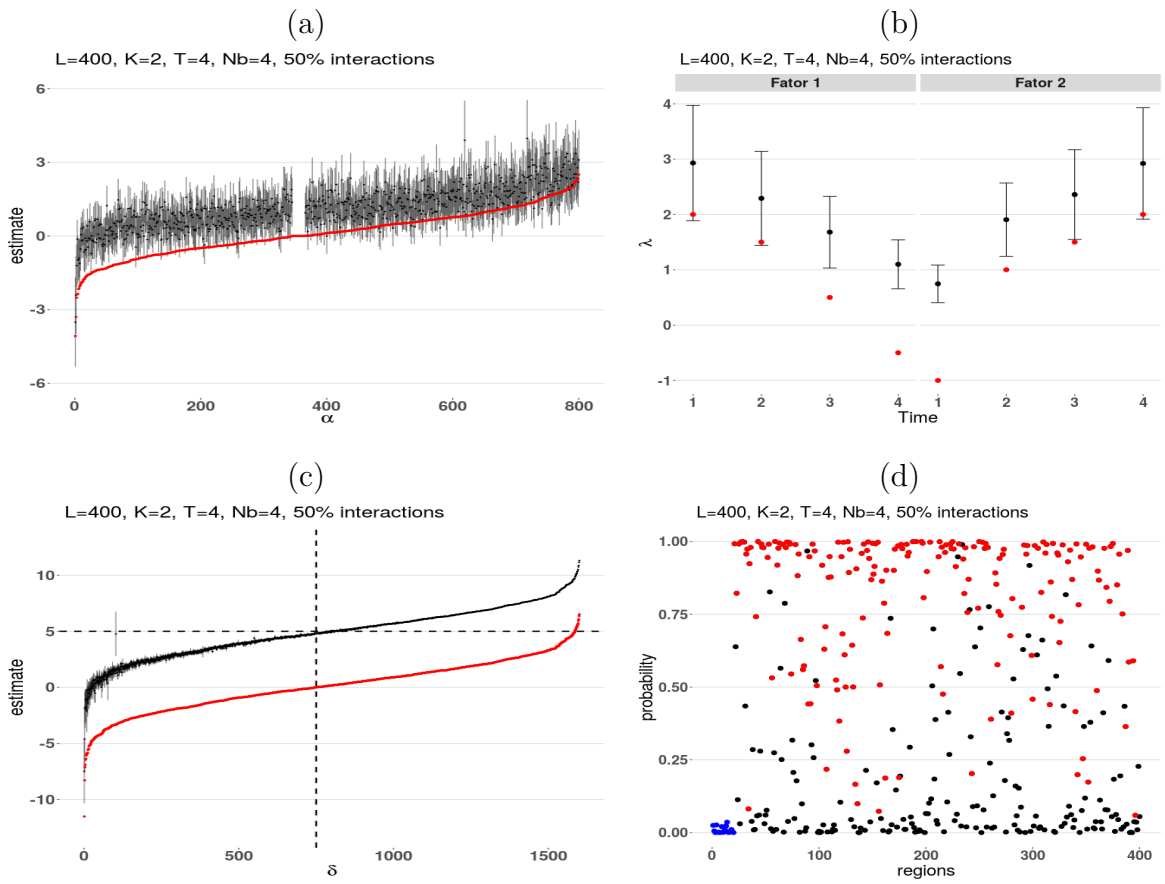


Figure 5.3: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais do grupo G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o caso Poisson: $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$, com $\approx 3\%$ de contagens zero em que o valor verdadeiro de $\beta_0 = 5$ e o chute inicial da cadeia de $\beta_0 = 0$, ou seja, $\beta_0^V = 5$ e $\beta_0^{(0)} = 0$.

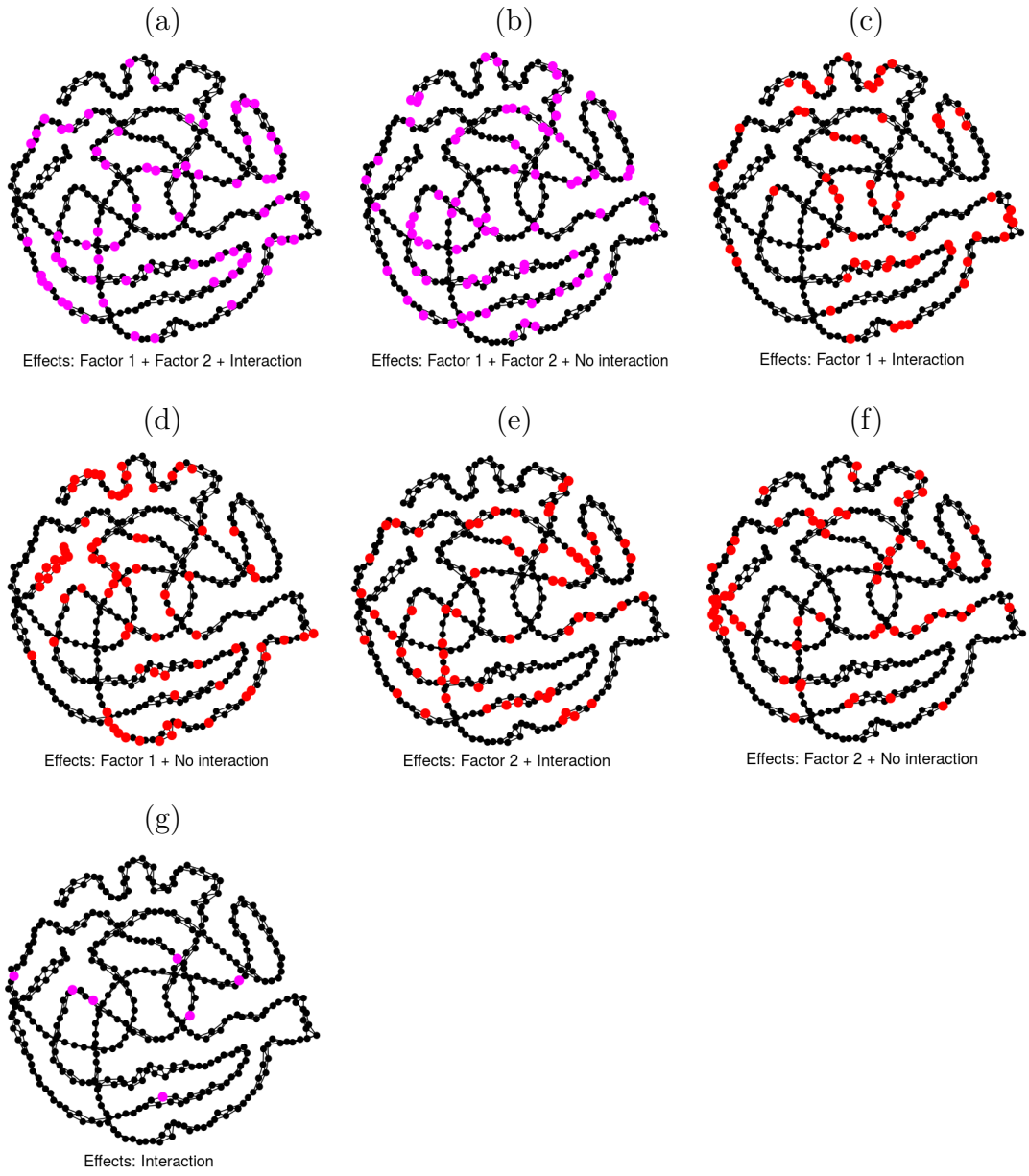


Figure 5.4: Grafos com 4 vizinhos por região imitando a estrutura espacial dos dados artificiais. Cada ponto representa um local. A cor vermelha (carga positiva) identifica os locais associados a algum efeito principal e/ou interação (Paineis c, d, e, f). Nos Paineis (a, b, g), a cor magenta indica os locais afetados por mais de um efeito principal ou somente por interação. Considere o caso Poisson: $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 3\%$ de contagens zero em que o valor verdadeiro de $\beta_0 = 5$ e o chute inicial da cadeia de $\beta_0 = 0$, ou seja, $\beta_0^V = 5$ e $\beta_0^{(0)} = 0$.

Ajuste com valor inicial da cadeia de β_0 igual ao valor verdadeiro

Nesta seção apresentaremos as análises das estimativas *a posteriori* dos parâmetros ao atribuirmos o valor verdadeiro como o chute inicial da cadeia de β_0, o que, neste estudo, equivale a configurar $\beta_0^{(0)} = 5$.

A Tabela 5.3 apresenta as estimativas dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α e da interação não linear η^*. Os desvios padrão para os elementos do vetor β são tão próximos de 0 que foi preciso configurar a exibição dos valores tabelados com 3 casas decimais. Veja como as estimativas para β foram praticamente iguais ao valor verdadeiro. Todos os verdadeiros valores dos parâmetros ficaram dentro do intervalo HPD de 95%. A Figura 5.5 destaca esse fato sobre η^* em que o valor verdadeiro (linha vermelha) está completamente dentro do envelope do HPD e a tendência global totalmente capturada (crescimento seguido por decrescimento).

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	5.000	5.150	5.150	0.002	5.145	5.155
β_1	-1.000	-0.999	-0.999	0.001	-1.000	-0.998
β_2	1.000	1.001	1.001	0.001	1.000	1.002
σ^2	0.800	0.849	0.846	0.048	0.757	0.945
τ_α	2.000	1.741	1.575	0.636	0.823	3.114
η_1^*	-2.000	-1.976	-1.977	0.205	-2.376	-1.580
η_2^*	1.500	1.708	1.712	0.162	1.386	2.016
η_3^*	0.750	0.944	0.944	0.142	0.666	1.224
η_4^*	-1.000	-0.744	-0.749	0.182	-1.103	-0.374

Tabela 5.3: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o caso Poisson: $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$, com $\approx 3\%$ de contagens zero em que o valor verdadeiro e o chute inicial da cadeia de β_0 são os mesmos ($\beta_0^V = \beta_0^{(0)} = 5$).

A Figura 5.6 ilustra os mapas de calor (verdadeiro vs estimado) para α, λ e δ. As

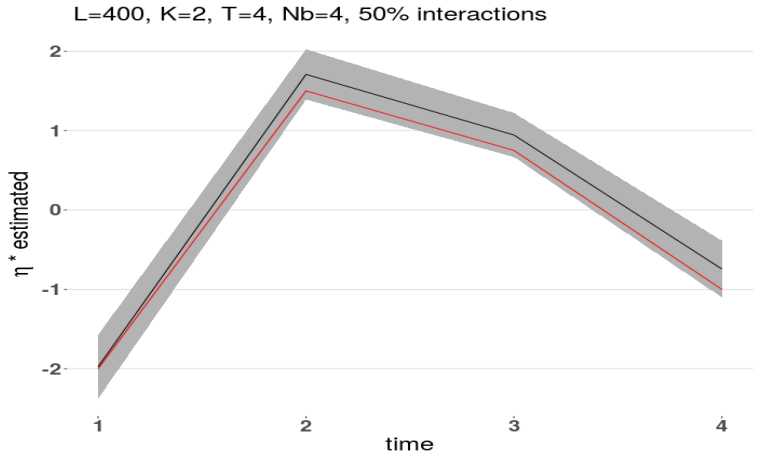


Figure 5.5: Gráfico da média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha) para o caso Poisson $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$, com $\approx 3\%$ de contagens zero em que o valor verdadeiro e o chute inicial da cadeia de β_0 são os mesmos ($\beta_0^V = \beta_0^{(0)} = 5$).

estimativas para todos esses parâmetros citados seguiram, visualmente, muito bem o padrão global. No caso de α temos algumas subestimações para as cargas associadas ao Fator 2 (veja locais próximos dos números 250, 270, 325 e 355). No entando, visualmente é difícil identificar alguma sub ou sobrestimação em λ e δ, Paineis (c) vs (d), e (e) vs (f), respectivamente. Essas informações podem ser comprovadas analisando os gráficos da Figura 5.7. Perceba que para α, Painel (a), subestimações ocorrem para valores acima de 1. Contrariamente, vemos que as estimativas de λ, Painel (b), e de δ, Painel (c), estão muito bem ajustadas. Destaque deve ser dado para o Painel (c), em que toda a subestimação, que ocorre quando iniciamos $\beta_0 = 0$ (Painel (c) da Figura 5.3), foi eliminada. Semelhante ao caso anterior, as probabilidades estimadas de interação para os locais gerados com interação (cor vermelha) ficaram próximas de 1 e as probabilidades daqueles gerados sem interação (cor preta e azul) ficaram próximas de 0, Painel (d).

A Figura 5.8 ilustra grafos que imitam a estrutura espacial dos dados artificiais quando cada região tem 4 vizinhos. Diferentemente do caso quando iniciamos $\beta_0 = 0$,

identificamos locais afetados por apenas 1 efeito principal (Fator 1 ou Fator 2) cujas cargas são positivas (cor vermelha) e negativas (cor azul), Paineis (c, d, e, f).
Outra diferença em relação ao caso anterior é a ocorrência de locais sem qualquer efeito (Painel h). Também temos mais locais afetados pelos 2 efeitos principais (Paineis a, b) e afetados apenas por interação (Painel g).

Concluindo esta seção, relativa à análise de contagens com poucos $0's$ na variável resposta, devido ao fato de termos identificado que o modelo Poisson, com a especificação proposta nesta tese, ser sensível ao valor inicial da cadeia MCMC de β_0, descrevemos, a seguir, um estudo comparativo das estimativas de δ para diferentes valores iniciais de β_0. Esse estudo envolve apenas o parâmetro δ, pois, lembramos ao leitor de que $\delta = \alpha\lambda + \eta + \epsilon$ e, conforme analisado anteriormente, as variações nas estimativas dos parâmetros α e λ são compensadas entre si, gerando estimativas satisfatórias para δ.

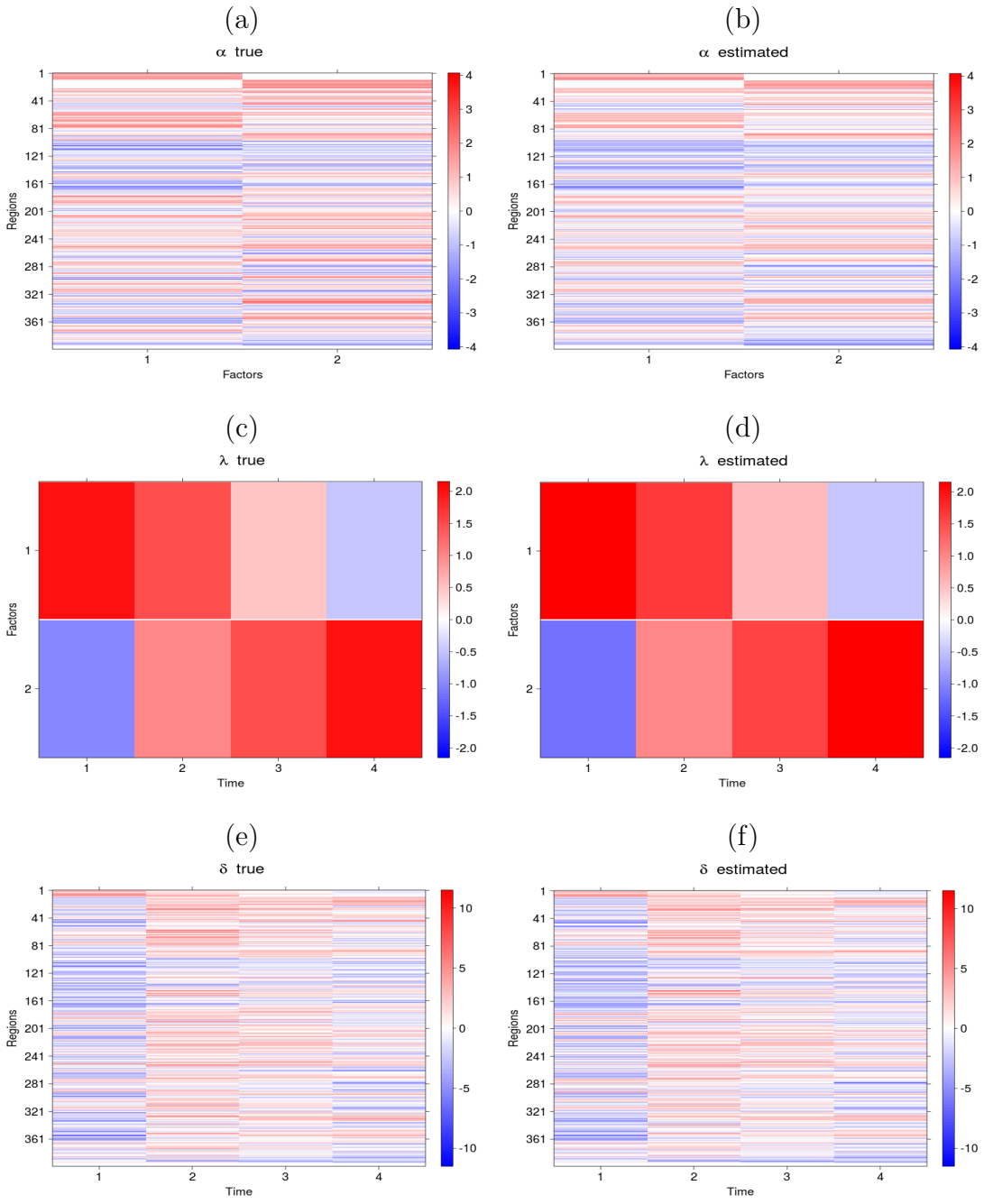


Figure 5.6: Mapas de calor comparando valores verdadeiros e estimados para o caso Poisson $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$, com $\approx 3\%$ de contagens zero em que o valor verdadeiro e o chute inicial da cadeia de β_0 são os mesmos ($\beta_0^V = \beta_0^{(0)} = 5$). Paineis: (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

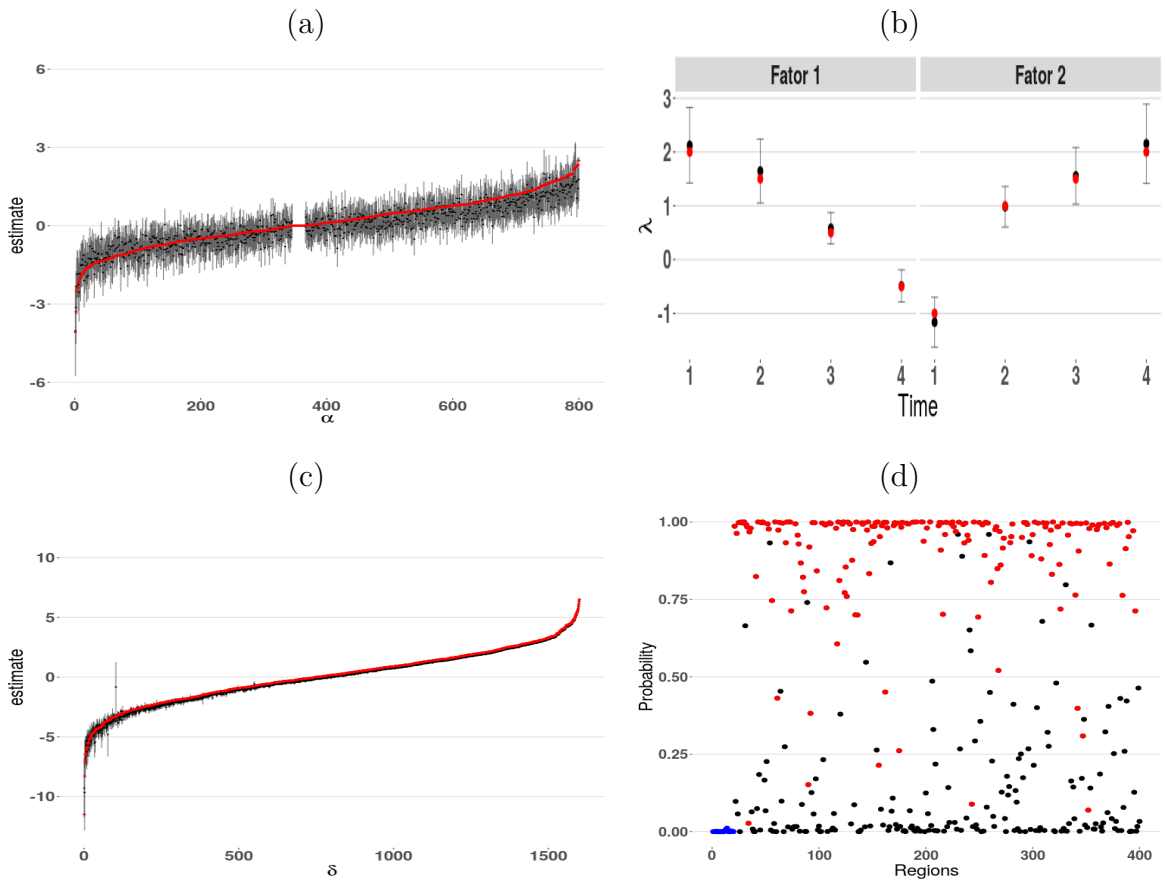


Figure 5.7: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais do grupo G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o caso Poisson: $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$, com $\approx 3\%$ de contagens zero em que o valor verdadeiro e o chute inicial da cadeia de β_0 são os mesmos ($\beta_0^V = \beta_0^{(0)} = 5$).

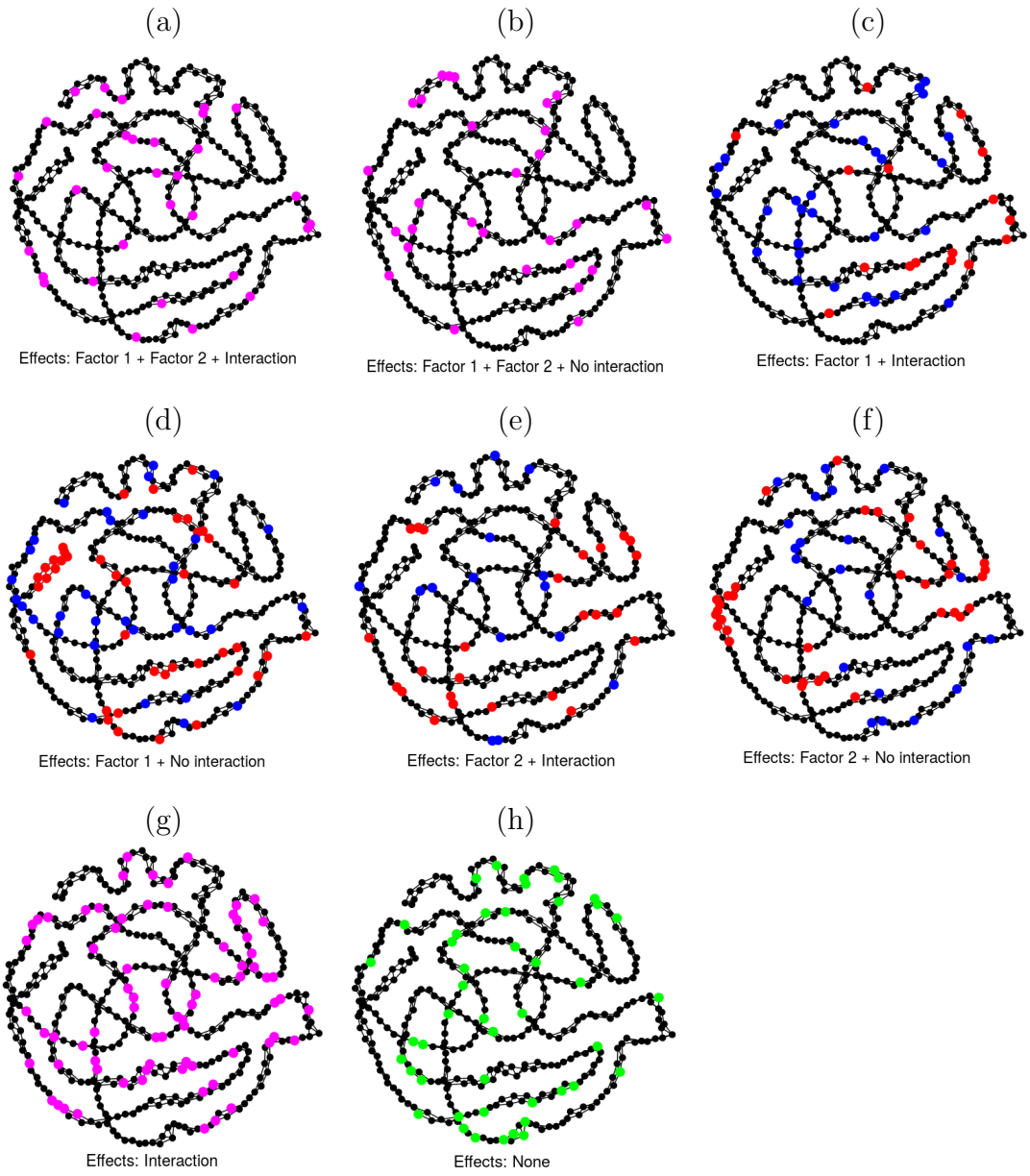


Figure 5.8: Grafos com 4 vizinhos por região imitando a estrutura espacial dos dados artificiais. Cada ponto representa um local. A cor vermelha (carga positiva) ou azul (carga negativa) identifica os locais associados a algum efeito principal e/ou interação (Paineis c, d, e, f). A cor verde (Painel h) denota as regiões não afetadas por qualquer efeito. Nos Paineis (a, b, g), a cor magenta indica os locais afetados por mais de um efeito principal ou somente por interação. Considere o caso Poisson: $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 3\%$ de contagens zero em que o valor verdadeiro e o chute inicial da cadeia de β_0 são os mesmos ($\beta_0^V = \beta_0^{(0)} = 5$).

Análise comparativa de δ para diferentes valores iniciais de β_0

Durante os ajustes do caso com poucas contagens zero ($\approx 3\%$) identificamos que ocorreu um problema de identificabilidade entre o β_0 e o δ. Analisando a Expressão (3.13) em que $\theta_i = \exp\{\beta_0 + X_{2i}\beta_1 + X_{3i}\beta_2 + \delta_{l_i^* t_i^*}\}, \quad i \in \{1, ..., n\}$, identificamos que δ é sobrestimado ou subestimado na magnitude de β_0. O problema de identificabilidade foi resolvido a partir da atribuição de valor adequado para $\beta_0^{(0)}$ no algoritmo MCMC. Mostramos que, para se obter poucas contagens zero, o valor verdadeiro de β_0 pode ser configurado próximo de 5. Neste tópico apresentamos uma análise comparativa para diferentes valores de $\beta_0^{(0)}$ nos casos em que $\beta_0^V = 5$. Analisamos, anteriormente, os casos em que $\beta_0^{(0)}$ é menor que o valor verdadeiro (0 versus 5), e quando $\beta_0^{(0)}$ é igual ao valor verdadeiro (5 versus 5). No presente tópico, vamos comparar esses casos com o cenário em que $\beta_0^{(0)} = 10$, ou seja, um valor inicial da cadeia de β_0 maior que o verdadeiro, e definir uma estratégia para solução do problema de identificabilidade no caso de contagens com poucos zeros.

A Figura 5.9 apresenta as médias *a posteriori* de δ (linha preta) versus o valor verdadeiro (linha vermelha). O Painel (a) ilustra a estimação de δ quando $\beta_0^{(0)} = 0$, o Painel (b) quando $\beta_0^{(0)} = 10$ e o Painel (c) quando $\beta_0^{(0)} = 5$. Veja que a curva do valor verdadeiro (cor vermelha) é cortada, pelo eixo horizontal = 0, aproximadamente ao meio. Perceba, também, que essa divisão é simétrica em valores negativos e positivos para δ, conferindo a δ, em conformidade com as Equações (3.3), (3.4), (3.5) e (3.10), o valor médio esperado igual a zero. Confira como a curva do valor estimado (cor preta) segue o mesmo traçado da curva do valor verdadeiro (cor vermelha) em todos os pontos e como a amplitude dos intervalos HPD são pequenos, inclusive para valores verdadeiros baixos. No entanto, apesar das curvas serem praticamente espelhadas, existe uma sobrestimação (Painel a) ou subestimação (Painel b) de ≈ 5 unidades ao longo de toda a curva. Esse valor pode ser calculado a partir do tamanho do segmento da reta tracejada vertical compreendido entre a curva vermelha e a curva preta. Ou seja, o segmento de reta que liga o ponto de cruzamento da curva vermelha com o eixo horizontal (zero) e o ponto da curva preta que corta a linha tracejada horizontal. Perceba que o tamanho desse

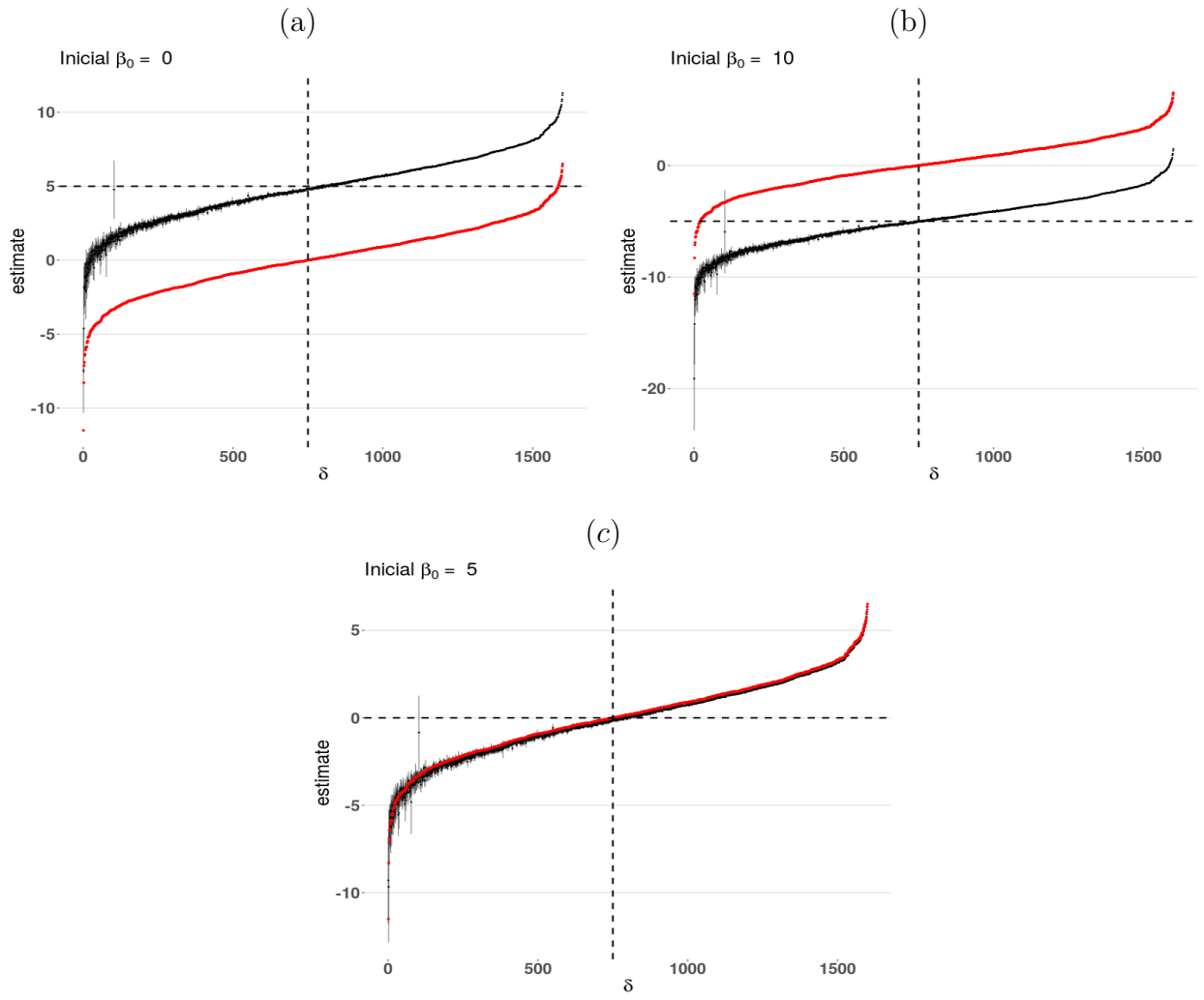


Figure 5.9: Média *a posteriori* (linha preta), intervalos HPD de 95% para η^* e λ (área sombreada) e valor verdadeiro (linha vermelha). Considere o caso Poisson: $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 3\%$ de contagens zero.

segmento (ponto na curva preta subtraído do ponto na curva vermelha) é ≈ 5 quando $\beta_0^{(0)} = 0$, Painel (a); ≈ -5 quando $\beta_0^{(0)} = 10$, Painel (b); e ≈ 0 quando $\beta_0^{(0)} = 5$, Painel (c).

Finalizamos esta seção concluindo que, no caso de contagens com poucos zeros, é preciso verificar se ocorreu o problema de identificabilidade entre β_0 e δ. Caso afirmativo, acertá-lo. O processo consiste em dividir ao meio o eixo horizontal do Painel (c), Figura 5.7, e traçar uma reta vertical neste ponto. Em seguida, localize o ponto de interseção entre essa reta vertical e a curva. Calcule a distância desse ponto até o eixo horizontal $= 0$. Caso essa distância não seja próxima de zero, atribua o seu valor absoluto ao chute inicial da cadeia de β_0 e execute o MCMC novamente. Na próxima seção apresentamos o ajuste do modelo no caso com muitas contagens zero ($\approx 40\%$).

5.2 Ajustes para muitas contagens zero

Nesta seção vamos apresentar as análises do cenário com muitas contagens zero ($\approx 40\%$) para a variável resposta no caso em que temos: $L = 400$ locais, $T = 4$ tempos e $K = 2$ fatores. As configurações dos parâmetros são aquelas descritas nas Tabelas 4.2, 4.3, 4.4 e 4.5. Para o parâmetro β, na Tabela 4.2, as análises se referem à configuração β_{C_1}. Analisaremos as situações em que $\approx 50\%$ e 30% das regiões são afetadas por interação.

Cenário com $\approx 50\%$ de locais em G_E afetados por interação

A Tabela 5.4 apresenta as estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. Podemos ver que apenas os parâmetros σ^2 e η_4^*, verdadeiros, ficaram um pouco fora do intervalo HPD de 95%, ambos sendo sobrestimados. A média e a mediana de todos eles foram muito próximas, indicando que as distribuições desses parâmetros são simétricas. Os parâmetros β_1 e β_2 foram estimados exatamente iguais ao valor verdadeiro e com um desvio padrão muito próximo de 0, assim como β_0 e σ^2. A estimação τ_α também ficou próxima do valor verdadeiro (2.00 versus 2.02). Analisando a Figura 5.10 vemos que, apesar de ter ocorrido uma sobrestimação de η_4^*, a tendência de crescimento e decrescimento de η^*, no tempo, foi bem capturada.

A Figura 5.11 ilustra mapas de calor de α, λ e δ comparando valores verdadeiros e estimados. Em todos esses mapas podemos ver, visualmente, que o padrão global foi capturado para todos os fatores em α e λ, e todos os tempos em δ. Entretanto, pela Figura 5.12, temos uma análise mais clara de que ocorre sobrestimação para valores de α abaixo de ≈ -1 e subestimação para valores acima de ≈ 1, Painel (a). Também tem-se subestimação para δ abaixo de ≈ -2, Painel (c). Além disso, verificamos que o intervalo HPD de 95% é bem maior, indicando uma grande incerteza na estimação nesse intervalo. Por outro lado, a estimação de δ para valores acima de -2 é muito boa, especialmente para os valores positivos.

Fazendo uma análise da expressão da taxa θ_i, na Equação (3.13), e lembrando que $\beta = (0.5, -1.0, 1.0)^\top$ e $X_{1i} = 1$, X_{2i} pode assumir 0 ou 1 e que $X_{3i} \sim U(-1, 1)$ tem média

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.50	0.46	0.46	0.04	0.39	0.53
β_1	-1.00	-1.00	-1.00	0.01	-1.01	-0.99
β_2	1.00	1.00	1.00	0.01	0.99	1.01
σ^2	0.80	0.96	0.96	0.07	0.83	1.11
τ_α	2.00	2.02	2.00	0.74	0.76	3.40
η_1^*	-2.00	-1.78	-1.79	0.29	-2.34	-1.23
η_2^*	1.50	1.41	1.41	0.22	0.97	1.84
η_3^*	0.75	0.96	0.96	0.19	0.58	1.32
η_4^*	-1.00	-0.34	-0.33	0.24	-0.81	0.14

Tabela 5.4: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o caso Poisson: $M_{L_{400}T_4V_4}^{K_2I_{50\%}}$ com $\approx 40\%$ de contagens zero.

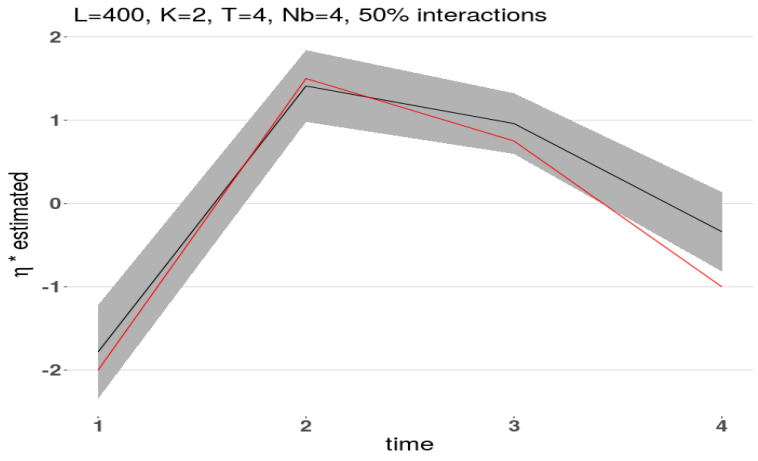


Figure 5.10: Gráfico da média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha) para o caso Poisson $M_{L_{400}T_4V_4}^{K_2I_{50\%}}$ e $\approx 40\%$ de contagens zero.

0, vemos que para valores de δ entre -4 e -12, teremos muitas taxas $\theta_i^{'}s = 0$. Nesse cálculo mencionado, considere os menores e maiores valores de $X_{\bullet i}^{\top}\beta = (-0.5, 0.5)$ e de $\delta = (-12, -3.5)$. Temos, então, θ_i variando de $e^{-12.5}$ e $e^{-3.5}$, que equivale a 0.000003726 e 0.0497, respectivamente. Isso significa a determinação de contagens zero. Concluimos que, quando temos muitos zeros na variável resposta, a inferência fica sobrestimada e com alta incerteza, englobando o valor verdadeiro perto do limite inferior do HPD (Figura 5.12).

O Painel (b) da Figura 5.12 ilustra os intervalos HPD de 95% para λ em cada um dos tempos. Podemos ver que todos os intervalos incluem o valor verdadeiro e, com isso, o padrão de decrescimento do Fator 1 e crescimento do Fator 2 é capturado pela estimação. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações. Considerando que cada ponto é um local, podemos ver que as regiões configuradas para não terem qualquer tipo de efeito ($L \in G_1$ ou G_2), realmente obtiveram probabilidades próximas de zero (cor azul). Os locais de G_E contendo interação na geração dos dados (cor vermelha) tiveram, em sua grande maioria, estimativas de probabilidade maiores que 0.5. Por outro lado, locais de G_E sem interação na geração (cor preta), ficaram com probabilidades abaixo de 0.5.

A Figura 5.13 ilustra mapas com 4 vizinhos por região através de grafos que imitam a estrutura espacial dos dados artificiais. Sendo cada ponto um local, podemos ver que alguns locais foram afetados por mais de um fator (Fator 1 e Fator 2), Paineis (a) e (b). Nos Paines (c, d, e, f) podemos ver os locais afetados por apenas um fator (Fator 1 ou Fator 2). Finalmente, o Painel (g) destaca as regiões que foram afetadas apenas pela interação e no Painel (h) temos, de cor verde, os locais que não foram afetados por qualquer tipo de efeito (principal ou interação). Pode-se verificar, pelos Paineis (c), (d), (e), (f), que existem mais regiões com efeito principal positivo (pontos vermelhos) do que os encontrados no caso anterior, Figura 5.8. Isso reflete o fato de se ter, neste caso, uma subestimação para valores positivos de α , ilustrada no Painel (a) da Figura 5.12. Lembrando ao leitor de que os locais são identificados com efeito principal se os intervalos HPD's de 95% das cargas não englobam o valor zero. Avaliando o efeito apenas de interação, Painel (g), e comparando-o com o mesmo painel da Figura 5.8 não se

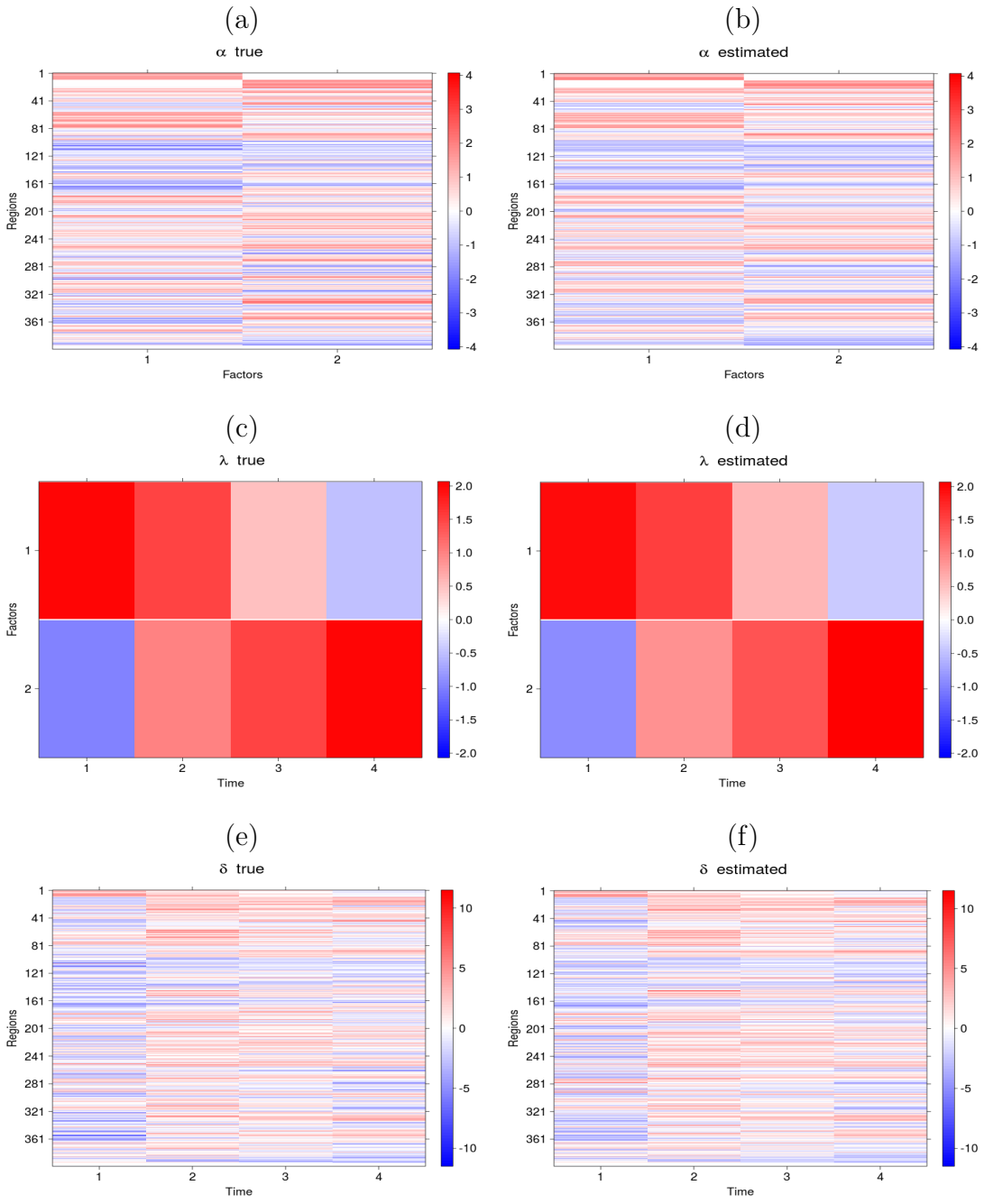


Figure 5.11: Mapas de calor comparando valores verdadeiros e estimados para o caso Poisson $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ e $\approx 40\%$ de contagens zero. Paineis: (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

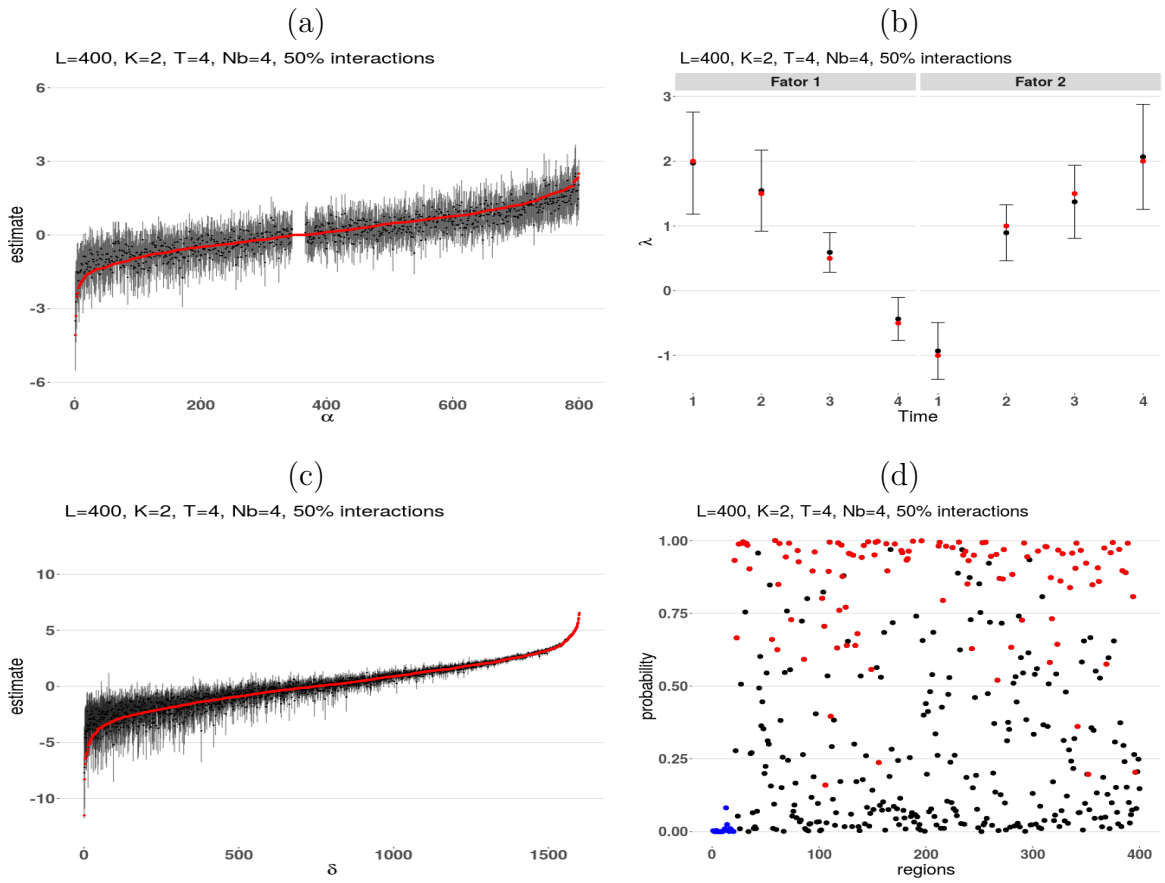


Figure 5.12: Análise gráfica dos intervalos HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais do grupo G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o caso Poisson: $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 40\%$ de contagens zero.

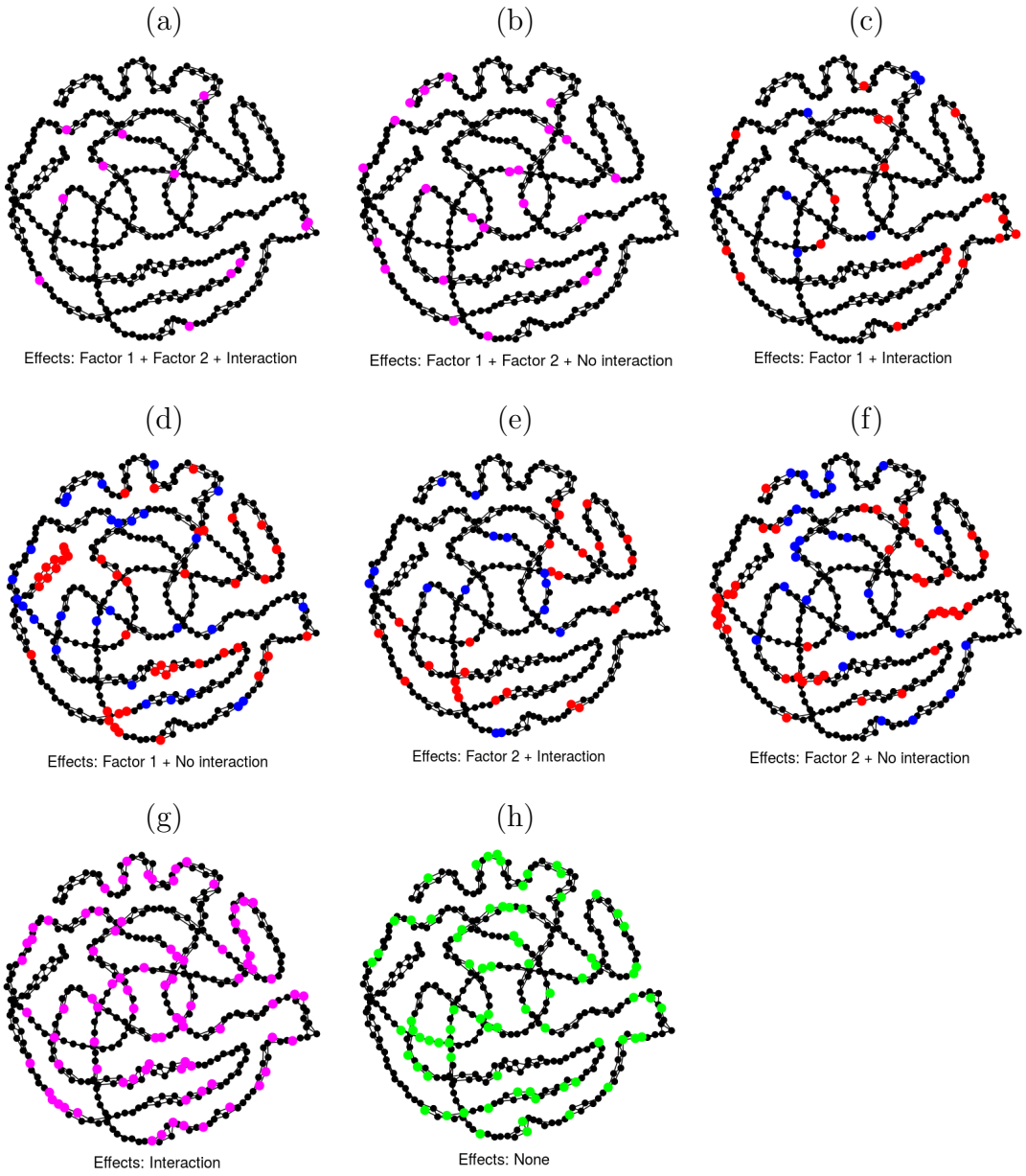


Figure 5.13: Grafos com 4 vizinhos por região imitando a estrutura espacial dos dados artificiais. Cada ponto representa um local. A cor vermelha (carga positiva) ou azul (carga negativa) identifica os locais associados a algum efeito principal e/ou interação (Paineis c, d, e, f). A cor verde (Painel h) denota as regiões não afetadas por qualquer efeito. Nos Paineis (a, b, g), a cor magenta indica os locais afetados por mais de um efeito principal ou somente por interação. Considere o caso Poisson: $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 40\%$ de contagens zero.

percebe grande diferença. Isso indica que a estimação do efeito de interação é semelhante nos casos com muitas ($\approx 40\%$) e poucas ($\approx 3\%$) contagens zero.

Terminamos, aqui, a análise com $\approx 50\%$ dos locais de G_E afetados por interação. A seguir vamos analisar a situação na qual existem $\approx 30\%$ de regiões afetadas por η^* em G_E, ainda para o caso $\approx 40\%$ de contagens zero.

Cenário com $\approx 30\%$ de locais em G_E afetados por interação

Completando o estudo de muitas contagens zero ($\approx 40\%$), neste tópico vamos analisar o cenário com 30% de locais de G_E afetados por interação. Na Tabela 5.5 temos as estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. Semelhante ao caso anterior, alguns parâmetros tiveram seus valores verdadeiros fora de seus intervalos HPD de 95%, a saber: β_0, η_3^* e η_4^*. Neste caso, β_0 foi subestimado e os demais foram sobrestimados. Mais uma vez temos que o desvio padrão para os coeficientes em β_1, β_2 e a variância σ^2 foram muito próximos de zero, indicando um bom ajuste, especialmente para β_1 e β_2 em que os valores estimados foram iguais aos verdadeiros. Veja ainda, pela Figura 5.14 como a estimativa de η^* foi satisfatória, com exceção do tempo 4, que saiu um pouco para fora do limite inferior do intervalo HPD. Apesar disso, a estimação de η^* seguiu a tendência de crescimento e decrescimento do valor verdadeiro.

A Figura 5.15 ilustra mapas de calor referentes à α, λ e δ, comparando valores verdadeiros e estimados. Visualmente podemos ver, novamente, que o padrão verdadeiro foi capturado pelas estimativas para todos esses parâmetros citados. Comparando os Paineis (a) das Figuras 5.16 e 5.12, pode-se perceber que a sobrestimação ocorrida em $\alpha < -1$ foi maior do que a estimação do caso 50% e para $\alpha > 0.5$, a subestimação foi também maior do que no caso anterior. Para o parâmetro δ, Painel (c) da Figura 5.16, percebemos que a sobrestimação é maior do que no caso 50% e os intervalos HPD de 95% são mais largos para valores de δ inferiores a -3. Boas estimativas são obtidas para valores acima de -3. Assim como na análise do caso 50% de locais de G_E afetados por interação, percebe-se uma estimação satisfatória e incerteza *a posteriori* baixa quando $\delta > 0$.

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.50	0.39	0.39	0.01	0.37	0.41
β_1	-1.00	-1.00	-1.00	0.01	-1.01	-0.99
β_2	1.00	1.00	1.00	0.01	0.99	1.01
σ^2	0.80	0.88	0.88	0.06	0.77	1.01
τ_α	2.00	1.40	1.25	0.51	0.63	2.38
η_1^*	-2.00	-2.01	-2.02	0.26	-2.50	-1.49
η_2^*	1.50	1.72	1.73	0.18	1.36	2.08
η_3^*	0.75	1.09	1.09	0.16	0.77	1.41
η_4^*	-1.00	-0.53	-0.53	0.22	-0.96	-0.11

Tabela 5.5: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o caso Poisson: $M_{L_{400}T_4V_4}^{K_2I_{30\%}}$ com $\approx 40\%$ de contagens zero.

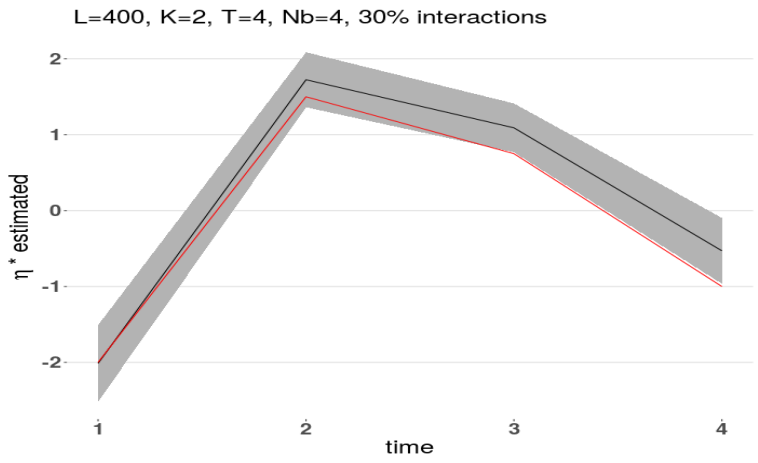


Figure 5.14: Gráfico da média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha) para o caso Poisson $M_{L_{400}T_4V_4}^{K_2I_{30\%}}$ com $\approx 40\%$ de contagens zero.

Ainda na Figura 5.16 podemos ver como as estimativas para λ, Painel (b), ficaram dentro do intervalo HPD de 95% e capturando todo o padrão do Fator 1, decrescimento, e do Fator 2, crescimento. Pelo Painel (d) avalia-se, mais uma vez, que as estimativas das probabilidades de interação (termo p^* do Passo 3 da Seção 3.1.1) para os locais gerados com efeito de interação (cor vermelha) estão, na sua maioria, acima de 0.5 e os locais de G_E gerados sem efeito de interação estão, em grande parte, com probabilidades < 0.5. Entretanto, veja que a quantidade de locais gerados com interação, mas que obtiveram probabilidades abaixo de 0.5, é maior do que no caso anterior, reforçando o fato de que ter mais locais contribuindo para a estimação do efeito de interação obtem-se melhores inferências.

Na Figura 5.17 temos grafos com 4 vizinhos por região. Esses grafos imitam a estrutura espacial dos dados artificiais. Os locais podem ser afetados ou não por algum efeito principal e/ou interação. Esses efeitos podem ter cargas (α) positivas (cor vermelha) ou negativas (cor azul); veja Paineis (c, d, e, f). As regiões não afetadas por qualquer efeito estão diferenciadas pela cor verde, Painel (h). Por fim, os locais afetados por mais de um efeito principal ou somente por interação (cor magenta) estão ilustrados nos Paineis (a, b, g). Verifique que ocorre mais locais afetados apenas por um efeito principal, Fator 1 ou Fator 2 nos Paineis (d) e (f), do que locais afetados por um efeito principal e interação, Paineis (c) e (e). Nesses paineis citados, também pode-se identificar que o número de locais cujas cargas são negativas (cor azul) se assemelha ao número de cargas positivas (cor vermelha), fato que pode constatado analisando o Painel (a) da Figura 5.16, em que, aproximadamente, o mesmo número de locais possuem cargas com $\alpha < -1$ e com $\alpha > 1$ para as quais o intervalo HPD de 95% não inclui o zero. Um boa quantidade de locais foi afetada apenas por interação, Painel (g), o que, mais uma vez, está coerente com os locais gerados com interação e identificados pela cor vermelha no Painel (d) da Figura 5.16, para os quais as probabilidades estimadas de serem afetados por interação estão, em sua grande maioria, maiores do que 0.5.

Finalizando a análise das estimativas para contagens com muitos zeros vamos traçar uma comparação entre os dois casos, 50% e 30% de locais de G_E afetados por interação. Pelas Tabelas 5.4 e 5.5 vemos que: em relação aos coeficientes em β, β_1 e β_2 foram

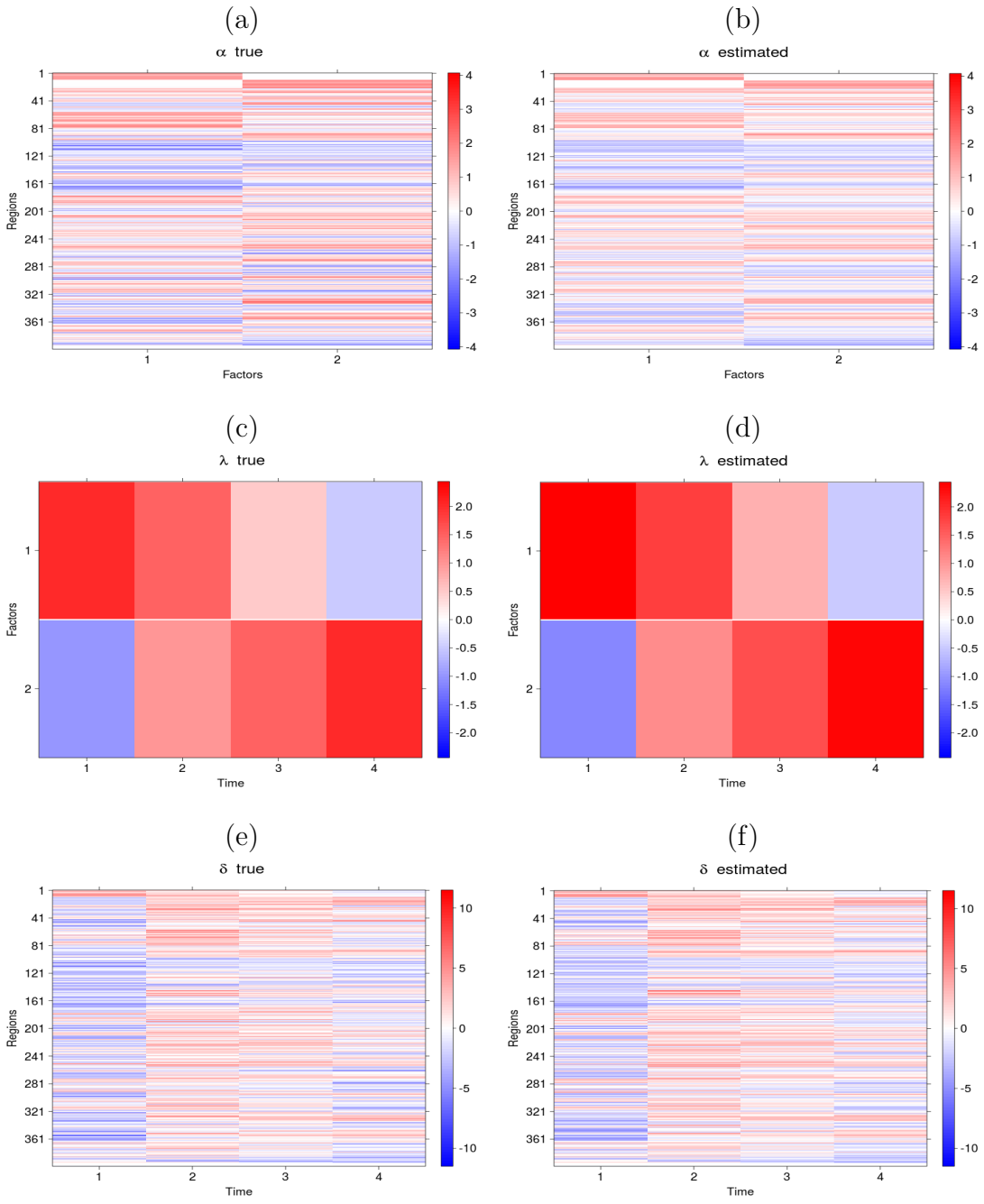


Figure 5.15: Mapas de calor comparando valores verdadeiros e estimados para o caso Poisson $M^{K_2 I_{30\%}}_{L_{400} T_4 V_4}$ com $\approx 40\%$ de contagens zero. Paineis: (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

igualmente bem estimados nos dois casos, e β_0, τ_α, η_2^* e η_3^* foram melhores estimados no caso 50%. Os termos η_1^* e η_4^* obtiveram melhores estimativas no caso 30%. Analisando as Figuras 5.11 e 5.15 é difícil, visualmente, identificar alguma diferença entre os paineis correspondentes de cada uma delas. Através das Figuras 5.12 e 5.16 fica claro, pelo Painel (d), como o fato de mais locais contribuirem para a estimação do efeito de interação induz à mais regiões terem as probabilidades de serem afetadas por η^* acima de 0.5 e, consequentemente, menos probabilidades abaixo de 0.5. Avaliando os Paineis (a), (b) e (c) dessas mesmas figuras, vemos α e λ obterem melhores ajustes no caso 50%, no entanto, as inferências para δ, nos dois casos, são parecidas. Com base nos Paineis (c) e (d) vemos que o caso 50% de locais de G_E afetados por interação obteve estimativas melhores do que o caso 30%, o que era de se esperar. Na próxima seção, semelhante ao caso logístico, traçamos um comparativo de ajustes do modelo para η^* e λ para os cenários com $L = 100,\ 200$ e 400 locais, $K = 2$ fatores, $T = 4$ tempos, 4 vizinhos por região e variando o número de locais de G_E afetados por interação (30% e 50%). Lembramos ao leitor que a análise comparativa levou em consideração apenas η^* e λ por serem os únicos parâmetros fixados durante a geração dos dados.

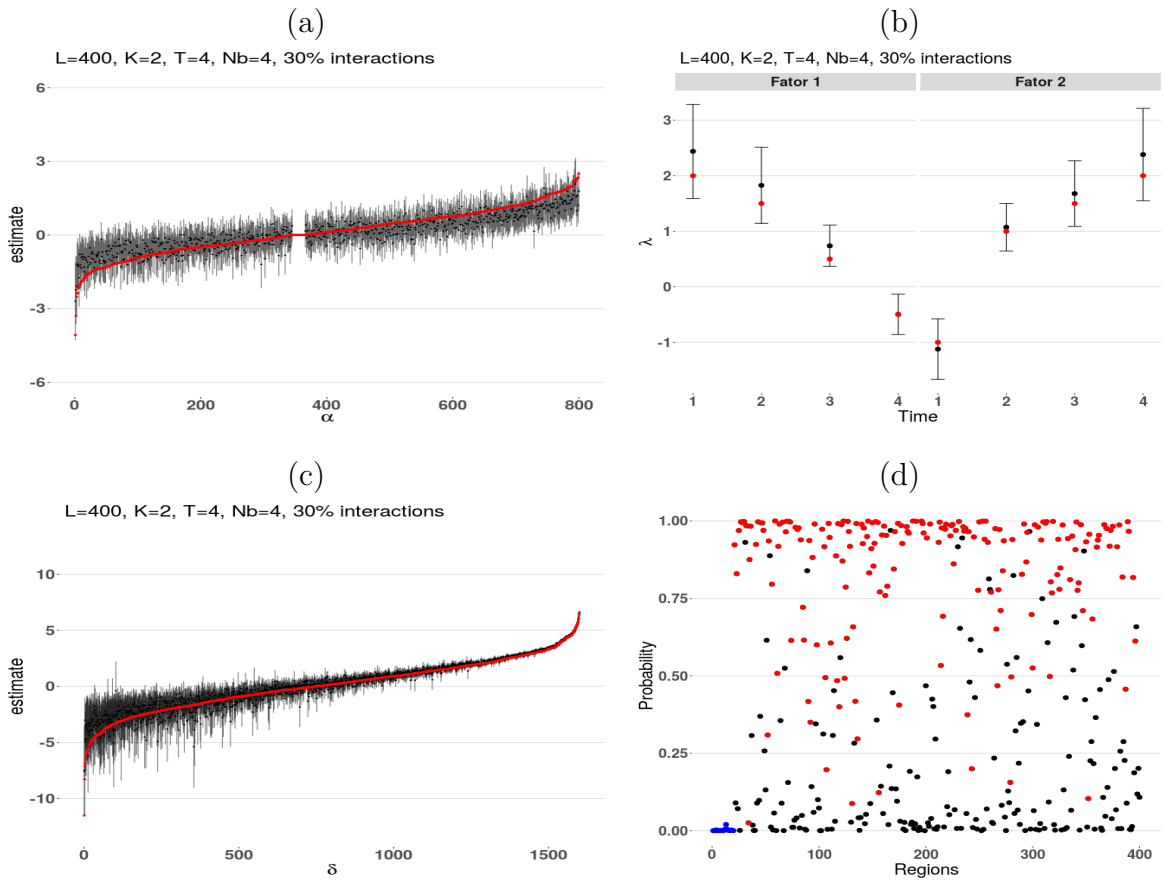


Figure 5.16: Análise gráfica dos intervalos HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais do grupo G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o caso Poisson: $M^{K_2 I_{30\%}}_{L_{400} T_4 V_4}$ com $\approx 40\%$ de contagens zero.

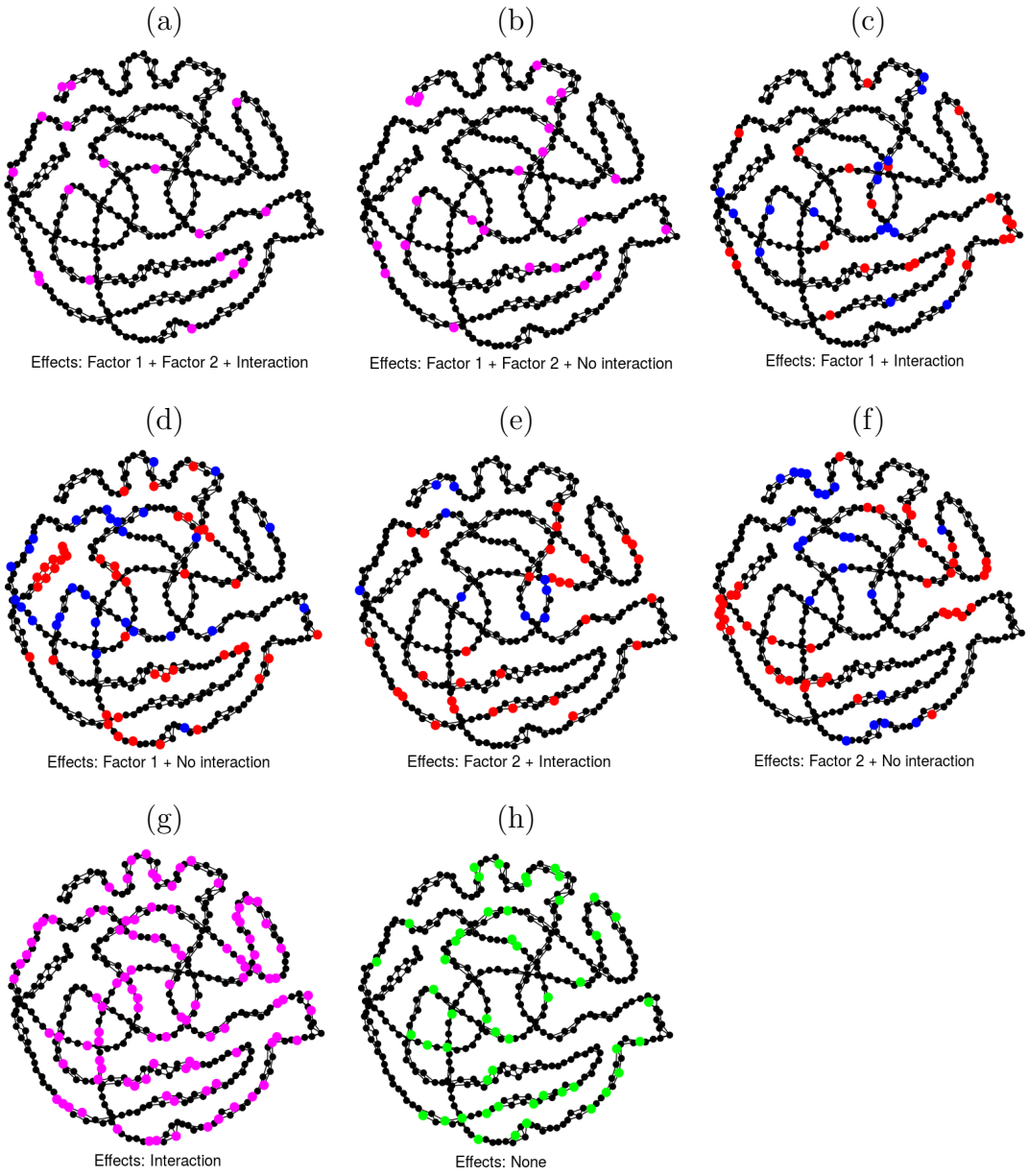


Figure 5.17: Grafos com 4 vizinhos por região imitando a estrutura espacial dos dados artificiais. Cada ponto representa um local. A cor vermelha (carga positiva) ou azul (carga negativa) identifica os locais associados a algum efeito principal e/ou interação (Paineis c, d, e, f). A cor verde (Painel h) denota as regiões não afetadas por qualquer efeito. Nos Paineis (a, b, g), a cor magenta indica os locais afetados por mais de um efeito principal ou somente por interação. Considere o caso Poisson: $M^{K_2 I_{30\%}}_{L_{400} T_4 V_4}$ com $\approx 40\%$ de contagens zero.

5.3 Comparação geral dos cenários

Conforme apresentado para o caso logístico, nesta seção desenvolvemos um estudo comparativo das estimativas de η^* e λ variando o número de locais. Consideramos $L = 100,\ 200$ e 400 locais, $K = 2$ fatores, $T = 4$ tempos, 4 vizinhos por região, $\approx 30\%$ e $\approx 50\%$ de locais de G_E afetados por interação. Lembrando que η^* e λ são os únicos parâmetros cujos valores são fixados durante a geração dos dados. O objetivo, aqui, também é avaliar o efeito sob as estimativas quando se tem um número maior de regiões L, de G_E, afetados por interação. Naturalmente, espera-se uma melhor estimação, mas queremos avaliar o tamanho desse ganho. Consideramos, para esse estudo, apenas o caso no qual temos $\approx 40\%$ de contagens zero, uma vez que, como pode ser visto nas Seções 5.1 e 5.2, as estimativas para η^* e λ foram satisfatórias tanto para contagens com poucos quanto para o caso com muitos zeros.

A Figura 5.18 mostra a média *a posteriori* para η^* nas configurações com 30% de locais afetados por interação (paineis da esquerda) e 50% (paineis da direita), sendo cada linha referente a um valor para L. Em conformidade com a inferência Bayesiana, as amplitudes dos intervalos HPD's de 95% (área sombreada) foram mais estreitas para os cenários com mais dados para estimar o efeito de interação (paineis da direita versus paineis da esquerda). Perceba também como a área sobreada do caso $L = 400$ é bem mais estreita do que no caso $L = 100$. Equivalente ao caso logístico, para $L = 100$ a diferença no tamanho dos envelopes HPD para as situações de 30% e 50% de locais de G_E com efeito η^* é mais clara. Isso pode ser justificado pela quantidade de regiões que contribuem para explicar η^* (30% de $80 = 24$ versus 50% de $80 = 40$ locais).

As Figuras 5.19, 5.20 e 5.21 ilustram as estimativas para λ nos casos $L = 100, 200$ e 400, respectivamente. Os paineis da esquerda se referem ao primeiro fator e os da direita, ao segundo. Vemos que as estimativas são satisfatórias em todos os cenários. Verifique como o comportamento, no tempo, dos escores dos Fatores 1 e 2 é bem capturado (linha preta versus linha vermelha). No caso do Fator 1 ocorre o decrescimento do escore, enquanto para o Fator 2, observa-se o crescimento no tempo. Mais uma vez, podemos observar que a incerteza *a posteriori* de λ é menor no caso $L = 400$. Para não

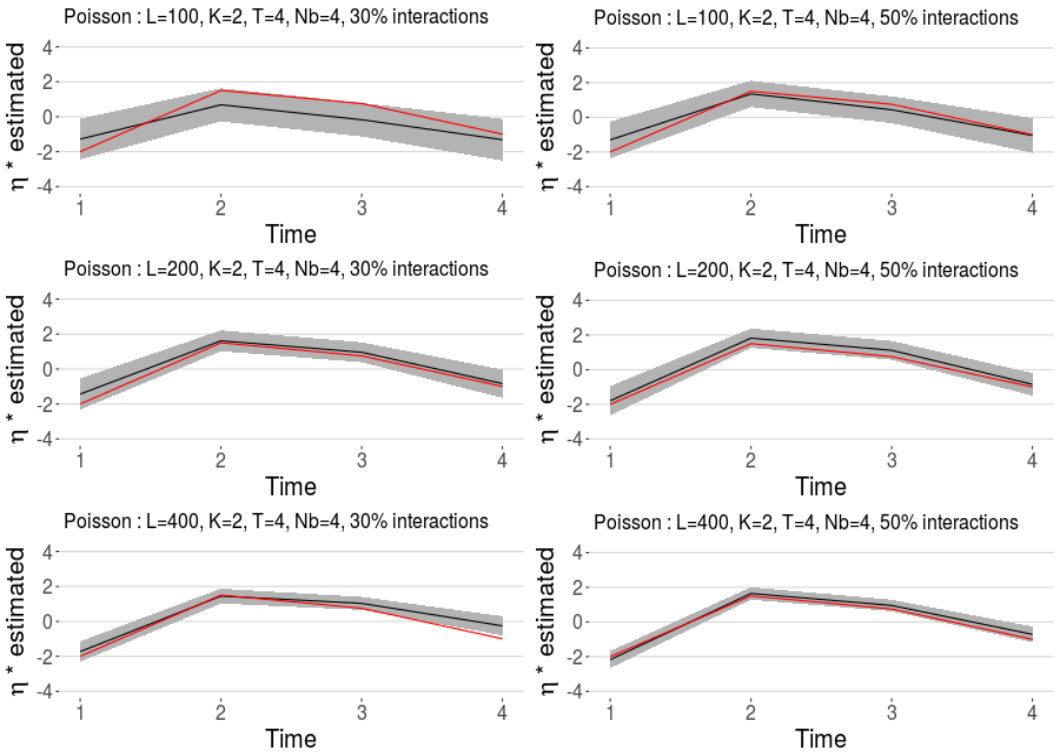


Figure 5.18: Média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha) em todas as configurações de números de regiões ($L = 100, 200$ e 400). Considere os casos Poisson: $M^{K_2 I_{30\%}}_{L_{100} T_4 V_4}$ versus $M^{K_2 I_{50\%}}_{L_{100} T_4 V_4}$, $M^{K_2 I_{30\%}}_{L_{200} T_4 V_4}$ versus $M^{K_2 I_{50\%}}_{L_{200} T_4 V_4}$ e $M^{K_2 I_{30\%}}_{L_{400} T_4 V_4}$ versus $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$, sendo todos eles com $\approx 40\%$ de contagens zero, ou seja, $\beta = (0.5, -1.0, 1.0)^\top$.

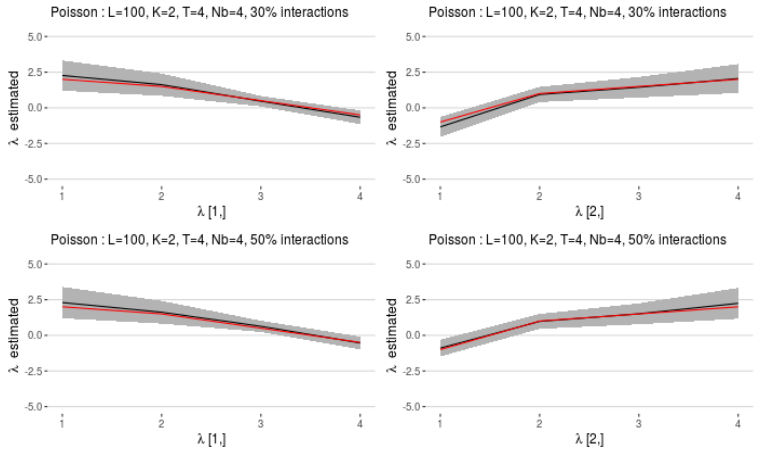


Figure 5.19: Média *a posteriori* (linha preta), intervalo HPD de 95% para λ (área sombreada) e valor verdadeiro (linha vermelha). Considere os casos Poisson: $M^{K_2 I_{30\%}}_{L_{100} T_4 V_4}$ e $M^{K_2 I_{50\%}}_{L_{100} T_4 V_4}$ com $\approx 40\%$ de contagens zero, ou seja, $\beta = (0.5, -1.0, 1.0)^\top$.

comprometer a leitura e não desviar a atenção do leitor quanto ao objetivo desta seção, transferimos para o Apêndice D os mapas de calor comparativos de valores verdadeiros versus estimados, e os gráficos dos intervalos HPD de 95% para α, λ e δ referentes a $L = 100$ e $L = 200$ locais.

Seguindo as análises desenvolvidas para o modelo logístico, o próximo passo é avaliar o comportamento do modelo Poisson com outros números de fatores e de tempos. Na próxima seção, apresentamos uma análise considerando $T = 10$ tempos e $K = 2$ fatores e, em seguida, a configuração de $K = 3$ fatores, mas mantendo $T = 4$. Em ambos os casos, restringimos o estudo a $L = 400$ locais, 4 vizinhos por região, $\approx 50\%$ de locais de G_E afetados por interação e $\approx 40\%$ de contagens zero.

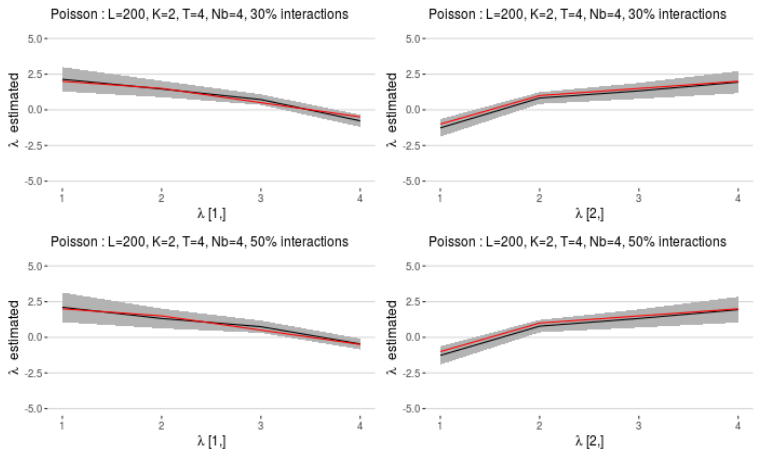


Figure 5.20: Média *a posteriori* (linha preta), intervalo HPD de 95% para λ (área sombreada) e valor verdadeiro (linha vermelha). Considere os casos Poisson: $M^{K_2 I_{30\%}}_{L_{200} T_4 V_4}$ e $M^{K_2 I_{50\%}}_{L_{200} T_4 V_4}$ com $\approx 40\%$ de contagens zero, ou seja, $\beta = (0.5, -1.0, 1.0)^\top$.

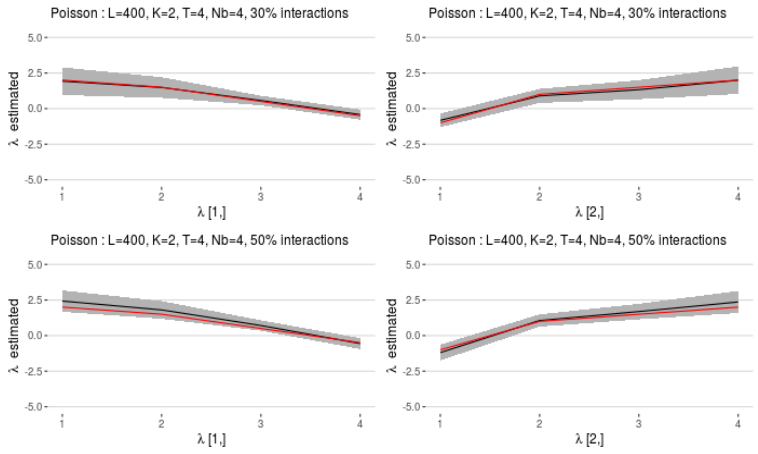


Figure 5.21: Média *a posteriori* (linha preta), intervalo HPD de 95% para λ (área sombreada) e valor verdadeiro (linha vermelha). Considere os casos Poisson: $M^{K_2 I_{30\%}}_{L_{400} T_4 V_4}$ e $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 40\%$ de contagens zero, ou seja, $\beta = (0.5, -1.0, 1.0)^\top$.

5.4 Análise com nova especificação de K e T

Nesta seção, desenvolvemos a avaliação do comportamento do modelo em outras configurações de fatores e tempos para o caso Poisson. Estudar o ajuste do modelo quando variamos o número de fatores está relacionado ao fato de, na prática, não sabermos o número mais adequado de fatores latentes a serem introduzido no modelo. Especificação diferente para o número de tempos é igualmente importante, pois novos dados são constantemente coletados por aplicações comerciais ou em um ambiente de pesquisa. Exemplo disso é o sistema de telediagnóstico do Centro de Telessaúde do Hospital das Clínicas da UFMG, origem dos dados utilizados nesta tese, que coleta e registra dados pessoais, clínicos e laudos de eletrocardiogramas de milhares de pacientes diariamente.

Da mesma forma que no caso logístico, avaliamos primeiro o caso com $K = 2$ e $T = 10$, e depois, $K = 3$ e $T = 4$. Lembramos ao leitor que o estudo considera $L = 400$ locais, 4 vizinhos por região, $\approx 50\%$ de locais de G_E afetados por interação e $\approx 40\%$ de contagens zero.

Análise para T igual a 10 tempos

O cenário com mais tempos, $T = 10$, ilustra casos em que novos dados são acrescentados à base de dados no decorrer dos anos. Situação muito comum, conforme citado anteriormente, em diversos tipos de aplicações reais.

A Tabela 5.6 apresenta as estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2 e do parâmetro de variância τ_α. Note que os valores para β são aqueles utilizados na geração de dados assumindo muitos zeros na variável resposta. Os valores para η^* são os apresentados na Tabela 4.7, resultado do produto dos $\lambda's$ descritos na Tabela 4.4. Os resultados indicam que os valores reais de β_0 e η_8^* ficaram ligeiramente fora do intervalo HPD de 95% e a estimativa de τ_α ficou distante do valor verdadeiro, apesar desse valor estar dentro do envelope HPD. Com exceção desses parâmetros citados, todos os demais obtiveram estimativas bem próximas do valor verdadeiro. Perceba que apesar do aumento no número de parâmetros, a estimação para η^* foi satisfatória. Além disso, houve a captura do padrão de crescimento seguido por decrescimento ao longo do

tempo (Figura 5.22). Seguindo os resultados anteriores, a média e a mediana obtiveram estimativas muito próximas, em todos os casos, indicando simetria na distribuição *a posteriori*. Novamente os desvios padrão para os coeficientes da regressão são bem pequenos mostrando que a incerteza *a posteriori* é baixa.

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.500	0.426	0.427	0.031	0.364	0.487
β_1	-1.000	-1.003	-1.004	0.009	-1.021	-0.985
β_2	1.000	0.996	0.996	0.008	0.981	1.011
σ^2	0.800	0.783	0.782	0.031	0.726	0.845
τ_α	2.000	1.305	1.189	0.521	0.531	2.560
η_1^*	-2.000	-1.823	-1.832	0.310	-2.442	-1.204
η_2^*	-1.190	-1.059	-1.073	0.261	-1.575	-0.544
η_3^*	-0.750	-0.742	-0.753	0.217	-1.160	-0.297
η_4^*	0.600	0.747	0.746	0.163	0.434	1.065
η_5^*	1.000	1.276	1.277	0.154	0.970	1.583
η_6^*	0.840	1.010	1.010	0.144	0.728	1.290
η_7^*	0.750	0.893	0.893	0.156	0.589	1.195
η_8^*	-0.160	0.262	0.266	0.172	-0.087	0.594
η_9^*	-0.540	-0.452	-0.448	0.218	-0.897	-0.035
η_{10}^*	-1.000	-0.863	-0.860	0.224	-1.323	-0.429

Tabela 5.6: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o caso Poisson: $M^{K_2 I_{50\%}}_{L_{400} T_{10} V_4}$ com $\approx 40\%$ de contagens zero, ou seja, $\beta = (0.5, -1.0, 1.0)^\top$.

Através dos gráficos de mapas de calor, apresentados na Figura 5.23 pode-se verificar que o padrão global foi capturado para α (Paineis a e b), λ (Paineis c e d) e δ (Paineis e e f). Especificamente para λ e δ veja que todos os 10 tempos (colunas) seguiram o padrão exibido pela matriz verdadeira que está à esquerda.

A Figura 5.24 mostra os intervalos HPD's de 95% para α, λ e δ. Perceba nos Paineis

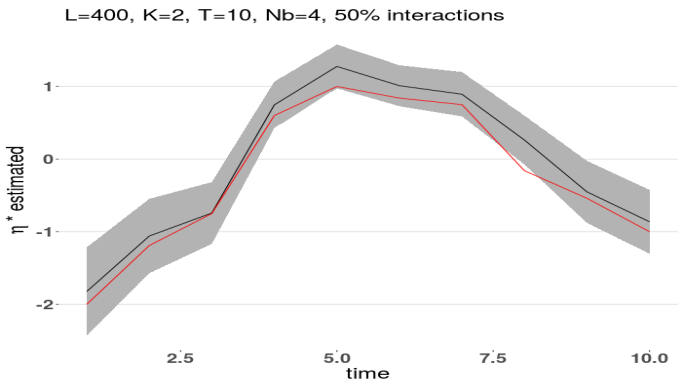


Figure 5.22: Média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha). Considere o caso Poisson: $M^{K_2 I_{50\%}}_{L_{400} T_{10} V_4}$ com $\approx$ 40% de contagens zero.

(a) e (b) o erro de estimação. No Painel (a) temos sobrestimação para valores de α menores que ≈ -0.5 e subestimação para valores maiores que ≈ 0.5. Podemos ver também, no Painel (b), que ocorre erro de estimação para λ em quase todos os tempos. Vê-se, claramente, que a maioria dos valores verdadeiros se encontram nas extremidades dos intervalos HPD de 95%. Verifica-se, também, que a maioria dos escores dos fatores são sobrestimados. Comparando o Painel (c) da Figura 5.24 com o caso $T = 4$ (Painel (c) da Figura 5.12), verificamos que a incerteza *a posteriori* se mantém maior para valores pequenos de δ. Mais uma vez, a estimação de δ é satisfatória em que o valor verdadeiro (linha vermelha) praticamente divide ao meio a núvem de intervalos, exceto quando δ assume valores muito pequenos. Ou seja, os desvios identificados no Painel (a) e no Painel (b) não se repetem no Painel (c), indicando que houve uma compensação entre as estimativas de α e λ, equilibrando e permitindo uma boa inferência de δ. Pelo Painel (d), veja que a maioria das probalilidades estimadas está alinhada com o valor verdadeiro, pois uma parcela maior de regiões de cor vermelha (regiões verdadeiramente afetadas pela interação) foi estimada com probabilidades acima de 0.5. O mesmo ocorre para as regiões geradas sem o efeito de interação (cor preta), registrando uma maioria de estimativas com probabilidades abaixo de 0.5.

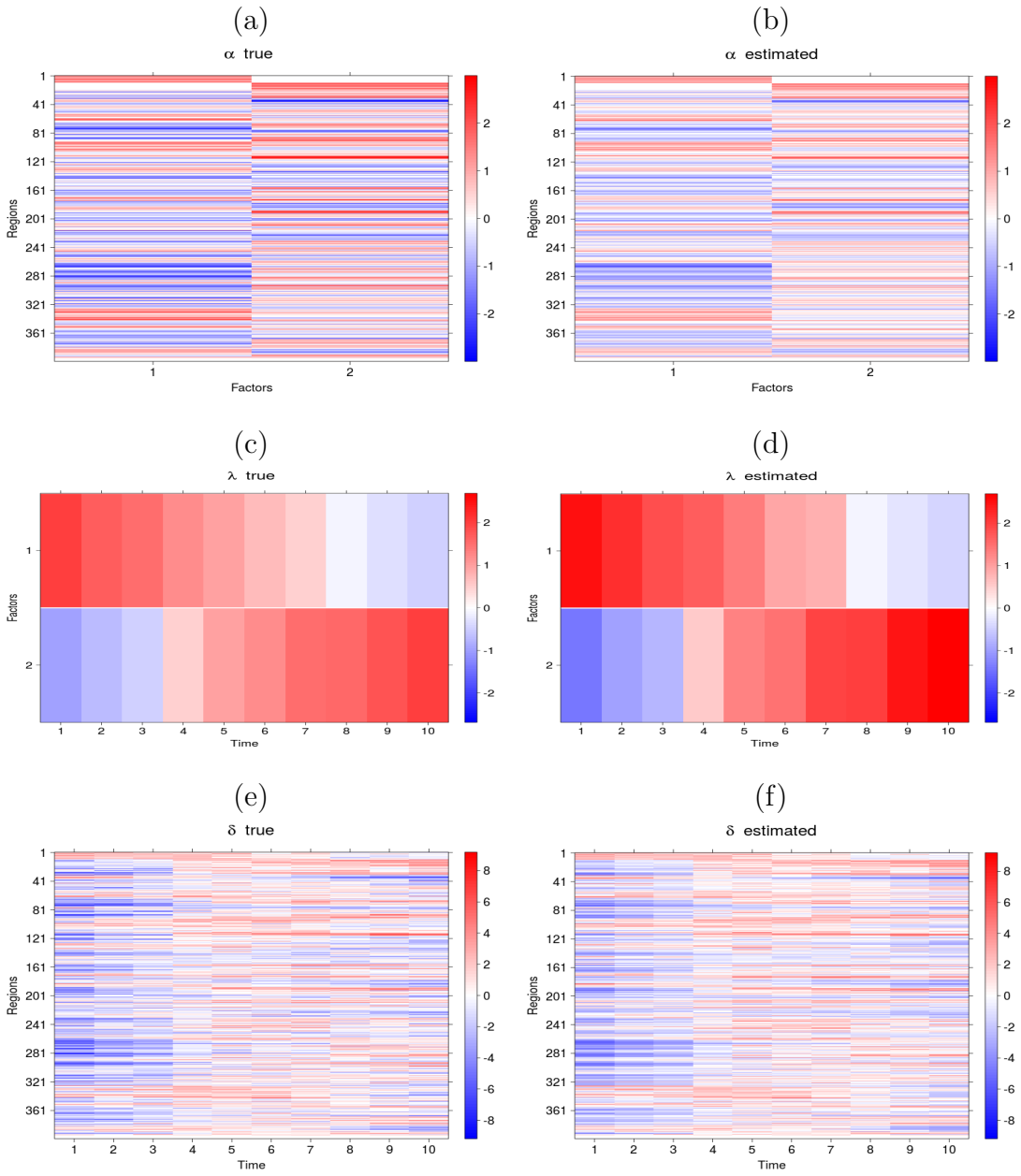


Figure 5.23: Mapas de calor comparando valores verdadeiros e estimados para o caso Poisson $M^{K_2 I_{50\%}}_{L_{400} T_{10} V_4}$ com $\approx 40\%$ de contagens zero. Paineis (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

Complementando a análise para $T = 10$, a Figura 5.25 ilustra mapas, através de grafos, em que pode-se analisar os conglomerados de locais sob algum tipo de efeito comum. Novamente, os Paineis (c, d, e, f) mostram os *clusters* subdivididos em cargas positivas (cor vermelha) e negativas (cor azul). Lembramos ao leitor que nesses paineis mencionados apenas as estimativas para as quais o intervalo HPD de 95% não inclui o zero são destacadas em cores vermelha e azul. Note que, a quantidade de locais identificados nos Paineis (g) e (h) são semelhantes a quantidade de locais vermelhos e pretos na Figura 5.24 (d), respectivamente. Lembrando que os locais sob efeito de interação são aqueles com $\mathrm{p}^*(z_l = 1|\bullet) > 0.5$ (veja o formulação dessa probabilidade no Passo 3 do algoritmo MH descrito na Seção 3.1.1). Os Paineis (a) e (b) ilustram os locais afetados pelos Fatores 1 e 2 (cor magenta), concomitantemente. Veja que menos locais foram afetados por esses dois efeitos no caso com $T = 4$ tempos (Paineis (a) e (b) da Figura 5.13).

Terminamos, aqui, a análise para $T = 10$. Novamente, podemos verificar que o modelo foi bem ajustado. Importante destacar que, neste caso, o maior valor de T implica em um aumento no número de observações muito maior do que o aumento no número de parâmetros. A cada novo tempo acrescentado temos a inserção de milhares de indivíduos novos contra a inclusão duas dezenas de parâmetros ($\sum_{l=1}^{L} n_l$ novos indivíduos contra $18 = 2 \times 6 + 6$ novos parâmetros, em que 2×6 significa o produto de 2 fatores e 6 tempos, e o segundo 6 somado equivale ao aumento de elementos em η^*). Ou seja, temos um crescimento muito maior do número de dados disponíveis para estimar os parâmetros. No próximo item iremos apresentar o estudo para o caso em que os dados foram gerados e estimados com $K = 3$ fatores, mas mantemos $T = 4$ tempos. As demais configurações ficam inalteradas, ou seja, $L = 400$ locais, 4 vizinhos por região, $\approx 50\%$ de locais de G_E afetados por interação e $\approx 40\%$ de contagens zero.

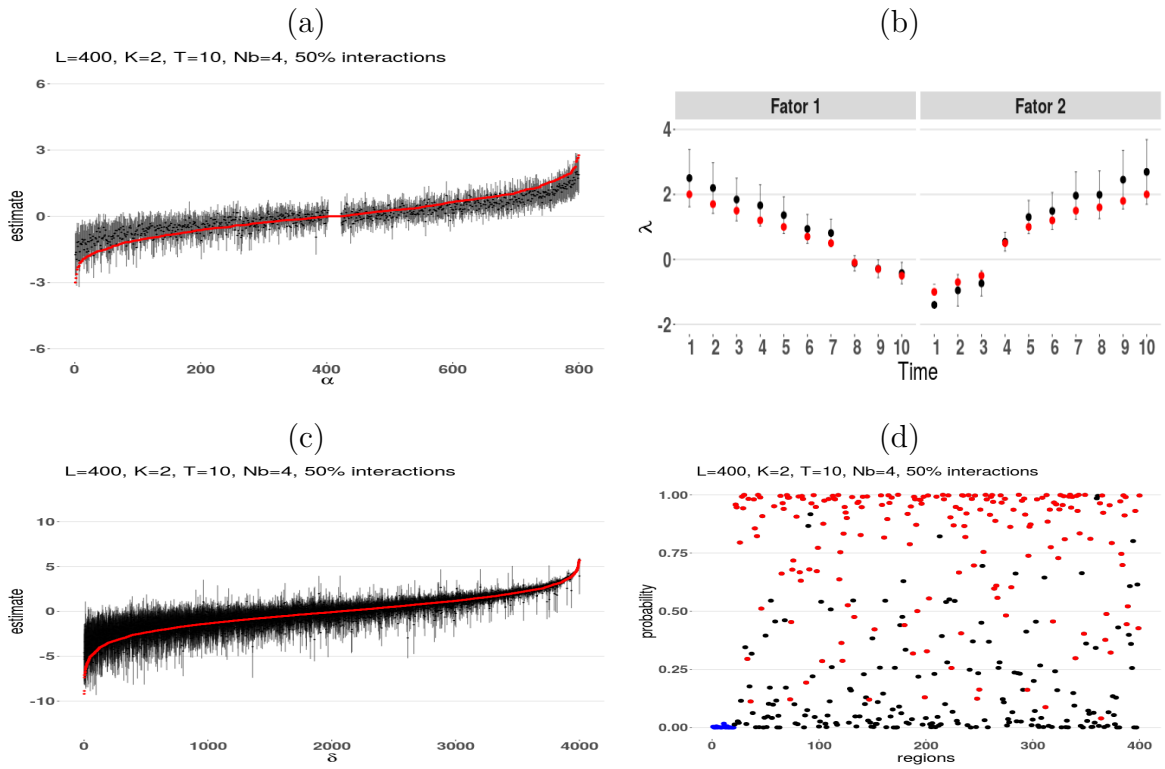


Figure 5.24: Análise gráfica dos intervalos HPD's de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais de G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o caso Poisson: $M^{K_2 I_{50\%}}_{L_{400} T_{10} V_4}$ com $\approx 40\%$ de contagens zero.

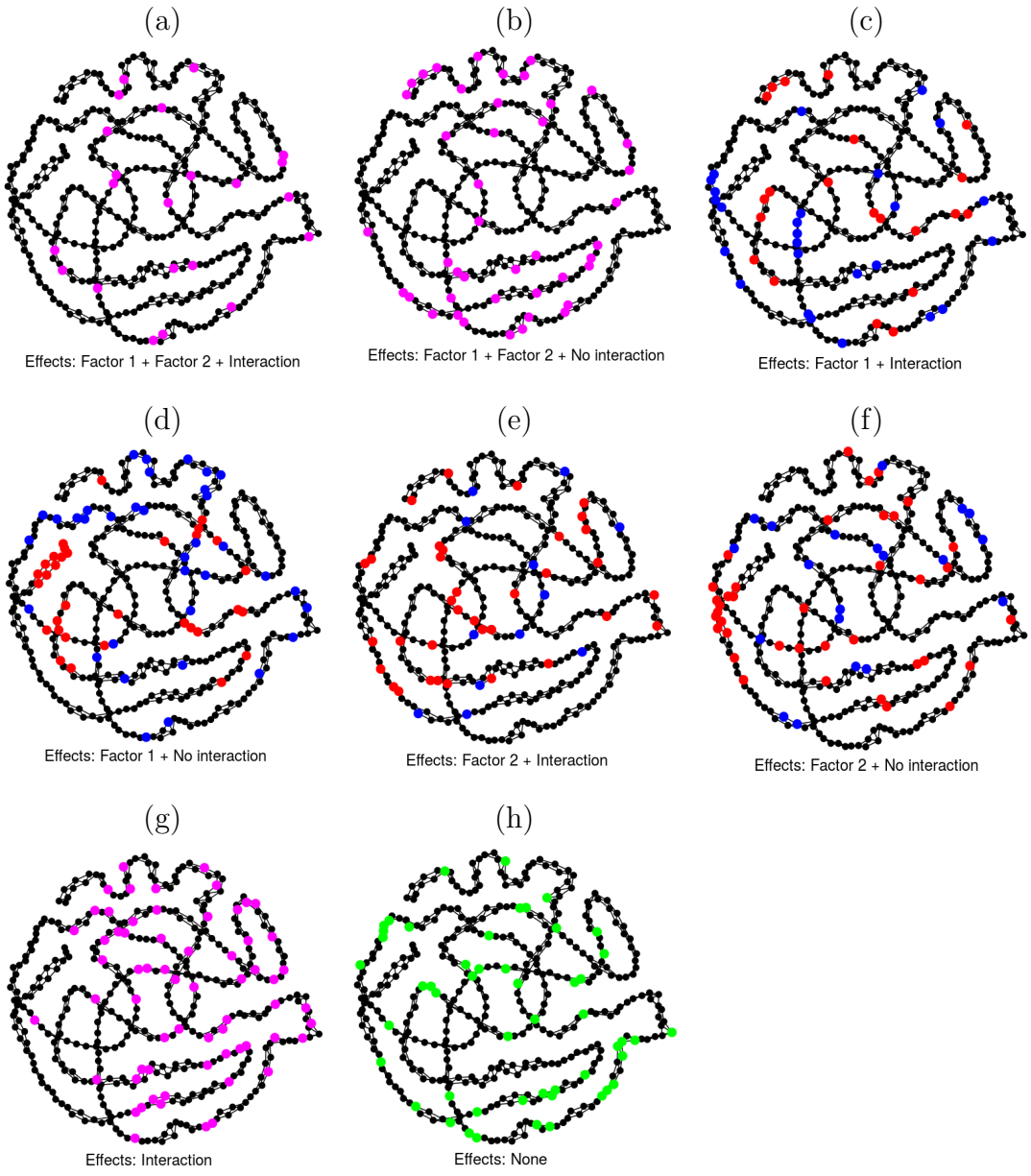


Figure 5.25: Grafos com 4 vizinhos por região imitando a estrutura espacial dos dados artificiais. Cada ponto representa um local. A cor vermelha (carga positiva) ou azul (carga negativa) identifica os locais associados a algum efeito principal e/ou interação (Paineis c, d, e, f). A cor verde (Painel h) denota as regiões não afetadas por qualquer efeito. Nos Paineis (a, b, g), a cor magenta identifica os locais afetados por mais de um efeito principal ou somente interação. Considere o caso Poisson: $M^{K_2 I_{50\%}}_{L_{400} T_{10} V_4}$ com $\approx 40\%$ de contagens zero.

Ajuste e geração dos dados com K igual a 3 fatores

Neste tópico, apresentamos a análise do ajuste do modelo Poisson, proposto nesta tese, para o cenário com $K = 3$ fatores na geração dos dados e na estimação. Considerando que o objetivo aqui é avaliar o comportamento do modelo em uma situação com mais de dois fatores, optamos por utilizar a configuração mais próxima dos dados reais no que diz respeito ao número de locais, tempos e vizinhos por região, ou seja: $L = 400$ locais, $T = 4$ tempos e 4 vizinhos por região. Para o percentual de locais de G_E afetados por interação mantivemos os mesmos 50% do tópico anterior e, também, o contexto de $\approx 40\%$ de contagens zero. O cenário $\approx 40\%$ de contagens zero foi utilizado por representar a configuração de β utilizada na geração dos dados artificiais do modelo logístico para dados balanceados ($\approx 50\%$ de $Y_i^{'}s = 1$).

Pela Tabela 5.7, podemos ver que apenas β_0 ficou fora do intervalo HPD de 95%. Novamente a média e a mediana são muito próximas indicando simetria da distribuição *a posteriori* para todos os parâmetros. Os desvios padrão para β_1 e β_2 são os menores e próximos de zero, indicando que os coeficientes das variáveis preditoras da regressão foram bem estimados. Em geral, pelas estimativas apresentadas na Tabela 5.7, temos que, em termos de estatística descritiva para esta amostra simulada, o modelo foi capaz de fornecer estimativas satisfatórias.

A Figura 5.26 ilustra a incerteza *a posteriori* sobre a estimação de η^* envolvendo três fatores latentes. Observe que o erro de estimação é maior para os tempos 1 e 2, mas para o tempo 3 a estimativa ficou muito próxima do valor verdadeiro. Além disso, perceba como a captura do valor verdadeiro pelo envelope HPD é visível para todos os tempos. Lembramos ao leitor que a interação utilizada na geração dos dados artificiais foi o produto entre os fatores, cujos valores para η^* se encontram na Tabela 4.5 e podem ser calculados pelos escores de λ apresentados na Tabela 4.4.

Pelos Paineis (a) e (b) da Figura 5.27, veja como o padrão global de α foi bem capturado, especialmente para os Fatores 2 e 3. Analisando os Paineis (c) e (d) de λ identifica-se que o padrão, também, foi bem capturado, apesar da sobrestimação ocorrida nos Fatores 2 e 3, Tempos 1 e 2, respectivamente. Baseado nas boas estimativas de α, λ

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.500	0.581	0.581	0.006	0.570	0.592
β_1	-1.000	-1.006	-1.007	0.004	-1.014	-0.998
β_2	1.000	0.996	0.997	0.004	0.989	1.004
σ^2	0.800	0.808	0.806	0.074	0.667	0.953
τ_α	2.000	2.570	2.546	0.664	1.347	3.764
η_1^*	-2.000	-1.665	-1.692	0.358	-2.320	-0.948
η_2^*	-1.500	-1.208	-1.250	0.377	-1.909	-0.464
η_3^*	0.750	0.764	0.770	0.309	0.151	1.386
η_4^*	1.000	0.849	0.864	0.327	0.183	1.459

Tabela 5.7: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o caso Poisson: $M^{K_3 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 40\%$ de contagens zero.

e η, era de se esperar que as estimativas de $\delta = \alpha\lambda + \eta + \epsilon$ também seriam satisfatórias, o que é verificado pelos Paineis (e) e (f).

Os Paineis (a) e (c), da Figura 5.28, ilustram que a maioria das médias *a posteriori* de α e de δ estão dentro do intervalo HPD de 95%. Perceba que nesses paineis citados as estimativas seguem a tendência dos valores verdadeiros com a linha de pontos na cor vermelha cortando os intervalos HPD's de 95% praticamente ao meio. A incerteza *a posteriori* para $\delta < -1$, Painel (c), é maior do que para os outros valores, fato já descrito no caso de $\approx 40\%$ de contagens zero ilustrado na Figura 5.16. No Painel (b) identifica-se que apenas para o Fator 3 no Tempo 2 temos que o valor verdadeiro (ponto vermelho) ficou fora do intervalo HPD, sendo que para a maior parte dos demais a estimativa foi bem próxima do valor verdadeiro. A maioria das probalilidades estimadas de que um local seja afetado por interação, Painel (d), estão, novamente, acima de 0.5. Comparando com os cenários anteriores ilustrados pelas Figuras 5.3, 5.7, 5.12, 5.16 e 5.24, identificamos apenas que poucas probabilidades estão próximas de 1 e o mesmo ocorrendo para as probabilidades abaixo de 0.5 (cor preta) em que poucas estão próximas de zero.

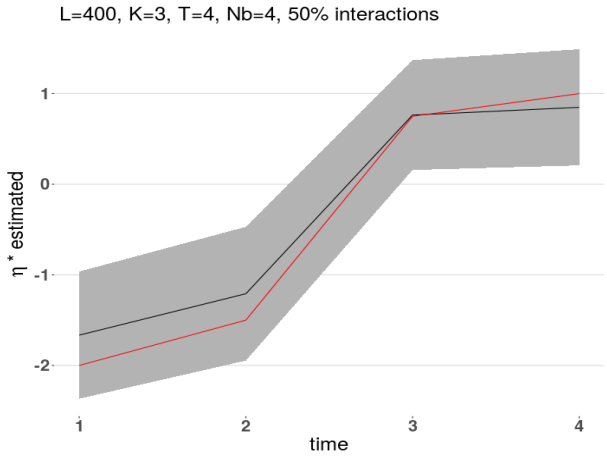


Figure 5.26: Média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha). Considere o caso Poisson: $M^{K_3 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx$ 40% de contagens zero.

Lembrando ao leitor que, na geração dos dados artificiais, os pontos vermelhos referem-se às regiões que sofrem de fato o efeito da interação e os pontos pretos representam regiões que não receberam efeito da interação.

Na Figura 5.29, igualmente como apresentado no caso logístico, por simplicidade e por considerarmos que não compromete as análises, optamos por não ilustrar os casos dos locais afetados por pares de fatores (1 e 2, 1 e 3, 2 e 3). Os grafos que imitam a estrutura espacial para os casos em que os locais são afetados por um único fator principal são apresentados nos Paineis de (a) a (f). O Painel (i) ilustra que três locais foram afetados pelos três fatores concomitantemente. No Painel (g) temos as regiões afetadas apenas por interação e no Painel (h) as regiões que não foram afetadas por qualquer efeito.

Finalizamos, aqui, as análises para o número de fatores $K = 3$. Concluimos que o modelo também capturou, satisfatoriamente, o padrão dos dados artificiais para este caso. Na próxima seção, seguindo o estudo desenvolvido para o caso logístico, vamos tratar de cenários nos quais avaliamos o impacto na estimação quando o analista comete o erro de especificar um fator acima ou abaixo da quantidade verdadeira para K usada

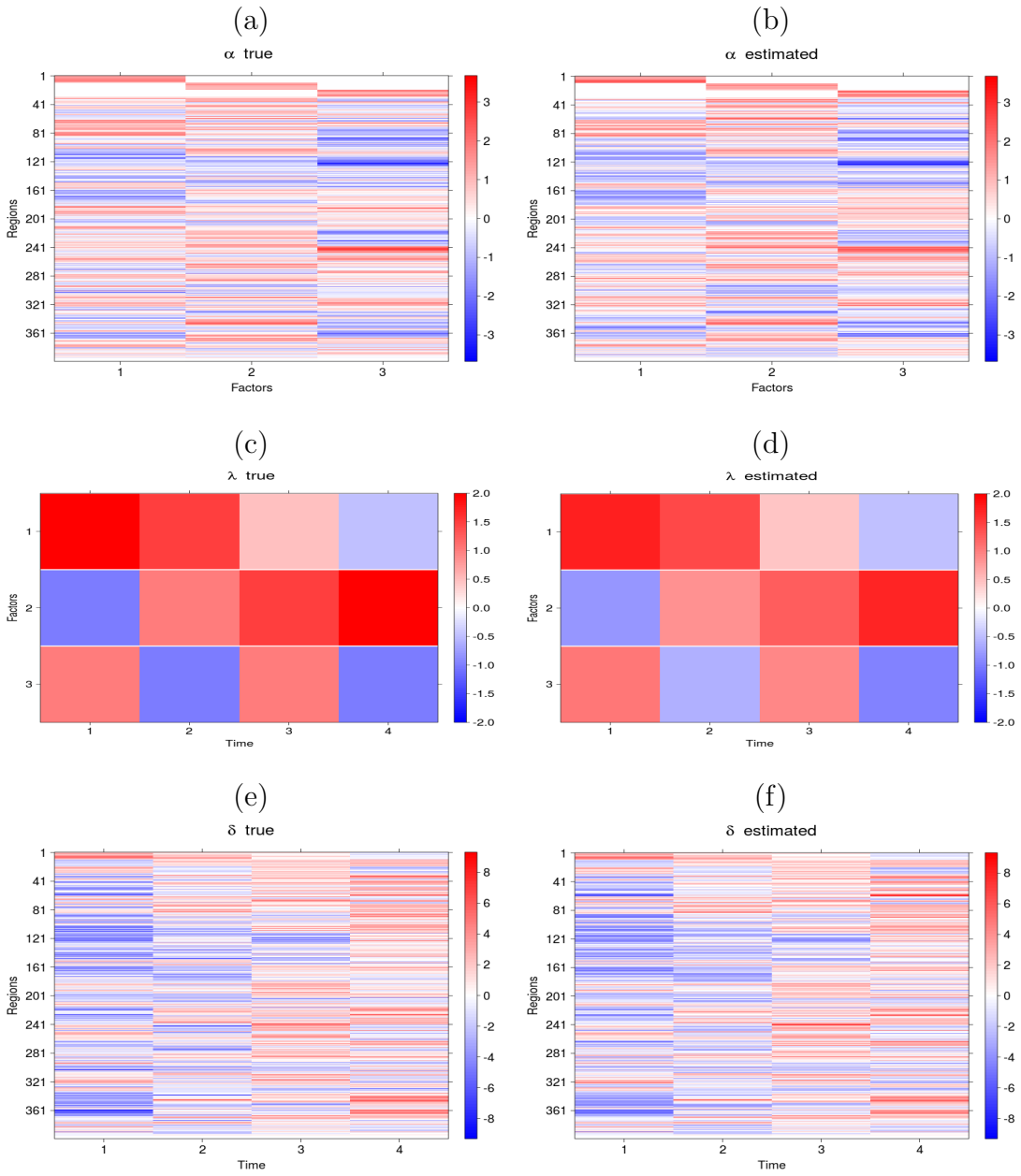


Figure 5.27: Mapas de calor comparando valores verdadeiros e estimados para o caso Poisson $M^{K_3 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 40\%$ de contagens zero. Paineis (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

na geração dos dados. Duas análises foram conduzidas. A primeira considera que existe 1 fator extra para ajustar, como por exemplo, $K_V = 2$ e $K_A = 3$, em que K_V equivale ao número verdadeiro de fatores e K_A o número ajustado. A segunda efetua a análise contrária com $K_V = 3$ e $K_A = 2$.

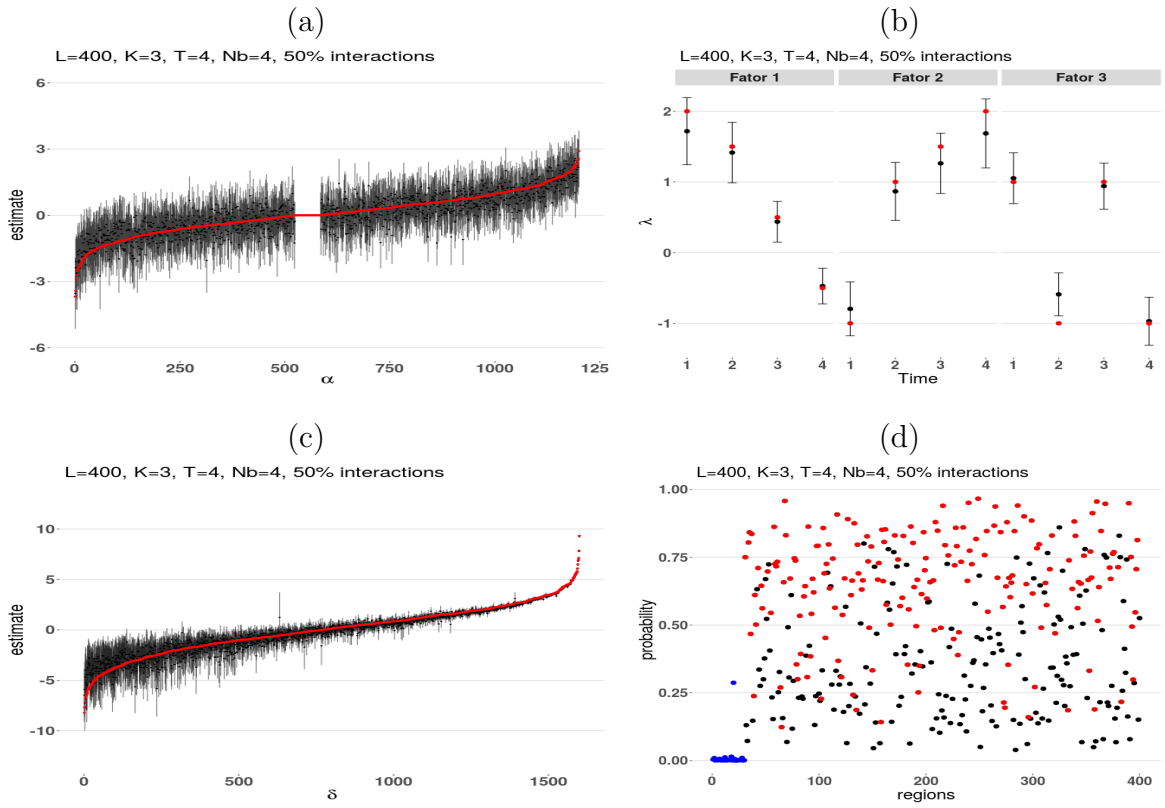


Figure 5.28: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1, G_2 e G_3, a cor vermelha representa locais de G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o caso Poisson: $M^{K_3 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 40\%$ de contagens zero.

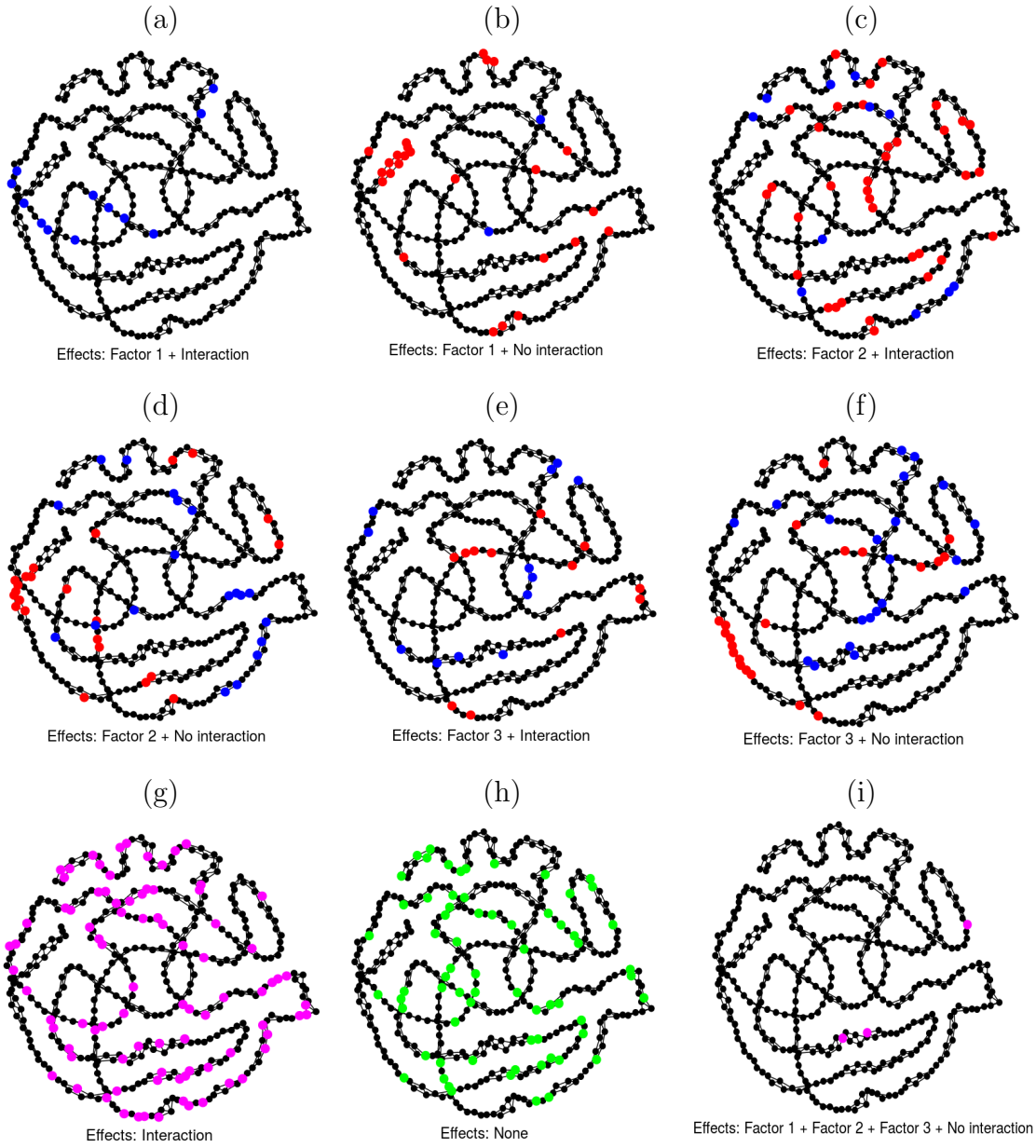


Figure 5.29: Grafos com 4 vizinhos por região imitando a estrutura espacial dos dados artificiais. Cada ponto representa um local. A cor vermelha (carga positiva) ou azul (carga negativa) identifica os locais associados a algum efeito principal e/ou interação (Paineis a, b, c, d, e, f). No Painel (g), a cor magenta indica os locais afetados somente por interação. A cor verde (Painel h) denota as regiões não afetadas por qualquer efeito. Considere o caso Poisson: $M_{L_{400}T_4V_4}^{K_3I_{50\%}}$ com $\approx 40\%$ de contagens zero.

5.5 Análise com erro de especificação de K

Quando ajustamos um modelo fatorial com dados reais, diferentemente do que acontece com dados artificiais, não sabemos *a priori* o número de fatores latentes mais adequado para uma análise mais representativa, sumarizando de forma eficiente todas as informações mais relevantes da base de dados. Nesta seção analisamos como o modelo Poisson, proposto nesta tese, se comporta em dois casos nos quais o número de fatores escolhido para ajuste do modelo é diferente do valor verdadeiro. Primeiramente, consideramos o caso em que $K_V = 2$ e $K_A = 3$, lembrando que K_V equivale ao valor verdadeiro de K e K_A é o valor ajustado. Em seguida analisamos a situação na qual temos $K_V = 3$ e $K_A = 2$, ou seja, quando o analista julga ser suficiente trabalhar com uma quantidade de fatores que está 1 unidade abaixo do que realmente foi considerado para gerar os dados.

Erro de especificação em que $K_V = 2$ e $K_A = 3$

A Tabela 5.8 apresenta as estimativas dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α e do termo de interação não linear η^*. Com exceção de β_0 e η_4^* todos os demais valores verdadeiros dos parâmetros estão dentro do intervalo HPD de 95%, indicando que o modelo foi capaz de proporcionar uma boa representação de várias informações da base de dados. Os desvios padrão dos coeficientes β_1 e β_2 são os menores na tabela indicando que há pouca incerteza quanto à estimação dos efeitos das variáveis preditoras incluídas na análise. A média e a mediana de todos os elementos são bem próximas mostrando que as distribuições *a posteriori* são simétricas. Os elementos em η^* apresentam os maiores desvios padrão, mas as estimativas para η_1^* e η_2^* foram praticamente iguais aos valores verdadeiros e, pela Figura 5.30, podemos ver que o padrão de crescimento e decrescimento foi capturado ocorrendo apenas sobrestimação para o tempo 4.

Analisando os Paineis (a, b) da Figura 5.31 vemos que as estimativas para α do Fator 1 seguiram, satisfatoriamente, o padrão verdadeiro, com sobrestimação mais evidente entre os locais 270 e 320. Comparando o Fator 2 verdadeiro com os Fatores 2 e 3 estimados, nota-se que houve uma diluição dos valores. Veja que as linhas de 10 a 30 da matriz de

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.500	0.261	0.256	0.034	0.206	0.330
β_1	-1.000	-0.998	-0.998	0.006	-1.010	-0.986
β_2	1.000	1.001	1.000	0.005	0.990	1.010
σ^2	0.800	0.993	0.990	0.071	0.846	1.127
τ_α	2.000	1.386	1.308	0.391	0.774	2.210
η_1^*	-2.000	-2.010	-2.020	0.295	-2.577	-1.416
η_2^*	1.500	1.582	1.586	0.212	1.142	1.991
η_3^*	0.750	1.052	1.059	0.187	0.670	1.408
η_4^*	-1.000	-0.393	-0.391	0.250	-0.883	0.096

Tabela 5.8: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o caso Poisson: $M_{L_{400}T_4V_4}^{K_2K_3I_{50\%}}$ com $\approx 40\%$ de contagens zero, em que o sobrescrito "K_2K_3" significa que o número de fatores verdadeiro é $K = 2$ e o número de fatores ajustados é $K = 3$.

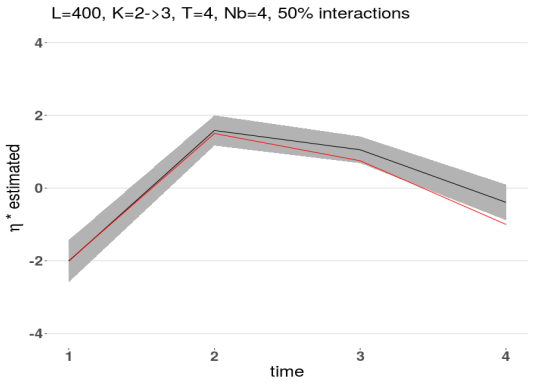


Figure 5.30: Média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha). Considere o caso Poisson: $M_{L_{400}T_4V_4}^{K_2I_{50\%}}$ com $\approx$ 40% de contagens zero, mas com ajuste de $K = 3$.

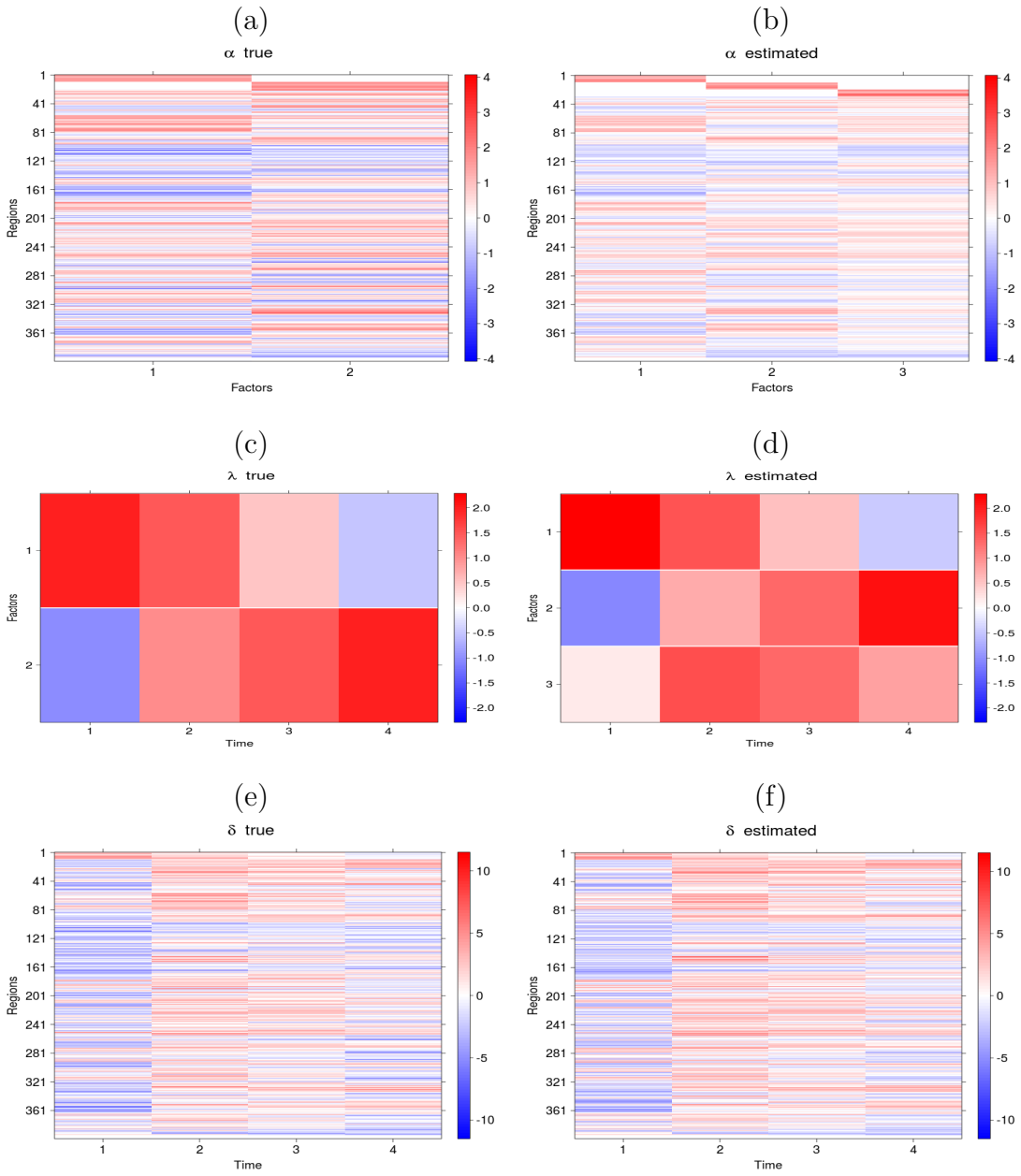


Figure 5.31: Mapas de calor comparando valores verdadeiros e estimados para o caso Poisson $M^{K_2K_3I_{50\%}}_{L_{400}T_4V_4}$ com $\approx 40\%$ de contagens zero, em que o sobrescrito "K_2K_3" significa que o número de fatores verdadeiro é $K = 2$ e o número de fatores ajustados é $K = 3$. Paineis à esquerda se referem aos valores verdadeiros e os à direita aos estimados.

valores reais parece ter sido quebrada em duas partes dentro da matriz estimada, sendo que uma delas foi incorporada ao Fator 2 e a outra ao Fator 3. Em geral, observe que o padrão principal exibido pelas colunas 1 e 2 da matriz verdadeira de α (Painel a) é mais fortemente capturado, respectivamente, pelas colunas 1 e 2 da matriz estimada (Painel b). A coluna referente ao Fator 3 da matriz estimada de α contém elementos tanto dos *loadings* da coluna 1 quanto da coluna 2 do α verdadeiro. Avaliando os Paineis (c) e (d) perceba que os escores estimados do Fator 1 seguiram o padrão verdadeiro de crescimento. É possível notar também que o padrão verdadeiro dos escores do Fator 2 (segunda linha da matriz do Painel c) foi melhor capturado pelo Fator 2 estimado (segunda linha da matriz do Painel d). O Fator 3 colocado no ajuste, mas que na verdade não existia na geração dos dados artificiais, será, desnecessariamente, interpretado pelo analista em sua análise com $K = 3$ fatores. Veja que esse fator possui um padrão de crescimento e depois, decrescimento, com os maiores escores ocorrendo nos Tempos 2 e 3. Apesar disso, os Fatores 1 e 2 estimados fornecerão conclusões compatíveis com os dois únicos fatores existentes na estrutura de dados verdadeira. Apesar da presença de um terceiro fator que inexistia, as estimativas para δ, Paineis (e) e (f), seguiram bem o padrão global para todos os tempos. Isso nos mostra que o erro na especificação de um fator a mais do que o verdadeiro não traz prejuízo para a estimação de δ, levando os coeficientes da regressão a serem bem estimados (veja Tabela 5.8).

No próximo tópico, analisamos a situação inversa em que o número verdadeiro de fatores é maior do que o número ajustado. Esclarecemos ao leitor que, devido ao fato do número de parâmetros verdadeiros ser diferente do número ajustado, nos tópicos desta seção não apresentamos os gráficos com intervalo HPD de 95%, pois tal análise intervalar não faz sentido nessas situações.

Erro de especificação em que $K_V = 3$ e $K_A = 2$

Neste tópico nosso estudo é direcionado para o caso com 3 fatores verdadeiros, mas o modelo é ajustado para 2 fatores. Pela Tabela 5.9, verificamos que os valores para β_0, σ^2 e η_3^* ficaram fora do intervalo HPD de 95%. Os desvios padrão para β_1 e β_2, como no caso anterior, foram os menores, indicando menor incerteza *a posteriori* para a

estimação desses coeficientes. O desvio padrão de τ_α, juntamente com os desvios padrão de η^*, apontam uma maior incerteza *a posteriori* sobre esses parâmetros. Apesar da maior incerteza, a Figura 5.32 mostra que o padrão de crescimento de η^* foi capturado, ficando fora do intervalo HPD de 95% (área sombreada) apenas para η^*_3.

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.500	0.343	0.348	0.013	0.314	0.360
β_1	-1.000	-1.006	-1.006	0.004	-1.014	-0.999
β_2	1.000	0.996	0.996	0.004	0.989	1.003
σ^2	0.800	1.267	1.264	0.091	1.097	1.445
τ_α	2.000	1.585	1.532	0.416	0.883	2.445
η^*_1	-2.000	-1.884	-1.894	0.280	-2.421	-1.318
η^*_2	-1.500	-1.401	-1.399	0.263	-1.915	-0.885
η^*_3	0.750	1.548	1.547	0.206	1.148	1.954
η^*_4	1.000	1.447	1.449	0.238	0.968	1.903

Tabela 5.9: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o caso Poisson: $M^{K_3K_2I_{50\%}}_{L_{400}T_4V_4}$ com $\approx 40\%$ de contagens zero, em que o sobrescrito "K_3K_2" significa que o número de fatores verdadeiro é $K = 3$ e que o número de fatores ajustados é $K = 2$.

Pela Figura 5.33 verificamos que o modelo teve mais dificuldade em capturar o padrão de α (Paineis a, b) e de λ (Paineis c, d) do que na configuração anterior ($K_V = 2$ e $K_A = 3$). No caso de α, o padrão de *loadings* da coluna 1 foi satisfatoriamente capturado para muitos locais, entretando é difícil encontrar alguma similaridade entre o padrão da segunda coluna de cargas estimadas com as colunas 2 e 3 da matriz verdadeira. A matriz estimada dos escores λ mostra uma configuração que reafirma a interpretação dada sobre os padrões das colunas de α. Veja que a primeira linha de λ estimado tem uma certa similaridade com a primeira linha da matriz verdadeira. Por outro lado, as linhas 2 e 3 de λ estimado não fornecem, individualmente, alguma similaridade clara com o padrão na linha 2 de λ verdadeiro. Apesar das estimativas para α e λ terem sido piores se

comparadas ao caso anterior, o padrão global de δ foi bem capturado, conforme ilustram os Paineis (e, f). Esse é um resultado importante, por nos indicar que se o analista ajustar o modelo com um K maior, ainda conseguirá estimar razoavelmente os efeitos em δ.

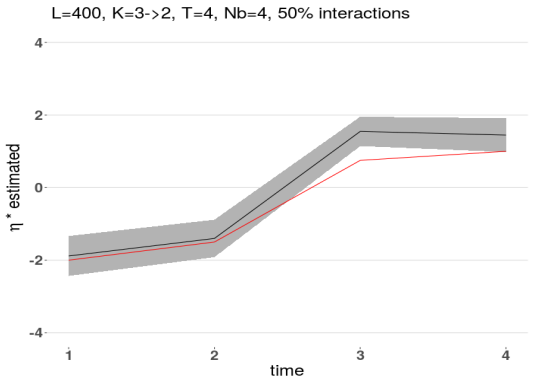


Figure 5.32: Média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha). Considere o caso Poisson: $M^{K_3 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx$ 40% de contagens zero, mas com ajuste de $K = 2$.

Terminamos aqui, a análise do erro de especificação de K. Da mesma forma que no caso logístico, concluimos que com variações de 1 unidade, para mais ou para menos, do número de fatores, a estimação de δ é satisfatória, não compremetendo a análise dos efeitos das covariáveis. Na próxima seção investigaremos o comportamento do modelo Poisson diante de uma sobreparametrização imposta ao modelo fatorial. Esse problema configura-se quando o número de fatores ajustados é igual ou superior ao número de tempos. Dois cenários serão analisados: o primeiro se refere ao caso em que o número de fatores é igual ao número de tempos ($K = T$) e o segundo quando o número de fatores é maior do que a quantidade de tempos ($K > T$). Especificamente, iremos analisar os casos: "$K = 4$, $T = 4$" e "$K = 5$, $T = 4$".

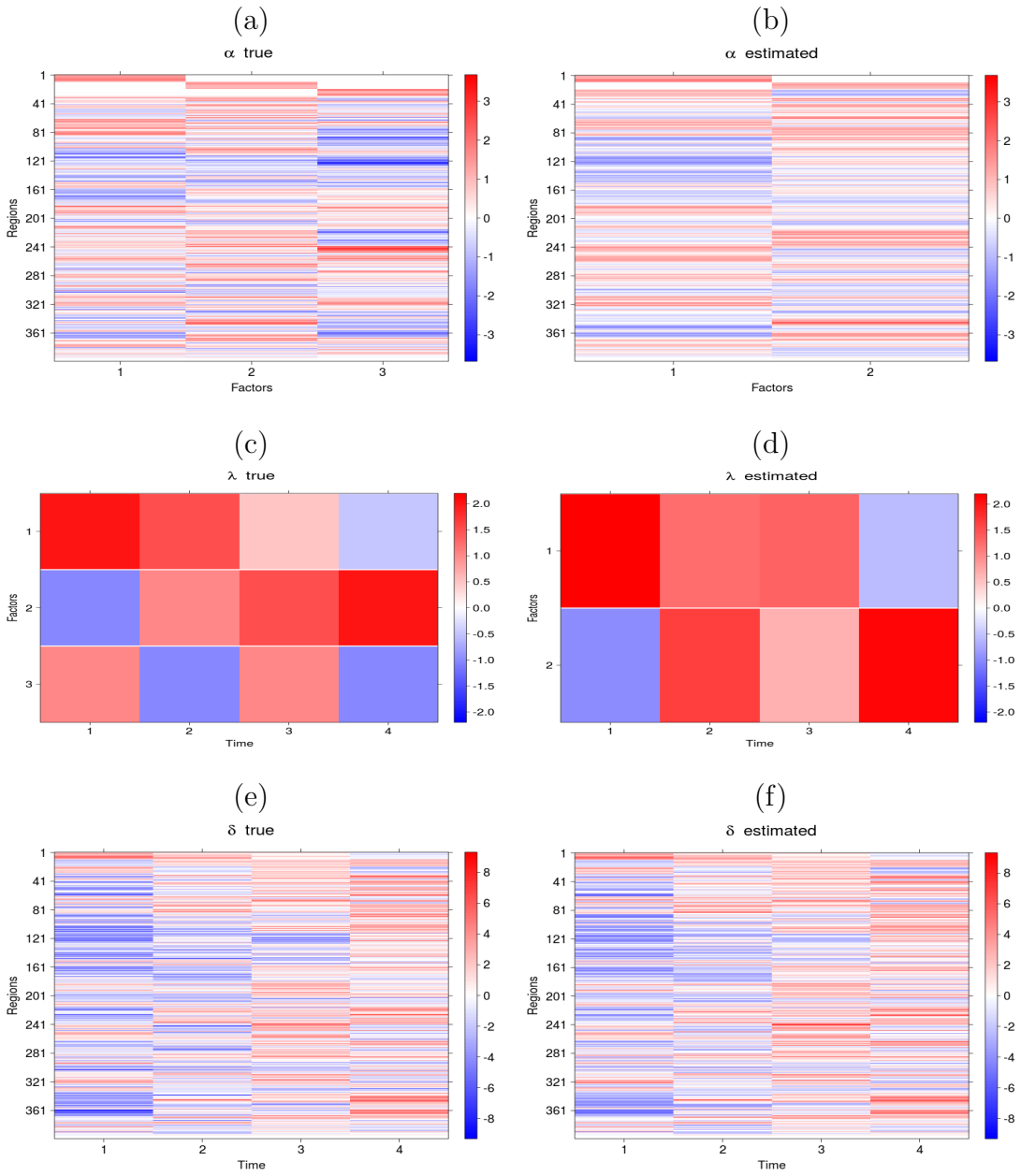


Figure 5.33: Mapas de calor comparando valores verdadeiros e estimados para o caso Poisson $M^{K_3K_2I_{50\%}}_{L_{400}T_4V_4}$ com $\approx 40\%$ de contagens zero. O sobrescrito "K_3K_2" significa que o número verdadeiro de fatores é $K = 3$ e que o número de fatores ajustados é $K = 2$. Paineis à esquerda se referem aos valores verdadeiros e os à direita aos estimados.

5.6 Sobreparametrização do modelo fatorial

Seguindo o estudo apresentado para o caso logístico, desenvolvemos, também, a análise do ajuste do modelo Poisson quando o número de fatores é igual ou maior do que o número de tempos, o que configura a sobreparametrização do modelo fatorial. Os mesmos dois casos foram considerados, a saber: "$K = 4$, $T = 4$" e "$K = 5$, $T = 4$". Para estas análises, consideramos apenas o cenário $\approx 40\%$ de contagens zero, pelo fato da configuração dos $\beta's$ na geração dos dados ter sido nesta situação a mesma do caso logístico.

Sobreparametrização em que $K = 4$ e $T = 4$

Mais uma vez temos a Tabela 5.10 com as estimativas dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α e da interação não linear η^*. Como tem ocorrido nos demais casos, os coeficientes em β apresentam menor desvio padrão, sendo eles bem próximos de zero, indicando menor incerteza *a posteriori* em relação aos demais. O valor verdadeiro de β_0 foi o único a ficar fora do intervalo HPD de 95% e, mais uma vez, os elementos em η^* apresentaram maiores desvios padrão em relação aos demais exibidos na tabela. E como também ocorreu nas outras situações, as estimativas de η^* seguiu o padrão verdadeiro (veja Figura 5.34).

Equivalente ao que ocorreu no modelo logístico, identificamos aqui que as estimativas de α e λ foram sobrestimadas ou subestimadas em alguns locais e fatores. No caso de locais, Painel (a) versus (b), podemos citar no Fator 1 do local ≈ 90 ao 110; no Fator 2 do local ≈ 170 ao 200, do ≈ 240 ao 260; no Fator 3 do ≈ 90 ao 150; e no Fator 4 do ≈ 130 ao 280. No que se refere a λ, podemos destacar as divergências ocorridas no Fator 1, Tempos 1, 2 e 4; no Fator 2, Tempo 3; no Fator 3, Tempo 2; e no Fator 4, Tempo 3. Essa análise pode ser melhor verificada pelo Painel (b) da Figura 5.36. Com relação à estimação de δ, podemos ver, pelos Paineis (e) e (f) da Figura 5.35, que ela foi satisfatória, identificando apenas, visualmente, divergências no Tempo 1 para faixas de locais entre os índices ≈ 110 e 180, em que verificamos tonalidades de azul mais fracas no Painel (f) em relação ao Painel (e).

Comparando os paineis da Figura 5.36 com os equivalentes dos cenários com $K = 2$ e

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.500	0.382	0.383	0.027	0.337	0.426
β_1	-1.000	-1.006	-1.006	0.006	-1.017	-0.994
β_2	1.000	0.995	0.995	0.005	0.985	1.005
σ^2	0.800	0.677	0.671	0.093	0.497	0.853
τ_α	1.000	1.966	1.921	0.488	1.023	2.929
η_1^*	-2.000	-1.762	-1.778	0.316	-2.385	-1.105
η_2^*	1.500	1.522	1.529	0.221	1.073	1.942
η_3^*	-0.750	-0.628	-0.653	0.265	-1.121	-0.062
η_4^*	1.000	0.603	0.593	0.330	-0.007	1.270

Tabela 5.10: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o caso Poisson: $M^{K_4 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 40\%$ de contagens zero.

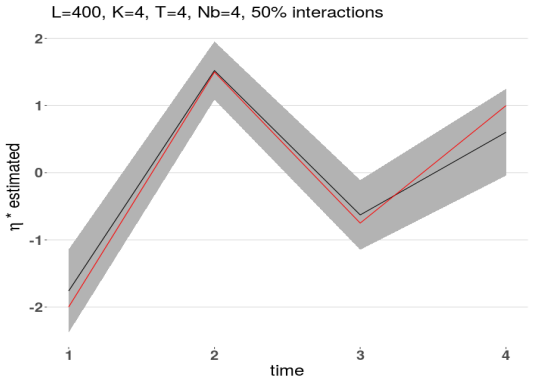


Figure 5.34: Média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha). Considere o caso Poisson: $M^{K_4 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 40\%$ de contagens zero.

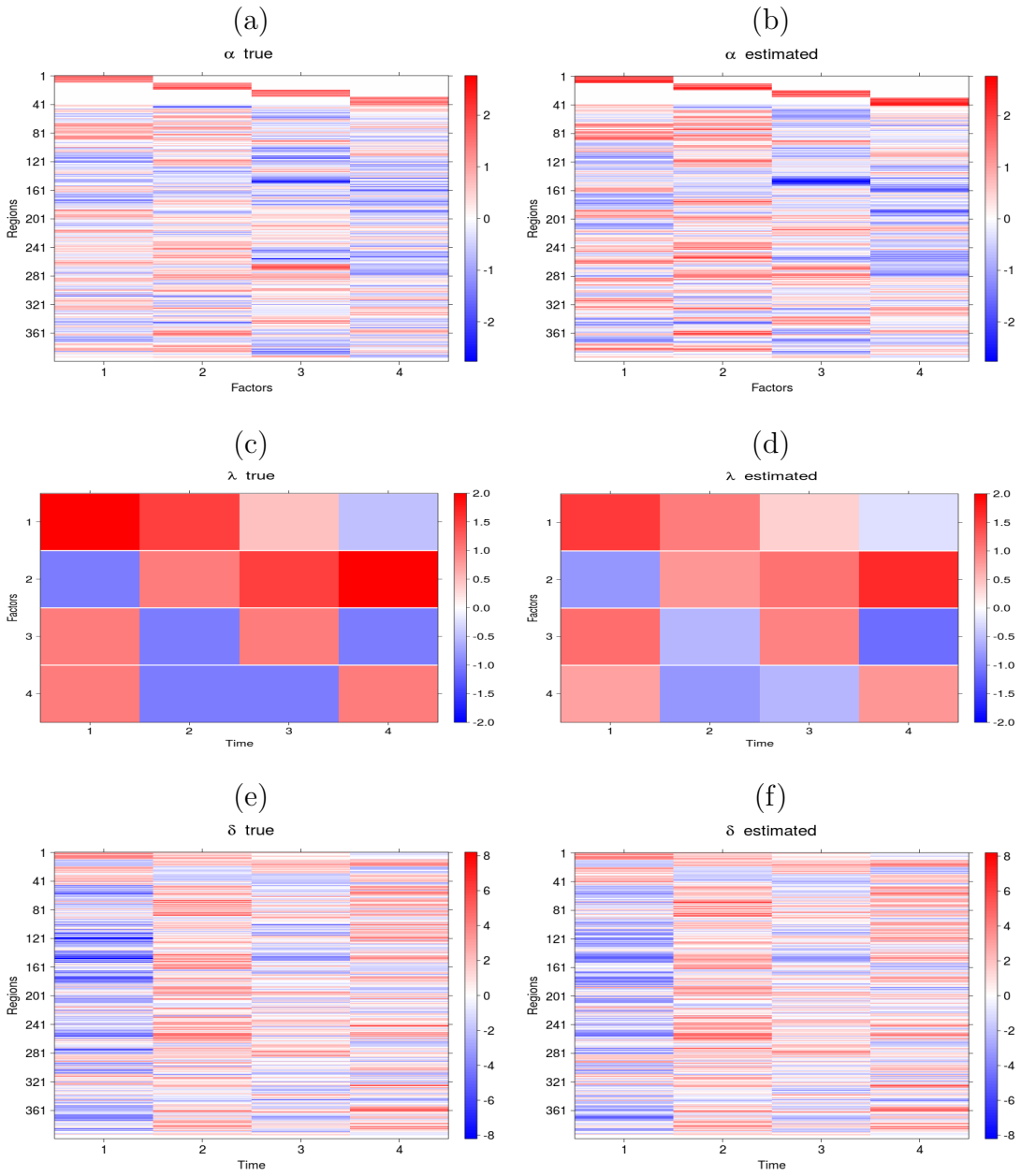


Figure 5.35: Mapas de calor comparando valores verdadeiros e estimados para o caso Poisson $M^{K_4 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 40\%$ de contagens zero. Paineis (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

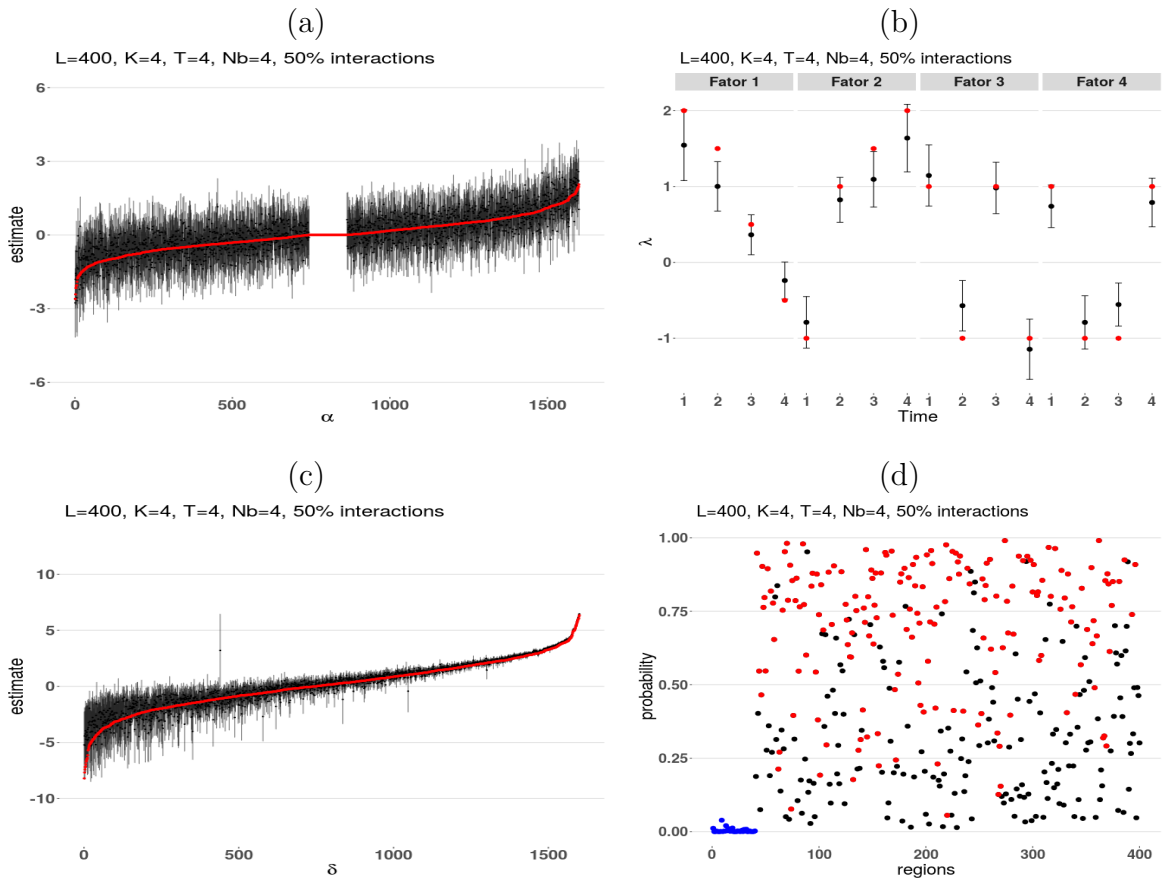


Figure 5.36: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2 , a cor vermelha representa locais de G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o caso Poisson: $M^{K_4 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 40\%$ de contagens zero.

$T = 4$, Figuras 5.12 e 5.16, vemos que: os intervalos HPD de 95% para α e δ são maiores, ver Paineis (a) e (c); as estimativas de λ são mais distantes do valor verdadeiro, veja isso no Painel (b) e, mais locias verdadeiramente afetados por interação (pontos vermelhos) foram detectados com probabilidade de interação abaixo de 0.5, veja isso no Painel (d). Para esse último caso, calculando a proporção $\text{prop}_{<0.5}$ (veja último parágrafo desse tópico para caso logístico na Seção 4.6) vemos que ocorre um aumento de 3.5% (média de $\text{prop}_{<0.5}$ nos cenários $M^{K_2 I_{30\%}}_{L_{400} T_4 V_4}$ e $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$) para 9.25%.

Finalizamos este tópico concluindo que apesar do modelo ter gerado estimativas satisfatórias, quando ocorre a sobreparametrização do modelo fatorial com $K = 4$ e $T = 4$, experimentamos uma perda no ajuste de α, λ e na probabilidade *a posteriori* dos locais serem afetados por interação, podendo comprometer um dos principais objetivos da proposta de modelo desta tese que é a identificação de conglomerados de locais baseado em fatores latentes a partir de dados que variam com o tempo.

Sobreparametrização em que $K = 5$ e $T = 4$

Mantendo o alinhamento das análises do modelo Poisson em relação ao modelo logístico, neste tópico descrevemos os resultados da sobreparametrização do modelo fatorial quando $K = 5$ e $T = 4$, ou seja, $K > T$. As mesmas observações do tópico anterior referentes às estimativas dos elementos constantes na Tabela 5.10 valem para as estimativas da Tabela 5.11. Uma diferença razoável entre esses dois casos está nas estimativas de η^*. Apesar do padrão verdadeiro ter sido capturado, aqui as estimativas estão bem mais distantes do valor verdadeiro, o que pode ser melhor visualizado pela Figura 5.37.

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.500	0.394	0.391	0.016	0.367	0.430
β_1	-1.000	-1.002	-1.002	0.006	-1.013	-0.991
β_2	1.000	0.996	0.996	0.005	0.987	1.006
σ^2	0.800	0.737	0.733	0.093	0.558	0.912
τ_α	1.000	1.869	1.789	0.590	0.948	2.976
η_1^*	2.000	1.652	1.661	0.331	1.003	2.314
η_2^*	-1.500	-0.986	-0.993	0.267	-1.502	-0.449
η_3^*	-0.750	-0.352	-0.355	0.253	-0.845	0.155
η_4^*	1.000	1.286	1.295	0.262	0.767	1.792

Tabela 5.11: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o caso Poisson: $M_{L_{400}T_4V_4}^{K_5 I_{50\%}}$ com $\approx 40\%$ de contagens zeros.

Novamente vemos, pela Figura 5.38, que existem divergências nas estimativas de α, Painel (a) versus (b), para vários locais nos 5 fatores. Analisando o Painel (c) versus o (d), também ocorrem divergências entre os 5 fatores nos 4 tempos, semelhante ao tópico anterior. Comparando o Painel (a) da Figura 5.36, do tópico anterior, com seu equivalente na Figura 5.39, parece haver mais sobrestimação e subestimação nos valores extremos

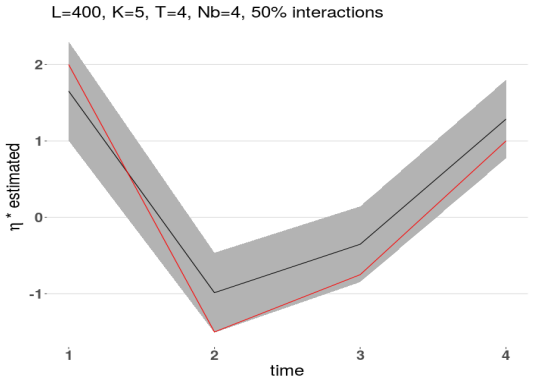


Figure 5.37: Média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha). Considere o caso Poisson: $M^{K_5 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx$ 40% de contagens zero.

negativos e positivos, respectivamente, além de intervalos HPD de 95% mais longos. O mesmo fenômeno de se ter mais sobrestimação nos extremos negativos, para o caso $K > T$, pode ser percebido nas estimativas de δ ilustradas pelo Painel (c) das mesmas figuras citadas. Finalmente, comparando o Painel (d) das duas figuras, visualmente elas são muito parecidas. De fato a diferença é de apenas 8 locais a menos no caso atual, sendo a Figura 5.36 (d) determinando $\text{prop}_{<0.5} = 9.25\%$ e a Figura 5.39 (d) fornecendo $\text{prop}_{<0.5} = 7.25\%$.

Finalizamos esta seção com a mesma conclusão obtida para o modelo logístico de que a configuração $K \geq T$ traz alguns prejuízos na parte de inferência e pode causar dificuldades na identificação de locais associados a fatores latentes e na construção de conglomerados formados por locais afetados por efeitos principais e interação não linear. A próxima seção é dedicada à análise de resíduos de Pearson com o objetivo de compararmos o ajuste do modelo em diversos cenários considerados para o caso Poisson.

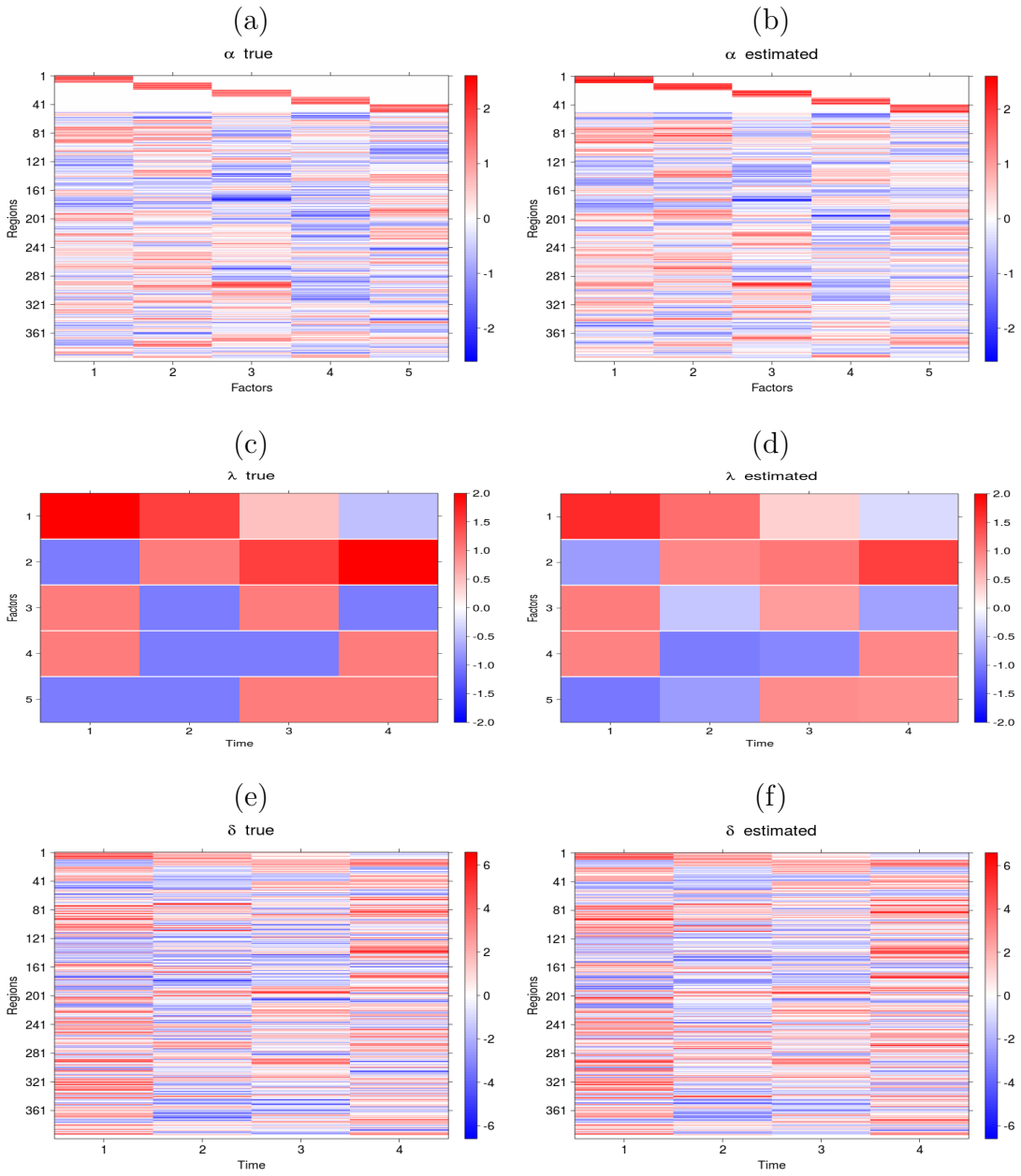


Figure 5.38: Mapas de calor comparando valores verdadeiros e estimados para o caso Poisson $M^{K_4 I_{50\%}}_{L_{400} T_5 V_4}$ com $\approx 40\%$ de contagens zero. Paineis (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

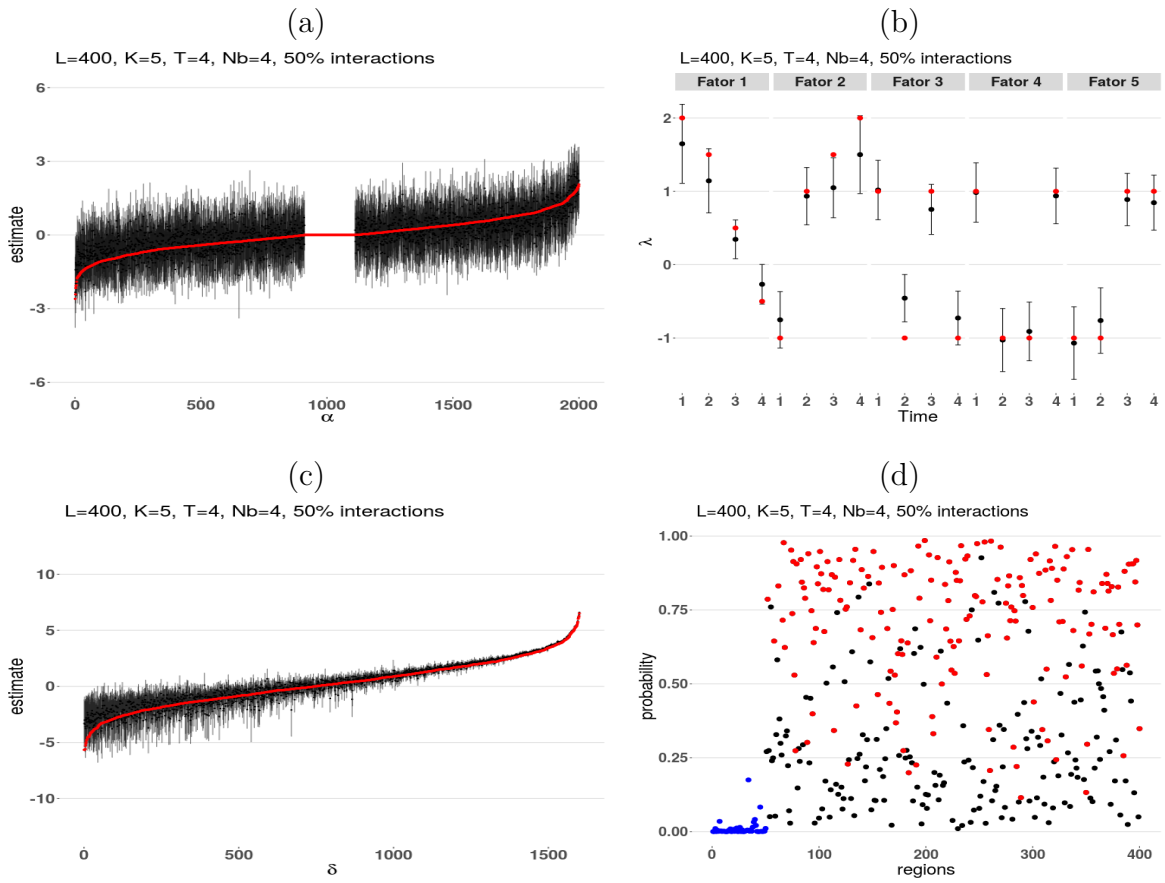


Figure 5.39: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais de G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o caso Poisson: $M^{K_4 I_{50\%}}_{L_{400} T_5 V_4}$ com $\approx 40\%$ de contagens zero.

5.7 Análise de resíduos

A análise dos resíduos é comumente utilizada para avaliar a adequação de modelos de regressão. Os resíduos são usados para identificar discrepâncias dos dados em relação aos valores estimados da variável resposta. Eles desempenham um papel muito importante na validação do ajuste do modelo (Cordeiro e Simas, 2009). Seguindo o procedimento realizado para o modelo logístico, nesta seção, estudamos os resíduos de Pearson para os cenários com $L = 100,\ 200$ e 400 locais, $K = 2$ fatores, $T = 4$ tempos, 4 e 6 vizinhos por região e $\approx 40\%$ de contagens zero em que temos $\approx 30\%$ e 50% de locais afetados por interação. O cenário com $\approx 3\%$ de contagens zero também foi considerado para a configuração $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$. Em termos de variação do número de fatores, incluimos a análise dos resíduos para $K = 3$, ou seja, $M^{K_3 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 40\%$ de contagens zero. Em todos os casos o número de fatores verdadeiros é igual à quantidade de fatores ajustados. No caso Poisson, os resíduos de Pearson são obtidos pela formulação $R_i = \frac{Y_i - \hat{\theta}_i}{\sqrt{\hat{\theta}_i}}, i \in \{1, ..., n\}$, em que Y_i é o valor da variável resposta para o indivíduo i e $\hat{\theta}_i$ representa o valor estimado de θ_i. A formulação para θ_i está definida na Equação (3.13). Para efetuarmos uma análise comparativa entre os cenários, calculamos a média quadrática dos resíduos, ou seja, $\sum_{i=1}^{n} = R_i^2/n$, em que n é número total de observações.

A Tabela 5.12 apresenta a média quadrática dos resíduos de Pearson para os cenários citados. Conforme destacado em alguns trechos desta tese, pode-se verificar, novamente, a similaridade dos resultados entre os cenários com 4 e 6 vizinhos por região, em que os resíduos para esses casos são muito próximos.

Seguindo a linha de estudo aplicada ao caso logístico, perceba que também, aqui, os resíduos para os cenários com 50% de locais afetados por interação são inferiores aos correspondentes cenários com 30%, exceto para $M^{K_2}_{L_{100} T_4}$ e $M^{K_2}_{L_{200} T_4}$ com 4 e 6 vizinhos por região, respectivamente. Esse resultado reforça o fato de que ter mais locais afetados pelo efeito η^* resulta em uma melhor estimação deste efeito de interação, influenciando na estimação de δ e, consequentemente, na estimação dos coeficientes β da regressão. Dentre os casos $\approx 40\%$ de contagens zero, $K = 2$ fatores e 6 vizinhos por região, coluna "Resíduo (6 viz.)", a média quadrática dos resíduos para o cenário $M^{K_2 I_{50\%}}_{L_{400} T_4}$ é menor do

Cenário	Num. Parâmetros	% Int.	Contagens zero	Resíduo (4 viz.)	Resíduo (6 viz.)
$M^{K_2}_{L_{100}T_4}$	317	$\approx 30\%$	$\approx 40\%$	0.8140	0.8337
		$\approx 50\%$		0.8543	0.8298
$M^{K_2}_{L_{200}T_4}$	617	$\approx 30\%$	$\approx 40\%$	0.8103	0.8317
		$\approx 50\%$		0.7993	0.8333
$M^{K_2}_{L_{400}T_4}$	1217	$\approx 30\%$	$\approx 40\%$	0.8260	0.8330
		$\approx 50\%$		0.8161	0.8078
$M^{K_3}_{L_{400}T_4}$	1621	$\approx 50\%$	$\approx 40\%$	0.8196	
$M^{K_2}_{L_{400}T_4}$	1217	$\approx 50\%$	$\approx 3\%$	0.8944	

Tabela 5.12: Média quadrática dos resíduos para os casos Poisson $M^{K_2}_{L_{100}T_4}$, $M^{K_2}_{L_{200}T_4}$, $M^{K_2}_{L_{400}T_4}$, $M^{K_3}_{L_{400}T_4}$ em diversas configurações: 4 e 6 vizinhos por região, $\approx 40\%$ e 3% de contagens zero, $\approx 30\%$ e 50% de locais de G_E afetados por interação não linear.

que todos os demais da mesma coluna e, pela coluna "Num. Parâmetros" podemos ver que o número de parâmetros é maior. O mesmo não acontece com esse cenário para os casos com 4 vizinhos, mas veja que a média quadrática residual, nesse caso, foi menor do que a do cenário $M^{K_2 I_{50\%}}_{L_{100}T_4V_4}$, destacando, ainda, a grande diferença do número de parâmetros entre esses cenários (1217 versus 317).

A Figura 5.40 ilustra a evolução da média quadrática do resíduo de Pearson (eixo vertical) com o aumento no número de parâmetros (eixo horizontal). Na análise da Tabela 5.12, coluna "Resíduo (6 viz.)", é fácil ver que a média quadrática dos resíduos para os cenários com 50% de locais afetados por interação são menores ou iguais, se considerarmos apenas 2 casas decimais para $L = 200$, do que os casos com 30%. Isso ocorre também para os casos com mais parâmetros a serem estimados. Veja como a média quadrática dos resíduos para a situação com maior número de parâmetros ($L = 400$) é menor do que aquela no cenário com menos parâmetros ($L = 100$). Tal fato reforça que, mantido as demais configurações constantes ($K = 2$ e $T = 4$), o aumento na quantidade de dados (aumento no número de locais) é maior do que o aumento no número de parâmetros e, consequentemente, contribui para uma melhor estimação *a posteriori*.

Finalizando este capítulo, a Seção 5.8, apresentada a seguir, compara os vícios relativos

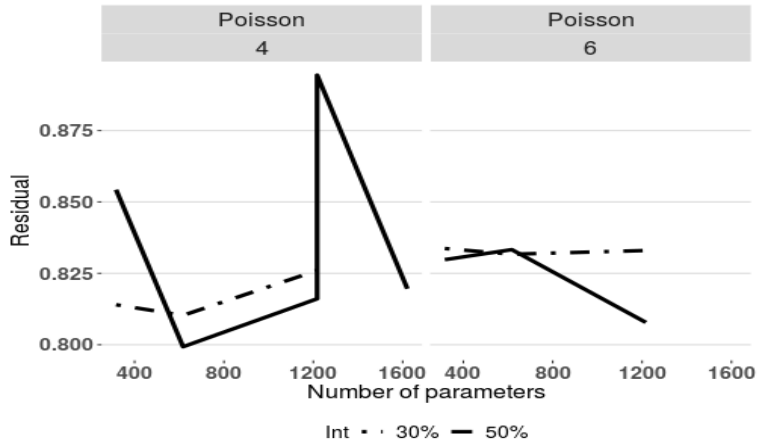


Figure 5.40: Média quadrática dos resíduos de Pearson por número de parâmetros considerando os cenários $M^{K_2}_{L_{100}T_4}$, $M^{K_2}_{L_{200}T_4}$, $M^{K_2}_{L_{400}T_4}$, $M^{K_3}_{L_{400}T_4}$ em diversas configurações: 4 e 6 vizinhos por região, $\approx 3\%$ ou 40% de contagens zero, $\approx 30\%$ e 50% de locais de G_E afetados por interação não linear, ou seja, os mesmos cenários listados na Tabela 5.12.

de todos os cenários explorando dados artificiais configurados com 4 tempos, 2 fatores e 4 vizinhos após a execução de 30 réplicas Monte Carlo. Diferentemente do caso logístico, em que analisamos um cenário artificial semelhante ao dos dados reais (último *boxplot* dos Paineis a, c, e, da Figura 4.37) não foi possível incluirmos o vício relativo de um cenário que simula uma aplicação real em que a variável resposta é uma contagem. O fato de não termos uma aplicação real motivadora não permitiu que identificássemos valores de referência, para os coeficientes da regressão, que direcionaria a geração de dados artificiais.

5.8 Análise de Monte Carlo

Todas as análises descritas nas seções anteriores, deste capítulo, foram baseadas em apenas uma réplica de Monte Carlo. Para que possamos ter uma ideia mais ampla sobre o comportamento do modelo é necessário que os ajustes sejam efetuados para vários bancos de dados gerados sob as mesmas condições. Esta seção é baseada na utilização do esquema Monte Carlo com 30 base de dados. Consideramos que 30 réplicas é uma quantidade satisfatória e viável, em termos de tempo computacional, para este propósito de um estudo mais abrangente. Concentramos, aqui, apenas na avaliação dos parâmetros α, λ e δ, por terem papel de maior destaque na comparação dos ajustes, uma vez que eles incorporam as estruturas espacial e temporal do modelo. Além disso, as matrizes associadas a esses parâmetros incorporam maior porcentagem do total de elementos a serem estimados no ajuste do modelo. O objetivo deste estudo é avaliar o impacto no ajuste do modelo quando temos diferentes quantidades de dados e parâmetros. Para isso, alguns elementos foram mantidos fixos. Especificamente, variamos os números de locais ($L = 100, 200$ e 400) e o percentual de regiões afetadas por interação ($\approx 30\%$ e 50%). Os elementos mantidos fixos são: $K = 2$ fatores, $T = 4$ tempos, 4 vizinhos por região e $\approx 40\%$ de contagens zero. Novamente não analisamos os casos com 6 vizinhos por região, por apresentarem resultados semelhantes aos de 4 vizinhos.

Equivalente ao caso logístico, analisamos o vício relativo, cuja formulação se encontra descrita na Seção 4.9. Para α e δ, em cada réplica, foi feita aleatoriamente e, sem reposição, uma seleção de 100 valores internos dessas matrizes, sendo que, para α consideramos apenas estimativas relacionadas a locais de G_E. O conjunto total de observações de α, que foi considerado na seleção da amostra de tamanho 100 em cada réplica, envolveu todos os fatores ($K = 2$), e de δ, todos os tempos ($T = 4$). Ou seja, a amostra de cada réplica de Monte Carlo poderia ter $\alpha^{'}s$ dos 2 fatores e $\delta^{'}s$ dos 4 tempos. Diferentemente de α e δ, os vícios relativos e a mediana *a posteriori* de λ foram calculados considerando toda a matriz de escores estimada.

A Figura 5.41 mostra o resultado do cálculo do vício relativo de α, Paineis (a) e (b); λ, Paineis (c) e (d), e δ, Paineis (e) e (f). Os paineis da esquerda apresentam uma visão

da dispersão e variabilidade das estimativas e os paineis da direita mostram de forma mais simplificada e clara o valor da mediana dos vícios nos diversos cenários. Analisando os Paineis (a) e (b), vemos que, para os casos $L = 200$ e $L = 400$, os cenários com maior número de locais afetados por interação tiveram a mediana dos vícios das cargas mais próxima de zero (linha horizontal tracejada). A diferença entre o número de regiões afetadas por interação quando se tem poucos locais ($L = 100$), não foi suficiente para que houvesse uma estimação melhor de α para o caso com mais locais afetados. Lembrando ao leitor que 30% e 50% de 80 é igual 24 e 40 locais que contribuem para estimação de η^*, respectivamente.

Analisando os Paineis (c), (d), (e) e (f) vemos que não há muita diferença entre os vícios nos diversos cenários. Os intervalos interquartis são bem pequenos e as medianas muito próximas do vício zero (rente à linha tracejada horizontal), indicando que o ajuste do modelo foi satisfatório, além de semelhante, para todos os cenários. Finalmente, perceba que os Paineis (a), (c) e (e) estão na mesma escala sendo difícil julgar a amplitude dos *boxplots* na comparação de 30% versus 50%. Veja que os *boxplots* de α possuem o intervalo interquatílico maior, sugerindo mais variação do vício relativo para as cargas.

Os Paineis (b), (d) e (f) da Figura 5.41 ilustram com mais precisão o valor da mediana que foi calculada para α e δ considerando a seleção aleatória de 100 elementos da matriz estimada em cada réplica Monte Carlo, totalizando 3000 observações para construir o *boxplot*. Para λ o cálculo do vício foi efetuado considerando toda a matriz estimada em todas as réplicas, totalizando 240 elementos. Perceba que para δ, Painel (f), a mediana dos vícios relativos quando se tem 50% de locais afetados por interação é mais próxima de zero (linha tracejada) do que no caso 30%, em todos os cenários. O mesmo não acontece para α e λ. Para α, Painel (b), a mediana *a posteriori* dos vícios relativos fica mais próxima de zero, quando temos 50% de locais afetados por interação, nos casos $L = 200$ e $L = 400$ e, para λ isso acontece apenas quando $L = 200$. O resultado observado no Painel (f) indica que o maior número de locais afetados por η^* contribuiu para uma melhor estimação de δ. Conforme ocorreu no caso logístico, vemos que a variação no número de regiões afetadas pelo efeito η^* se mostrou relevante para o ajuste do modelo.

A Figura 5.42 ilustra o vício relativo de λ para cada fator em cada tempo. Os Paineis

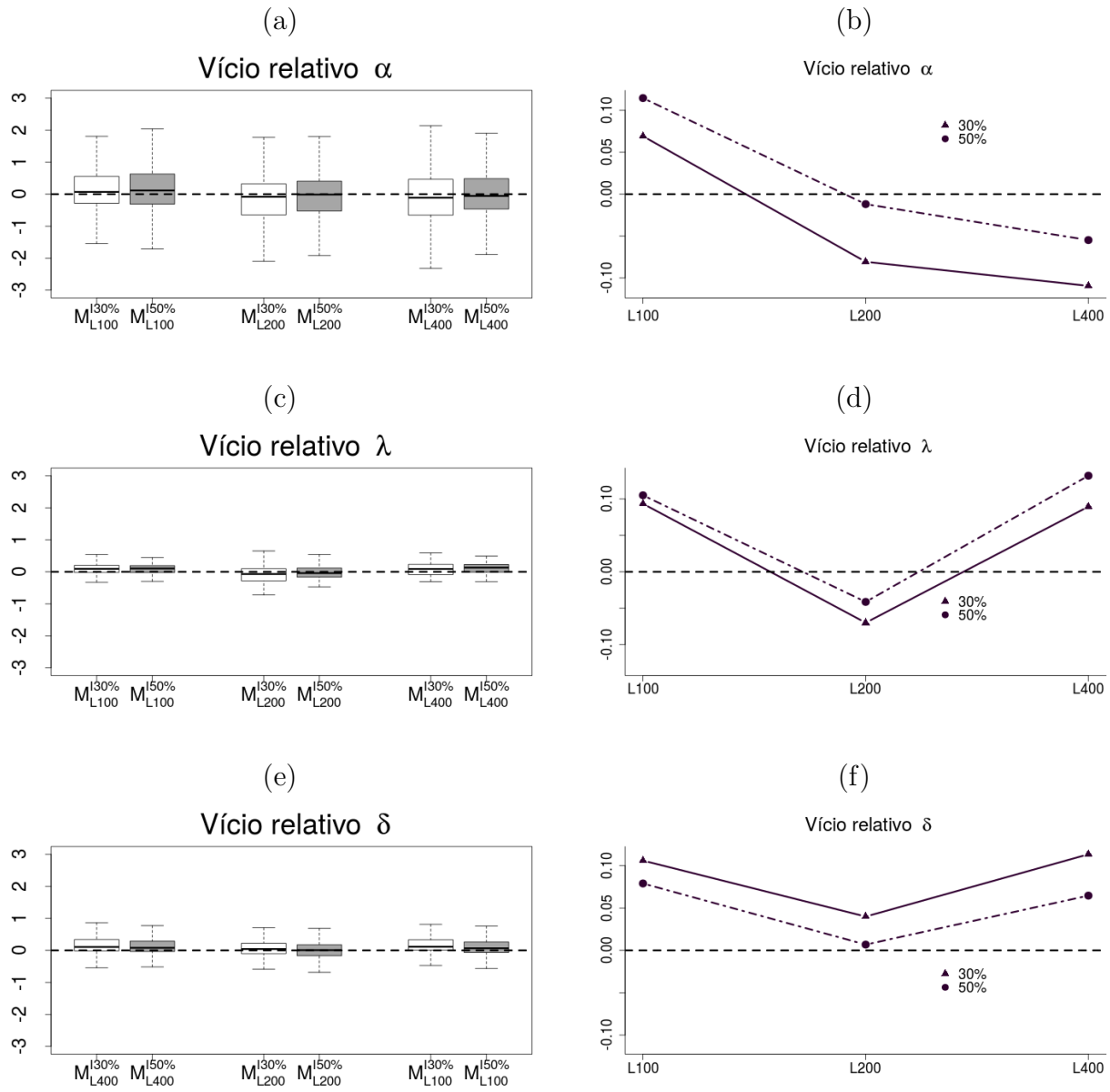


Figure 5.41: Gráficos do vício relativo, cuja formulação se encontra descrita na Seção 4.9. Análise baseada em 30 réplicas de um esquema Monte Carlo. Considere uma seleção aleatória de 100 elementos em α e em δ. Todos os escores de λ são explorados aqui. Estão considerados os casos Poisson com $K = 2$ fatores, $T = 4$ tempos, 4 vizinhos por região, número de locais $L \in \{100, 200, 400\}$, $\approx 30\%$ e 50% de locais de G_E afetados pela interação não linear. Os paineis da esquerda ilustram a dispersão e variabilidade dos vícios relativos e os paineis da direita os valores das medianas.

(a), (b), (c) e (d) se referem ao Fator 1 para os tempos 1, 2, 3 e 4, respectivamente, e os Paineis (e), (f), (g) e (h) são relativos ao Fator 2. Apenas para o Fator 1, no Tempo 1, o padrão do vício em $L = 100, 200$ e 400 difere entre os casos 30% e 50% de locais afetados por interação. Veja no Painel (f) o paralelismo entre as linhas indicando um padrão praticamente igual. A Tabela 5.13 resume a comparação da mediana dos vícios relativos entre os casos 30% e 50% de locais de G_E afetados por interação. Perceba que das 24 medianas estimadas para λ (2 fatores em 4 tempos e 3 configurações de locais), o caso 50% tem valor mais próximo de zero em 8 delas. O caso 30% obteve uma estimativa melhor em 7 e as medianas são praticamente iguais em 9 delas. Ou seja, de forma geral, quando se tem $\approx 50\%$ de locais afetados por interação parece não fazer diferença para a estimação de λ em comparação com o caso 30%.

Cenário	50% é melhor	30% é melhor	50% e 30% próximos
L_{100}	a, e	f, g	b, c, d, h
L_{200}	d, f, g, h	a, c	b, e
L_{400}	b, h	d, e, f	a, c, g
		Total de Paineis	
$L_{100,200,400}$	8	7	9

Tabela 5.13: Resumo dos paineis nos quais a porcentagem de locais de G_E afetados por interação obteve mediana dos vícios mais perto de zero para λ. Cada painel representa um tempo de um determinado fator. Considere os casos Poisson com $L = 100, 200$ e 400 locais, $K = 2$ fatores e $T = 4$ tempos, totalizando 24 medianas. A proximidade de zero da mediana do vício indica um desempenho melhor.

Analisando, agora, a Figura 5.43, vemos que a mediana *a posteriori* dos vícios relativos de δ, em todos os tempos, segue o mesmo padrão ao longo de $L = 100, 200$ e 400 para os casos com 30% e 50% de regiões afetadas pela interação. Perceba que as medianas para o caso 50% são consistentemente mais próximas de zero do que as medianas de 30% para todas as configurações de L e em todos os tempos. Diferente do que ocorre para λ, parece que ter mais regiões afetadas por η^* faz diferença na estimação de δ.

Encerramos, aqui, a seção referente à análise do vício relativo calculado com base

em 30 réplicas de Monte Carlo. Vimos que a estimação das cargas α não sofre grande influência quando se tem "muitos" ou "poucos" locais afetados por interação para o caso $L = 100$ (poucos locais). A estimação das cargas apresentou maior variação do vício relativo em comparação com a estimação de λ e δ. Identicamos que os vícios relativos dos escores dos fatores, λ, e do efeito aleatório, δ, obtiveram estimações semelhantes para os casos com 30% e 50% de locais afetados pela interação. Na análise das medianas *a posteriori* dos vícios relativos, verificamos que apenas na estimação de δ o caso 50% de locais afetados por interação obteve, em todos os cenários, resultados mais próximos de zero do que o caso 30%. Concluimos que, também para o caso Poisson, o maior número de locais afetados pela interação não linear contribui para uma melhor estimação de δ.

Com esta seção, estamos finalizando também, o capítulo referente ao estudo simulado Poisson. No próximo capítulo apresentaremos um estudo realizado a partir de dados que nunca foram analisados em um estudo científico, com modelagem estatística, em outros trabalhos na literatura. Esses dados foram coletados pelo sistema de telediagnóstico do Centro de Telessaúde do Hospital das Clínicas da UFMG e organizados por uma equipe multidisciplinar composta por cientistas da computação, estatísticos e médicos cardiologistas. Tais dados são referentes a indivíduos residentes no estado de Minas Gerais que realizaram exames eletrocardiológicos através do Centro de Telessaúde e que sofreram infarto agudo do miocárdio (IAM). Através de uma base de mortalidade, mantida pelo Ministério da Saúde do Brasil, foi identificado qual desses pacientes veio a falecer. Essa informação define a variável resposta de nosso modelo que é o indicativo de morte ($Y_i = 1$) ou não ($Y_i = 0$). As variáveis idade e sexo compõem o restante dos dados de cada indivíduo, que reside em um local e para o qual foi selecionado um exame cardiológico referente a um determinado ano.

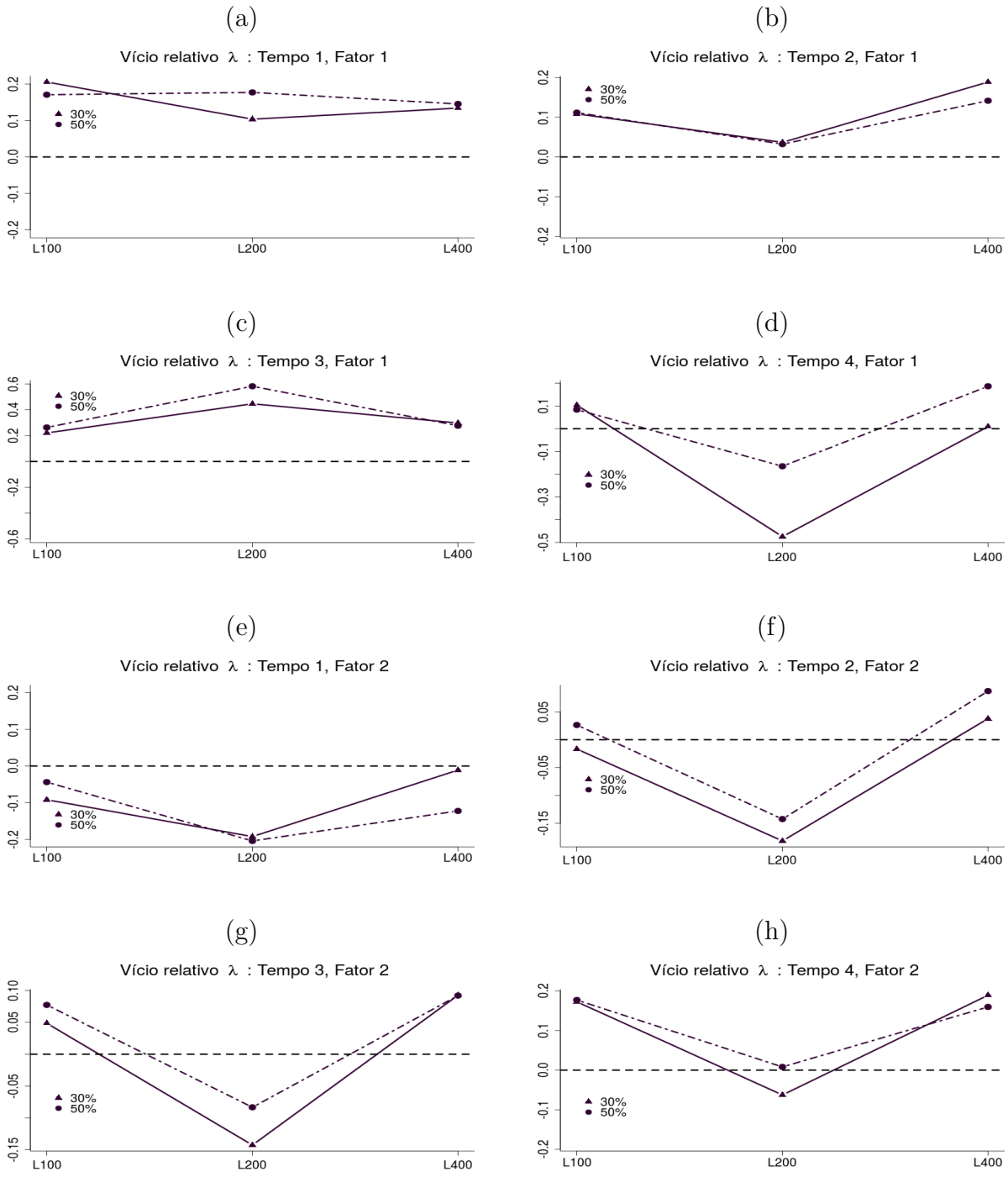


Figure 5.42: Medianas dos vícios relativos de λ para cada fator em cada tempo. A formulação do vício se encontra descrita na Seção 4.9. Os cálculos foram baseados nas amostras completas das 30 réplicas de Monte Carlo. Considere os casos Poisson com $K = 2$ fatores, $T = 4$ tempos, 4 vizinhos por região, número de locais $L \in \{100, 200, 400\}$, $\approx 30\%$ e 50% de locais de G_E afetados por η^*.

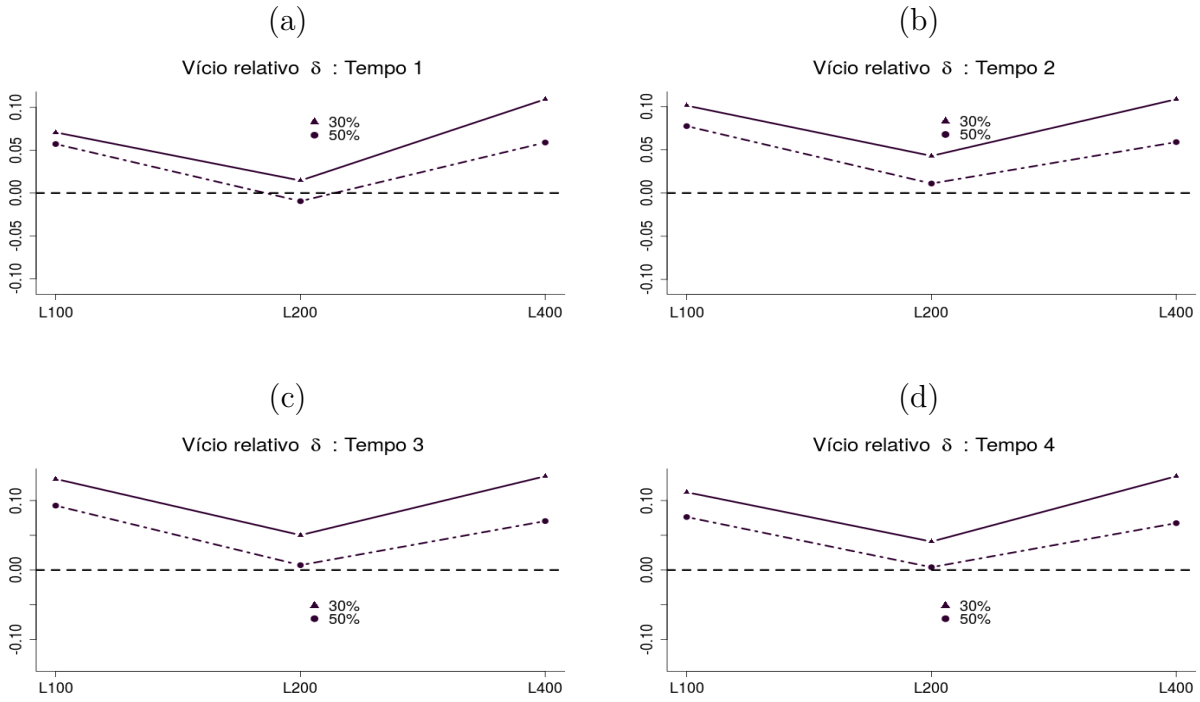


Figure 5.43: Medianas dos vícios relativos de δ para cada tempo. Em contraste com o Painel (f) da Figura 5.41, nos casos analisados aqui, foram selecionados 100 locais aleatoriamente em cada uma das 30 réplicas de Monte Carlo para que o tamanho das amostras de cada tempo fosse a mesma. A fórmula de cálculo do vício relativo está descrita na Seção 4.9. Considere os cenários com $K = 2$ fatores, $T = 4$ tempos, 4 vizinhos por região, número de locais $L \in \{100, 200, 400\}$, $\approx 30\%$ e 50% de locais de G_E afetados por η^*.

Capítulo 6

Análise de dados reais

Origem dos dados

Conforme descrito no Capítulo 2, a motivação inicial para o desenvolvimento do presente trabalho surgiu da necessidade apresentada pelo professor Antônio Ribeiro, coordenador do Centro de Telessaúde do Hospital das Clínicas da UFMG (`https://telessaude.hc.ufmg.br`) e do grupo de pesquisa CODE (*Clinical Outcomes in Digital Electrocardiography*), de se verificar o impacto na saúde dos pacientes de municípios de Minas Gerais após a implantação do sistema de telediagnóstico de eletrocardiogramas (ECGs). Uma das maneiras de se avaliar esse impacto, de acordo com Ribeiro, é a partir da análise do desfecho do paciente (morte ou não) após a ocorrência de infarto agudo do miocárdio (IAM). Para que se possa fazer essa análise é necessário efetuar um mapeamento, tecnicamente chamado de pareamento, entre a base de eletrocardiogramas e de mortalidade. O desfecho pode ser rastreado a partir de bases de dados de mortalidade mantidas pelo DATASUS (Departamento de Informática do Sistema Único de Saúde). Para se ter acesso a esses dados, em que se tem a identificação do indivíduo, é necessário efetuar uma solicitação formal ao Ministério da Saúde e de assinar um termo de confidencialidade. Para este estudo, a base de dados foi obtida a partir da Secretaria de Saúde do Estado de Minas Gerais, após a assinatura, por todos os integrantes da equipe do CODE, de um acordo de sigilo na manipulação dos dados. No site do DATASUS (`http://www2.datasus.gov.br/ DATASUS/index.php`) é possível baixar dados consolidados. O processo de preparação e organização da base de dados para a análise proposta envolve várias etapas que estão descritas no Capítulo 2 desta tese.

Organização dos dados e covariáveis

A base de dados existente possui o registro dos eletrocardiogramas de pacientes de 811 municípios além de várias informações pessoais e clínicas dos indivíduos. Os dados utilizados no ajuste do modelo são de 15835 pacientes que sofreram IAM, no período de 2013 a 2016, dos quais 1286 morreram. Os 811 municípios foram agrupados em 441 regiões de forma que todas elas possuem pelo menos um indivíduo para o qual o exame eletrocardiográfico registrou a ocorrência de IAM em algum dos anos do período citado. Cada indivíduo da base possui apenas uma ocorrência que foi registrada em um determinado ano. Caso exista mais de um ECG para o paciente, apenas o primeiro foi selecionado para compor a base de dados final, ou seja, a informação de cada indivíduo é relativa a apenas um exame em determinado local/ano. As covariáveis consideradas foram sexo (X_{2i}; 1 = Masculino, 0 = Feminino) e idade/100 (X_{3i}), por se tratarem das variáveis mais utilizadas neste tipo de estudo (Medeiros et al., 2018). A variável resposta Y_i é binária com o 1 indicando morte, para $i = 1, ..., 15835$. Maiores detalhes dessa base e sobre os procedimentos necessários para sua estruturação estão descritos no Capítulo 2.

Objetivos

Lembramos ao leitor de que os principais objetivos desse estudo são aprimorar a abordagem da análise fatorial assumindo a existência de efeito de interação entre os fatores, e determinar conglomerados de municípios com comportamentos semelhantes em relação ao atendimento de pacientes que sofreram IAM ao longo dos anos. Especificamente, estamos interessados em identificar se existem regiões que foram impactadas positivamente após a implantação do telediagnóstico de eletrocardiogramas e se há regiões que não sofreram qualquer efeito ou que experimentaram efeitos negativos depois da disponibilização desse serviço. Considerando que estamos tratando com dados geolocalizados em determinados anos, a análise desenvolvida é baseada em um modelo fatorial espaço-temporal com a inclusão de um termo que capta uma interação não linear entre os fatores. Considerando que a variável resposta é o indicativo de morte do indivíduo que sofreu um IAM, podemos acrescentar a estimação da probabilidade desse evento como mais um objetivo importante deste trabalho.

Configuração para identificabilidade do modelo

Conforme explicado e demonstrado anteriormente, o modelo requer a separação, *a priori*,

das regiões em grupos. A cada grupo deve ser atribuído um fator, com exceção de um grupo, denominado grupo extra (G_E), que não será forçado a se associar a qualquer dos fatores. A associação, *a priori*, de cada grupo de regiões a determinado fator é uma estratégia para resolver um problema de identificabilidade do modelo fatorial (veja a Seção 3.1). Conforme sugestão de especialistas da área da saúde, o índice de desenvolvimento humado (IDH) dos municípios de Minas Gerais é uma variável interessante a ser levada em conta para definir os fatores latentes a serem tratados neste estudo. O IDH é um índice disponível livremente na Internet. Neste trabalho, utilizamos o IDH de 2010 que pode ser encontrado no *site* do Atlas do Desenvolvimento Humano no Brasil (`http://www.atlasbrasil.org.br/2013/pt/ranking`) onde estão registrados 4 tipos de IDH: Geral, Educacional, Longevidade e Renda. Desenvolvemos análises prévias para a escolha do tipo de IDH a ser tratado nesta tese. Os resultados mais interessantes, do ponto de vista prático, são relacionados ao IDH Renda. Portanto, desenvolvemos as análises tomando como base essa opção.

Nesta aplicação, escolhemos trabalhar com dois grupos de regiões, ou seja, $K = 2$ fatores. Cada grupo foi composto por 10 regiões. Um grupo foi montado pelos municípios de maior IDH Renda que explicam o Fator 1 e o outro pelos de menor, que impactam o Fator 2. Outra alternativa para escolha do número de regiões para cada grupo seria a definição de limiares que idenficassem IDH's bons e ruins. Essa alternativa levaria a grupos com muitas regiões podendo perder a interpretabilidade dos fatores. De acordo com o professor Antônio Ribeiro, essa divisão em dois grupos (IDH alto e baixo) é a que faz mais sentido em uma análise prática. Incluir uma terceira categoria não seria tão atrativo. A Figura 6.1 ilustra o posicionamento das regiões selecionadas para compor os grupos G_1 e G_2. Perceba que os locais de pior IDH Renda estão localizados nas regiões Norte e Nordeste de Minas Gerais (vale do Jequitinhonha e Mucuri), ou seja, as regiões mais pobres do estado. Os locais de melhor IDH, por outro lado, estão situados nas regiões Central e Triângulo Mineiro.

Formação das regiões

Relebrando ao leitor o processo de união de municípios descrito no Capítulo 2, existem municípios que não possuem indivíduos que sofreram IAM em todos os anos. Com isso, foi necessário unir os municípios em regiões de forma que houvessem pacientes em todos os períodos selecionados em todas as regiões. As regiões foram formadas analisando os dados dos indivíduos por município, ano-a-ano. A cada ano aqueles municípios que não possuíam pacientes foram

unidos a outros, com os quais fazem fronteira e recursivamente, até que se obtivesse todas as regiões classificadas como válidas, ou seja, com dados. A seleção do município fronteiriço para se formar a região foi realizada de forma arbitrária, ou seja, não foi levado em consideração nenhuma característica específica do município, além do fato de fazer fronteira, para sua seleção. No final da análise de todos os anos, obtivemos uma única configuração de regiões para todos os períodos, totalizando 441 áreas, as quais são compostas por municípios vizinhos.

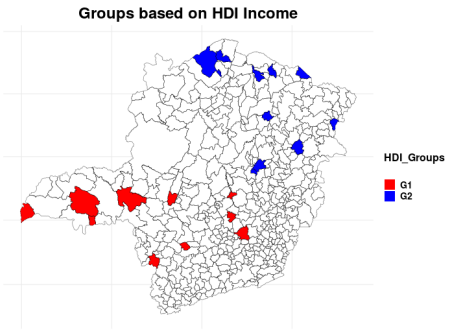


Figure 6.1: Disposição espacial dos grupos de IDH Renda 2010 por local. G_1 representa o grupo de 10 regiões com IDH's mais altos e G_2 o grupo com IDH's mais baixos.

Análise dos coeficientes em β

A Tabela 6.1 apresenta as estimativas dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α e da interação não linear η^*. Equivalente ao identificado nos cenários artificiais, vemos que a média e a mediana são muito próximas, indicando distribuições simétricas *a posteriori*. Veja como o desvio padrão do coeficiente da covariável sexo, β_1, e da variância do erro, σ^2, foram bem próximos de zero, indicando uma baixa incerteza *a posteriori*. Perceba como todos os elementos de η^* obtiveram desvios padrão menores que em todas as simulações do modelo logístico apresentadas no Capítulo 4 (Tabelas 4.8, 4.9, 4.10, 4.11, 4.12, 4.13, 4.14, 4.15, 4.16).

Com $\beta_2 = 5.11$ (coeficiente da covariável 'idade/100'), o modelo aponta que, se tomarmos

homens ou mulheres separadamente, quanto mais idosa for a pessoa que sofreu infarto, maior será a probabilidade de morte, o que faz sentido. A análise de $\beta_1 = 0.32$ (coeficiente da covariável 'sexo') indica que, considerando homens e mulheres da mesma idade e residentes no mesmo município, os homens tem mais chances de morrer após a detecção de IAM pelo exame eletrocardiográfico. Considere, por exemplo, homens e mulheres de 50 anos, indicando que $X_{3i} = 50/100 = 0.5$. Nessa situação, a diferença relativa (do masculino em relação ao feminino) entre as probabilidades de morte são 34.24%, 36.33% e 37.19% para residentes em cidades com o componente aleatório δ igual a 1.0, 0 e -1.0, respectivamente (veja a escala de δ no Painel (c) da Figura 6.3). Para maior esclarecimento do leitor, mostraremos, aqui, o cálculo do percentual de 34.24%, que se refere a $\delta = 1$. Dado que $\hat{\theta}_i = \frac{\exp\{X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}\}}{1+\exp\{X_{\bullet i}^\top \beta + \delta_{l_i^* t_i^*}\}}$, veja Equação (3.2), temos $\hat{\theta}_i = \frac{\exp\{\beta_0 X_{1i} + \beta_1 X_{2i} + \beta_2 X_{3i} + \delta_{l_i^* t_i^*}\}}{1+\exp\{\beta_0 * X_{1i} + \beta_1 X_{2i} + \beta_2 X_{3i} + \delta_{l_i^* t_i^*}\}} => \hat{\theta}_i^M = \frac{\exp\{-6.14+0.32+5.11\times 0.5+1\}}{1+\exp\{-6.14+0.32+5.11\times 0.5+1\}}$ $= 0.0941$ e $\hat{\theta}_i^F = \frac{\exp\{-6.14+5.11\times 0.5+1\}}{1+\exp\{-6.14+5.11\times 0.5+1\}} = 0.0701$, em que M = masculino, e F = feminino. A conta é finalizada por $100 \times (\hat{\theta}_i^M - \hat{\theta}_i^F)/|\hat{\theta}_i^F| = 100 \times (0.0941 - 0.0701)/|0.0701| = 34.24\%$. Note que se utilizarmos a formulação $100 \times (\hat{\theta}_i^M - \hat{\theta}_i^F)$ podemos dizer que a probabilidade de óbito do homem é $100 \times (0.0941 - 0.0701) = 2.4\%$ maior do que a da mulher. Para $\delta = 0$ teremos $\hat{\theta}_i^M = \frac{\exp\{-6.14+0.32+5.11\times 0.5\}}{1+\exp\{-6.14+0.32+5.11\times 0.5\}} = 0.037$ e $\hat{\theta}_i^F = \frac{\exp\{-6.14+5.11\times 0.5\}}{1+\exp\{-6.14+5.11\times 0.5\}} = 0.027$ o que gera uma estimativa de 1% a mais de probabilidade de falecimento dos homens em relação às mulheres. E para $\delta = -1$ chegamos aos valores $\hat{\theta}_i^M = \frac{\exp\{-6.14+0.32+5.11\times 0.5-1\}}{1+\exp\{-6.14+0.32+5.11\times 0.5-1\}} = 0.014$ e $\hat{\theta}_i^F = \frac{\exp\{-6.14+5.11\times 0.5-1\}}{1+\exp\{-6.14+5.11\times 0.5-1\}} = 0.01$ indicando uma diferençã de 0.4% a mais de probabilidade de morte para os homens.

Análise das cargas dos fatores

O Painel (a), da Figura 6.3, ilustra as cargas associadas aos Fatores 1 e 2. Apenas por esse gráfico fica difícil avaliar a relação de cargas positivas e negativas associadas a cada fator. Calculando o percentual para se ter uma análise melhor da proporção de cargas positivas e negativas, respectivamente, chega-se a seguinte proporção: 49.21% versus 48.53% para o Fator 1 e 56.46% versus 41.27% para o Fator 2 (veja a tonalidade mais avermelhada da coluna 2). Diferentemente do que ocorre com os dados artificiais em que sabemos o sinal de α e λ, o que nos permite identificar a ocorrência, ou não, da troca de sinais entre eles em suas estimativas, com os dados reais não temos controle sobre essa troca. O resultado apresentado aqui considera a configuração de sinal que faz mais sentido do ponto de vista de interpretação. Importante destacar que a inversão de sinal das cargas e dos fatores, desde que efetuada para todas as cargas

	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	-6.14	-6.13	0.19	-6.53	-5.77
β_1	0.32	0.32	0.06	0.19	0.44
β_2	5.11	5.10	0.24	4.63	5.60
σ^2	0.15	0.14	0.04	0.08	0.22
τ_α	0.29	0.26	0.12	0.10	0.54
η_1^*	0.43	0.43	0.16	0.10	0.73
η_2^*	0.01	0.00	0.19	-0.35	0.36
η_3^*	-0.51	-0.52	0.20	-0.87	-0.07
η_4^*	-0.78	-0.78	0.22	-1.26	-0.33

Tabela 6.1: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o cenário de dados reais em que $L = 441$ locais, $T = 4$ tempos e $K = 2$ fatores.

e seus correspondentes fatores em todos os tempos, a matriz de covariâncias da distribuição de η^* não é afetada. Lembramos ao leitor que a distribuição *a priori* de η^* é dada pela Equação (3.7), contendo $\kappa(\lambda)$ que é a função de covariância exponencial quadrática dependente da norma Euclidiana $||\lambda_{\bullet t_1} - \lambda_{\bullet t_2}||$. Ou seja, a inversão do sinal de λ não afeta a distância entre $\lambda_{\bullet t_1}$ e $\lambda_{\bullet t_2}$. A Figura 6.2 mostra a situação na qual o sinal de $\lambda_{1\bullet}$ é invertido, sendo o Painel (a) referente à configuração de λ dos dados artificiais (veja a Tabela 4.4 para $K = 2$ e $T = 4$) e o Painel (b) referente à λ estimado aqui, em que os pontos vermelhos são um espelhamento dos pontos pretos em relação ao eixo horizontal.

Analisando com detalhe o Painel (a) da Figura 6.4, identifica-se que ocorre mais cargas negativas do que positivas até, aproximadamente, o ponto 300 do eixo horizontal, e o contrário (mais cargas negativas do que positivas) depois do ponto 550. A linha azul equivale às médias *a posteriori* das cargas associadas aos dois fatores e classificada em ordem ascendente. Veja que elas estão concentradas dentro do intervalo -0.5 e 0.5. Perceba que a amplitude dos intervalos HPD's de 95% incorporam o zero em todos os casos. Esse fato remete à análise de que as cargas não são significativas, mas prosseguiremos com a interpretação dos resultados obtidos.

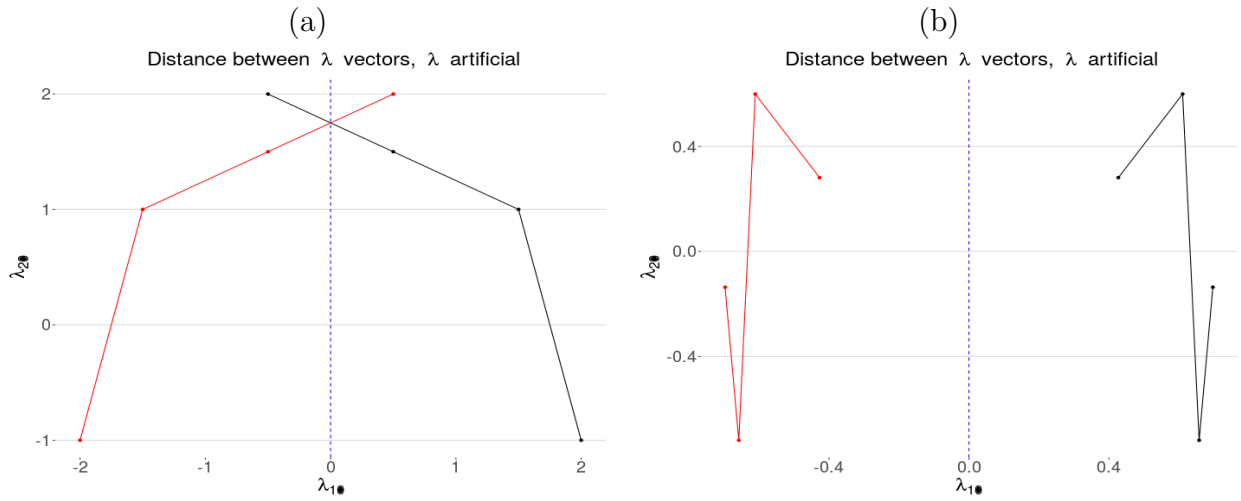


Figure 6.2: Gráfico de $\lambda_{1\bullet}$ versus $\lambda_{2\bullet}$. A cor vermelha ilustra o resultado quando trocamos o sinal dos escores de $\lambda_{1\bullet}$ dos elementos em cor preta. Veja como ocorre um espelhamento em relação ao zero (linha tracejada azul) de forma que a distância entre os pontos permanece inalterada e, consequentemente, não mudando a matriz de covariâncias da distribuição de η^*. O Painel (a) refere-se aos dados artificiais (veja a Tabela 4.4 para $K = 2$ e $T = 4$) e o Painel (b) às estimativas de λ no ajuste do modelo com os dados reais.

Análise dos fatores

Pelo Painel (b), Figura 6.3, vemos que a situação dos pacientes permaneceu, praticamente, inalterada para as regiões associadas ao Fator 1 (municípios com melhor IDH Renda). Verifica-se uma pequena mudança no escore do ano 3, mas retornando à situação anterior no ano seguinte. No entando, para as regiões com pior IDH Renda (Fator 2), ocorreu um decrescimento do escore ano-a-ano. No Painel (b), Figura 6.4, pode-se ver com mais precisão que os escores do Fator 1 se mantém, praticamente, inalterados no decorrer dos anos. Por outro lado, pode-se verificar o decrescimento persistente dos escores do Fator 2.

Impacto na probabilidade de morte

O efeito global da utilização do serviço de telediagnóstico de eletrocardiogramas pode ser avaliado pelo Painel (c) da Figura 6.3. Veja que, ao longo do tempo, o efeito δ apresenta um decrescimento na maioria das regiões, pois a tonalidade das linhas caminha da cor vermelha para a cor azul (positivo para negativo). Para avaliarmos como a variação nos valores de δ influencia na estimação da probabilidade de morte, desenvolveremos uma análise da *odds* de θ_i quando acrescentamos e diminuimos 1 unidade no valor de δ. Efetuando as contas chegamos a seguinte expressão $\frac{\theta_i}{1-\theta_i} = \exp\{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}\} \exp\{\delta_{l_i^* t_i^*}\}$. Semelhante à análise da probabilidade de morte entre homens e mulheres desenvolvida anteriormente, considere pacientes homens de 50 anos, residentes no mesmo município e para os quais o exame foi realizado no mesmo ano. Primeiramente vamos calcular a *odds* considerando a não exitência do efeito δ ($\delta = 0$). Temos, então, $\frac{\theta_i}{1-\theta_i} = \exp\{-6.14 + 0.32 + 5.11 \times 0.5\} \exp\{0\} = 0.038$. Fazendo, agora, $\delta = 1$ e $\delta = -1$, obtemos $\frac{\theta_i}{1-\theta_i} = 0.104$ e 0.014, respectivamente. Isto é, a probabilidade de morte de homens de 50 anos aumenta ou diminui se o local de residência tem o efeito δ, em determinado ano, acrescido ou diminuído, respectivamente. Concluimos que a diminuição do efeito δ, mantido todos os demais elementos fixos, causa uma redução na probabilidade de morte. A redução da probabilidade de morte dos pacientes pode estar relacionada à uma melhora no atendimento às pessoas que sofreram infarto no decorrer dos 4 anos analisados, provavelmente devido ao fato de se ter acesso mais rápido ao diagnóstico, via o sistema de telediagnóstico. Embora seja razoável essa associação, não se pode atribuir esse resultado apenas como conseqüência do sistema de telediagnóstico, uma vez que outras políticas públicas de saúde foram implementadas no mesmo período (`https://www.saude.gov.br/atencao-primaria`).

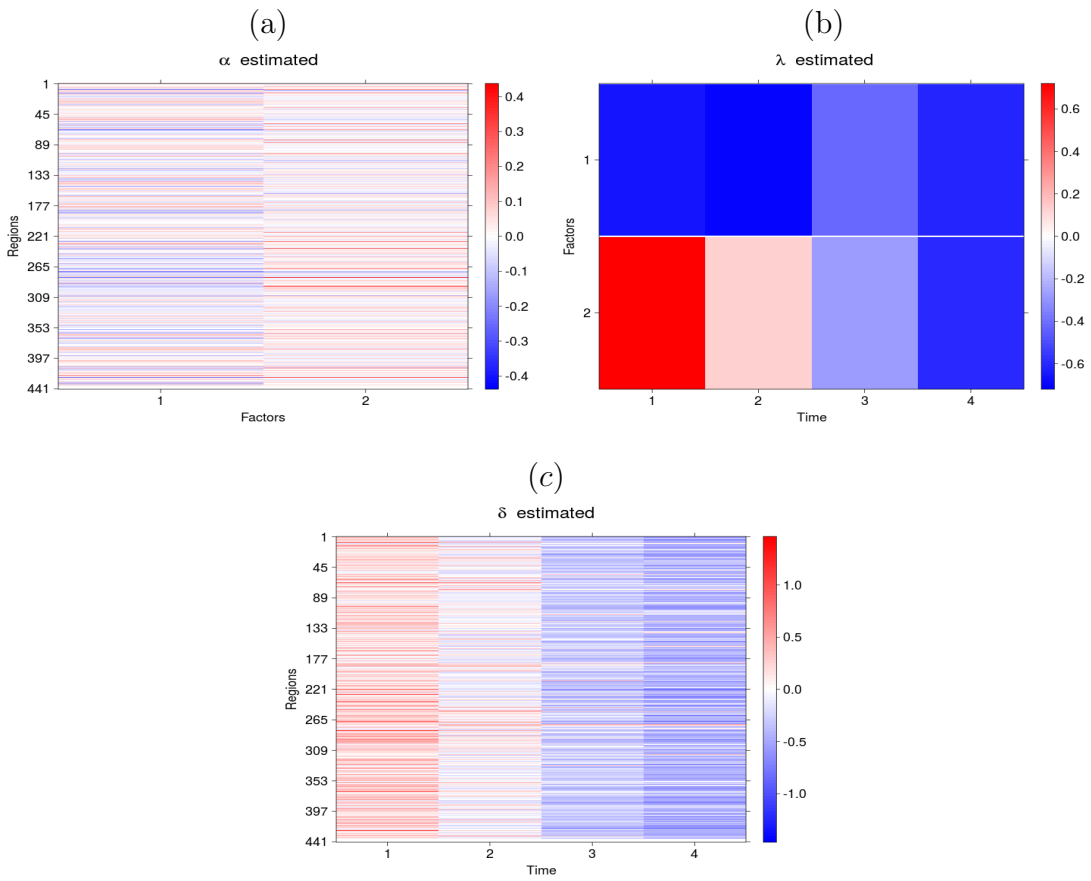


Figure 6.3: Mapas de calor da média *a posteriori* dos parâmetros α, Painel (a); λ, Painel (b) e δ, Painel (c). Considere o cenário de dados reais em que $L = 441$ locais, $T = 4$ tempos e $K = 2$ fatores. O Fator 1 é relacionado a regiões com alto IDH e o Fator 2 está conectado a regiões de baixo IDH.

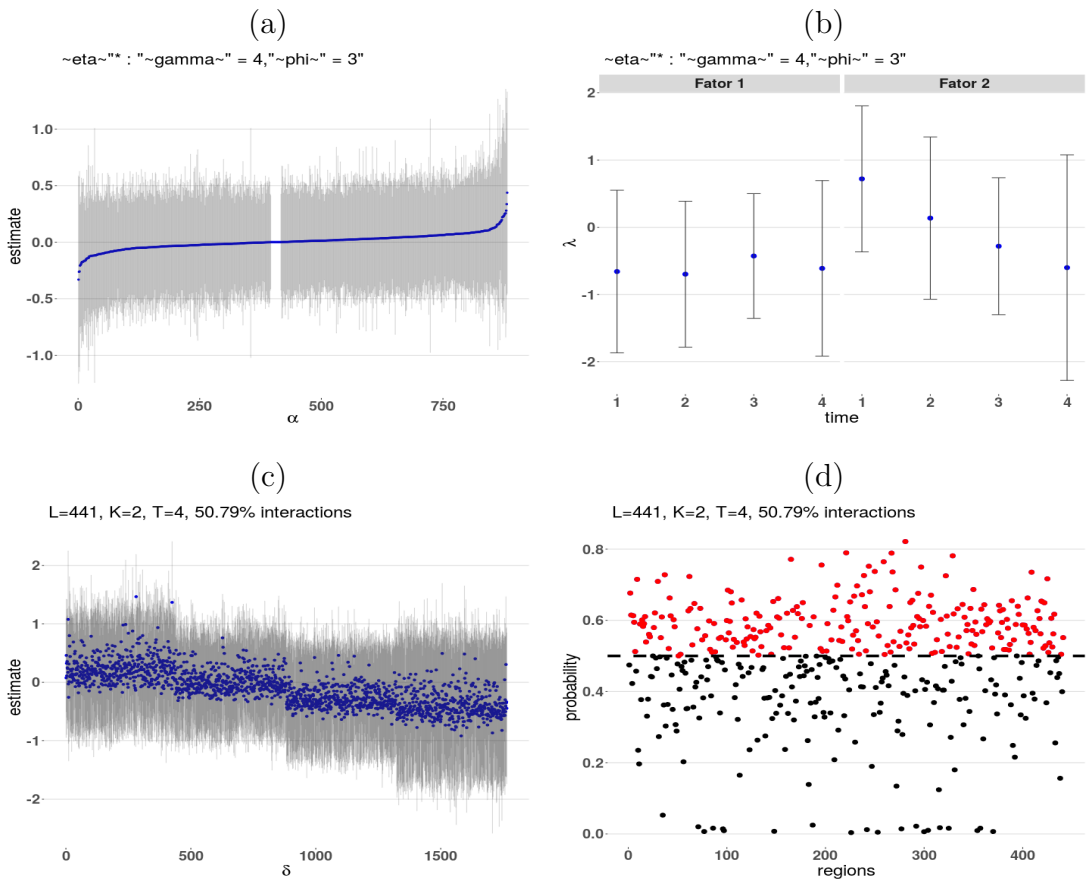


Figure 6.4: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). O Painel (d) apresenta as probabilidades das regiões serem afetadas pela interação; cada ponto é um local. A linha traceja demarca o limiar 0.5 e os pontos vermelhos identificam as regiões com probabilidade de interação acima do limiar sugerido. Considere o cenário de dados reais em que $L = 441$ locais, $T = 4$ tempos e $K = 2$ fatores. A probabilidade de haver interação não linear foi calculada baseada no critério $\mathrm{p}^*(z_l = 1|\bullet) = \frac{\mathrm{p}(z_l=1|\bullet)}{\mathrm{p}(z_l=1|\bullet)+\mathrm{p}(z_l=0|\bullet)} > 0.5$ (veja Etapa 3 do algoritmo MCMC descrito na Seção 3.1.1).

Seguindo a estratégia de avaliar o impacto de δ na probabilidade de morte quando ocorre uma variação no seu valor, prosseguiremos com a análise do impacto de λ quando ocorre uma alteração do valor do escore. Sendo assim, seja $\delta_{lt} = \alpha_{l\bullet}\lambda_{\bullet t} + \eta_{lt} + \epsilon_{lt}$. Para avaliarmos apenas a influência de λ, precisamos assumir algumas suposições. Considere $\eta_{lt} = 0$, $\epsilon_{lt} = 0$ e configure $\alpha_{l\bullet} = (1, 0)$ ou $\alpha_{l\bullet} = (0, 1)$. Com essas suposições chegamos à seguinte formulação para a *odds* de θ_i : $\frac{\theta_i}{1-\theta_i} = \exp\{\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}\}\exp\{\lambda_{kt}\}$. Utilizando o mesmo cenário para a avaliação da variação de δ (homens de 50 anos de um determinado local e tempo) obtemos o mesmo resultado padrão (quando $\lambda_{kt} = 0$) obtidos para δ, isto é: $\frac{\theta_i}{1-\theta_i} = \exp\{-6, 14+0.32+5.11\times 0.5\}\exp\{0\} = 0.038$. Como a magnitude dos escores de λ é muito pequena, avaliaremos o efeito de λ_{kt} na *odds* em termos percentuais. Para isso, avalie a expressão $100 \times (e^{\lambda_{kt}} - 1)$, em que 1 representa o efeito padrão quando $\lambda_{kt} = 0$. Considere, agora, $\lambda_{kt} = 0.5$ e $\lambda_{kt} = -0.5$ (veja a magnitude de λ no Painel (b) da Figura 6.3). Os resultados obtidos são $100 \times (e^{0.5} - 1) = 64.87\%$ e $100 \times (e^{-0.5} - 1) = -39.35\%$. Ou seja, o acréscimo de 0.5 no escore de λ_{kt}, mantido todos os demais elementos fixos e de acordo com as suposições assumidas, ocorre um aumento na probabilidade de morte de $\approx 64.87\%$. De modo contrário, a redução de -0.5, acarreta uma diminuição da probabilidade de morte de $\approx 39.35\%$. Ou seja, os efeitos de λ e δ na probabilidade de morte quando aumentamos/diminuimos os valores desses parâmetros são similares.

A Figura 6.5 ilustra a variação de λ_{kt} para G_1 e G_2 considerando a expressão $100\times(e^{\lambda_{kt}}-1)$ e as suposições descritas anteriormente, ou seja, $\alpha_{l\bullet} = (1, 0)$ ou $\alpha_{l\bullet} = (0, 1)$. Para esse cálculo foi utilizado as estimativas de λ obtidas no ajuste do modelo. Veja como a variação em G_1 (locais com IDH alto) permaneceu, praticamente, constante e sempre negativa sugerindo uma redução (em relação ao caso $\lambda_{kt} = 0$) em ritmo constante ao longo dos anos. Esse comportamento indica que o sistema de telediagnóstico parece não ter contribuido para que houvesse uma aceleração dessa redução. A variação em G_2 (locais com IDH baixo), por outro lado, registrou decaimento dos escores ano-a-ano, sugerindo que houve uma grande contribuição do telediagnóstico do exame eletrocardiográfico para a redução da probabilidade de morte nos locais com baixo IDH Renda.

Retornando à Figura 6.4, o Painel (c), conforme já identificado na Figura 6.3, ilustra o decaimento da média *a posteriori* de δ. Perceba que pode-se identificar 4 blocos de estimativas semelhantes ($[0, \approx 450]$,$[\approx 450, \approx 900]$,$[\approx 900, \approx 1300]$ e $[\approx 1300, \approx 1700]$) que estão associados aos 4 tempos do Painel (c) da Figura 6.3. Finalmente, o Painel (d) ilustra as probabilidades dos locais serem afetados por interação com os pontos em vermelho identificando aqueles com

probabilidade acima do limiar de 0.5. O percentual de locais nessa situação foi de $\approx 52.83\%$, o que equivale a 233 regiões.

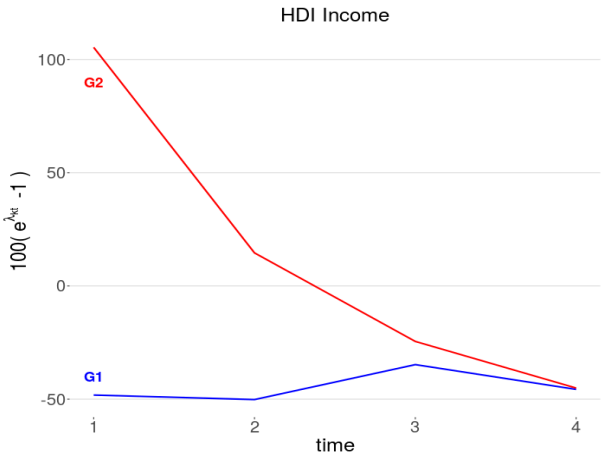


Figure 6.5: Gráfico do efeito de λ_{kt} na *odds* $\frac{\theta_i}{1-\theta_i}$ em termos percentuais para G_1 e G_2 utilizando as estimativas de λ obtidas no ajuste do modelo, lembrando que para G_1 as cargas associadas ao Fator 2 são iguais a zero. O mesmo acontecendo para as cargas associadas ao Fator 1 de G_2. Considere a formulação $100 \times (e^{\lambda_{kt}} - 1)$, em que 1 representa o efeito padrão quando $\lambda_{kt} = 0$.

Análise IDH Renda e SMR

A Figura 6.6 apresenta a distribuição do IDH Renda, Painel (a), e a taxa padrão de mortalidade (SMR = *Standardized Mortality Ratio*) por sexo para os locais de Minas Gerais, Paineis (c) e (d). A SMR é a razão entre o número observado e o número esperado de mortes do local (Banerjee et al., 2004), sendo que o número esperado é dado por $E_l = n_l\overline{r} = n_l\left(\frac{\sum_{l=1}^{441} y_l}{\sum_{l=1}^{441} n_l}\right)$, em que y_l é o número de mortes e n_l é o número de indivíduos que sofreram IAM no local l. Ou seja, $SMR_l = \frac{y_l}{E_l}$. Perceba como, no Painel (a), a tonalidade de cor azul (IDH Renda baixo) predomina nas regiões Norte e Nordeste. As regiões Central, Sul e Triângulo Mineiro, por outro lado, apresentam tonalidades da cor vermelha, indicando IDH's Renda mais elevados. No Painel (b) temos o mapa da taxa SMR global, ou seja, não estratificada por sexo. Veja que

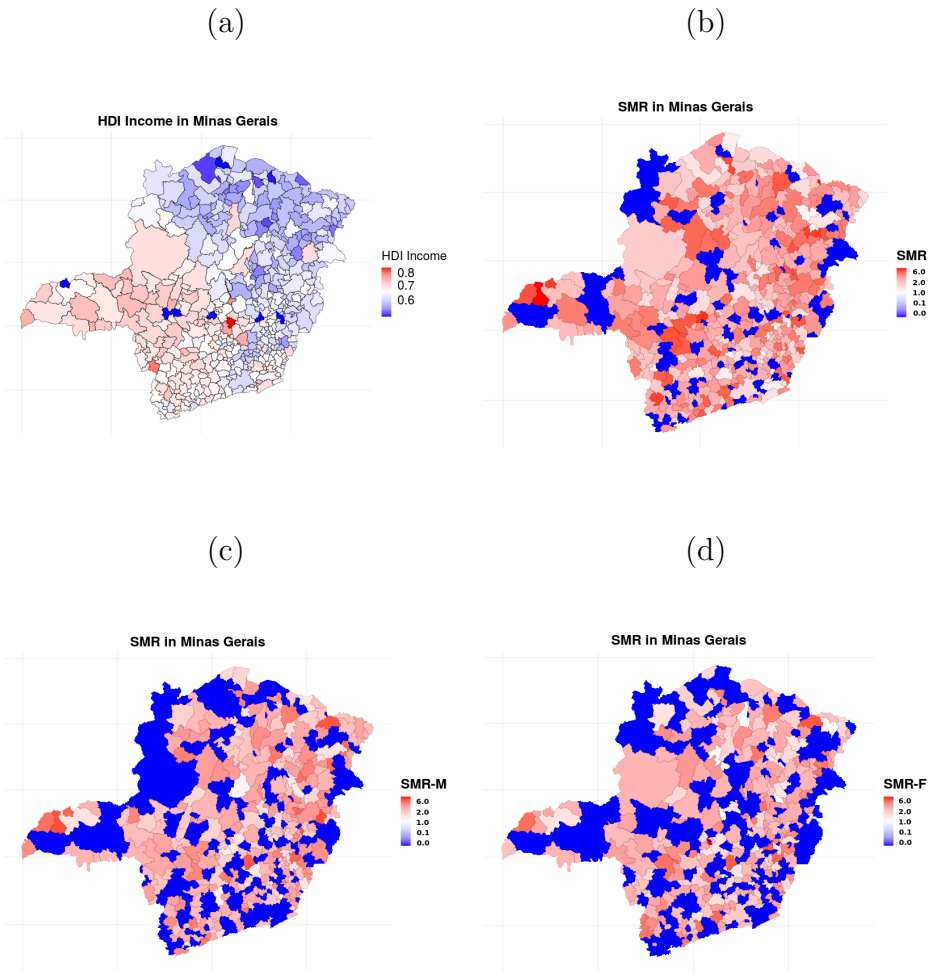


Figure 6.6: Disposição espacial do IDH Renda 2010 e do SMR (*Standardized Mortality Ratio*) por sexo em Minas Gerais. Painel (a): IDH Renda de 2010. Painel (b): SMR global. Paineis (c) e (d): SMR por sexo, masculino e feminino, respectivamente.

a maioria dos locais com maior SMR (cor vermelha mais escura), se encontra nas regiões Norte e Nordeste. Perceba que a taxa de SMR alta significa que o valor observado é maior do que o valor esperado nessas regiões. Comparando as taxas de SMR de homens e mulheres, Paineis (c) e (d) respectivamente, vemos que há diferenças em poucas regiões, mostrando que as taxas de mortalidade entre homens e mulheres são semelhantes na maioria dos locais. Esse resultado é equivalente ao identificado pelo modelo, pois quando $\delta = 1$ (valor alto em relação ao valor de referência, $\delta = 0$) a diferença entre as probabilidades de morte entre homens e mulheres foi baixa, sendo $\approx 2.4\%$, indicando uma coerência entre o ajuste do modelo proposto nesta tese e os dados reais, embora no cálculo das probabilidades de morte consideramos apenas a situação de homens e mulheres de 50 anos de idade. A distinção clara, entre regiões de Minas Gerais, que vemos no Painel (a) referente ao IDH Renda não ocorre nos paineis referentes ao SMR (Paineis b, c, d). Importante destacar que o taxa de SMR apresentada considera o total de mortes e de indivíduos que sofreram IAM em todo o período analisado (2013-2016).

Configuração dos hiperparâmetros do termo de interação η^*

A distribuição *a priori* da interação não linear, dada pela Equação (3.7), utiliza a função de covariância exponencial quadrática (Banerjee et al., 2004), em que os parâmetros γ e ϕ são fixos. Conforme descrito na Seção 3.1, o elemento ϕ é um parâmetro de escala que controla a proximidade dos pontos λ_{t_1} e λ_{t_2} de forma a considerar a existência de uma associação entre eles e γ define a variância quando $t_1 = t_2$. Para $\gamma = 1$, temos uma função de correlação, pois $\kappa(\lambda)_{t_1 t_2}$ variaria entre 0 e 1. Considerando $\phi = 1$, a proximidade dos pontos λ_{t_1} e λ_{t_2} é definida pela distância entre eles. Após alguns testes, decidimos por considerar $\gamma = \phi = 1$. Para avaliarmos o efeito de η^* na probabilidade de morte, lembre que $\delta_{lt} = \alpha_{l\bullet}\lambda_{\bullet t} + \eta_{lt} + \epsilon_{lt}$ e considere que não há nenhum efeito principal e que o erro aleatório é igual a zero, ou seja, $\alpha_{l\bullet}\lambda_{\bullet t} = 0$ e $\epsilon_{lt} = 0$ para todo l e t. Com isso, temos que $\frac{\theta_i}{1-\theta_i} = \exp\{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}\}\exp\{\eta^*_{lt}\}$. Considerando o mesmo cenário utilizado na análise dos impactos de δ e λ (homens de 50 anos de um determinado local e tempo) chegamos ao mesmo resultado padrão, isto é: $\frac{\theta_i}{1-\theta_i} = \exp\{-6.14 + 0.32 + 5.11 \times 0.5\}\exp\{0\} = 0.038$. Da mesma forma que λ, a magnitude de η^*_{lt} também é muito pequena.Desta forma, avaliaremos o efeito de η^*_{lt} na *odds* em termos percentuais. Considere a expressão $100 \times (e^{\eta^*_{lt}} - 1)$, em que 1 representa o efeito padrão quando $\eta^*_{lt} = 0$. Fazendo, agora, $\eta^*_{lt} = 0.2$ e $\eta^*_{lt} = -0.2$, chegamos aos seguintes resultados : $100 \times (e^{0.2} - 1) = 22.14\%$ e $100 \times (e^{-0.2} - 1) = -18.13\%$. Ou seja, o acréscimo de 0.2 em η^*_{kt}, considerando as suposições

estabelecidas, acarreta um aumento na *odds* de morte de $\approx$ 22.14%. O decréscimo de -0.2, por outro lado, determina uma redução na *odds* de morte de $\approx$ 18.13%. A Figura 6.7 ilustra o resultado da expressão $100 \times (e^{\eta^*_{lt}} - 1)$ em que utilizamos as estimativas de η^*_{lt}. Veja como ocorre um decaimento acentuado do efeito e^{η^*} em relação ao efeito padrão e^0 ao longo do tempo. Verificamos, então, que os efeitos de δ, λ e η na probabilidade de morte quando aumentamos ou diminuimos os valores desses parâmetros seguem o mesmo padrão de interpretação.

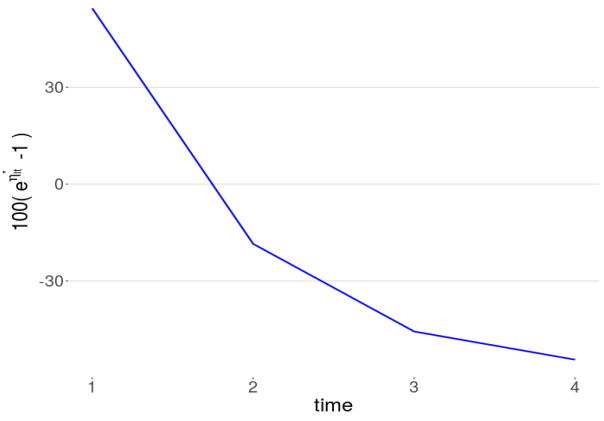


Figure 6.7: Gráfico do efeito de η^*_{lt} na *odds* $\frac{\theta_i}{1-\theta_i}$ em termos percentuais utilizando as estimativas de η^* obtidas no ajuste do modelo. Considere a formulação $100 \times (e^{\eta^*_{lt}} - 1)$, em que 1 representa o efeito padrão quando $\eta^*_{lt} = 0$.

Análise espacial dos conglomerados

Vamos efetuar, agora, análises espaciais das estimativas obtidas com o ajuste do modelo. A Figura 6.8 ilustra a disposição espacial das estimativas de α para os 441 locais. O mapa da esquerda destaca as cargas associadas ao Fator 1 e o mapa da direita as associadas ao Fator 2. Visualmente, o painel da esquerda, cargas associadas ao Fator 1, parece conter mais regiões com $\alpha's$ negativos do que positivos. Mas, conforme mostrado anteriormente, a proporção é muito próxima (49.21% versus 48.53%). O painel da direita, referente às cargas associadas ao Fator 2, apresenta 56.46% de cargas positivas e 41.27% de negativas, o que, nesse caso, pode ser verificado visualmente, ocorrendo uma predominância da cor vermelha. Pela Figura 6.9

podemos identificar o padrão de decrescimento de δ no tempo, conforme já havíamos destacado na análise da Figura 6.3, mas com a diferença de que, pelo mapa, podemos avaliar a evolução de local por local. Nesse sentido, veja como o local em vermelho na região Norte do mapa referente a δ no Tempo 1 (Painel $\delta_{\bullet 1}$) obtem uma redução perceptível, visualmente, ficando com a estimativa no Tempo 2 (Painel $\delta_{\bullet 2}$) bem próximo de zero (cor branca). Esse mesmo local apresenta δ negativo no Tempo 3 (Painel $\delta_{\bullet 3}$), mantendo, praticamente inalterado, no Tempo 4 (Painel $\delta_{\bullet 4}$).

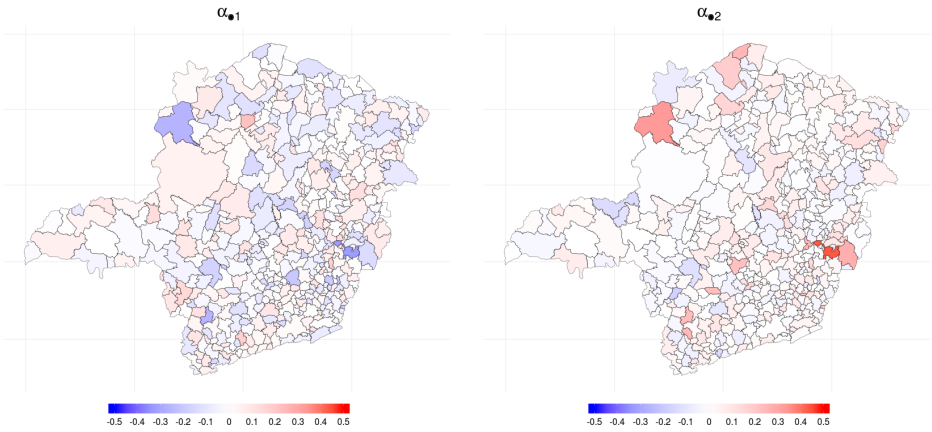


Figure 6.8: Disposição espacial das estimativas de α para os $L = 441$ locais. O mapa da esquerda ilustra as cargas associadas ao Fator 1 e o mapa da direita as associadas ao Fator 2.

A Figura 6.10 ilustra os efeitos pelos quais os locais foram afetados. Essa análise foi realizada a partir da verificação das cargas para as quais o intervalo HPD não inclui o zero. Inicialmente avaliamos o intervalo HPD de 95%, mas, como pode ser visto pelo Painel (a) da Figura 6.4, todos os intervalos HPD incluem o zero. Desta forma, foi necessário reduzir o tamanho do envelope HPD para identificar cargas estimadas em regiões mais distantes de zero referentes a um dos fatores. Para isso, avaliamos os intervalos HPD dos seguintes tamanhos: $90\%, 85\%, 80\%, \cdots, 40\%$. Apenas com o intervalo HPD de 40% foi possível identificar algumas cargas para as quais o zero está fora da região de probabilidade ao redor da estimativa. A Figura 6.11 apresenta o gráfico dos intervalos HPD's de 40% para as cargas classificadas em ordem crescente. Então, a partir do HPD de 40% avaliamos os locais afetados por algum

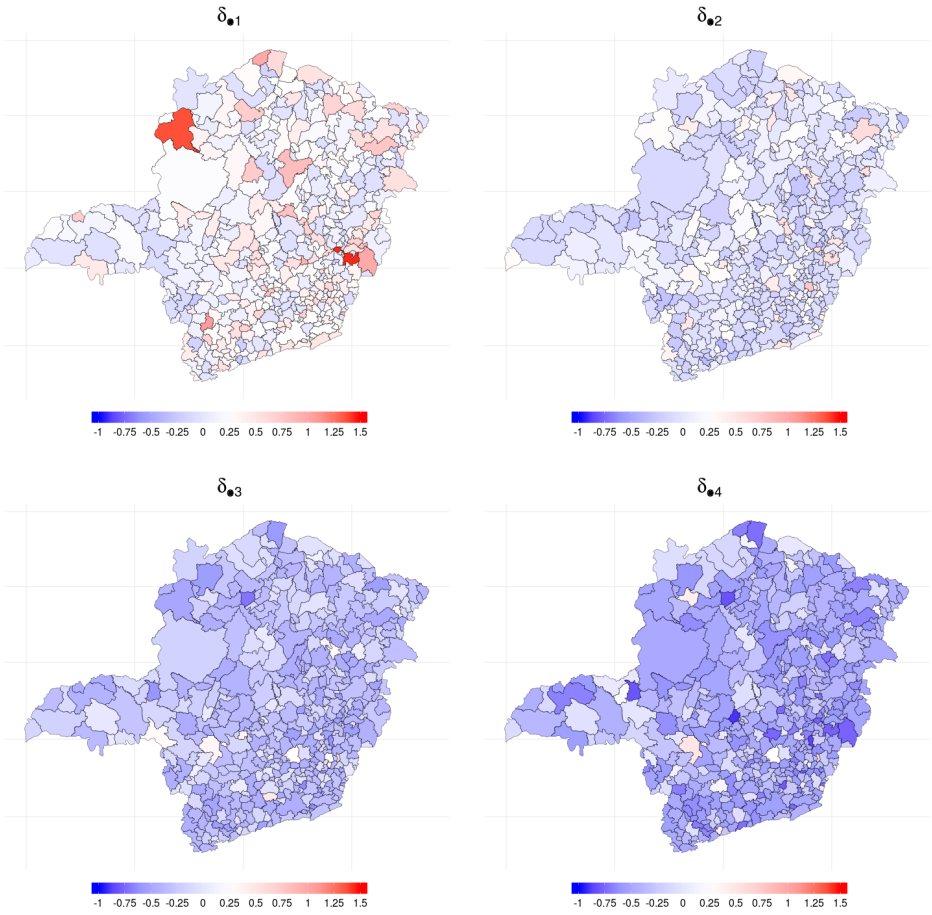


Figure 6.9: Disposição espacial das estimativas de δ para os $L = 441$ locais e $T = 4$ tempos.

efeito principal e/ou pela interação não linear. A Figura 6.10 ilustra a disposição espacial dos locais afetados pelos Fatores 1 e 2, concomitantemente no Painel (a), daqueles locais apenas afetados pelo Fator 1 ou pelo Fator 2, individualmente (Paineis b, c), e dos locais que não apresentaram nenhum dos efeitos principais, mas foram afetados apenas por interação ou não sofreram qualquer tipo de efeito (Painel d). O efeito de interação foi atribuído aos locais com probabilidade $\mathrm{p}^*(z_l = 1|\bullet) = \frac{\mathrm{p}(z_l=1|\bullet)}{\mathrm{p}(z_l=1|\bullet)+\mathrm{p}(z_l=0|\bullet)}$ (veja Etapa 3 do algoritmo MCMC descrito na Seção 3.1.1), conforme apresentado no Painel (d) da Figura 6.4.

Análise preditiva de mortalidade e conclusão

Finalmente, apresentamos as curvas ROC e a AUC, Figura 6.12, referentes à estimativa de

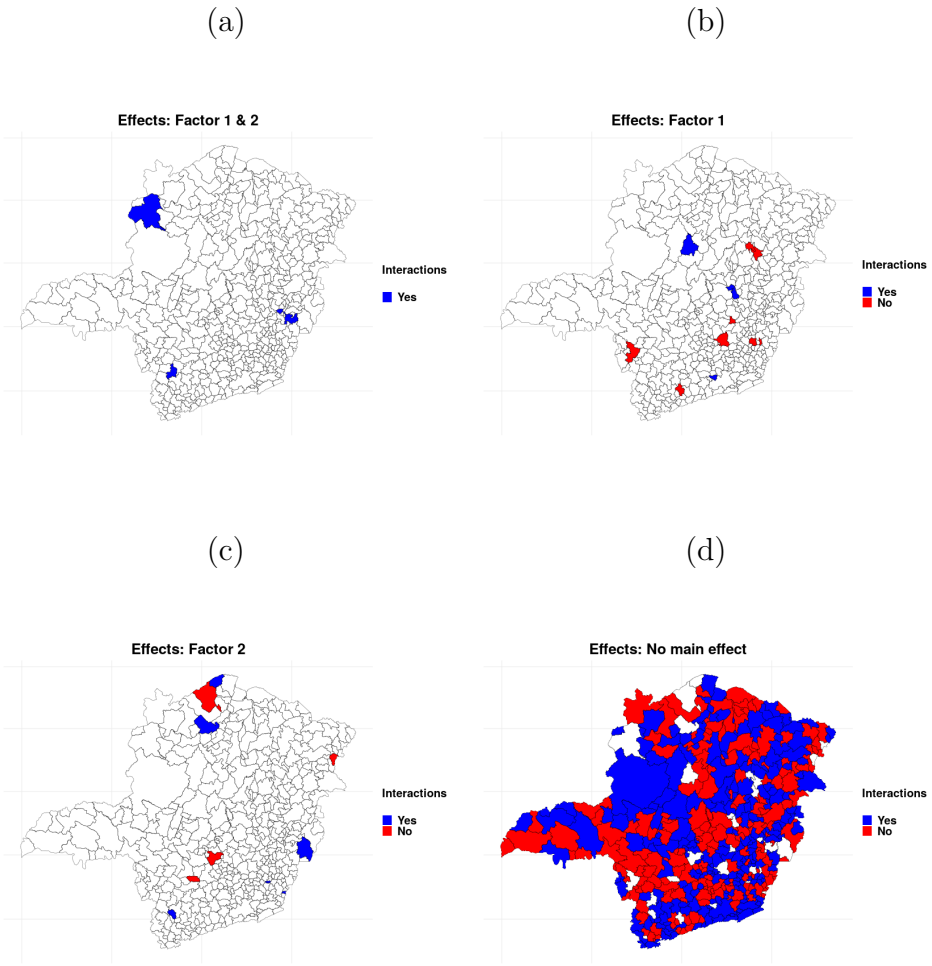


Figure 6.10: Disposição espacial dos efeitos de α para os $L = 441$ locais. Foram consideradas as cargas para as quais o intervalor HPD de 40% não inclui o zero. Iniciamos essa análise considerando o intervalo HPD de 95%, mas com essa configuração todos os intervalos HPD incluem o zero. Para podermos avaliar os efeitos nos locais reduzimos o tamanho do envelope HPD considerando os seguintes percentuais: 90%, 85%, 80%,...,40%. Desses, o que identificou mais locais afetados pelo efeito principal e pela interação foi o intervalo de 40%.

probabilidade a partir da base real de pacientes do sistema de telessaúde. O gráfico mostra os resultados para os cenários sem o efeito δ (vermelho), com o efeito δ, mas considerando o efeito de interação η^* ausente (verde) e presente (azul). Podemos ver que os ajustes com o efeito δ obtiveram resultados, em termos preditivos, melhores do que o cenário com apenas as

covariáveis sexo e idade. Da mesma forma que no caso simulado dividimos a base de dados real em duas : treino e teste. A base de treino ficou com 95.88% (15185) e a de teste com 4.11% das observações (650 indivíduos). Novamente, a base de teste foi gerada tomando-se o cuidado de manter, em cada local, indivíduos que sofreram IAM em todos os 4 anos existentes nos dados originais. Veja que a AUC obtida foi de quase 70% para uma base bem desbalanceada onde temos $\approx 8\%$ de mortes contra $\approx 92\%$ de indivíduos que sobreviveram após sofrerem o infarto agudo do miocárdio. Esse resultado, além dos demais já apresentados, indica que o modelo proposto nesta tese está com bom comportamento em termos preditivos.

Encerramos o capítulo sobre a ilustração que motivou o desenvolvimento desta tese. Em termos de interpretação prática destacamos que a conclusão mais interessante para os pesquisadores da Telessaúde, e para os pacientes em MG, é a identificação da contribuição que o sistema trouxe, através da análise integrada de ECG's, para melhorar o acompanhamento de pacientes provenientes de regiões com baixo IDH no estado de Minas Gerais. A análise de *clusters* mostrada na Figura 6.10 é outro ponto de destaque, servindo para identificar os locais associados aos grupos de municípios com IDH Alto e Baixo, bem como de municípios que são afetados apenas pelo efeito de interação (Painel d), o que pode apoiar na aplicação de políticas públicas comuns para esses municípios.

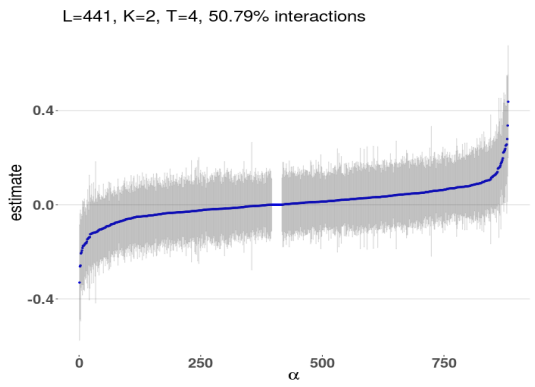


Figure 6.11: Análise gráfica do intervalo HPD de 40% *a posteriori* para α. As estimativas foram ordenadas em relação a média *a posteriori*.

Figure 6.12: Curvas ROC e AUC para o ajuste do modelo utilizando dados reais em que $L = 441$ locais, $T = 4$ tempos e $K = 2$ fatores. Cenários: sem o efeito δ (vermelho), com o efeito δ, mas sem (verde) e com (azul) o efeito de interação η^*. Para construção desta curva, a base de dados original foi dividida em treino e teste. A base de treino ficou com 95.88% e a de teste com 4.11% das observações.

Capítulo 7

Conclusões e trabalhos futuros

A motivação para o desenvolvimento do modelo desta tese surgiu a partir de duas questões igualmente importantes : o trabalho de Mayrink e Lucas (2013) e a disponibilidade dos dados do sistema de telediagnóstico de eletrocardiogramas do Centro de Telessaúde do Hospital das Clínicas (HC) da UFMG. Mayrink e Lucas (2013) propuseram, no contexto da análise fatorial padrão, a novidade de se trabalhar com interações não lineares entre os fatores através do Processo Gaussiano. Os dados disponibilizados pelo Centro de Telessaúde do HC-UFMG relacionam pacientes que sofreram infarto agudo do miocárdio (IAM), identificados a partir do laudo de eletrocardiograma, com o desfecho de morte ou não. Com base nisso, identificamos a possibilidade de avaliar resultados de eletrocardiogramas realizados em tempos diferentes e espacialmente localizados (pacientes residentes em municípios de Minas Gerais), bem como, testar a suposição de que essas localidades são afetadas por interação não linear que representam interrelações existentes entre municípios não capturáveis através de métodos simples. O ineditismo de nosso trabalho está no fato de incorporarmos essas interações em modelos fatoriais construídos para aplicações do tipo espaço-temporal e de utilizar uma base de dados que nunca foi analisada em outros trabalhos na literatura com modelagem estatística. Esta tese, então, apresenta uma aplicação real do modelo fatorial espaço-temporal com interações não lineares. A presença de interações não lineares visa melhorar a explicação das complexas interrelações entre as regiões do espaço e, também, permitir a detecção de conglomerados (*clusters*). Novos tipos de *clusters* podem surgir como uma combinação dos efeitos principais dos fatores e do efeito da interação. Desenvolvemos dois modelos lineares generalizados, logístico e Poisson, com efeito aleatório estruturado pela modelagem fatorial para explorar dados coletados no espaço e no tempo.

No Capítulo 1, introduzimos os modelos que precederam e serviram de base para este trabalho. Lopes et al. (2008) apresentam o desenvolvimento do modelo fatorial espaço-temporal e em Lopes et al. (2011) eles estendem esse modelo e propoem o modelo fatorial espaço-temporal generalizado. Complementando, Mayrink e Lucas (2013) incorporam ao modelo fatorial padrão ($\delta = \alpha\lambda + \epsilon$) o componente de interação não linear ($\delta = \alpha\lambda + \eta + \epsilon$). Esses estudos formam o suporte metodológico para nossa proposta do modelo fatorial espaço-temporal generalizado com interação não linear. Inspirado no trabalho de Lopes et al. (2011), a dependência espacial entre regiões é estabelecida através das colunas da matriz de cargas, para a qual assumimos a configuração do modelo autoregressivo condicional (CAR) (Besag, 1974). A dependência temporal é considerada na associação das colunas da matriz de escores dos fatores, também inserida por meio do modelo CAR.

No Capítulo 2 apresentamos a base de dados real, um dos principais motivadores para o desenvolvimento desta tese. O fato dos dados serem referentes à pacientes que executaram exames eletrocardiológicos reforçaram nossa motivação, pois estudar problemas cardíacos é um assunto relevante, visto que estudos mostram que o infarto agudo do miocárdio é a principal causa de morte em pacientes com problemas do coração. Descrevemos que a origem dos dados é o sistema de telediagnóstico do Centro de Telessaúde do Hospital das Clínicas da UFMG a partir do qual mais de 2500 exames são laudados diariamente, a partir de eletrocardiogramas transmitidos de 811 municípios de Minas Gerais e armazenados em um banco de dados central. Médicos cardiologistas registram os laudos dos exames, pela internet, agilizando o diagnóstico de doenças. Destacamos o processo complexo de tratamento dos dados para se alcançar a qualidade e formato adequados para a análise estatística. De um total de 2470424 de ECGs, 1773689 pacientes foram identificados. Após excluir os ECGs com problemas técnicos e pacientes com menos de 16 anos, um total de 1558415 pacientes foram considerados. Em seguida, filtramos apenas os exames cujos laudos indicavam a ocorrência de IAM, consideramos os anos de 2013 a 2016, para os quais o número de pacientes era adequado para a realização do nosso estudo. Com isso, a base de dados final totalizou 15835 indivíduos, dos quais 1286 morreram. Aqueles municípios que não possuiam pacientes que sofreram IAM, em todos os anos selecionados, foram unidos a municípios fronteiriços de forma que as regiões resultantes dessa união possuissem registros de pacientes com IAM em todos os 4 anos selecionados.

No Capítulo 3 apresentamos as formulações dos modelos logístico e Poisson, os dois modelos integrantes da família de modelos lineares generalizados mistos (GLMMs) avaliados neste estudo.

Detalhamos a estrutura hierárquica desses modelos que são uma extensão dos modelos lineares generalizados (GLMs) em que a ideia principal é a incorporação de correlação através da modelagem contendo um ou mais termos de efeitos aleatórios. Neste estudo, acrescentamos aos preditores lineares apenas um termo aleatório o qual incorpora a estrutura espaço-temporal, o efeito de interação não linear e o componente de erro ($\delta = \alpha\lambda + \eta + \epsilon$). Nesta tese, os modelos foram trabalhados sob o ponto de vista da inferência Bayesiana. Com isso, além das funções de verossimilhança, especificamos as distribuições a priori relacionadas aos parâmetros e descrevemos o algoritmo Metropolis-Hasting, juntamente com os valores iniciais e os parâmetros necessários a sua execução (número de iterações, *burn-in* e *lag*).

Nos Capítulos 4 e 5 apresentamos as análises a partir de dados artificiais. A utilização desses tipos de dados, para os quais conhecemos o modelo gerador, é muito conveniente na verificação da validação do modelo proposto, uma vez que os parâmetros são conhecidos. Uma boa aproximação das estimativas dos parâmetros em relação aos dados verdadeiros (artificiais) sugere um bom comportamento do modelo. O Capítulo 4 é dedicado às análises do modelo logístico e o Capítulo 5, do modelo Poisson. Vários cenários com dados artificiais foram preparados para validação dos modelos. Os cenários considerados envolveram número de locais iguais a 100, 200 e 400, sendo este último semelhante aos dados reais (441); 4 e 10 tempos (4 porque é o número de anos da base real e 10 para simular uma situação futura dos dados reais); 4 e 6 vizinhos por local, sendo 6 o número médio de vizinhos por região da base de ECGs; 30% e 50% de locais afetados por interação não linear. No caso logístico, essas análises citadas consideram que os dados são balanceados em relação ao número de $Y_i^{'}s = 1$ e $Y_i^{'}s = 0$, ou seja, aproximadamente 50% dos $Y_i^{'}s = 1$. Também mostramos que, para avaliarmos situações mais parecidas com os dados reais de ECG's, consideramos o cenário desbalanceado, isto é, com $\approx$ 20% de $Y_i's = 1$. Várias configurações variando o número de locais, número de fatores, número de tempos e percentual de locais afetados pela interação não linear foram avaliadas e concluimos que o ajuste para o modelo logístico é, em geral, satisfatório e, com isso, a estrutura hierárquica proposta parece atender bem as diversas situações. Para o caso Poisson, desenvolvemos análises para poucas e muitas contagens zero. Identificamos que o ajuste para o cenário com poucas contagens zero é sensível ao chute inicial da cadeia MCMC de β_0, mas mostramos uma estratégia de configuração de valores iniciais para direcionar o pesquisador a obter uma boa estimação dos parâmetros do modelo. Conseguimos estimar muito bem as cargas (*loadings*), os escores dos fatores, o efeito aleatório, a interação não linear e os coeficientes da regressão. Para os cenários

desbalanceado (logístico) e de contagens com alta ocorrência de zeros (Poisson) identificamos algum vício para os parâmetros com valores verdadeiros extremos. Para as cargas verificamos a ocorrência de subestimação ou sobrestimação. Também observamos sobrestimação para valores extremos negativos de δ e intervalos HPD longos, mas para valores positivos as estimativas foram muito boas e com intervalos curtos. No que se refere às probabilidades dos locais serem afetados por interação constatamos muita coerência das estimativas em relação à indicação verdadeira dos locais terem ou não sido afetados por interação.

No Capitulo 6 descrevemos os resultados com os dados reais do sistema de telediagnóstico de exames eletrocardiológicos. Essa análise considerou 2 fatores e 4 tempos. Selecionamos as 10 regiões de Minas Gerais com maior IDH Renda e as 10 com menor associando-os ao Fator 1 e 2, respectivamente. Através da especificação de distribuições *a priori* informativas, esses locais foram configurados para não serem afetados por interação e nem por qualquer outro efeito principal. Verificamos uma estabilização dos escores do Fator 1 no tempo, sugerindo nenhuma mudança para os locais que já apresentavam IDH Renda alto. Por outro lado, os escores do Fator 2 diminuiram a cada ano, apontando uma redução da probabilidade de morte nos locais de pior IDH Renda, provavelmente decorrente da melhoria no atendimento aos pacientes devido à disponibilização do sistema de telediagnóstico. Embora seja razoável essa associação, não se pode atribuir esse resultado apenas à disponibilidade do sistema de telediagnóstico, pois outras políticas públicas de saúde foram implementadas no mesmo período (`https://www.saude.gov.br/atencao-primaria`). Observamos a ocorrência de interação não linear decrescente entre os fatores das regiões com IDH Renda alto e baixo, indicando um efeito de redução de interação no decorrer do anos. Finalmente, detectamos *clusters* de municípios de Minas Gerais afetados por 1 fator, 2 fatores, por interação, combinações desses efeitos e nenhum efeito. A identificação desses conglomerados é ponto chave neste estudo, inclusive mensionado no título da tese, pois através deles é possível identificar as regiões que são mais parecidas com aquelas de IDH Renda mais alto ou mais baixo e, também, as regiões que sofrem influência de ambas, ou seja, aquelas que possuem efeitos parecidos com as duas regiões. Com essa informação pode-se avaliar estratégias de direcionamento de políticas de atuação do sistema de telediagnóstico para torná-lo ainda mais eficiente. Os resultados obtidos na aplicação real podem alterar se utilizarmos distribuições *a priori* mais informativas, mas para isso será necessário uma análise mais profunda com especialistas.

Trabalhos futuros

Esta tese se concentrou em dois modelos da família de modelos lineares generalizados: logístico e Poisson. Nossa proposição é válida, também, para outros modelos dessa família. Sendo assim, avaliar outras distribuições para a variável resposta é uma das possibilidades de trabalho futuro. Dentre elas podemos destacar a distribuição Gama, a Beta e, até mesmo, a Normal. Além disso, outra proposta interessante é trabalharmos com dados reais em que a variável resposta é contagem, pois assim poderemos avaliar o comportamento do modelo Poisson nessa situação. Outra alternativa de estudo futuro é voltar a focar no modelo de resposta binária, morte ou não, utilizando os dados reais considerados neste trabalho, porém alterando a função de ligação de logit para probit. Recentemente, uma nova base de dados foi disponibilizada com mais dados que compreendem os anos de 2006 a 2012, 2017 e 2018. Então, uma nova análise aplicada utilizando dados de 2006 a 2018 é outra atividade para desenvolvimento futuro. Infelizmente, não houve tempo hábil para elaborar essa análise com mais anos, mas foi interessante termos avaliado o ajuste do modelo com os dados atuais, mostrando que o modelo obteve resultados satisfatórios com poucos anos (2013-2016). Conforme informado no Capitulo 2, o sistema de telediagnóstico armazem várias variáveis referentes a dados clínicos dos pacientes. Então o desenvolvimento de um processo de seleção de variáveis pode ser desenvolvido. Finalmente, outra atividade que consideramos interessante a ser desenvolvida é a implementação de um pacote em R para que outras pessoas possam aplicar nosso modelo.

Apêndice

Apêndice A: Modelo Logístico - Configurações de α para os conjuntos de dados nos contextos de K = 4 e K = 5.

Apresentamos nesta seção as configurações do modelo logístico para a matriz de cargas utilizadas na geração dos dados artificiais nos casos com $K = 4$ e $K = 5$ fatores. As Tabelas 1 e 2 complementam a Tabela 4.3.

Configurações de α para geração das bases de dados nos contextos de K = 4.			
Parâmetro	Índice	Dimensão	Valor real
α_{lk}	$l \in \{1, ..., 10\},\ k = 1$	10×1	$U(1, 2)$
	$l \in \{1, ..., 10\},\ k = 2$		$\mathbf{0}$
	$l \in \{1, ..., 10\},\ k = 3$		$\mathbf{0}$
	$l \in \{1, ..., 10\},\ k = 4$		$\mathbf{0}$
α_{lk}	$l \in \{11, ..., 20\}, k = 1$	10×1	$\mathbf{0}$
	$l \in \{11, ..., 20\},\ k = 2$		$U(1, 2)$
	$l \in \{11, ..., 20\},\ k = 3$		$\mathbf{0}$
	$l \in \{11, ..., 20\},\ k = 4$		$\mathbf{0}$
α_{lk}	$l \in \{21, ..., 30\}, k = 1$	10×1	$\mathbf{0}$
	$l \in \{21, ..., 30\},\ k = 2$		$\mathbf{0}$
	$l \in \{21, ..., 30\},\ k = 3$		$U(1, 2)$
	$l \in \{21, ..., 30\},\ k = 4$		$\mathbf{0}$
α_{lk}	$l \in \{31, ..., 40\}, k = 1$	10×1	$\mathbf{0}$
	$l \in \{31, ..., 40\},\ k = 2$		$\mathbf{0}$
	$l \in \{31, ..., 40\},\ k = 3$		$\mathbf{0}$
	$l \in \{31, ..., 40\},\ k = 4$		$U(1, 2)$
α_{G_E}	$l \in \{41, ..., L\},\ k \in \{1, 2, 3, 4\}$	$(L - 40) \times 4$	$\mathrm{N}_{(L-40)}(\mathbf{0}, [D_\alpha - \rho_\alpha W_\alpha]^{-1})$

Tabela 1: Valores da matriz de cargas (*loadings*), utilizados na geração dos diferentes conjuntos de dados baseados nos cenários apresentados na Tabela 4.1 e considerando os números de fatores $K = 4$.

Configurações de α para geração das bases de dados nos contextos de K = 5.			
Parâmetro	Índice	Dimensão	Valor
α_{lk}	$l \in \{1, ..., 10\},\ k = 1$	10×1	$U(1, 2)$
	$l \in \{1, ..., 10\},\ k = 2$		$\mathbf{0}$
	$l \in \{1, ..., 10\},\ k = 3$		$\mathbf{0}$
	$l \in \{1, ..., 10\},\ k = 4$		$\mathbf{0}$
	$l \in \{1, ..., 10\},\ k = 5$		$\mathbf{0}$
α_{lk}	$l \in \{11, ..., 20\}, k = 1$	10×1	$\mathbf{0}$
	$l \in \{11, ..., 20\},\ k = 2$		$U(1, 2)$
	$l \in \{11, ..., 20\},\ k = 3$		$\mathbf{0}$
	$l \in \{11, ..., 20\},\ k = 4$		$\mathbf{0}$
	$l \in \{11, ..., 20\},\ k = 5$		$\mathbf{0}$
α_{lk}	$l \in \{21, ..., 30\},\ k = 1$	10×1	$\mathbf{0}$
	$l \in \{21, ..., 30\},\ k = 2$		$\mathbf{0}$
	$l \in \{21, ..., 30\},\ k = 3$		$U(1, 2)$
	$l \in \{21, ..., 30\},\ k = 4$		$\mathbf{0}$
	$l \in \{21, ..., 30\},\ k = 5$		$\mathbf{0}$
α_{lk}	$l \in \{31, ..., 40\},\ k = 1$	10×1	$\mathbf{0}$
	$l \in \{31, ..., 40\},\ k = 2$		$\mathbf{0}$
	$l \in \{31, ..., 40\},\ k = 3$		$\mathbf{0}$
	$l \in \{31, ..., 40\},\ k = 4$		$U(1, 2)$
	$l \in \{31, ..., 40\},\ k = 5$		$\mathbf{0}$
α_{lk}	$l \in \{41, ..., 50\},\ k = 1$	10×1	$\mathbf{0}$
	$l \in \{41, ..., 50\},\ k = 2$		$\mathbf{0}$
	$l \in \{41, ..., 50\},\ k = 3$		$\mathbf{0}$
	$l \in \{41, ..., 50\},\ k = 4$		$\mathbf{0}$
	$l \in \{41, ..., 50\},\ k = 5$		$U(1, 2)$
α_{G_E}	$l \in \{51, ..., L\},\ k \in \{1, 2, 3, 4, 5\}$	$(L - 50) \times 5$	$\mathrm{N}_{(L-50)}(\mathbf{0}, [D_\alpha - \rho_\alpha W_\alpha]^{-1})$

Tabela 2: Valores da matriz de cargas (*loadings*), utilizados na geração dos diferentes conjuntos de dados baseados nos cenários apresentados na Tabela 4.1 e considerando os números de fatores $K = 5$.

Apêndice B: Modelo Logístico - Análises para 100 e 200 locais

Esta seção é dedicada aos cenários com $L = 100, 200$ locais, $T = 4$ tempos, $K = 2$ fatores, 4 vizinhos por região, e $\approx 30\%$ e 50% de locais do Grupo extra com interações e $\approx 50\%$ de $Y_i^{'} s = 1$.

Estimativas para $L = 100$ locais e $\approx 30\%$ locais de G_E afetados pela interação não linear

	Verdadeiro	Média	Mediana	DP	HPD (inf)	HPD (sup)
β_0	0.50	0.61	0.62	0.19	0.21	0.98
β_1	-1.00	-0.82	-0.82	0.09	-1.01	-0.64
β_2	1.00	1.10	1.10	0.09	0.93	1.27
σ^2	0.80	0.87	0.85	0.22	0.45	1.28
τ	2.00	2.21	2.05	1.06	0.45	4.20
η_1	-2.00	-1.04	-1.03	0.73	-2.46	0.36
η_2	1.50	0.71	0.70	0.55	-0.36	1.79
η_3	0.75	-0.02	-0.03	0.57	-1.14	1.06
η_4	-1.00	-0.91	-0.93	0.70	-2.31	0.52

Tabela 3: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Cenário: $M^{K_2 I_{30\%}}_{L_{100} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

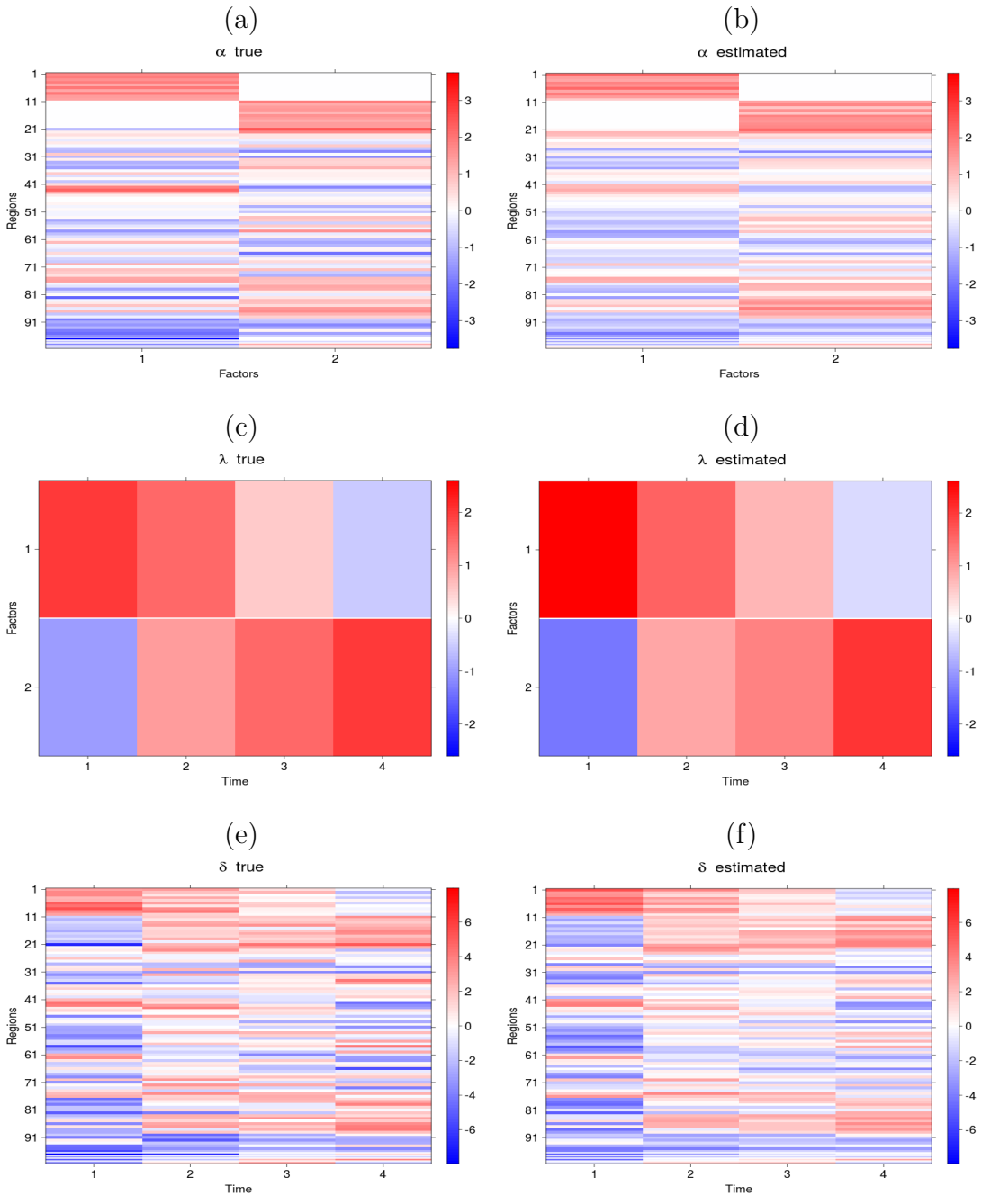


Figure B.1: Mapas de calor comparando valores verdadeiros e estimados para o cenário $M^{K_2 I_{30\%}}_{L_{100} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$. Paineis: (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

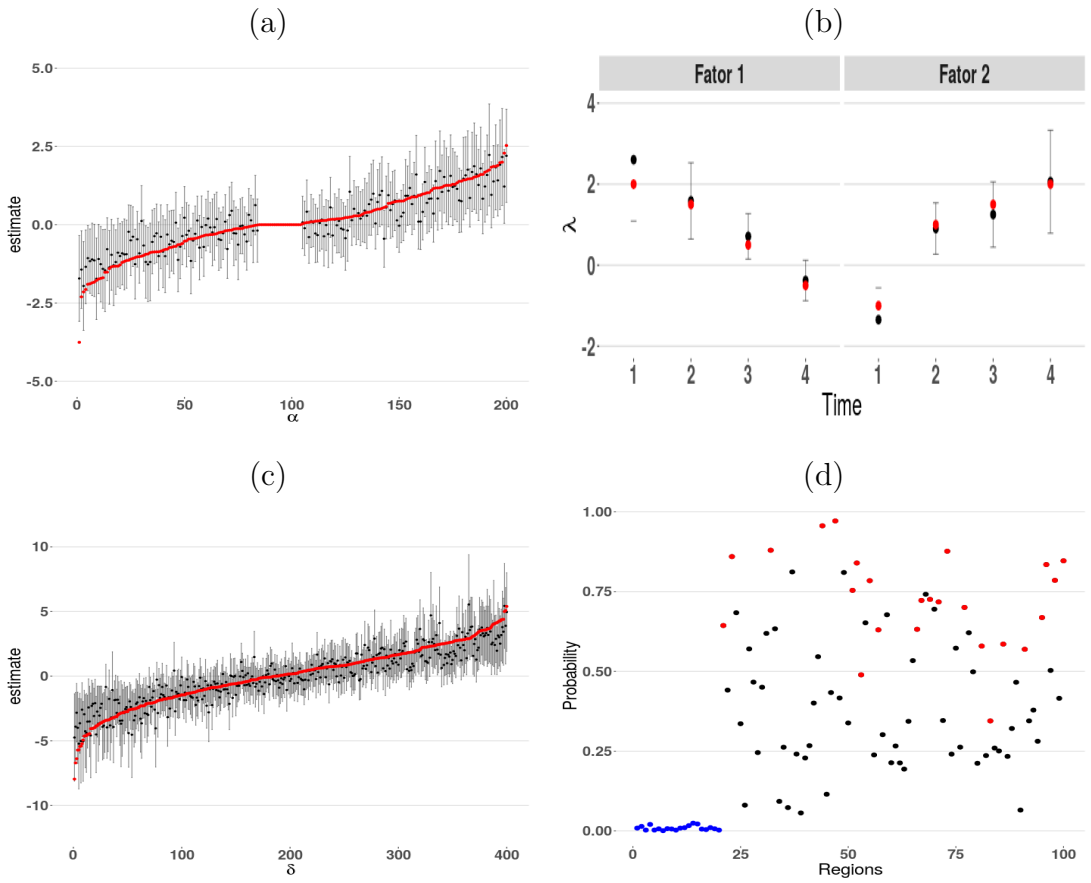


Figure B.2: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais do grupo G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o cenário: $M^{K_2 I_{30\%}}_{L_{100} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

Estimativas para $L = 100$ locais e $\approx 50\%$ locais de G_E afetados pela interação não linear

	Verdadeiro	Média	Mediana	DP	HPD (inf)	HPD (sup)
β_0	0.50	0.47	0.50	0.23	-0.03	0.89
β_1	-1.00	-0.86	-0.86	0.10	-1.05	-0.68
β_2	1.00	1.12	1.12	0.09	0.95	1.31
σ^2	0.80	0.69	0.67	0.18	0.37	1.07
τ	2.00	1.89	1.65	0.91	0.64	3.89
η_1	-2.00	-1.23	-1.20	0.69	-2.63	0.06
η_2	1.50	1.09	1.09	0.43	0.21	1.96
η_3	0.75	0.22	0.22	0.45	-0.68	1.12
η_4	-1.00	-1.07	-1.07	0.51	-2.08	-0.06

Tabela 4: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Cenário: $M^{K_2 I_{50\%}}_{L_{100} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

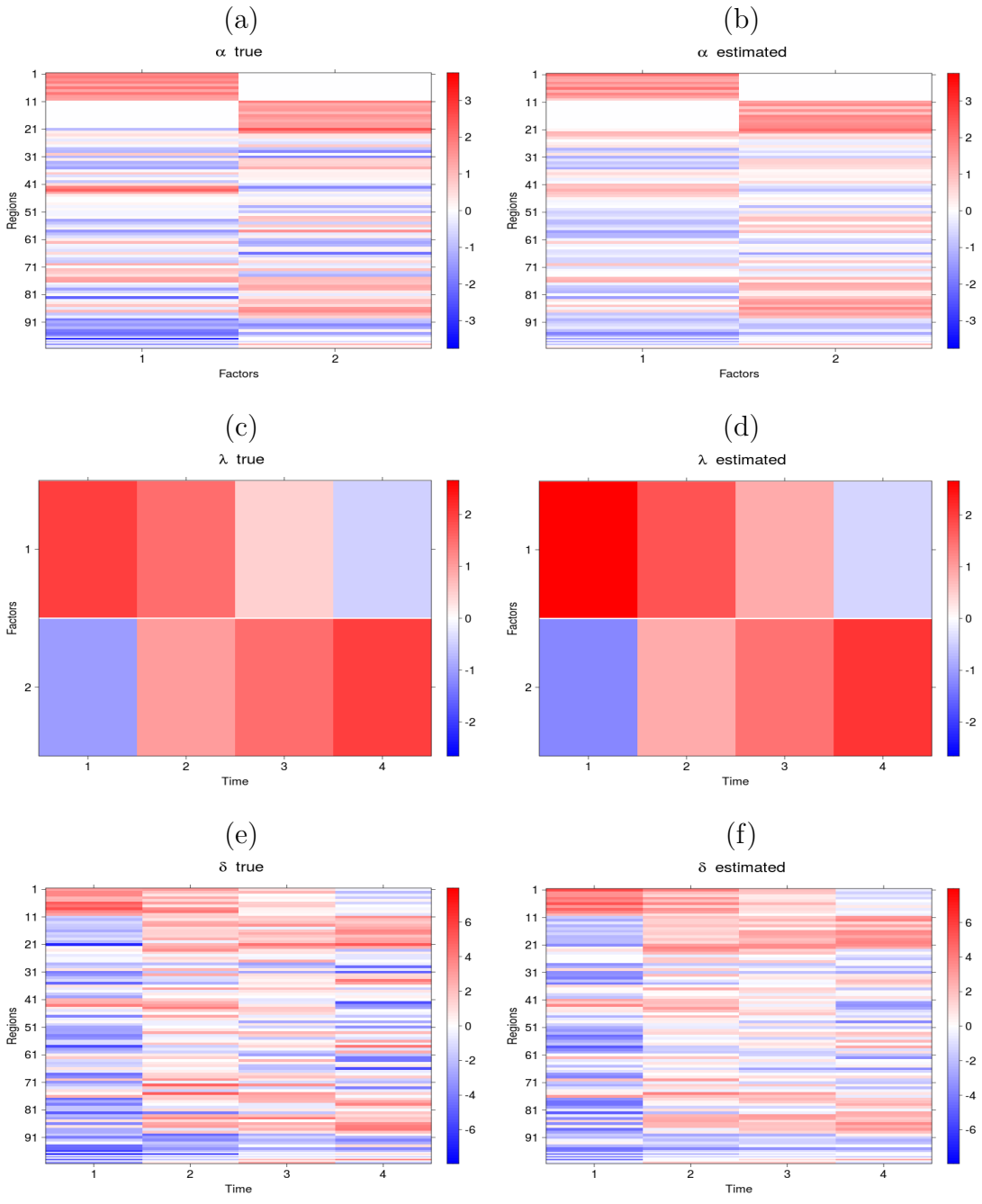


Figure B.3: Mapas de calor comparando valores verdadeiros e estimados para o cenário $M^{K_2 I_{50\%}}_{L_{100} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$. Paineis: (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

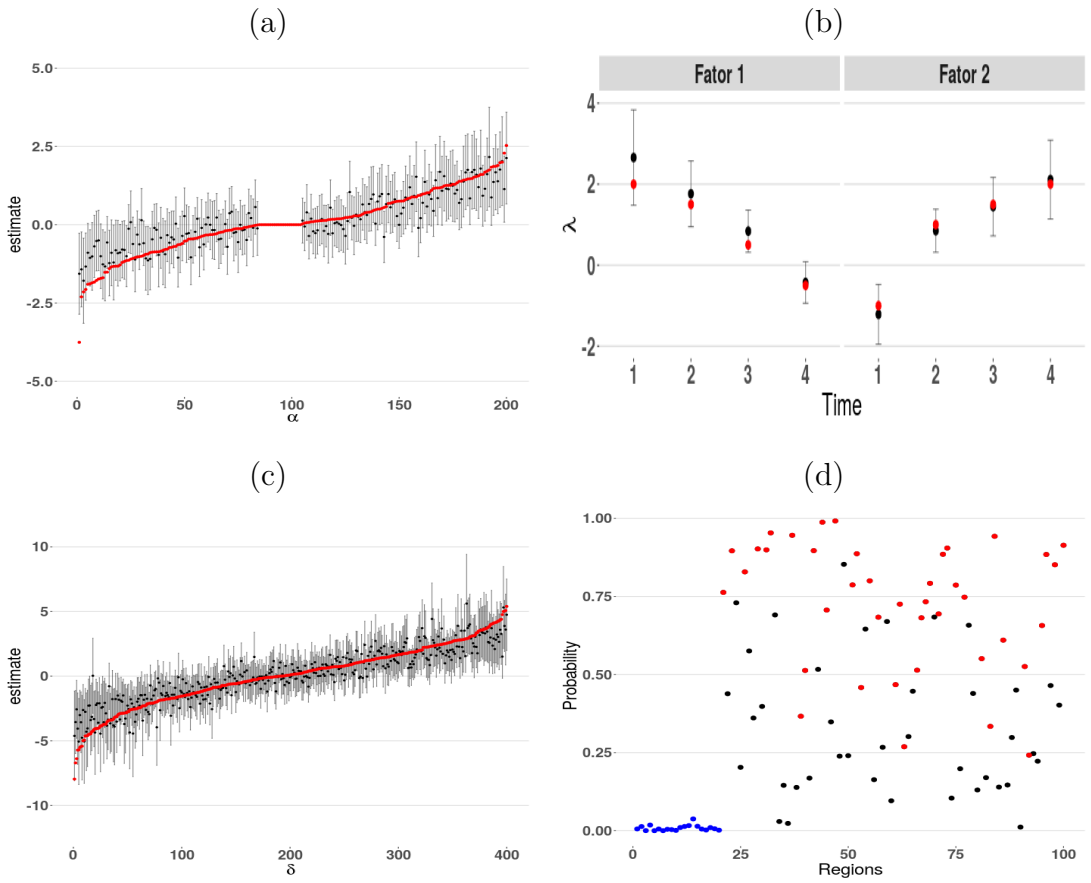


Figure B.4: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais do grupo G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o cenário: $M^{K_2 I_{50\%}}_{L_{100} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

Estimativas para $L = 200$ locais e $\approx 30\%$ locais de G_E afetados pela interação não linear

	Verdadeiro	Média	Mediana	DP	HPD (inf)	HPD (sup)
β_0	0.50	0.44	0.44	0.15	0.14	0.76
β_1	-1.00	-0.94	-0.93	0.07	-1.07	-0.80
β_2	1.00	1.03	1.03	0.06	0.91	1.15
σ^2	0.80	1.11	1.10	0.19	0.75	1.49
τ	2.00	1.94	1.78	0.80	0.70	3.51
η_1	-2.00	-1.37	-1.38	0.45	-2.27	-0.49
η_2	1.50	1.58	1.59	0.39	0.82	2.38
η_3	0.75	1.26	1.27	0.39	0.50	2.02
η_4	-1.00	-0.29	-0.28	0.46	-1.21	0.60

Tabela 5: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Cenário: $M^{K_2 I_{30\%}}_{L_{200} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

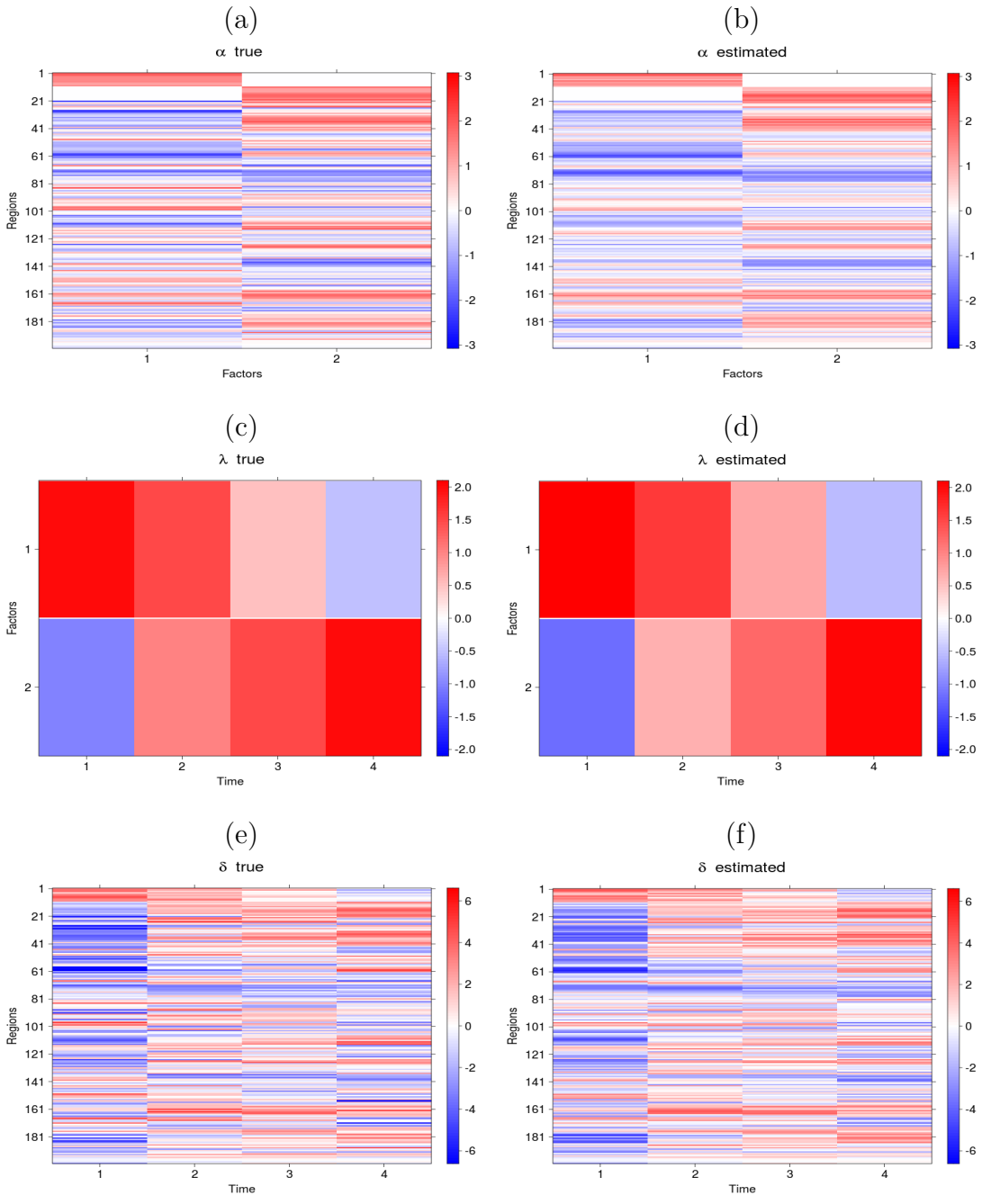


Figure B.5: Mapas de calor comparando valores verdadeiros e estimados para o cenário $M^{K_2 I_{30\%}}_{L_{200} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$. Paineis: (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

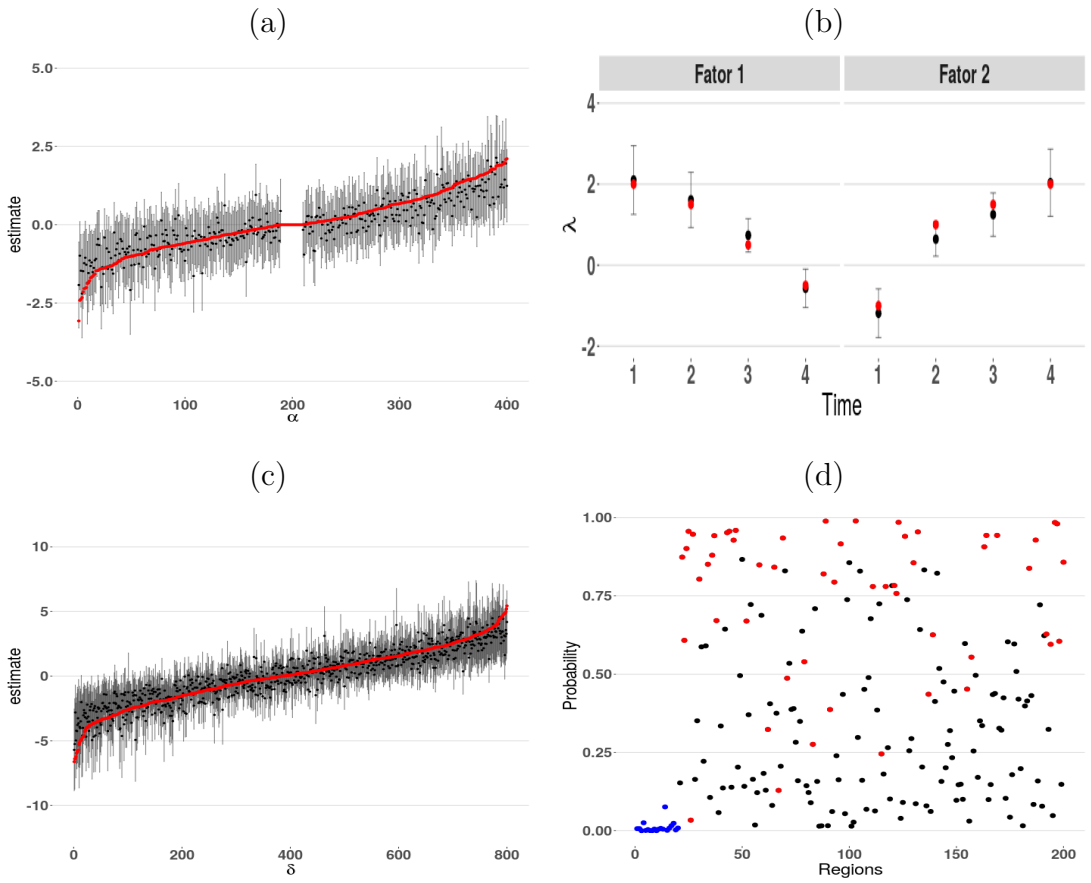


Figure B.6: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais do grupo G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o cenário: $M^{K_2 I_{30\%}}_{L_{200} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

Estimativas para $L = 200$ locais e $\approx 50\%$ locais de G_E afetados pela interação não linear

	Verdadeiro	Média	Mediana	DP	HPD (inf)	HPD (sup)
β_0	0.50	0.30	0.31	0.17	-0.03	0.62
β_1	-1.00	-0.92	-0.92	0.07	-1.06	-0.78
β_2	1.00	0.94	0.94	0.06	0.82	1.06
σ^2	0.80	0.97	0.96	0.17	0.65	1.32
τ	2.00	1.86	1.74	0.71	0.76	3.30
η_1	-2.00	-1.65	-1.64	0.45	-2.55	-0.75
η_2	1.50	1.85	1.85	0.32	1.21	2.47
η_3	0.75	1.48	1.49	0.32	0.82	2.10
η_4	-1.00	-0.48	-0.47	0.38	-1.25	0.24

Tabela 6: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Cenário: $M^{K_2 I_{50\%}}_{L_{200} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

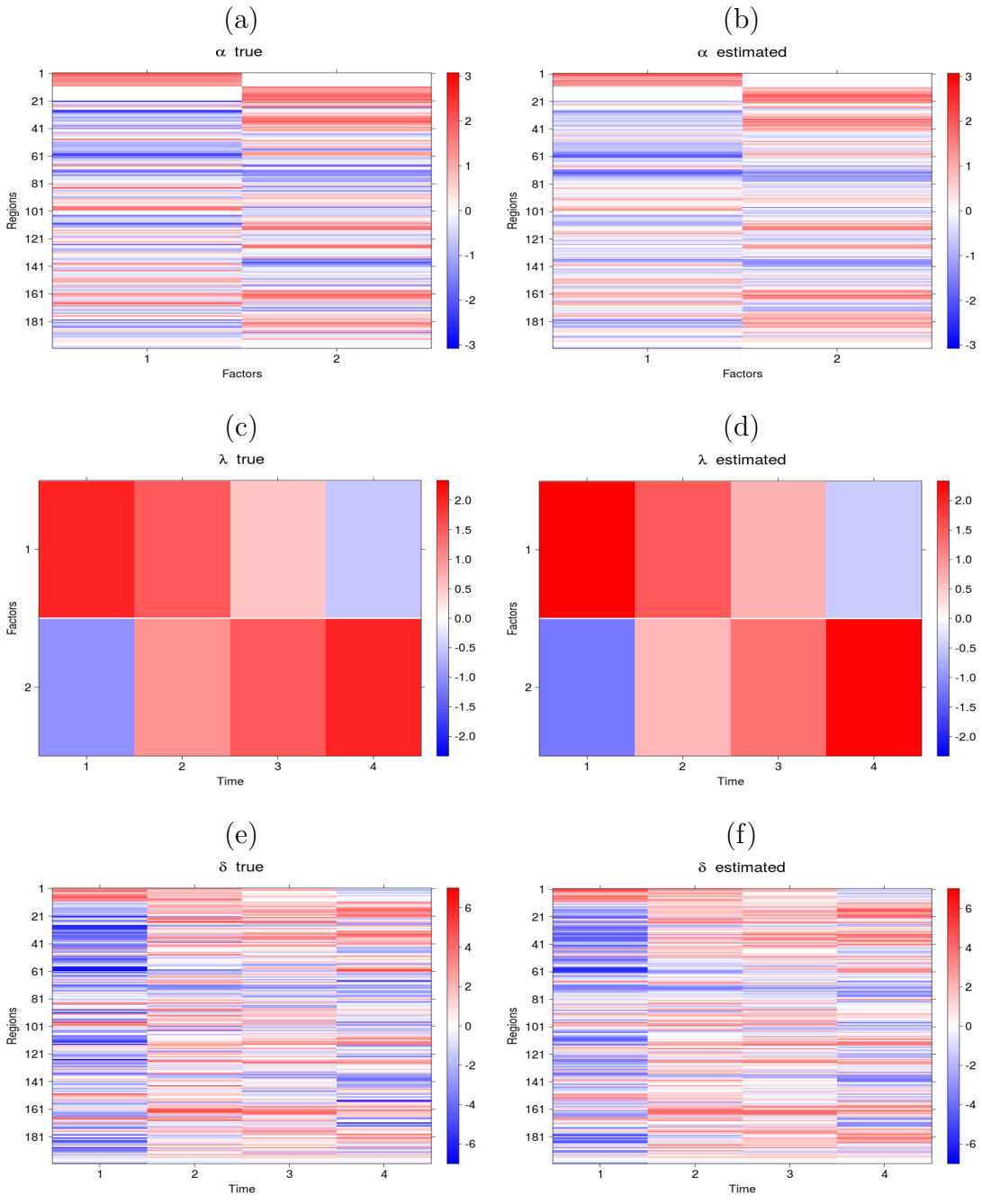


Figure B.7: Mapas de calor comparando valores verdadeiros e estimados para o cenário $M^{K_2 I_{50\%}}_{L_{200} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$. Paineis: (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

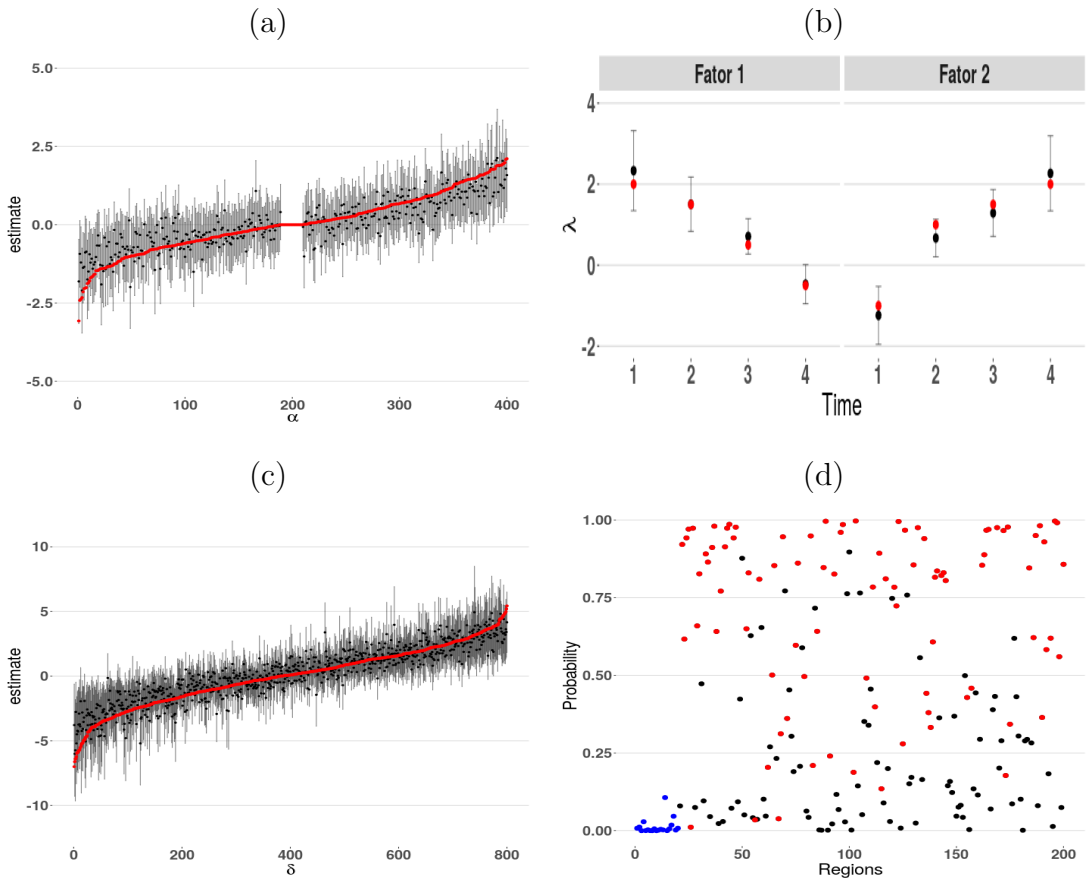


Figure B.8: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais do grupo G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o cenário: $M^{K_2 I_{50\%}}_{L_{200} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

Apêndice C: Modelo Logístico - Estimativas para 6 vizinhos por região

Conforme descrito no texto principal desta tese, as estimativas para os parâmetros nos cenários para 6 vizinhos por região foram muito semelhantes às estimativas para 4 vizinhos. Apresentamos, aqui, gráficos comparativos entre os casos com $L = 100, 200$ e 400 locais, $K = 2$ fatores, $T = 4$ tempos, $\approx 30\%$ e 50% de locais de G_E afetados por interação. Considere em todas as situações $\approx 50\%$ de $Y_i^{'}s = 1$. Também optamos por ilustrar alguns gráficos comparativos do intervalo HPD de 95% para os parâmetros η^* e λ. Em relação ao vício relativo a partir da execução de 30 réplicas de Monte Carlo apresentamos os resultados para α, λ e δ.

Intervalo HPD de 95% para η^* e λ

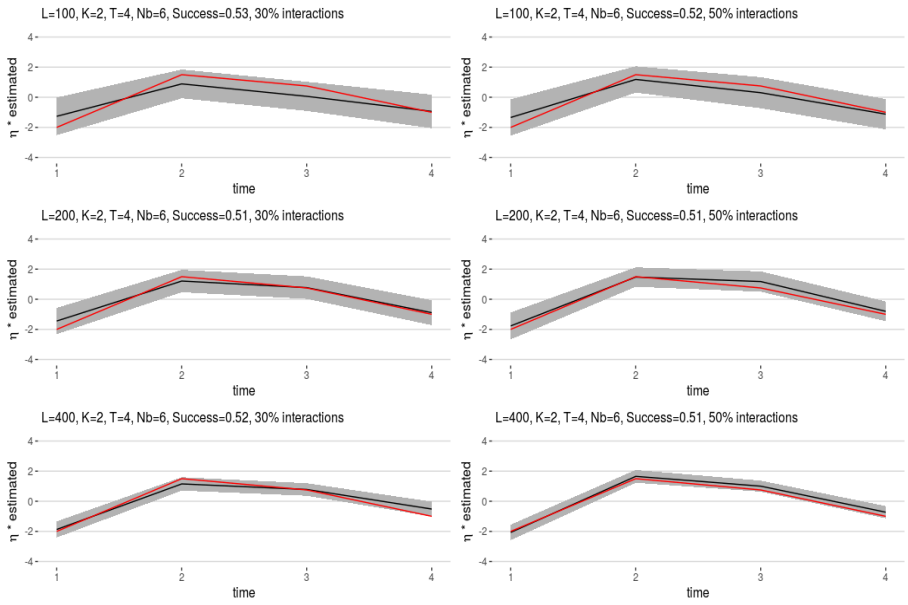


Figure C.1: Média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha) em todas as configurações de números de regiões (L = 100, 200 e 400). Considere os cenários: $M_{L100T4V6}^{K2I30\%}$ e $M_{L100T4V6}^{K2I50\%}$, $M_{L200T4V6}^{K2I30\%}$ e $M_{L200T4V6}^{K2I50\%}$, $M_{L400T4V6}^{K2I30\%}$ e $M_{L400T4V6}^{K2I50\%}$ com $\approx 50\%$ de $Y_i^{'}s = 1$, ou seja, $\beta = (0.5, -1.0, 1.0)$

.

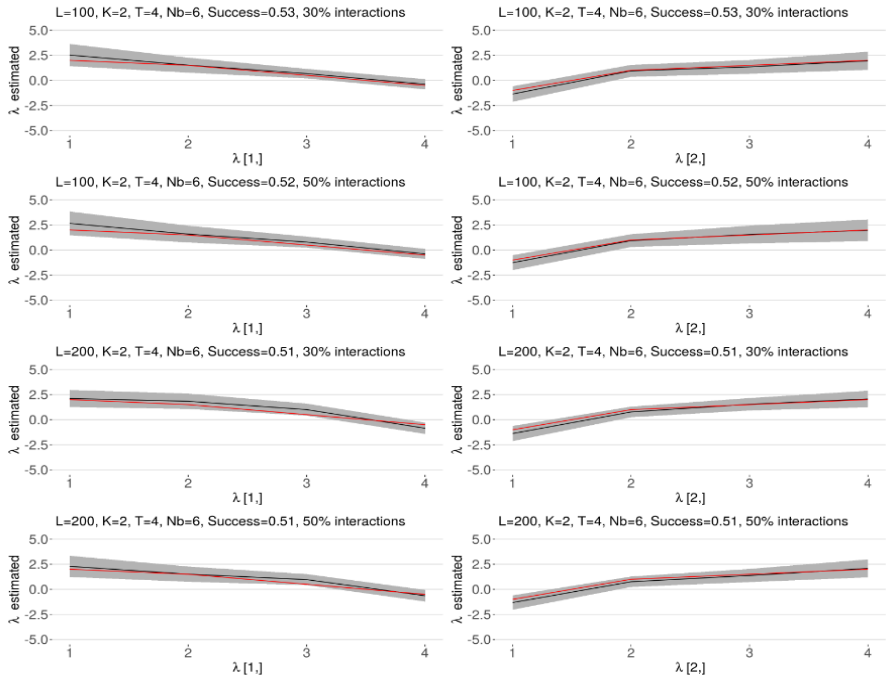


Figure C.2: Média *a posteriori* (linha preta), intervalo HPD de 95% para λ (área sombreada) e valor verdadeiro (linha vermelha). Considere os cenários: $M^{K_2 I_{30\%}}_{L_{100} T_4 V_6}$, $M^{K_2 I_{50\%}}_{L_{100} T_4 V_6}$ e $M^{K_2 I_{30\%}}_{L_{200} T_4 V_6}$, $M^{K_2 I_{50\%}}_{L_{200} T_4 V_6}$ com $\approx 50\%$ de $Y_i^{'} s = 1$, ou seja, $\beta = (0.5, -1.0, 1.0)$

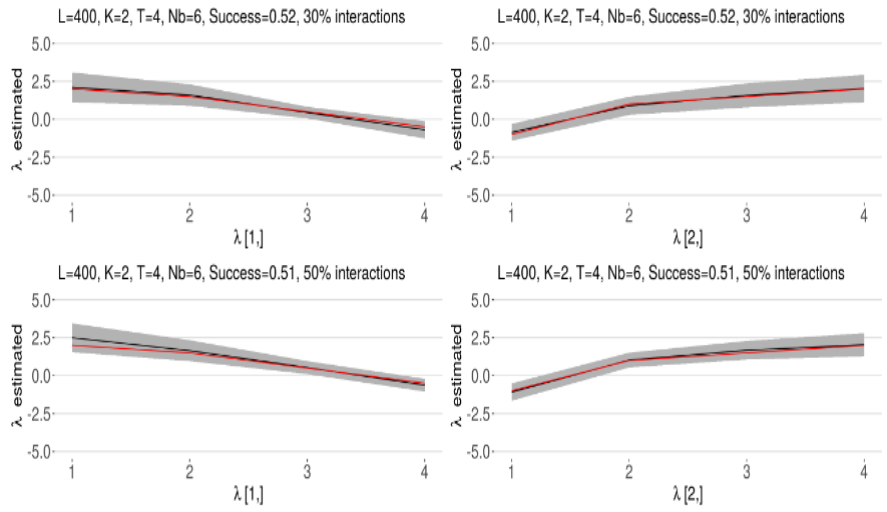


Figure C.3: Média *a posteriori* (linha preta), intervalo HPD de 95% para λ (área sombreada) e valor verdadeiro (linha vermelha). Considere os cenários: $M^{K_2 I_{30\%}}_{L_{400} T_4 V_4}$ e $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$, ou seja, $\beta = (0.5, -1.0, 1.0)$

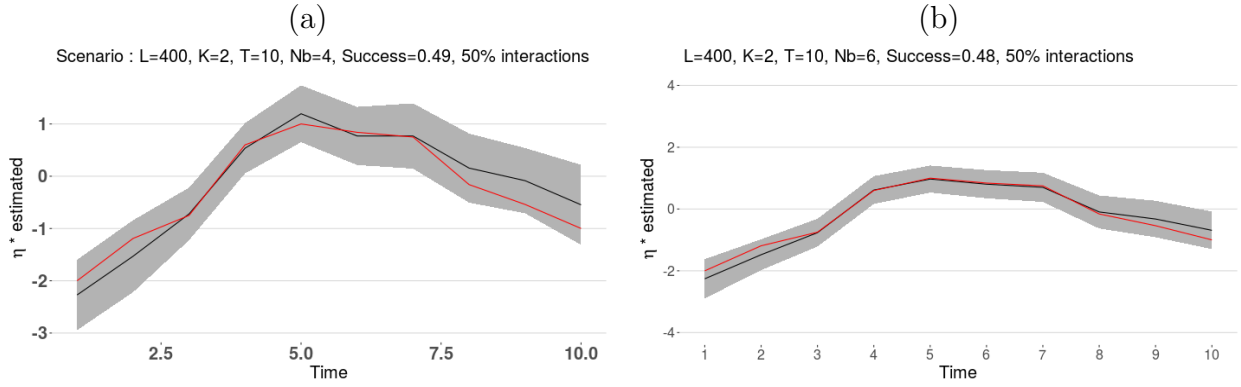


Figure C.4: Gráfico da média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha) para os cenários $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ e $M^{K_2 I_{50\%}}_{L_{400} T_4 V_6}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

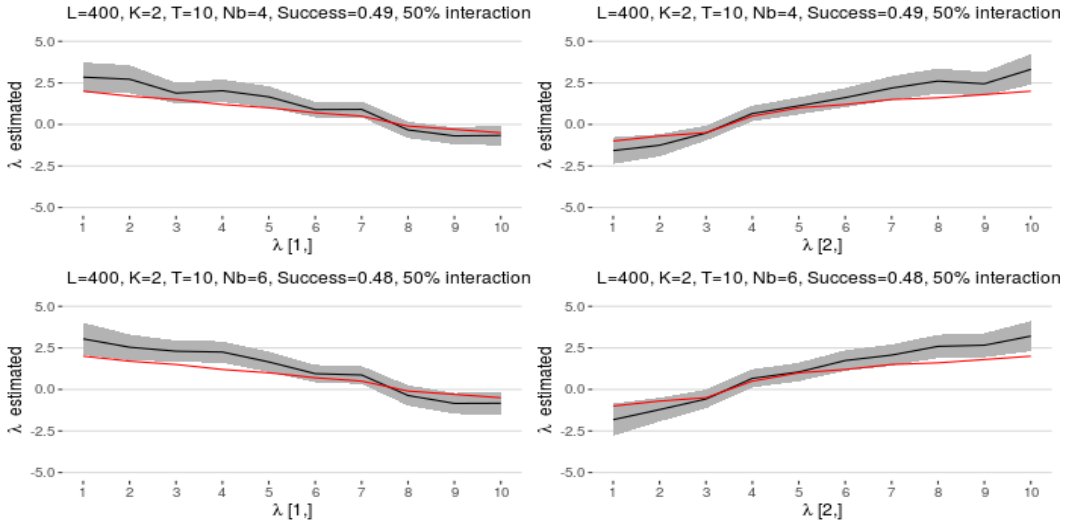


Figure C.5: Gráfico da média *a posteriori* (linha preta), intervalo HPD de 95% para λ (área sombreada) e valor verdadeiro (linha vermelha) para os cenários $M^{K_2 I_{50\%}}_{L_{400} T_{10} V_4}$ e $M^{K_2 I_{50\%}}_{L_{400} T_{10} V_6}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

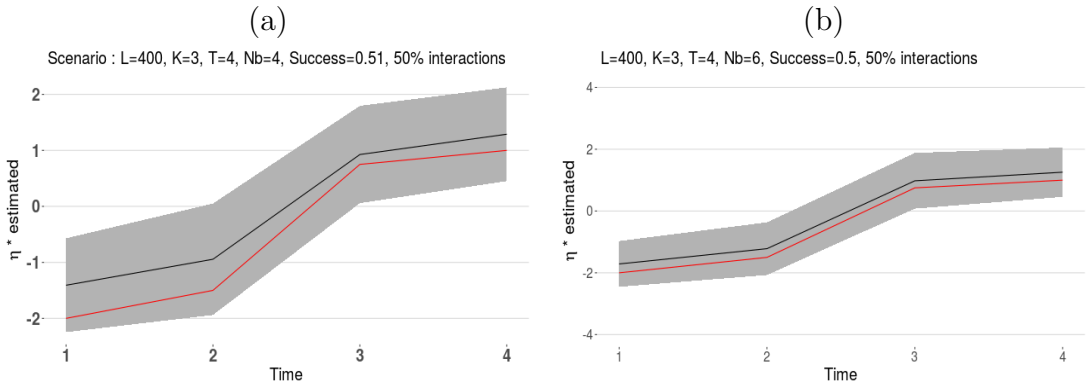


Figure C.6: Gráfico da média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha) para os cenários $M^{K_3 I_{50\%}}_{L_{400} T_4 V_4}$ e $M^{K_3 I_{50\%}}_{L_{400} T_4 V_6}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

Análise Monte Carlo do vício relativo para α, λ e δ

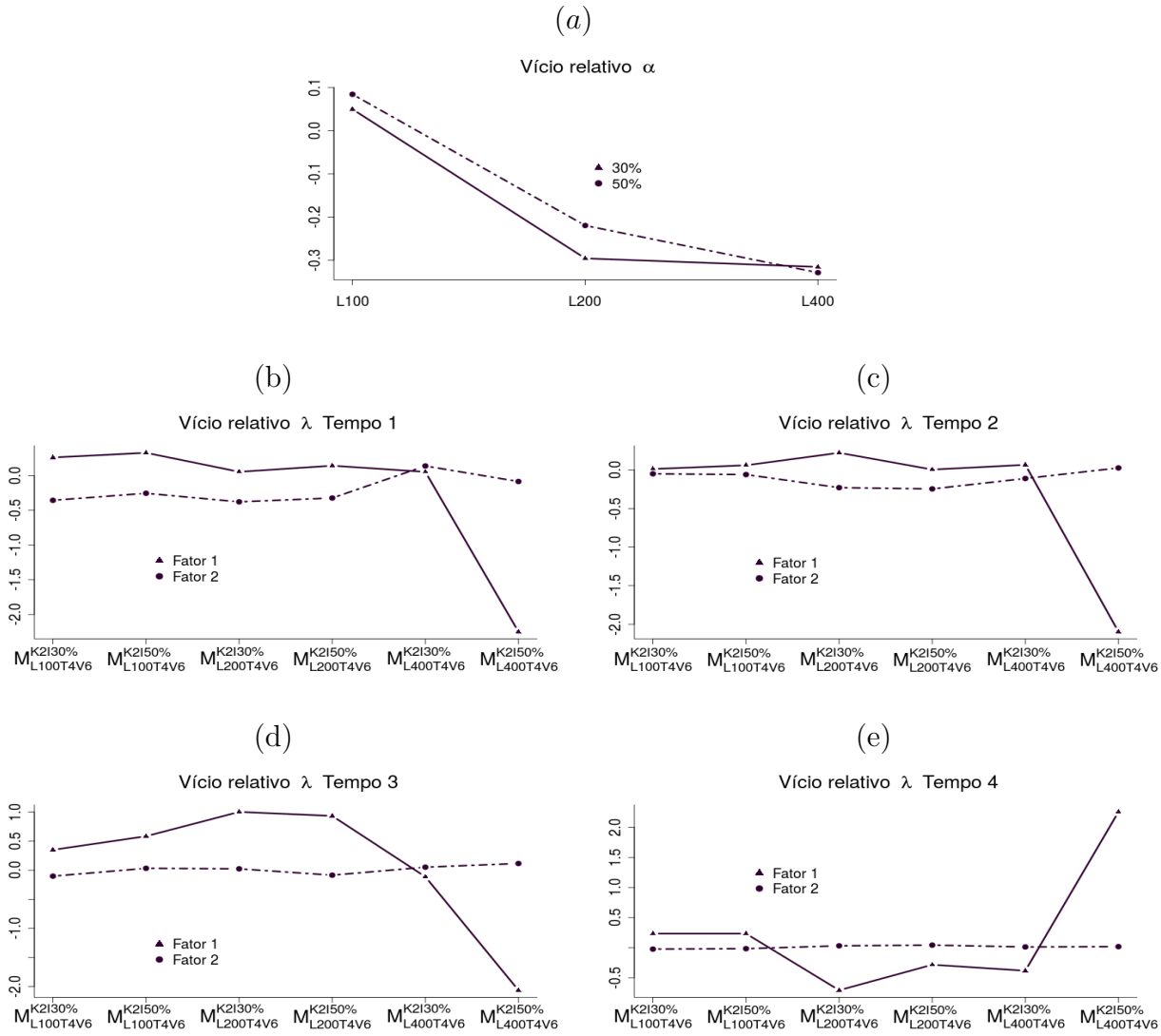


Figure C.7: Mediana do vício relativo de α e λ calculada a partir de amostras de tamanho 100 (α) de cada uma das 30 réplicas de Monte Carlo. A expressão do vício é dada por $\frac{(\hat{\zeta}-\zeta)}{|\zeta|}$, em que $\hat{\zeta}$ representa, genericamente, o valor estimado e ζ, o verdadeiro, e $|\zeta|$ simboliza o valor verdadeiro absoluto. Considere os cenários com $K = 2$ fatores, $T = 4$ tempos, 6 vizinhos por região, número de locais $L \in \{100, 200, 400\}$, $\approx 30\%$ e 50% de locais de G_E afetados pela interação não linear.

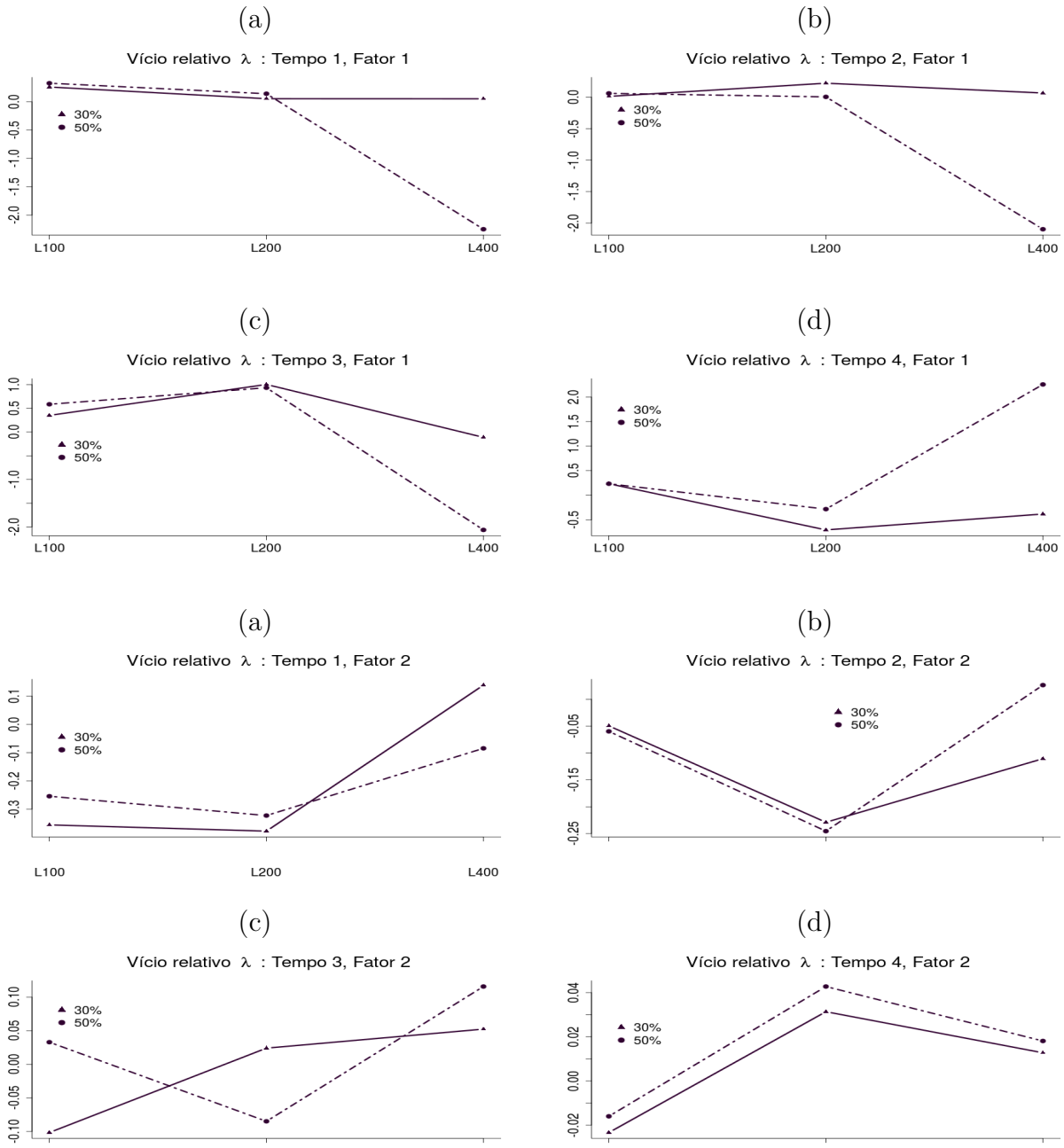


Figure C.8: Mediana do vício relativo de $\lambda_{1\bullet}$ e $\lambda_{2\bullet}$ para os 4 tempos. O valor foi calculado a partir de cada uma das 30 réplicas de Monte Carlo. A expressão do vício é dada por $\frac{(\hat{\zeta}-\zeta)}{|\zeta|}$, em que $\hat{\zeta}$ representa, genericamente, o valor estimado e ζ, o verdadeiro, e $|\zeta|$ simboliza o valor verdadeiro absoluto. Considere os cenários com $K = 2$ fatores, $T = 4$ tempos, 6 vizinhos por região, número de locais $L \in \{100, 200, 400\}$, $\approx 30\%$ e 50% de locais de G_E afetados pela interação não linear.

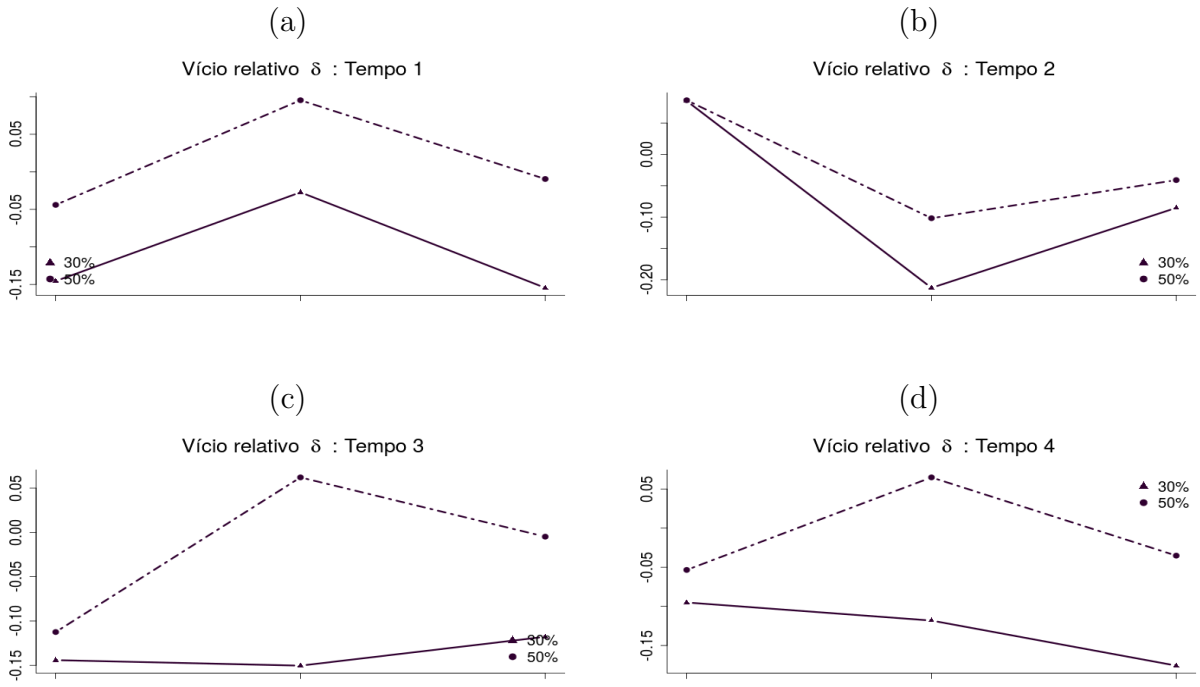


Figure C.9: Mediana do vício relativo de δ para cada tempo calculado a partir de amostras de tamanho 100 de cada uma das 30 réplicas de Monte Carlo. O vício é calculado pela expressão $\frac{(\hat{\delta}-\delta)}{|\delta|}$. Considere os cenários com $K = 2$ fatores, $T = 4$ tempos, 6 vizinhos por região, número de locais $L \in \{100, 200, 400\}$, 30% e 50% de locais de G_E afetados por η^*.

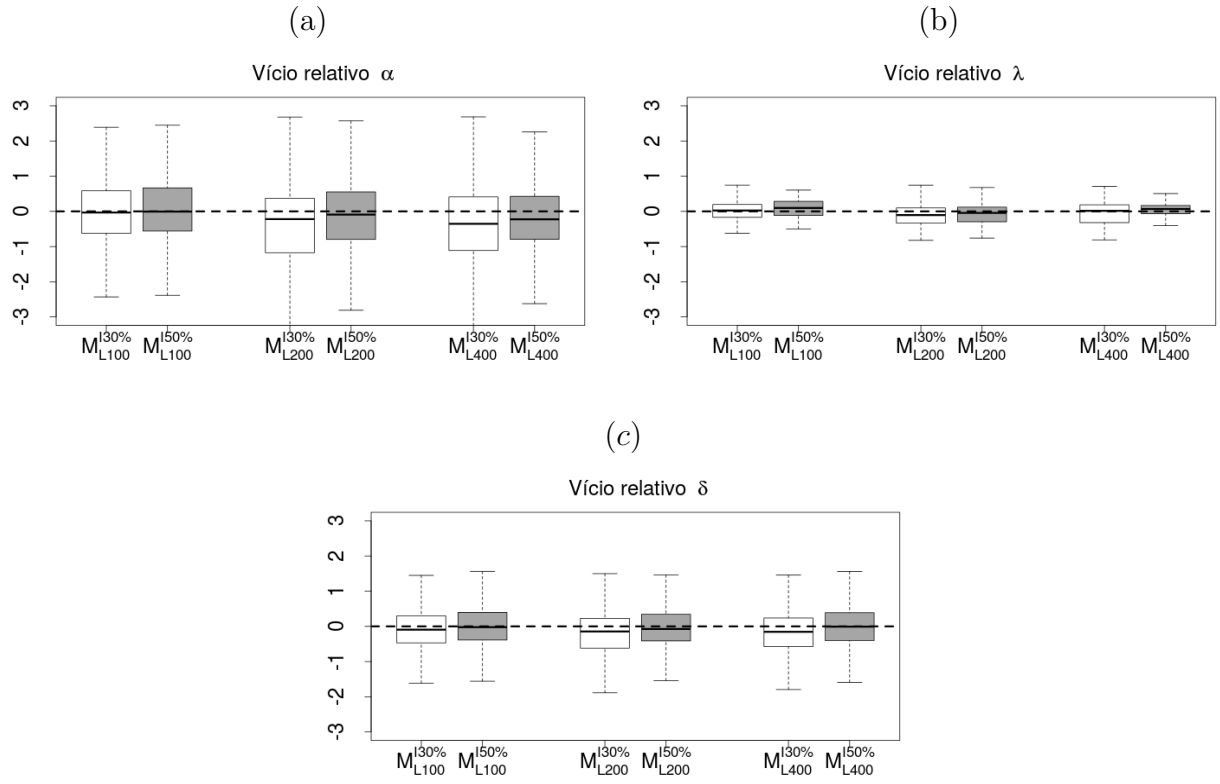


Figure C.10: Mediana do vício relativo de α, λ e δ calculada a partir de amostras de tamanho 100 (α e δ) de cada uma das 30 réplicas de Monte Carlo. A expressão do vício é dada por $\frac{(\hat{\zeta}-\zeta)}{|\zeta|}$, em que $\hat{\zeta}$ representa, genericamente, o valor estimado e ζ, o verdadeiro, e $|\zeta|$ simboliza o valor verdadeiro absoluto. Considere os cenários com $K = 2$ fatores, $T = 4$ tempos, 6 vizinhos por região, número de locais $L \in \{100, 200, 400\}$, 30% e 50% de locais de G_E afetados pela interação não linear.

Apêndice D: Modelo Poisson - Estimativas para 100 e 200 locais

Apresentamos, aqui, as estimativas para os casos Poisson com $L = 100$ e 200 locais, $T = 4$ tempos, $K = 2$ fatores, $\approx 40\%$ de contagens zero para a variável resposta e com locais de G_E afetados por interação de $\approx 30\%$ e 50%.

Cenário: $L = 100$ locais e $\approx 30\%$ de locais em G_E afetados por η^*.

	Verdadeiro	Média	Mediana	DP	HPD.Linf	HPD.Lsup
β_0	0.50	0.50	0.47	0.10	0.36	0.66
β_1	-1.00	-0.98	-0.98	0.01	-1.00	-0.95
β_2	1.00	0.98	0.98	0.01	0.96	1.00
σ^2	0.80	0.90	0.89	0.13	0.66	1.15
τ_α	2.00	1.80	1.65	0.74	0.69	3.20
η_1^*	-2.00	-1.11	-1.10	0.59	-2.26	0.04
η_2^*	1.50	0.89	0.89	0.48	-0.03	1.83
η_3^*	0.75	0.13	0.12	0.50	-0.81	1.15
η_4^*	-1.00	-0.98	-0.99	0.62	-2.25	0.25

Tabela 7: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o caso Poisson : $M_{L_{100}T_4V_4}^{K_2I_{30\%}}$ com $\approx 40\%$ de contagens zero.

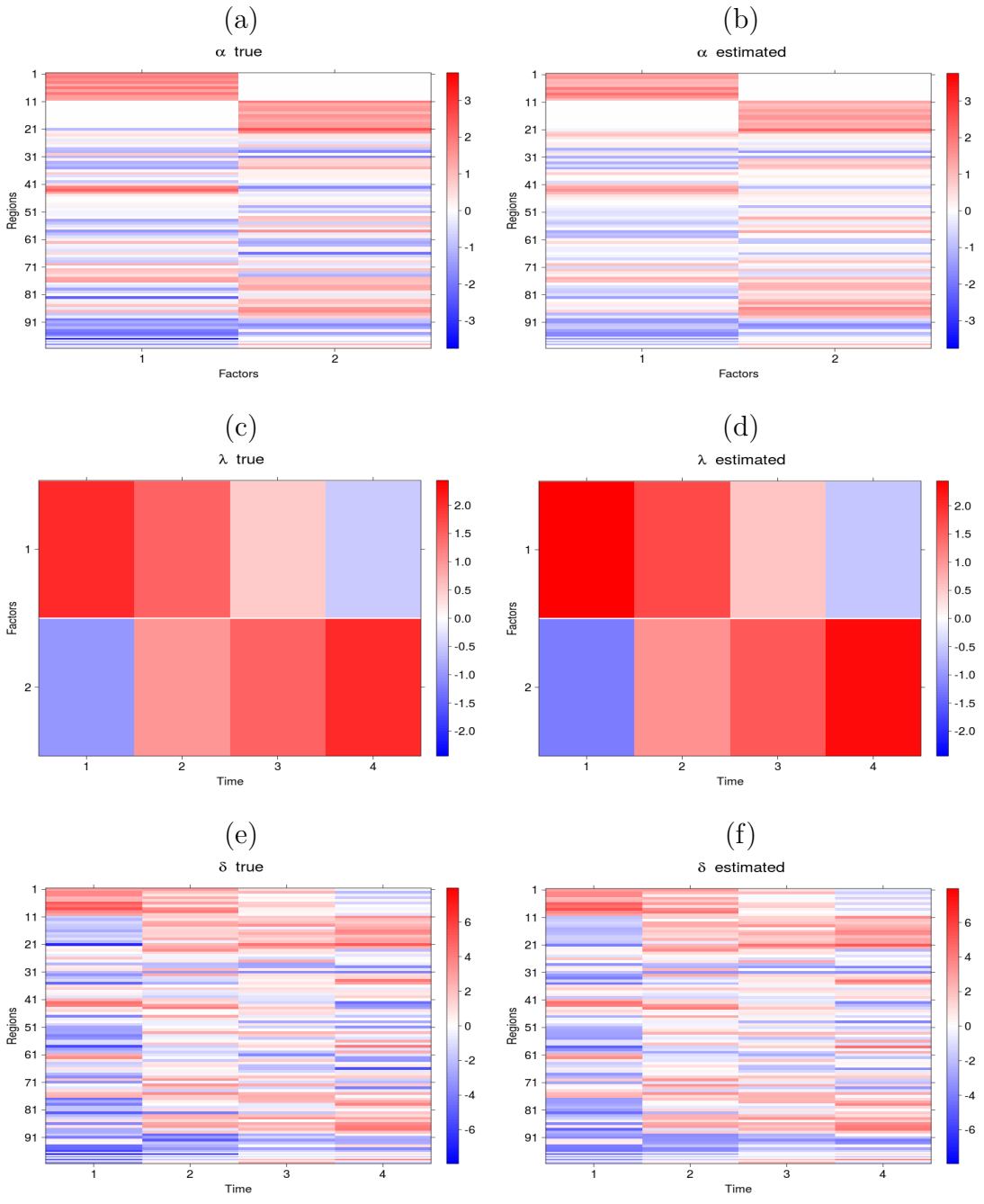


Figure D.1: Mapas de calor comparando valores verdadeiros e estimados para o caso Poisson $M^{K_2 I_{30\%}}_{L_{100} T_4 V_4}$ e $\approx 40\%$ de contagens zero. Paineis : (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

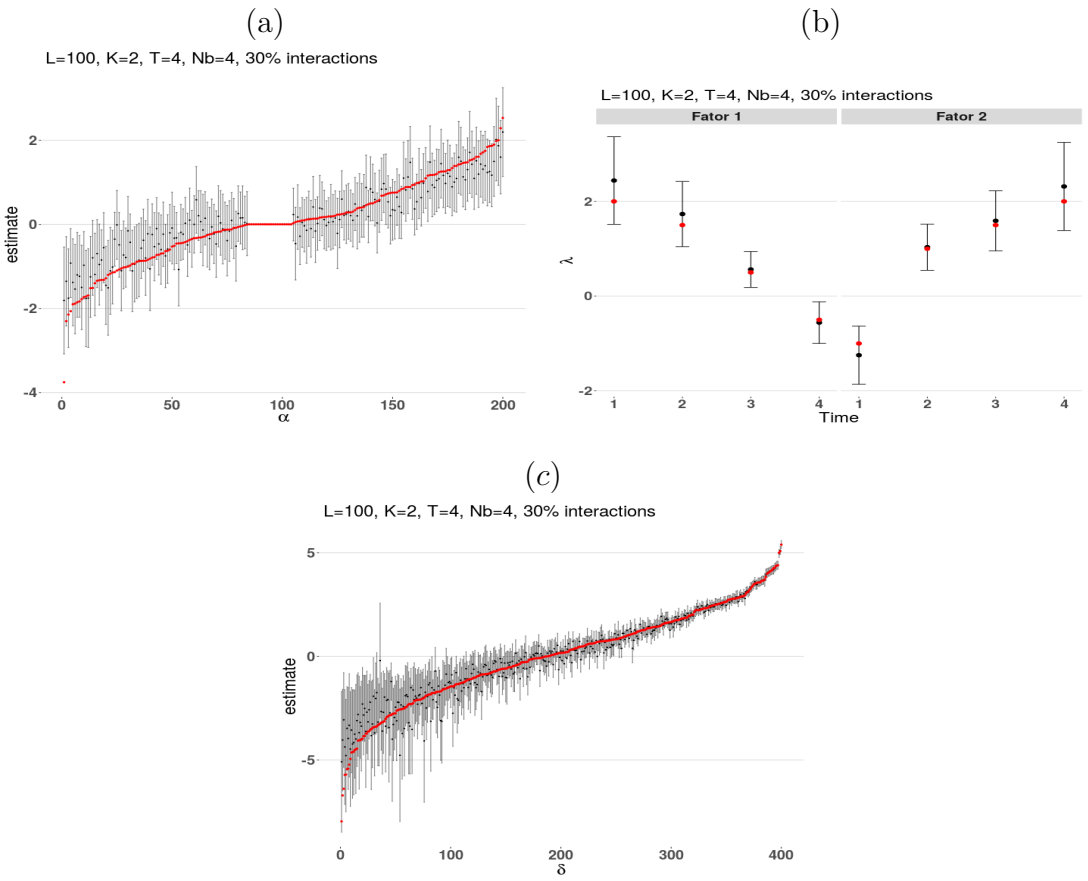


Figure D.2: Análise gráfica dos intervalos HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais do grupo G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o caso Poisson : $M^{K_2 I_{30\%}}_{L_{100} T_4 V_4}$ com $\approx 40\%$ de contagens zero.

Cenário: $L = 100$ locais e $\approx 50\%$ de locais em G_E afetados por η^*.

	Verdadeiro	Média	Mediana	DP	HPD.Linf	HPD.Lsup
β_0	0.50	0.57	0.57	0.06	0.47	0.68
β_1	-1.00	-0.99	-0.99	0.01	-1.01	-0.96
β_2	1.00	1.01	1.01	0.01	0.98	1.03
σ^2	0.80	0.89	0.88	0.12	0.66	1.14
τ_α	2.00	1.57	1.42	0.65	0.63	2.89
η_1^*	-2.00	-1.30	-1.31	0.56	-2.41	-0.18
η_2^*	1.50	1.31	1.32	0.38	0.56	2.06
η_3^*	0.75	0.33	0.34	0.39	-0.43	1.08
η_4^*	-1.00	-1.18	-1.19	0.51	-2.17	-0.17

Tabela 8: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o caso Poisson : $M_{L_{100}T_4V_4}^{K_2I_{50\%}}$ com $\approx 40\%$ de contagens zero.

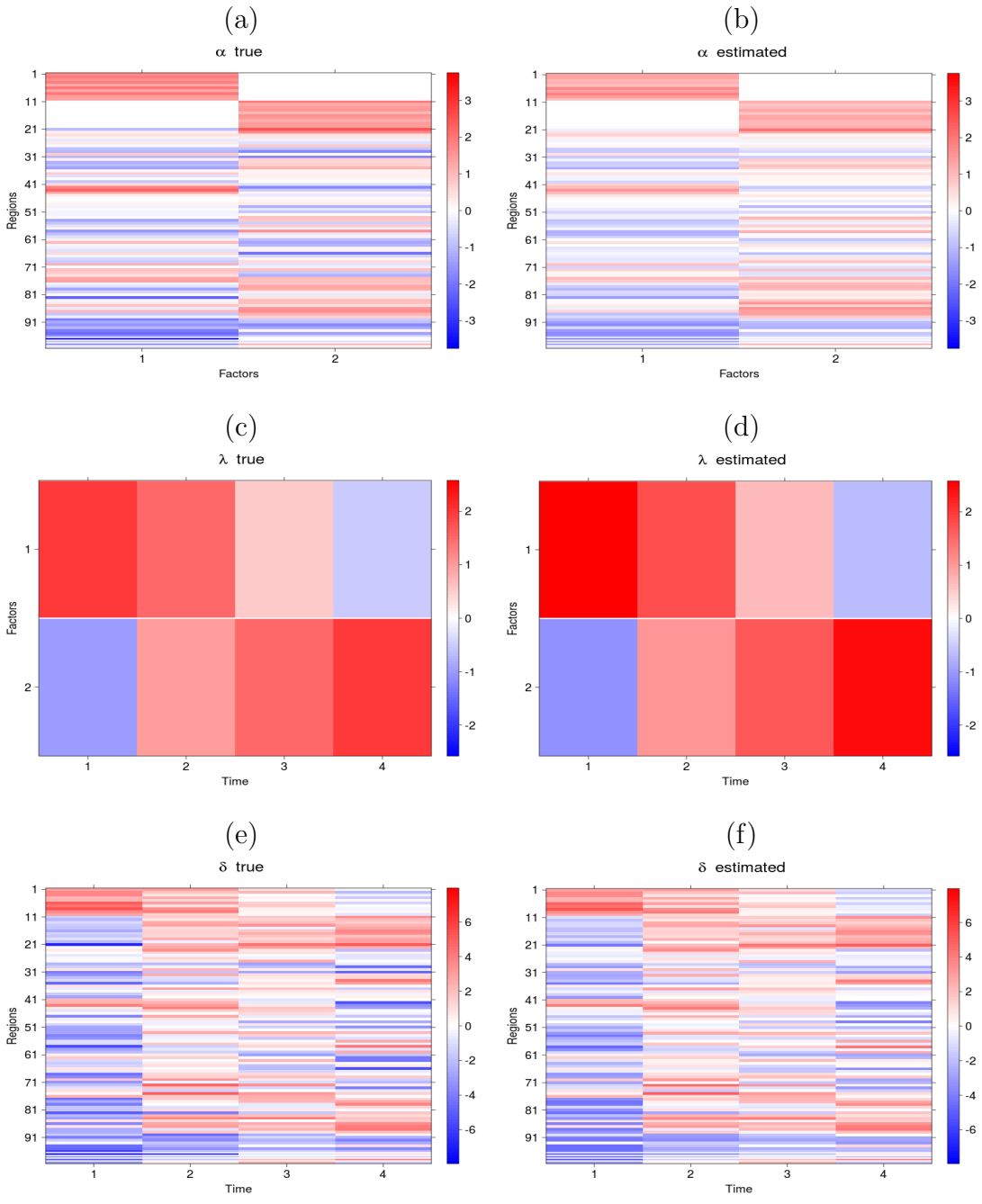


Figure D.3: Mapas de calor comparando valores verdadeiros e estimados para o caso Poisson $M^{K_2 I_{50\%}}_{L_{100} T_4 V_4}$ e $\approx 40\%$ de contagens zero. Paineis : (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

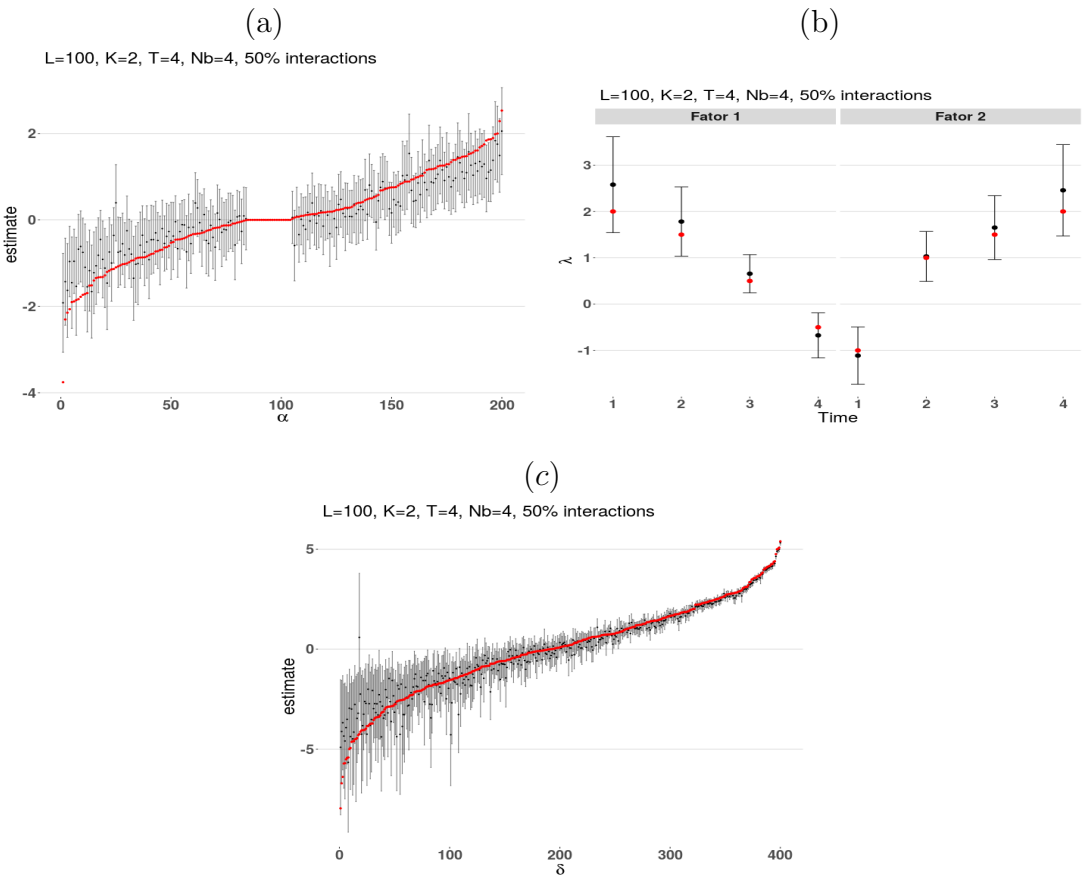


Figure D.4: Análise gráfica dos intervalos HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais do grupo G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o caso Poisson : $M^{K_2 I_{50\%}}_{L_{100} T_4 V_4}$ com $\approx 40\%$ de contagens zero.

Cenário: $L = 200$ locais e $\approx 30\%$ de locais em G_E afetados por η^*.

	Verdadeiro	Média	Mediana	DP	HPD.Linf	HPD.Lsup
β_0	0.50	0.69	0.68	0.05	0.62	0.78
β_1	-1.00	-1.01	-1.01	0.01	-1.03	-1.00
β_2	1.00	1.01	1.01	0.01	0.99	1.02
σ^2	0.80	0.90	0.89	0.10	0.71	1.09
τ_α	2.00	2.50	2.02	1.40	0.82	5.69
η_1^*	-2.00	-1.48	-1.48	0.47	-2.43	-0.58
η_2^*	1.50	1.60	1.61	0.30	1.02	2.18
η_3^*	0.75	0.97	0.97	0.28	0.39	1.51
η_4^*	-1.00	-0.81	-0.82	0.42	-1.60	0.05

Tabela 9: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o caso Poisson : $M_{L_{200}T_4V_4}^{K_2I_{30\%}}$ com $\approx 40\%$ de contagens zero.

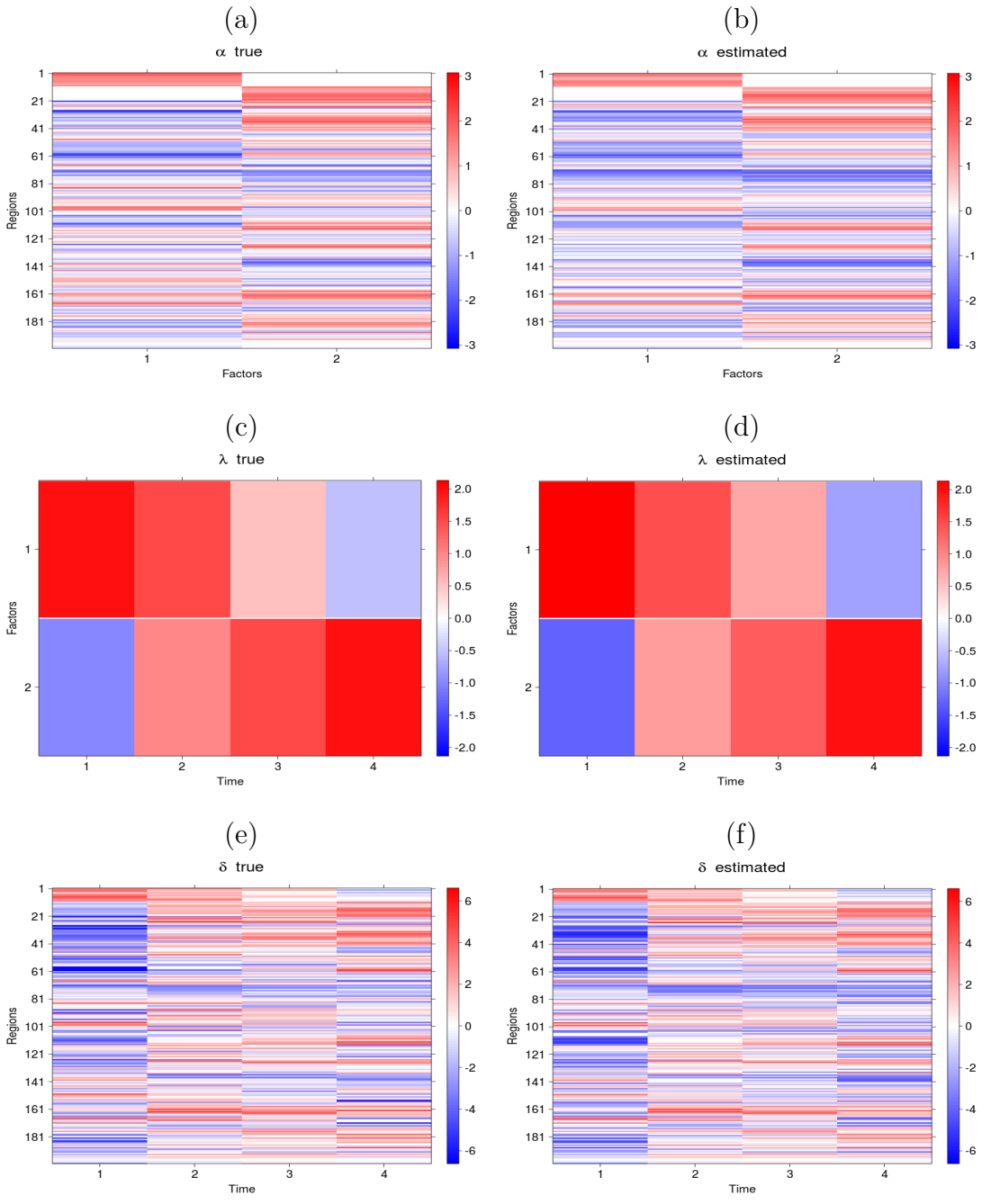


Figure D.5: Mapas de calor comparando valores verdadeiros e estimados para o caso Poisson $M^{K_2 I_{30\%}}_{L_{200} T_4 V_4}$ e $\approx 40\%$ de contagens zero. Paineis : (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

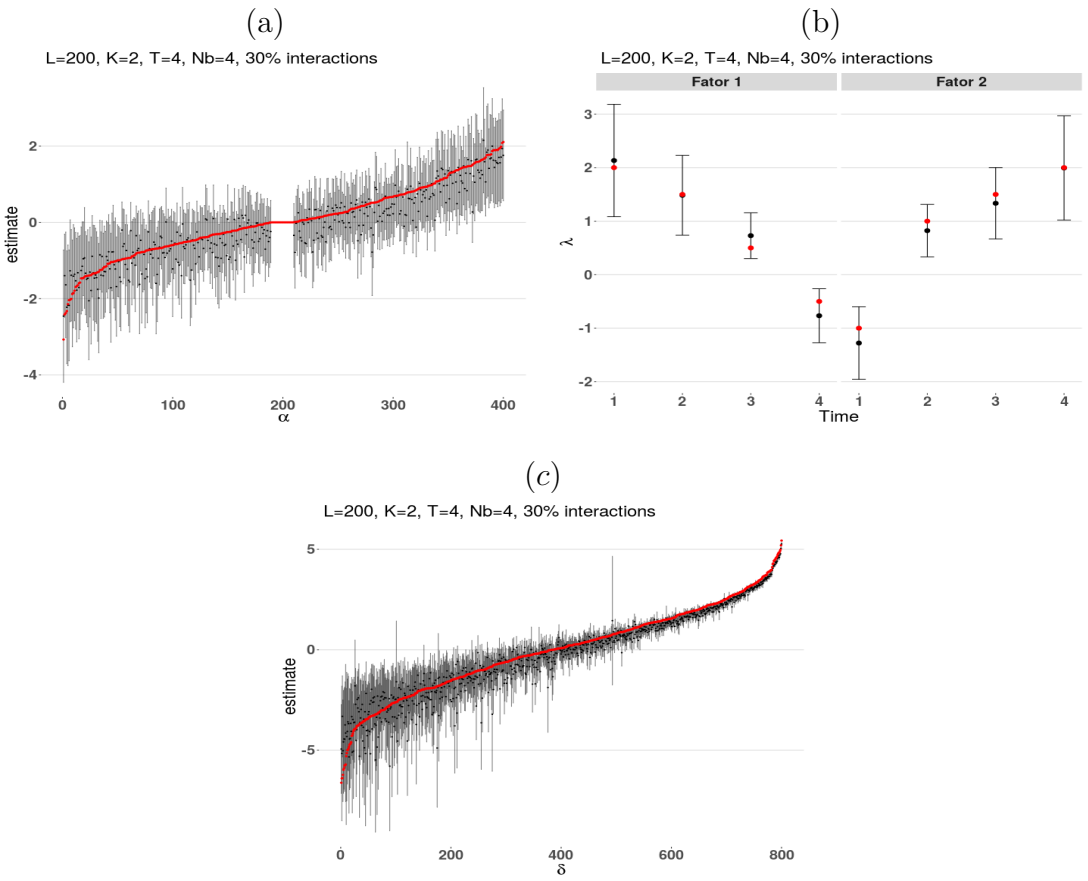


Figure D.6: Análise gráfica dos intervalos HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais do grupo G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o caso Poisson : $M^{K_2 I_{30\%}}_{L_{200} T_4 V_4}$ com $\approx 40\%$ de contagens zero.

Cenário: $L = 200$ locais e $\approx 50\%$ de locais em G_E afetados por η^*.

	Verdadeiro	Média	Mediana	DP	HPD.Linf	HPD.Lsup
β_0	0.50	0.45	0.45	0.03	0.40	0.50
β_1	-1.00	-1.00	-1.00	0.01	-1.01	-0.98
β_2	1.00	0.98	0.98	0.01	0.97	1.00
σ^2	0.80	0.79	0.79	0.08	0.64	0.96
τ_α	2.00	2.56	2.23	1.40	0.53	5.64
η_1^*	-2.00	-1.79	-1.77	0.47	-2.75	-0.90
η_2^*	1.50	1.83	1.84	0.27	1.28	2.36
η_3^*	0.75	1.12	1.13	0.27	0.59	1.65
η_4^*	-1.00	-0.86	-0.86	0.33	-1.52	-0.18

Tabela 10: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Considere o caso Poisson : $M_{L_{200}T_4V_4}^{K_2I_{50\%}}$ com $\approx 40\%$ de contagens zero.

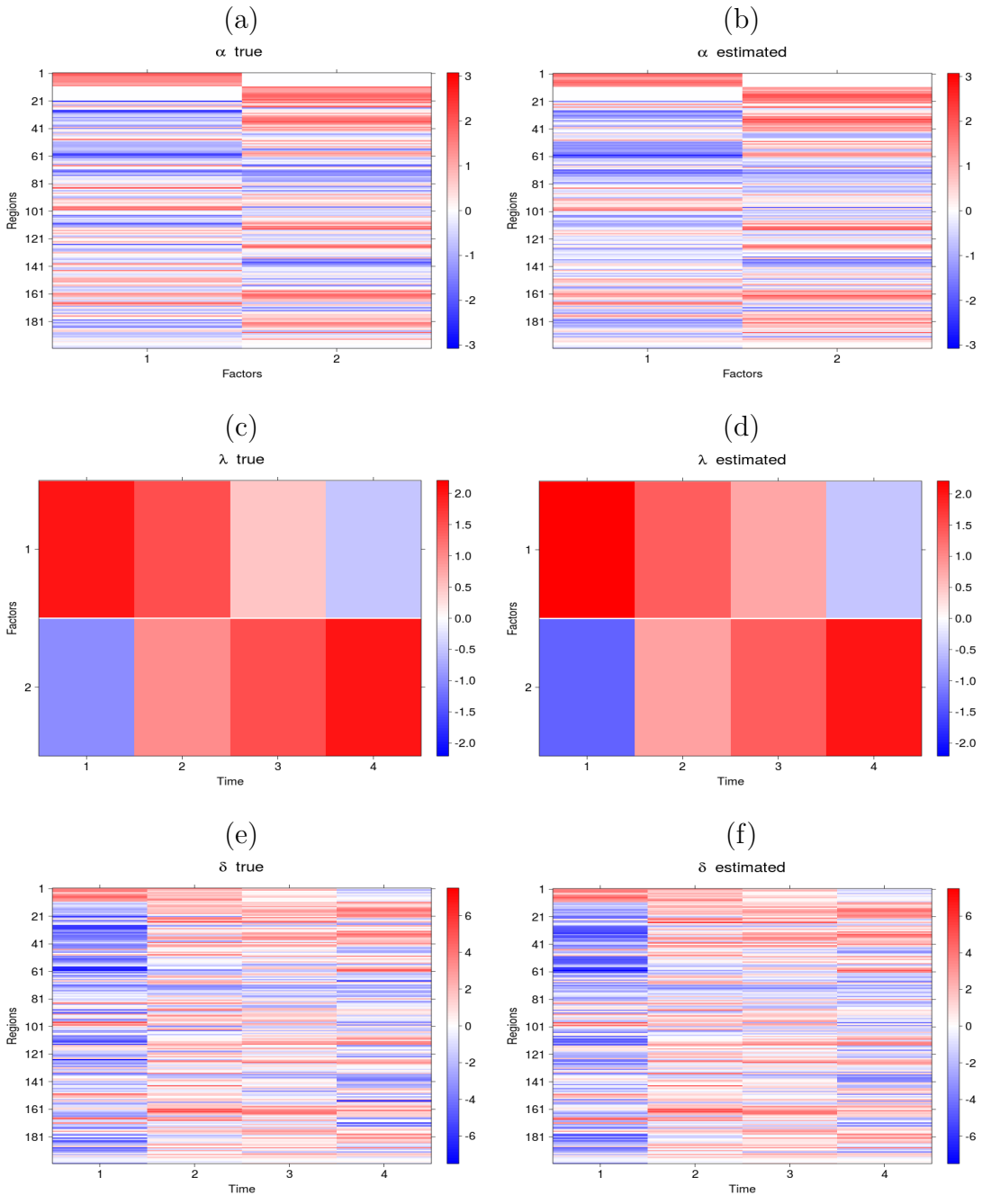


Figure D.7: Mapas de calor comparando valores verdadeiros e estimados para o caso Poisson $M^{K_2 I_{50\%}}_{L_{200} T_4 V_4}$ e $\approx 40\%$ de contagens zero. Paineis : (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

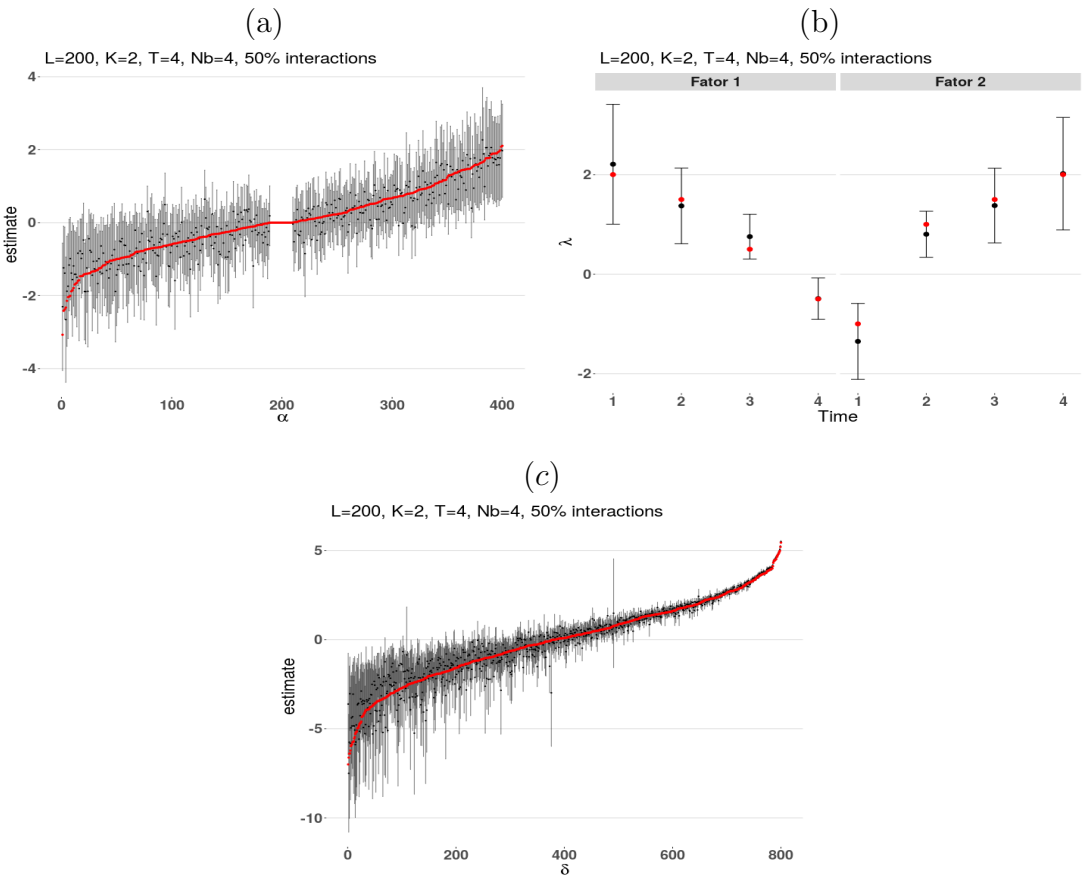


Figure D.8: Análise gráfica dos intervalos HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais do grupo G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o caso Poisson : $M^{K_2 I_{50\%}}_{L_{200} T_4 V_4}$ com $\approx 40\%$ de contagens zero.

Apêndice E: Modelo Logístico - Dados simulados SEM η e ajuste COM η.

Neste apêndice, apresentamos o resultado do ajuste do modelo logístico com dados simulados sem o efeito de interação entre os locais, mas ajustando com a presença do η^*. O cenário analisado considerou L = 400 locais, K = 2 fatores, T = 4 tempos, 4 vizinhos por região, 50% de locais afetados por interação e 50% de $Y_i^{'}s = 1$.

Podemos ver tanto pela Tabela 11 quanto pela Figura E.1 que o intervalo HPD de 95% engloba o valor verdadeiro (zero) do η. Importante destacar que os demais parâmetros continuaram sendo bem estimados, com o valor estimado próximo do valor verdadeiro e o intervalo HPD de 95% pegando o valor verdadeiro, sem prejuíso nos parâmetros da regressão.

As Figuras E.2 e E.3 mostram que as cargas, α, e os fatores, λ, foram bem estimados. Pelo Painel (d) da Figura E.3, vemos como a probabilidade de ocorrência do efeito de interação ficou próximo de 0.5 para a maioria dos locais.

	Verdadeiro	Média	Mediana	DP	HPD (inf.)	HPD (sup.)
β_0	0.50	0.79	0.79	0.17	0.39	1.08
β_1	-1.00	-1.03	-1.03	0.05	-1.13	-0.93
β_2	1.00	1.07	1.07	0.04	0.98	1.16
σ^2	0.80	0.89	0.89	0.11	0.69	1.10
τ_α	2.00	2.90	2.66	1.55	0.57	5.58
η_1^*	0.00	-0.60	-0.63	0.52	-1.55	0.39
η_2^*	0.00	-0.05	-0.07	0.50	-0.98	0.97
η_3^*	0.00	-0.26	-0.30	0.45	-1.07	0.68
η_4^*	0.00	-0.35	-0.39	0.61	-1.45	0.86

Tabela 11: Estimativas *a posteriori* dos coeficientes em β, da variância dos erros σ^2, do parâmetro de variância τ_α, e da interação não linear η^*. DP significa Desvio Padrão e o intervalo HPD informado é de 95% de probabilidade. Cenário: $M_{L_{400}T_4V_4}^{K_2I_{50\%}}$ com $\approx$ 50% de $Y_i^{'}s = 1$.

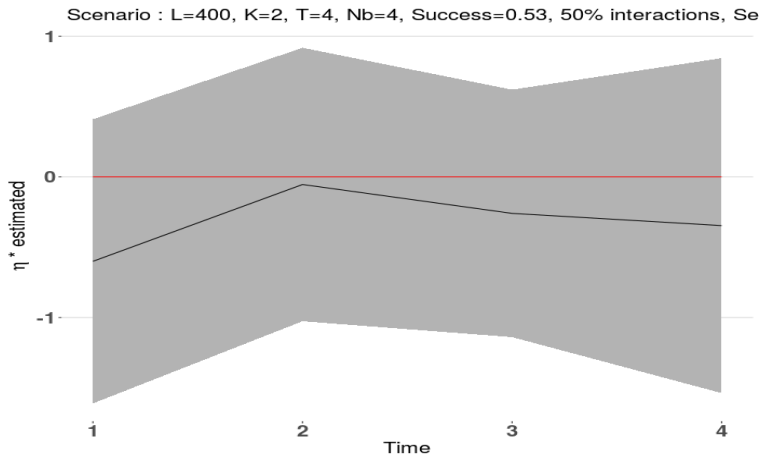


Figure E.1: Gráfico da média *a posteriori* (linha preta), intervalo HPD de 95% para η^* (área sombreada) e valor verdadeiro (linha vermelha) para o cenário $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

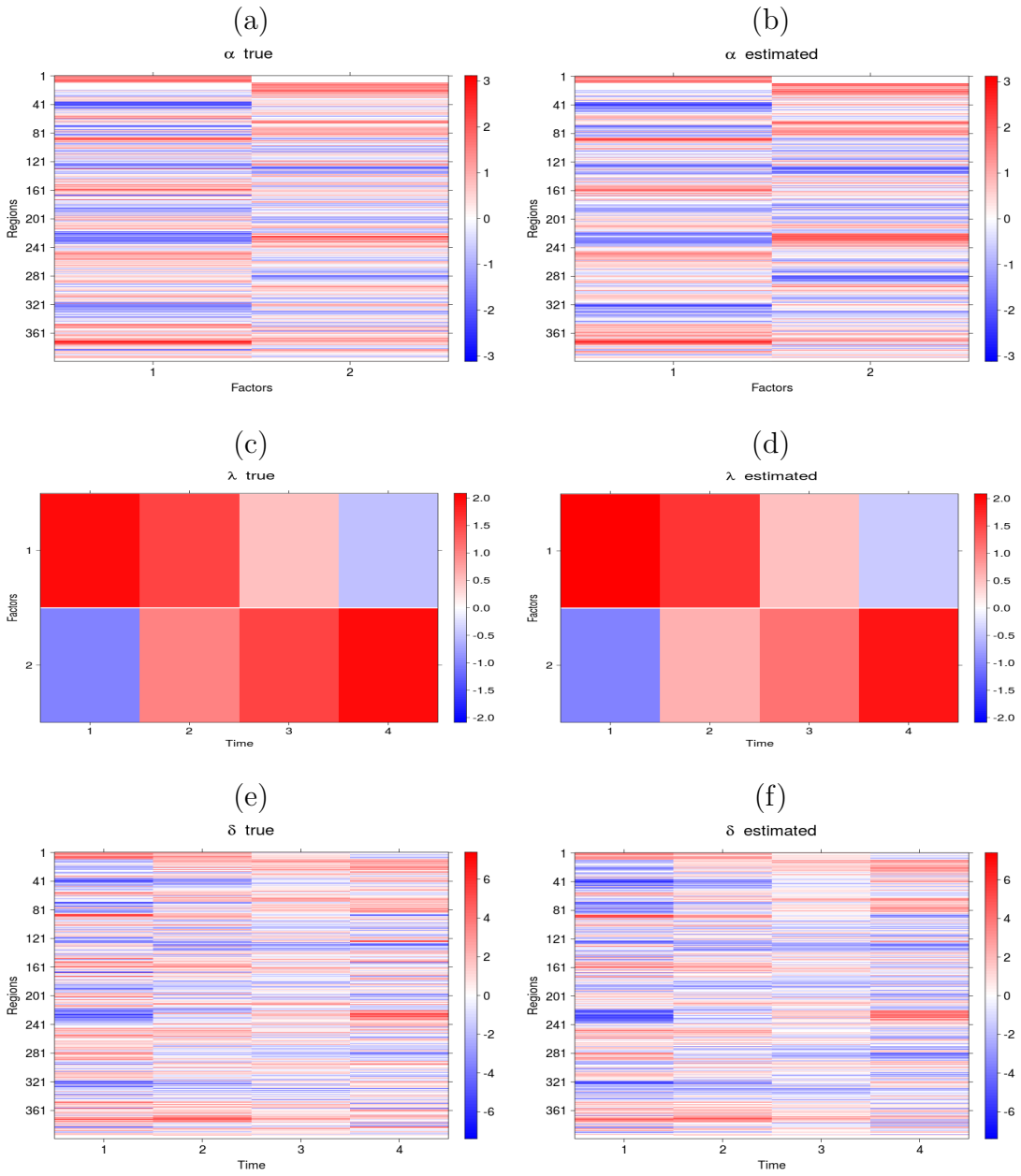


Figure E.2: Mapas de calor comparando valores verdadeiros e estimados para o cenário $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i's = 1$. Paineis: (a) e (b) são referentes à α, (c) e (d) referem-se à λ e (e) e (f) representam δ.

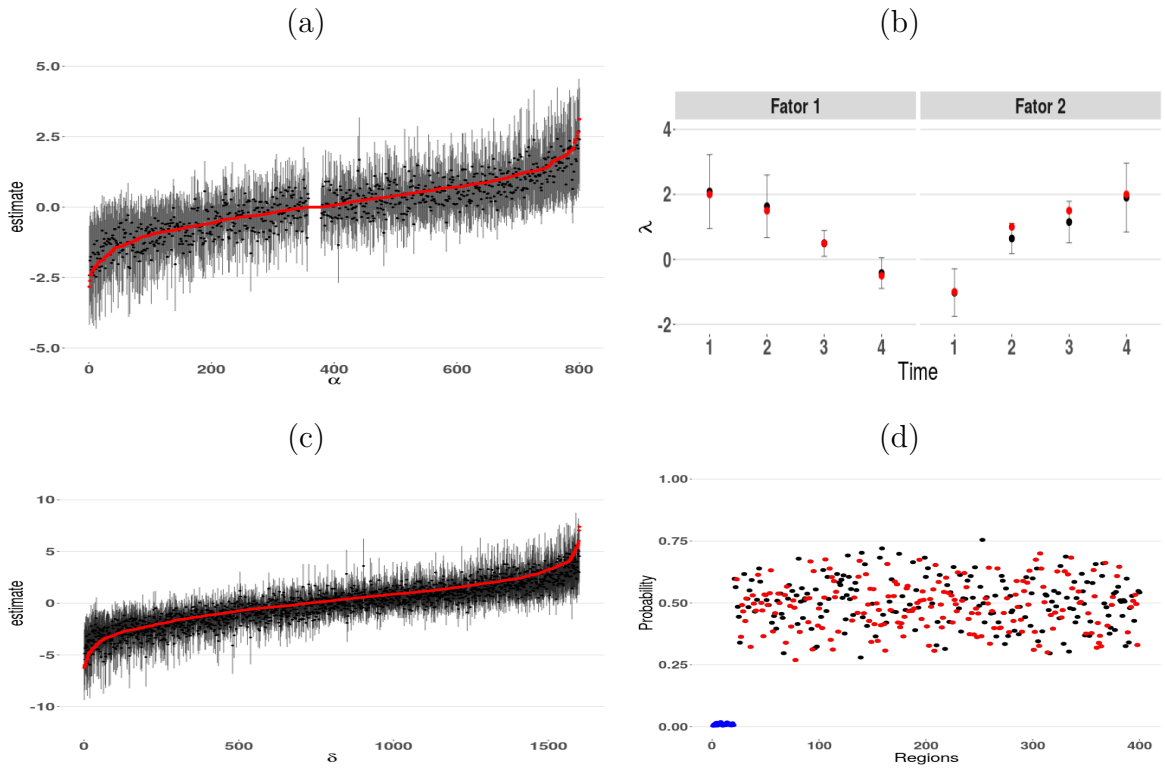


Figure E.3: Análise gráfica do intervalo HPD de 95% *a posteriori* para α (a), λ (b) e δ (c). A cor vermelha, nesses paineis mencionados, se refere ao valor verdadeiro. O Painel (d) apresenta as probabilidades das regiões serem afetadas por interações; cada ponto é um local. A cor azul indica locais de G_1 e G_2, a cor vermelha representa locais do grupo G_E com interação na geração dos dados. A cor preta denota locais de G_E que não tiveram interação na geração. Considere o cenário: $M^{K_2 I_{50\%}}_{L_{400} T_4 V_4}$ com $\approx 50\%$ de $Y_i^{'} s = 1$.

Apêndice F: Dados reais: Teste I-Moran para δ.

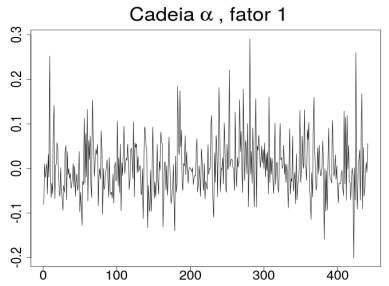
A Tabela 12 mostra o teste I-Moran para δ nos 4 tempos. Podemos ver que, em todos os casos, o p-valor foi praticamente zero, rejeitando a hipótese de que não há dependência espacial entre os locais.

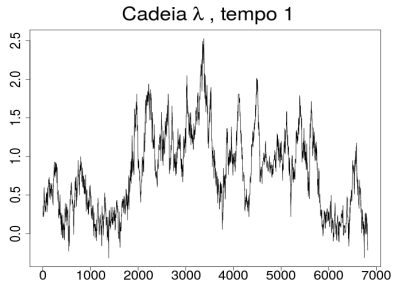
	Moran I statistic	Expectation	Variance	p-value
$\delta_{\bullet 1}$	0.278	-0.0023	0.00855	4.507881e-22
$\delta_{\bullet 2}$	0.246	-0.0023	0.00860	1.303889e-17
$\delta_{\bullet 3}$	0.115	-0.0023	0.00858	2.958159e-05
$\delta_{\bullet 4}$	0.174	-0.0023	0.00859	9.507140e-10

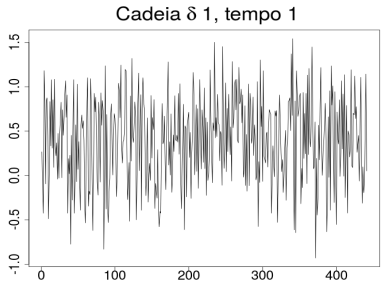
Tabela 12: Teste I-Moran para δ nos 4 tempos.

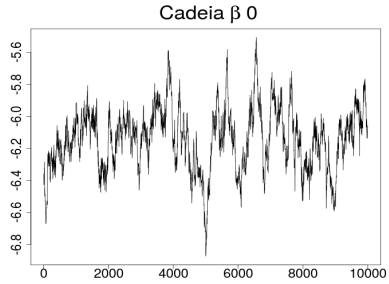
Apêndice G: Dados reais: Análise de convergência das cadeias.

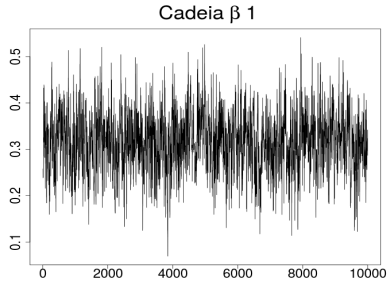
Os gráficos a seguir apresentam a análise da convergência das cadeias dos parâmetros após o ajuste da base de dados real de ECG.

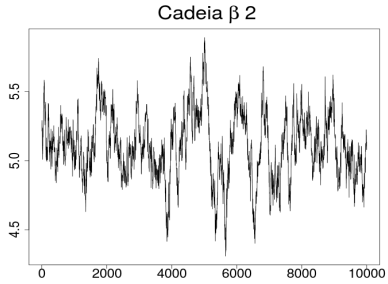


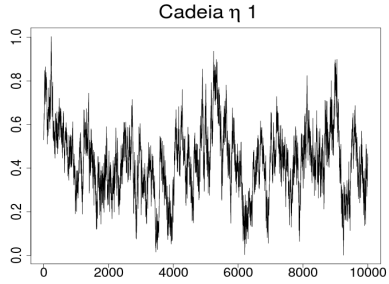


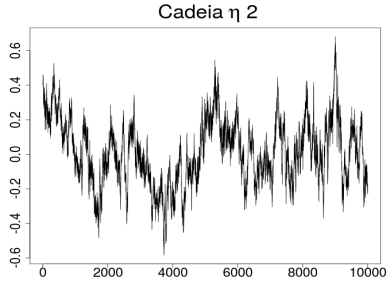


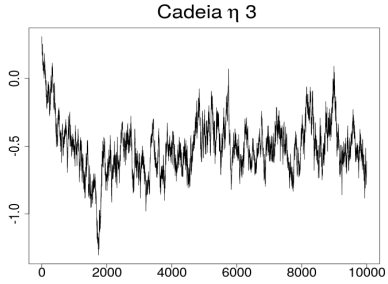


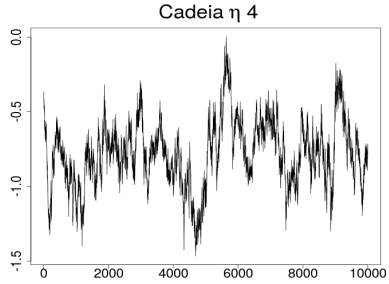












Bibliografia

Aguilar, O. e West, M. (2000), "Bayesian dynamic factor models and variance matrix discounting for portfolio allocation," *Journal of Business e Economic Statistics*, 18, 338–357.

Alkmim, M. B., Figueira, R. M., Marcolino, M. S., Cardoso, C. S., Pena, A. M., e Cunha, L. R. (2012), "Improving patient access to specialized health care: the Telehealth Network of Minas Gerais, Brazil." *Bull World Health Organ*, 90, 373–378.

Andrade, M., Maia, A., Cardoso, C., Alkmim, M., e Ribeiro, A. (2011), "Cost-benefit of the telecardiology service in the state of Minas Gerais: Minas Telecardio Project," *Arquivos Brasileiros de Cardiologia*, 97,4, 307–316.

Assuncao, R. e Krainski, E. (2009), "Neighborhood dependence in Bayesian spatial models," *Biometrical Journal*, 51(5), 851–869.

Banerjee, S., Carlin, B. P., e Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, CRC Press, New York, 2 edn.

Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society, Series B*, 36, 192–236.

Boulesteix, A. L. e Strimmer, K. (2006), "Partial least squares: a versatile tool for the analysis of high-dimensional genomic data," *Briefings in Bioinformatics*, 8, 32–44.

Box, G. E., Jenkins, G. M., e Reinsel, G. C. (2011), *Time series analysis: forecasting and control*, vol. 734, John Wiley & Sons.

Bradley, A. P. (1997), "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Science*, 30, 1145–1159.

Brown, H. e Prescott, R. (1999), *Applied Mixed Models in Medicine*, John Wiley and Sons, New York.

Browne, W. e Draper, D. (2006), "A comparison of Bayesian and likelihood-based methodsfor fitting multilevel models," *Bayesian Analysis*, 1,3, 473–514.

Brunet, J. P., Tamayo, P., Golub, T. R., e Mesirov, J. P. (2004), "Metagenes and molecular pattern discovery using matrix factorization," *PNAS - Proceedings of the National Academy of Sciences of the United States of America*, 101, 4164–4169.

Casella, G. e Berger, R. L. (2002), *Statistical Inference*, Duxbury, Pacific Grove, 2 edn.

Cordeiro, G. M. e Simas, A. B. (2009), "The distribution of Pearson residuals in generalized linear models," *Computational Statistics and Data Analysis*, 53, 3397–3411.

Demidenko, E. (2004), *Mixed Models: Theory and Application*, John Wiley and Sons, New York.

Egan, J. (1975), *Signal detection theory and ROC analysis*, Academic Press, New York.

Fye, W. (1994), "A history of the origin, evolution, and impact of electrocardiography," *Journal of the American College of Cardiology*, 73,13, 937–49.

Gamerman, D. e Lopes, H. F. (2006), *Markov Chain Monte Carlo - stochastic simulation for Bayesian inference*, Chapman and Hall/CRC, London.

Gamerman, D. e Salazar, E. (2013), "Hierarchical modeling in time series: the factor analytic approach," in *Bayesian Theory and Applications*, eds. N. G. P. Paul Damien, Petros Dellaportas e D. A. Stephens, pp. 167–182, Oxford University Press, Oxford.

Gelfand, A. E. e Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Geman, S. e Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

George, E. I. e McCulloch, E. (1993), "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

Geweke, J. F. (1977), "The dynamic factor analysis of economic time-series models," in *Latent Variables in Socio-Economic Models*, eds. D. Aigner e A. Goldberger, North-Holland, Amsterdam.

Geweke, J. F. e Zhou, G. (1996), "Measuring the pricing error of the arbitrage pricing theory," *The Review of Financial Studies*, 9, 557–587.

Green, P. J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82(4), 711–732.

Hanley, J. A. e McNeil, B. J. (1982), "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, 143, 29–36.

Hastings, W. (1970), "Monte Carlo sampling methods using Markov chains and their application," *Biometrika*, 57, 97–109.

Howell, J. (1991), "A history of the origin, evolution, and impact of electrocardiography," *Southern California Law Review*, 65,1, 529–64.

Johnson, R. A. e Wichern, D. W. (2007), *Applied Multivariate Statistical Analysis*, Pearson, Upper Saddle River, 6 edn.

Kim, P. M. e Tidor, B. (2003), "Subsystem identification through dimensionality reduction of large-scale gene expression data,," *Genome Research*, 13, 1706–1718.

Kligfield, P., Gettes, L., Bailey, J., Childers, R., Deal, B., e Hancock, E. (2007), "Recommendations for the standardization and interpretation of the electrocardiogram: Part I: the electrocardiogram and its technology: a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society: endorsed by the International Society for Com- puterized Electrocardiology." *Circulation*, 115,10, 1306–1324.

Lopes, H. F. e West, M. (2004), "Bayesian model assessment in factor analysis," *Statistica Sinica*, 14, 41–67.

Lopes, H. F., Salazar, E., e Gamerman, D. (2008), "Spatial dynamic factor analysis," *Bayesian Analysis*, 3, 759–792.

Lopes, H. F., Gamerman, D., e Salazar, E. (2011), "Generalized spatial dynamic factor analysis," *Computational Statistics and Data Analysis*, 55, 1319–1330.

Macfarlane, P. e Latif, S. (1996), "Automated serial ECG comparison based on the Minnesota code," *Journal of Electrocardiology*, 29, 29–34.

Macfarlane, P., Devine, B., Latif, S., McLaughlin, S., Shoat, D., e Watts, M. (1990), "Methodology of ECG interpretation in the Glasgow program," *Methods of Information in Medicine*, 29,4, 354–361.

Martin, C. D. e Porter, M. A. (2012), "The Extraordinary SVD," *The American Mathematical Monthly*, 119, 838–851.

Mayrink, V. D. e Gamerman, D. (2009), "On computational aspects of Bayesian spatial models: influence of the neighboring structure in the efficiency of MCMC algorithms," *Computational Statistics*, 24, 641–669.

Mayrink, V. D. e Lucas, J. E. (2013), "Sparse latent factor models with interactions: analysis of gene expression data," *The Annals of Applied Statistics*, 7, 799–822.

McCulloch, C. E. e Neuhaus, J. M. (2005), "Generalized linear mixed models," *International Encyclopedia of the Social and Behavioral Sciences*, doi:10.1016/B978-0-08-097086-8.42017-9.

Medeiros, T. L. F., Andrade, P. C. N. S., Davim, R. M. B., e Santos, N. M. G. (2018), "Mortality by an acute myocardial infarction," *Journal of Nursing UFPE online*, 12, 565–573.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., e Teller, E. (1953), "Equation of state calculations by fast computing machines," *Journal of Chemical Physics*, 21, 1087–1092.

Nelder, J. A. e Wedderburn, R. W. M. (1972), "Generalized linear models," *Journal of the Royal Statistical Society, Series A*, 135, 370–384.

Nguyen, D. V. e Rocke, D. M. (2002), "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, 18, 39–50.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro, A. L. P., Paixao, G. M. M., Gomes, P. R., Ribeiro, M. H., Ribeiro, A. H., Canazart, J. A., Oliveira, D. M., Ferreira, M. P., Lima, E. M., Moraes, J. L., Castro, N., Ribeiro, L. B., e Macfarlane, P. W. (2019), "Tele-electrocardiography and bigdata: the CODE (Clinical Outcomes in Digital Electrocardiography) study," *Journal of Electrocardiology*, 57, S75–S78.

Roberts, G. O. e Sahu, S. K. (1997), "Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler," *Journal of the Royal Statistical Society, Series B*, 59, 291–317.

Roth, G., Johnson, C., Abajobir, A., Abd-Allah, F., Abera, S., e Abyu, G. (2017), "Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015," *Journal of the American College of Cardiology*, 70,1, 1–25.

Rue, H. e Held, L. (2005), *Gaussian Markov random fields: theory and applications*, CRC press.

Sargent, T. J. e Sims, C. A. (1977), "Business cycle modeling without pretending to have too much a priori economic theory," Working Papers 55, Federal Reserve Bank of Minneapolis.

Sing, T., Sander, O., Beerenwinkel, N., e Lengauer, T. (2005), "ROCR: visualizing classifier performance in R," *Bioinformatics*, 21, 3940–3941.

Swets, J. (1988), "Measuring the accuracy of diagnostic systems," *Science*, 240, 1285–1293.

Swets, J., Dawes, R., e Monahan, J. (2000), "Better decisions through science," *Scientific American*, 283, 82–87.

Thurstone, L. L. (1931), "Multiple factor analysis," *Psychological Review*, 38, 406–427.

van de Leur, R. R., Blom, L. J., Gavves, E., Hof, I. E., van der Heijden, J. F., Clappers, N. C., Doevendans, P. A., Hassink, R. J., e van Es, R. (2020), "Automatic Triage of 12-Leads ECGs Using Deep Convolutional Neural Networks," *Journal of the American Heart Association*, 9, e015138.

Veloso, A., Meira, W., e Zaki, M. (2006), "Lazy associative classification," *Sixth international conference on data mining*, 18-22 Dec.

Yeung, K. Y. e Ruzzo, W. L. (2001), "Principal component analysis for clustering gene expression data," *Bioinformatics*, 17, 763–774.

Zellner, D., Keller, F., e Zellner, G. E. (2004), "Variable selection in logistic regression models," *Communications in Statistics - Simulation and Computation*, 33,3, 787–805.

Zhang, S., Zhang, L., Qiu, K., Lu, Y., e Cai, B. (2015), "Variable selection in logistic regression model," *Chinese Journal of Electronics*, 24,4, 813–817.

Zhao, Y., Staudenmayer, J., e Coull, B. (2006), "General design Bayesian generalized linear mixed models," *Statistical Science*, 21,1, 35–51.

Zou, K. (2002), "Receiver operating characteristic (ROC) literature research," *On-line bibliography available from: http://www.spl.harvard.edu/archive/spl-pre2007/pages/ppl/zou/roc.html.*