UNIVERSIDADE FEDERAL DO PARÁ
NÚCLEO DE MEDICINA TROPICAL
PROGRAMA DE PÓS-GRADUAÇÃO EM DOENÇAS TROPICAIS

RODRIGO RODRIGUES VIRGOLINO

ANÁLISE MULTIVARIADA DE CARACTERÍSTICAS CLÍNICAS DE PET/MAH E NÍVEIS DE EXPRESSÃO GÊNICA E DERIVAÇÃO DE MODELOS DE PREDIÇÃO DIAGNÓSTICA EM PACIENTES INFECTADOS COM O HTLV-1.

BELÉM
2017

RODRIGO RODRIGUES VIRGOLINO

ANÁLISE MULTIVARIADA DE CARACTERÍSTICAS CLÍNICAS DE PET/MAH E NÍVEIS DE EXPRESSÃO GÊNICA E DERIVAÇÃO DE MODELOS DE PREDIÇÃO DIAGNÓSTICA EM PACIENTES INFECTADOS COM O HTLV-1.

Dissertação de Mestrado apresentada à banca examinadora do Programa de Pós-graduação em Doenças Tropicais, do Núcleo de Medicina Tropical, da Universidade Federal do Pará, para obtenção do título de Mestre em Doenças Tropicais.
Orientador: Prof. Dr. Anderson Raiol Rodrigues

BELÉM
2017

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema de Bibliotecas da UFPA

Virgolino, Rodrigo Rodrigues

Análise multivariada de características clínicas de PET/MAH e níveis de expressão gênica e derivação de modelos de predição diagnóstica em pacientes infectados com o HTLV-1 / Rodrigo Rodrigues Virgolino; orientador, Anderson Raiol Rodrigues. — 2017.

Dissertação (Mestrado) - Universidade Federal do Pará. Núcleo de Medicina Tropical. Programa de Pós-Graduação em Doenças Tropicais. Belém, 2017
.

1.HTLV-1 (Vírus). 2. Infecção por HTLV – Complicações e sequelas. 3. Paraparesia Espática Tropical .I. Rodrigues, Anderson Raiol, orient. II. Título.

CDD - 22. ed. 616.9188

UNIVERSIDADE FEDERAL DO PARÁ
NÚCLEO DE MEDICINA TROPICAL
PROGRAMA DE PÓS-GRADUAÇÃO EM DOENÇAS TROPICAIS

RODRIGO RODRIGUES VIRGOLINO

ANÁLISE MULTIVARIADA DE CARACTERÍSTICAS CLÍNICAS DE PET/MAH E NÍVEIS DE EXPRESSÃO GÊNICA E DERIVAÇÃO DE MODELOS DE PREDIÇÃO DIAGNÓSTICA EM PACIENTES INFECTADOS COM O HTLV-1.

Dissertação de Mestrado apresentada para obtenção do título de Mestre em Doenças Tropicais.

Aprovada em:
Conceito:

Banca Examinadora

_____
Prof. Dr. Anderson Raiol Rodrigues
Orientador – NMT/UFPA

_____
Profa. Dra. Hellen Thais Fuzii
Membro – NMT/UFPA

_____
Prof. Dr. Cláudio Eduardo Corrêa Teixeira
Membro – CESUPA

_____
Prof. Dr. George Alberto da Silva Dias
Membro – CCBS/UEPA

_____
Prof. Dr. Juarez Antônio Simões Quaresma
Suplente – NMT/UFPA

AGRADECIMENTOS

Muitas pessoas contribuíram direta ou indiretamente para a realização desse sonho que é cursar um Mestrado num país carente de oportunidades deste tipo. Certamente os nomes não caberiam neste espaço.

Agradeço inicialmente à instituição Universidade Federal do Pará que possibilitou essa oportunidade e ao Núcleo de Medicina Tropical e ao Programa de Pós-Graduação em Doenças Tropicais (PPGDT) que foram a porta de estrada para este meu sonho e para o sonho de tantas outras pessoas de cursar uma Pós-Graduação no nível de mestrado e doutorado. Ao Dr. Anderson Raiol Rodrigues que se disponibilizou a me orientar e me permitiu desenvolver um trabalho relacionado à análise estatística, a qual tanto me empolga, desde a época da graduação em Biologia passando pela especialização em Bioestatística. Também pelas importantes dicas que me ajudaram a amadurecer o pensamento acadêmico. À Dra. Hellen Thais Fuzii e ao Dr. George Alberto da Silva Dias por ceder o banco de dados utilizado nesta pesquisa e pelas relevantes contribuições para o desenvolvimento da parte escrita. Aos colegas da turma 2015 e professores sempre solícitos do PPGDT, com os quais passei boa parte destes dois anos, os quais certamente mudaram minha forma de pensar a ciência, e consequentemente a vida.

Aos meus familiares, que suportaram vários adiamentos de compromissos por conta do vencimento de mais esta etapa, em especial aos meus pais Maria Rodrigues Bahia e Abdon Serrão Virgolino que por muitas vezes passarem por restrições para que eu e minhas irmãs Rosecleide Rodrigues Virgolino e Rosiane Rodrigues Virgolino pudéssemos chegar até onde chegamos, com dignidade. À minha esposa Rúbia Monteiro dos Santos e à minha filhinha Maria Eduarda Monteiro Virgolino pelos mesmos motivos, e por permanecerem comportadas enquanto eu ficava férias, feriados e fins de semana inteiros até altas horas escrevendo códigos esquisitos na frente do computador, na tela do R, sem saber pra que aquilo servia (risos).

Também, não posso deixar de registrar um agradecimento especial aos colegas de trabalho técnico-administrativos do Núcleo de Ciências Agrárias e Desenvolvimento Rural da UFPA (NCADR) Jacqueline Moraes, Kátia Cristina, Naiara Soraia e Moacir Pereira por terem "segurado as pontas" enquanto eu estava em aula ou em reuniões com o Prof. Anderson. Também aos meus chefes Drs. William Santos de Assis e Flávio Bezerra Barros pela concordância e pelo suporte às minhas atividades de mestrado, e aos estagiários de trabalho que são sempre indispensáveis para as atividades diárias da Universidade. Como o colega

Moacir costuma dizer, nós passamos a maior parte do dia com os colegas de trabalho do que com os nossos próprios familiares.

O sábio ouvirá e crescerá em conhecimento, e
o entendido adquirirá sábios conselhos.
Bíblia Sagrada, Provérbios 1, 5

RESUMO

A Paraparesia Espástica Tropical/Mielopatia associada ao Vírus Linfotrópico de Células T Humanas Tipo 1 (PET/MAH) é uma condição debilitante causada pela inflamação do tecido nervoso medular, por ação do HTLV-1. Neste trabalho objetivou-se avaliar a classificação de indivíduos infectados pelo vírus e propor um modelo de predição clínica para a ocorrência de PET/MAH. Foi utilizado um banco de dados composto de 63 indivíduos infectados, sendo 23 com diagnóstico de PET/MAH pelos critérios da OMS, e variáveis preditoras funcionais (variáveis ordinais), níveis de expressão gênica (variáveis contínuas) e a variável demográfica sexo. Em um primeiro momento uma técnica de análise de componentes principais mista foi empregada, seguida de análise de conglomerados hierárquica, para investigar a associação dos indivíduos em grupos, de forma não supervisionada, e comparar com a classificação definida pelos profissionais clínicos. Em um segundo momento, foram derivados modelos de predição diagnóstica com base em regressão logística binária penalizada, a qual é adequada nos contextos de tamanho amostral reduzido. A análise não supervisionada mostrou que os pacientes se organizavam em três grupos, sendo um grupo de pacientes com PET/MAH, um grupo de pacientes sem PET/MAH e um grupo intermediário, que comporta indivíduos com e sem a doença. Na modelagem estatística foram derivados dois modelos, um com critério de penalização 0,032 e outro com critério 0,1, mais extremo, sendo ambos avaliados por validação interna usando validação cruzada de 10 vezes. As variáveis que compuseram os modelos finais foram: grau de alteração na marcha, escore de Tinetti, tônus do músculo adutor direito e esquerdo e tônus do tríceps sural esquerdo. O uso de métodos estatísticos de predição pode ser útil como ferramenta de apoio à decisão diagnóstica de PET/MAH, especialmente em contextos de recursos limitados.

Palavras-chave: PET/MAH. HTLV-1. Variáveis Funcionais. Níveis de Expressão Gênica. Inflamação. Modelo de Predição Diagnóstica.

ABSTRACT

Human T-cell lymphotropic virus type 1 (HTLV-1)- associated myelopathy/tropical spastic paraparesis (HAM/TSP) is a debilitating condition resulting from inflammation of the nerve tissue of the spinal cord caused by the action of HTLV-1. The aim of the present study was to evaluate the classification of individuals infected with HTLV-1 and propose a clinical prediction model for the occurrence of HAM/TSP. A database composed of 63 infected individuals was used, 23 of whom were diagnosed with HAM/TSP using the criteria recommended by the World Health Organization. Functional predictors (ordinal variables), gene expression levels (continuous variables) and sex (demographic variable) were also used. A mixed principal component analysis was employed, followed by hierarchical cluster analysis to determine the allocation of individuals into groups in an unsupervised fashion and compare the results to the classifications defined by clinicians. Diagnostic prediction models were then derived based on penalized binary logistic regression, which is suitable when the sample size is small. The unsupervised analysis showed that the patients were arranged into three groups: patients with HAM/TSP, patients without HAM/TSP and an intermediate group composed of individuals with and without the disease. Two models were derived from the statistical modeling – one with a penalization criterion of 0.032 and another with a criterion of 0.1 (more extreme). Both models were evaluated by internal validation using 10-fold cross-validation. The variables that composed the final models were degree of gait alteration, derived Tinetti score, left and right adductor muscle tone and left triceps surae muscle tone. Statistical prediction methods may constitute a useful tool to support the diagnoses of HAM/TSP, especially in settings with limited resources.

Keywords: HAM/TSP. HTLV-1. Functional Variables. Gene Expression Levels. Inflammation. Diagnostic Prediction Model.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

ALCAM	*Activated Leukocyte Cell Adhesion Molecule*
AP	*Activator Protein*
ATLL	Adult T-cell Leukemia/Lymphoma
ATLV	*Adult T-cell Leukemia Virus*
AUC	*Area Under the ROC Curve*
CART	*Classification and Regression Trees*
CBP	*CREB Binding Protein*
CD	*Cluster of Differentiation*
CREB	*cAMP-Response Element Binding Protein*
CYLD	*Cylindromatosis Protein*
DNA	*Deoxyribonucleic acid*
ELISA	*Enzyme-Linked Immunosorbent Assay*
EPV	Eventos por Variável
FOXP3	*Forkhead Box P3*
GAPDH	*Glyceraldeyde 3-Phosphate Dehydrogenase*
GLM	*Generalized Linear Models*
GLUT	*Glucose Transporter*
HAM/TSP	*HTLV-1-Associated Myelopathy/Tropical Spastic Paraparesis*
HBZ	HTLV-1 *Basic ZIP Factor*
HEMOPA	Centro de Hemoterapia e Hematologia do Pará
HIV	*Human Immunodeficiency Virus*
HSPG	*Heparan Sulfate Proteoglycan*
HTLV	Vírus Linfotrópico de Células T Humanas (*Human T-cell Lymphotropic Virus*)
IκB	Inibidores de κB
ICF	*International Classification of Functioning, Disability and Health*

IFN	Interferon
Ig	Imunoglobulina
IKK	IκB quinase
IL	Interleucina
LASSO	*Least Absolute Shrinkage and Selection Operator*
LLcTA	Leucemia/Linfoma de Células T do Adulto
LTR	*Long Terminal Repeat*
MRC	*Medical Research Council*
mRNA	RNA mensageiro
NEMO	*NFκB Essential Modulator*
NFκB	*Nuclear Factor κ Beta*
OMS	Organização Mundial da Saúde
PCA	*Principal Component Analysis*
PCR	*Polymerase Chain Reaction*
PD	*Programmed Cell Death*
PET/MAH	Paraparesia Espástica Tropical/Mielopatia Associada ao HTLV-1
PSSE	*Penalized Sum of Squared Errors*
RNA	Ribonucleic Acid
ROC	*Receiver Operating Characteristic*
SNC	Sistema Nervoso Central
SPARC	*Secreted Protein Acidic and Rich in Cystein*
SSE	*Sum of Squared Errors*
SUMO	*Small Ubiquitin-like Modifier*
TGF	*Transforming Growth Factor*

SUMÁRIO

1 INTRODUÇÃO

O Vírus Linfotrópico de Células T Humanas (HTLV ou *Human T-cell Lymphotropic Virus*) é um vírus da família *Retroviridae*. Há descritos atualmente 4 tipos (HTLV 1-4), sendo que o HTLV-1 é o vírus mais representativo e estudado em razão de sua maior prevalência e maior associação a doenças em relação aos demais tipos. O HTLV-1 é o agente etiológico principalmente da Leucemia/Linfoma de Células T do Adulto (LLcTA), doença maligna agressiva relacionada a células T transformadas, e da Paraparesia Espástica Tropical/Mielopatia Associada ao HTLV-1 (PET/MAH), condição debilitante do Sistema Nervoso Central (SNC) caracterizada por uma manifestação inflamatória crônica, progressiva e incurável. Estimou-se inicialmente que a prevalência mundial de infecção pelo HTLV-1 era da ordem de 10 a 20 milhões de pessoas, conforme dados relativos a regiões endêmicas, as quais compreendiam um bilhão de habitantes à época (THÉ; BOMFORD, 1993; BRASIL, 2013). Dos infectados, até 5% desenvolviam alguma das manifestações clínicas mencionadas. Gessain e Cassar (2012) revisaram essa estimativa para 5 a 10 milhões de indivíduos infectados, considerando apenas as regiões globais reconhecidas como endêmicas conforme a literatura disponível, compreendendo 1,5 bilhões de habitantes. Há de se considerar ainda as dificuldades relacionadas à estimativa real de prevalência, uma vez que a grande maioria dos casos de infecção ocorre de forma latente e assintomática (HLELA et al., 2009).

A PET/MAH é uma das possíveis manifestações associadas à infecção pelo HTLV-1, onde o paciente apresenta alterações de progressão lenta em suas funções motoras, sensoriais e autonômicas, podendo estar presentes disfunções esfincterianas, sexuais e da bexiga. Frequentemente o paciente com PET/MAH apresenta aspectos clínicos e funcionais alterados, como tônus muscular aumentado, equilíbrio deficiente, necessidade de auxílio para deambulação e alterações de força muscular, especialmente nos membros inferiores, em decorrência dos sinais neurológicos observados no comprometimento medular inferior (FACCHINETTI et al., 2013; DIAS et al., 2016b; DIAS et al., 2016a).

Têm-se demonstrado que a fisiopatologia típica da PET/MAH resulta de alterações na resposta inflamatória induzidas pelo HTLV-1: células T $CD4^+$ são a principal reserva do vírus, e sofrem alteração em seu perfil de liberação de citocinas bem como proliferam de forma espontânea *in vivo*. As células T $CD4^+$ são moduladas de forma a favorecer a produção de citocinas pró-inflamatórias (como IFN-γ e TNF-α) típicas de uma resposta Th1, ao mesmo tempo em que o perfil de resposta Th2 relacionado à liberação de IL-4 é suprimido. Por outro lado, células T reguladoras (Treg), as quais possuem função de suprimir a resposta

inflamatória exacerbada pela liberação da citocina anti-inflamatória IL-10, diminuem a expressão de IL-10 durante a infecção (FUZII et al., 2014; BANGHAM et al., 2015). Porém ainda não foram esclarecidos os mecanismos que levam ao desequilíbrio nessa relação parasito-hospedeiro, após anos ou décadas do contágio.

Dadas as manifestações da PET/MAH, tanto no nível de características clínicas como de aspectos relacionados ao processo inflamatório intrínseco à doença, seria possível utilizar-se de métodos estatísticos multivariados de forma a acessar relações complexas entre estas variáveis e, a partir da análise multivariada, classificar pacientes com base em associações que não são imediatamente perceptíveis em um contexto prático? Seria possível modelar estas relações de forma a predizer a classificação de futuros pacientes, com base nas associações já conhecidas em termos do banco de dados inicial? Estas questões serão endereçadas no presente trabalho.

2 REFERENCIAL TEÓRICO

2.1 BREVE HISTÓRICO

Até a década de 70, sabia-se que retrovírus eram capazes de induzir leucemia em animais como gato, primatas (GALLO et al., 1978), bovinos (FERRER et al., 1975), e outros animais, mas havia muita controvérsia sobre a possibilidade de que vírus fossem causadores de tumores em humanos. No entanto, a elucidação dos mecanismos leucemogênicos envolvendo retrovírus em animais possibilitou a investigação de mecanismos similares na leucemia humana, culminando com a descoberta do primeiro retrovírus associado a patologias humanas: o HTLV.

O HTLV foi descrito em 1980 (POIESZ et al., 1980; POIESZ et al., 1981; GALLO et al., 1982), quando partículas virais foram isoladas de linfócitos de um paciente com linfoma cutâneo ligado a células T, nos Estados Unidos. Identificou-se nessas células atividades de Transcriptase Reversa não relacionada a DNA polimerases humanas, nem à transcriptase reversa encontrada em retrovírus conhecidos, de outros animais. Também foram descritos alguns aspectos morfológicos do vírus, por meio de microscopia eletrônica. Ainda em 1977, no Japão, foram reportados casos localizados de uma forma distinta de leucemia de evolução rápida e com características únicas, a qual foi denominada de Leucemia de Células T do Adulto (ATL - *Adult T-cell Leukemia*) (UCHIYAMA et al., 1977). Em 1982 foi caracterizado no Japão o vírus causador da ATL (ATLV - *Adult T-cell Leukemia Virus*) (YOSHIDA; MIYOSHI; HINUMA, 1982).

Posteriormente, estudos comparativos mostraram que o ATLV e o HTLV descrito em 1980 eram o mesmo vírus, prevalecendo este segundo nome em referência à descrição pioneira do vírus pelo grupo americano e o nome Leucemia de Células T do Adulto (ATL) em referência ao trabalho japonês que destacou a doença como uma entidade autônoma em relação às formas até então conhecidas.

Em 1982 foi descrito o HTLV-2 (KALYANARAMAN et al., 1982) com características de resposta imunológica diferentes do HTLV-1 e associado a menor virulência em relação a este. Outros estudos se seguiram, demonstrando a ocorrência do HTLV em outras regiões do mundo, como região do Caribe, Índia, Nigéria, sul da Europa e América Latina (LEVINE et al., 1988; LEVINE et al., 1989). Observou-se assim que o HTLV obedecia a um padrão de distribuição geográfica circunscrito regionalmente.

Em 1985, na região francesa da Martinica, na região do mar caribenho, Gessain et al. (1985) identificaram que pacientes apresentando Paraparesia Espástica Tropical (TSP ou *Tropical Spastic Paraperesis*) eram positivos para anticorpos anti-HTLV-1, numa proporção além do observado no grupo controle. Estudos posteriores confirmaram a associação do HTLV-1 a esta condição. Verificou-se também que esta condição ocorria em regiões não tropicais (OSAME et al., 1986), assumindo-se a denominação *HTLV-1 Associated Myelopathy/Tropical Spastic Paraperesis* (HAM-TSP) e no Brasil Paraparesia Espástica Tropical/Mielopatia Associada ao HTLV-1 (PET/MAH).

O HTLV-3 foi descrito em 2005 (CALATTINI et al., 2005), em Camarões, na África. Este diferia do HTLV-1 e HTLV-2 sorologicamente e em termos de sequência genética. Na mesma região e no mesmo ano foi isolado o HTLV-4 (WOLFE et al., 2005) de um caçador de 48 anos. Estes dois novos tipos até o momento não foram encontrados em outras regiões do mundo e não foram associados a doenças em humanos. Essa diversidade local do vírus é relacionada ao comportamento das populações rurais que se expõem ao contato com primatas não humanos, e é origem provável de transmissões interespecíficas destes retrovírus ocorridas no passado (FILIPPONE et al., 2015; MAHIEUX; GESSAIN, 2011).

2.2 CARACTERIZAÇÃO DO HTLV-I

O HTLV pertence à família *Retroviridae* (cuja principal característica é possuir material genético composto de RNA), gênero *Deltaretrovirus*. A partícula viral possui 100nm de diâmetro e é formada por um capsídeo, dentro do qual há duas fitas de RNA, transcriptase reversa e enzimas integrases. Penetra na célula alvo por meio de fusão entre a membrana plasmática e o envelope viral, seguida da liberação do material interno no citoplasma da célula. Então, a transcriptase reversa transcreve o RNA viral em dupla fita de DNA, que entra no núcleo e se insere no DNA da célula hospedeira como um provírus, que é a forma integrada do vírus (QUARESMA et al., 2015; FUZII et al., 2014; HOSHINO, 2012).

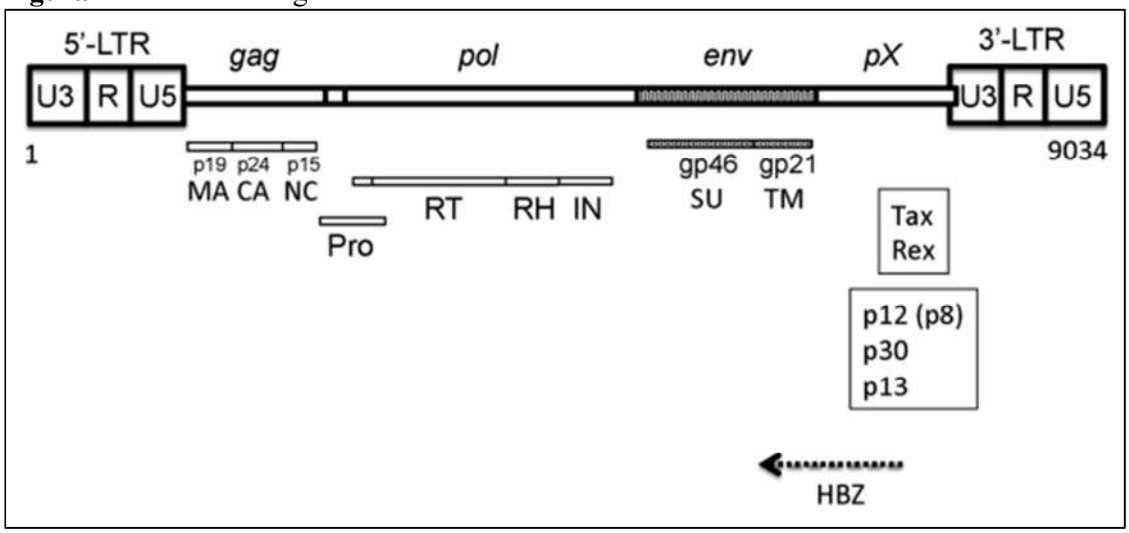
O provírus possui sequências de bases repetidas nas duas extremidades, chamadas regiões LTR (*Long Terminal Repeats*). A região LTR 5’ é a região promotora de transcrição dos genes virais estruturais *gag*, *pol* e *env*. Essa direção de transcrição é chamada senso. A transcrição chamada anti-senso ocorre na direção oposta, iniciando na região LTR 3’ onde fica a região promotora do gene pX. A partir da região pX são codificadas duas proteínas regulatórias muito importantes: a Tax e a Rex. A Tax é um potente transativador da região

LTR 5', e é bem expressa *in vitro* por não haver pressão do sistema imunológico. Rex modula vários comportamentos da célula hospedeira, como a exportação de mRNA viral a partir do núcleo (MIYAZATO et al., 2016; SEIKI et al., 1983).

Outra proteína importante codificada a partir da região pX é o *HTLV-1 bZIP factor* (HBZ), que é capaz de modular o crescimento celular e é necessária para o desenvolvimento da LLcTA. Recentemente, têm sido confirmada a importância da HBZ na modulação da proliferação de células T $CD4^+$ ativadas, pois ela altera a expressão de receptores da superfície dessas células infectadas pelo HTLV-1. Por exemplo, o aumento da expressão dos receptores *T Cell Immunoglobulin and ITIM Domain* (TIGIT) e *Programmed Cell Death 1* (PD-1) na superfície da célula T, leva à supressão da proliferação da célula T infectada se houver acionamento por meio da ligação de seu ligante específico. Logo, eles possuem função supressora da resposta imune. Porém, o HBZ perturba a sinalização que leva à ativação desses co-receptores, e desta forma promove a proliferação e sobrevivência da célula infectada *in vivo*, indiretamente, ao modificar a função de co-inibidores da superfície da célula T ativada (KINOSADA et al., 2017).

A Figura 1 esquematiza a estrutura do genoma viral. Perceba que as regiões LTR possuem três componentes: uma região única 3' (U3), uma região de repetição (R) e uma região única 5' (U5). O gene *gag* codifica três proteínas: a proteína viral do nucleocapsídeo (NC, proteína p15), a proteína do capsídeo (CA, proteína p24) e a proteína de matriz (MA, proteína p19). O gene *pol* codifica as proteínas: protease (Pro), transcriptase reversa (RT), RNase H (RH) e integrase (IN), todas com função enzimática. O gene *env* codifica as proteínas do envelope viral: unidade de superfície (SU, proteína gp46) e a unidade transmembrana (TU, proteína gp21), associadas à ligação do vírus à célula alvo e fusão à membrana plasmática durante a entrada do vírus, respectivamente. A região pX também codifica proteínas regulatórias acessórias (p12/p8, p13 e p30), que possuem função na transmissão e persistência viral, escape do sistema imune e transformação celular, principalmente *in vivo*. (PISE-MASISON et al., 2014; BAI; NICOT, 2012; LAIRMORE et al., 2011).

Figura 1 - Estrutura do genoma do HTLV-1.



Fonte: Hoshino (2012).

O genoma do HTLV-1 é altamente conservado, favorecendo a ideia de que a replicação viral é obtida pela expansão clonal das células infectadas, mais do que infecção de outras células por partículas virais produzidas *de novo*. Se este último caso fosse verdadeiro, o genoma do vírus estaria sujeito a erros (mutações genéticas) de transcrição da transcriptase reversa (é isso o que ocorre no HIV, por exemplo). No entanto, têm sido identificadas mutações pontuais em vários genes como o *tax* em pacientes com LLcTA (FURUKAWA et al., 2001).

A proteína Tax induz transcrição a partir da região LTR 5' do provirus, aumentando a produção das proteínas estruturais e outros antígenos virais. Porém, a expressão aumentada de Tax e de outros antígenos expõe a célula infectada à ação do sistema imune do hospedeiro, fazendo com que essas células sejam eliminadas. Por isso, há mecanismos de supressão da atividade promotora da região LTR 5', como: mutações do gene *tax*, deleção da região LTR 5' e alterações epigenéticas nesta região, causando silenciamento. Por outro lado, a transcrição da HBZ permanece *in vivo* na direção anti-senso, pois se trata de uma proteína de baixa imunogenicidade (GIAM; SEMMES, 2016). Logo, ao mesmo tempo em que há repressão transcricional do provirus, é mantida a sua capacidade de reativação para que ele se perpetue.

A regulação epigenética da célula hospedeira é dinâmica e depende de fatores intra e extracelulares. A hipermetilação de DNA é um caso de regulação epigenética onde regiões promotoras de genes (por exemplo, as porções LTR) são moduladas, regulando a transcrição através da adição de grupos metil ao DNA. No caso da região LTR 5', a hipermetilação de ilhas CpG está associada ao silenciamento, enquanto a região LTR 3' é metilada raramente,

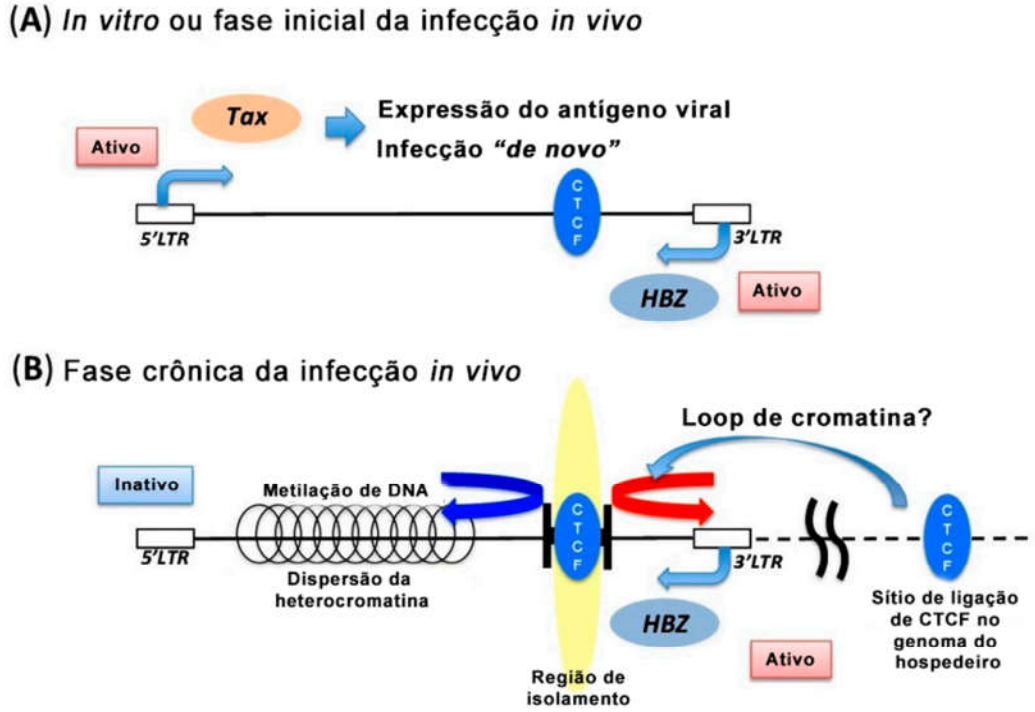
apesar de ambas serem idênticas quanto à sequência de nucleotídeos. Este achado é consistente com a supressão da transcrição senso e ativação da transcrição anti-senso, observadas constitutivamente.

Modificações de histonas também são mecanismos epigenéticos determinantes da expressão gênica (WANG et al., 2016). Histonas são proteínas que formam o nucleossomo. Ao redor de octâmeros de histonas estão porções de DNA de 147pb que se enovelam, formando a unidade básica da cromatina. Modificações das caudas das histonas (acetilação, metilação, fosforilação ou ubiquitinação) são críticas para o recrutamento de proteínas efetoras, como fatores de transcrição e RNA polimerase II, que determinam o início da transcrição. Tax é capaz de recrutar proteínas co-ativadoras como a CBP e a p300 para o LTR 5', contribuindo para uma forte ativação transcricional no sentido senso, conforme visto em experimentos *in vitro* onde não há pressão do sistema imune.

Regiões de isolamento (*insulator regions*) são regiões genômicas que delimitam bordas epigenéticas entre locais ativos e locais inativos transcricionalmente. A proteína CTCF se liga diretamente ao DNA proviral do HTLV-1 e auxilia a manutenção da estrutura da cromatina. Modificações de histonas ocorrem nas proximidades das regiões de isolamento de forma constitutiva e acredita-se que estejam envolvidas na manutenção da infecção persistente pelo HTLV-1 (MIYAZATO et al., 2016).

A Figura 2 ilustra esses comportamentos do vírus. Em "A" observamos a expressão constitutiva de Tax e consequente função promotora da região LTR 5', promovendo a transcrição senso de proteínas estruturais. Esse padrão é geralmente observado na fase inicial da infecção ou em experimentos *in vitro*, situações em que não há a pressão do sistema imune contra os antígenos virais. Em "B", na fase crônica da infecção in vivo, a região de isolamento parece prevenir a difusão da heterocromatina entre o LTR 5' e o LTR 3', mantendo o padrão de transcrição distinto observado a partir das duas extremidades (transcrição senso reprimida e transcrição anti-senso permitida).

Figura 2 - Mecanismos de regulação proviral do HTLV-1. A) Figura esquemática do provirus na fase inicial da infecção ou *in vitro*, quando há expressão de proteínas estruturais a partir do terminal LTR 5'. B) Figura esquemática do provirus na fase crônica, quando a CTCF mantém a região de isolamento com padrões distintos de expressão de sentido senso e anti-senso.



Fonte: adaptado de Miyazato et al. (2016).

O HTLV-1 infecta primariamente células T $CD4^+$ ativadas, porém, outros tipos celulares podem ser infectados, como células T $CD8^+$, linfócitos B, monócitos, macrófagos, células dendríticas, células sinoviais e células da glia (micróglia e astrócitos). Isso em parte se deve à ampla presença do receptor *glucose transporter 1* (GLUT-1), importante para a entrada do vírus na célula (MANEL et al., 2003). Células dendríticas são células apresentadoras de antígenos que parecem importantes na transmissão, disseminação e persistência viral *in vivo*, uma vez que podem capturar partículas virais livres e transferi-las para células T $CD4^+$ ou, uma vez infectadas, transferir novas partículas virais replicadas para células T saudáveis (JONES et al., 2008). Com exceção das células dendríticas, partículas livres do HTLV-1 são pouco infeccionsas aos demais tipos celulares. A infecção pode ocorrer das seguintes formas: pela indução da formação de uma sinapse viral entre a célula infectada e a célula não infectada, ou pela formação de uma estrutura similar a biofilme na superfície externa da célula infectada, pela qual o vírion é transferido ao linfócito não infectado (HOSHINO, 2012).

2.3 FORMAS DE TRANSMISSÃO

São reconhecidas três formas de transmissão do HTLV-1: a) transmissão vertical da mãe ao filho; b) transmissão sexual; e c) transmissão por sangue contaminado. Todas ocorrem principalmente pela transferência de linfócitos contaminados:

A possibilidade de transmissão vertical, da mãe para o filho, é demonstrada, por exemplo, pela presença de alta carga proviral de HTLV-1 encontrada no leite materno (neste contexto, antígenos virais e anticorpos anti-HTLV-1 são frequentemente encontrados no leite materno de mães soropositivas) e pelo fato do tempo de amamentação com leite materno (amamentação prolongada) representar fator de risco para a infecção pelo HTLV-1 (PERCHER et al., 2016). A transmissão sexual é a via menos eficiente de transmissão, e ocorre em especial do homem para a mulher por sêmen contaminado, o que tem sido relacionado à maior soroprevalência em mulheres com o avançar da idade. De forma geral, quanto mais prolongada a exposição e quanto maior a carga proviral maior o risco de infecção pela rota sexual, tanto do HTLV-1 quanto do HTLV-2 (característica esta que é compartilhada com outro retrovírus bem conhecido - o HIV) (KAPLAN et al., 1996).

A transmissão por meio de sangue contaminado ocorre em transfusões de sangue ou entre usuários de drogas intravenosas, quando principalmente linfócitos infectados são transferidos de pessoa a pessoa (o HTLV-2 parece ser mais prevalente que o HTLV-1 entre usuários de drogas intravenosas - Khabbaz et al., 1992). Em muitos países é realizada triagem de doadores com base na anamnese do candidato à doação e na sororreatividade para antígenos do HTLV. Neste caso se usa um ensaio imunoenzimático para a detecção de anticorpos, seguido da confirmação por meio de Western Blot e, persistindo a dúvida, procede-se a uma técnica de diagnóstico molecular, como a Reação em Cadeia da Polimerase (*Polymerase Chain Reaction* ou PCR). Em alguns países (principalmente desenvolvidos) é empregada a técnica de leucodepleção, que consiste em remover leucócitos do sangue do doador por meio de filtros especiais, permanecendo hemocomponentes onde o vírus não é encontrado (hemácias, plasma).

Porém, há regiões, algumas consideradas não endêmicas, onde não é realizada a triagem de doadores para o HTLV, como China (XIE et al., 2015), Alemanha, Itália, Rússia, Espanha. Essas decisões são geralmente baseadas em estudos de custo efetividade que consideram a prevalência/incidência da infecção, o risco de transmissão e os custos associados a todos estes fatores. Além disso, muitos países não contam com uma política apropriada ou sistema nacional de doação de sangue, em especial no continente africado (MARANO et al., 2016).

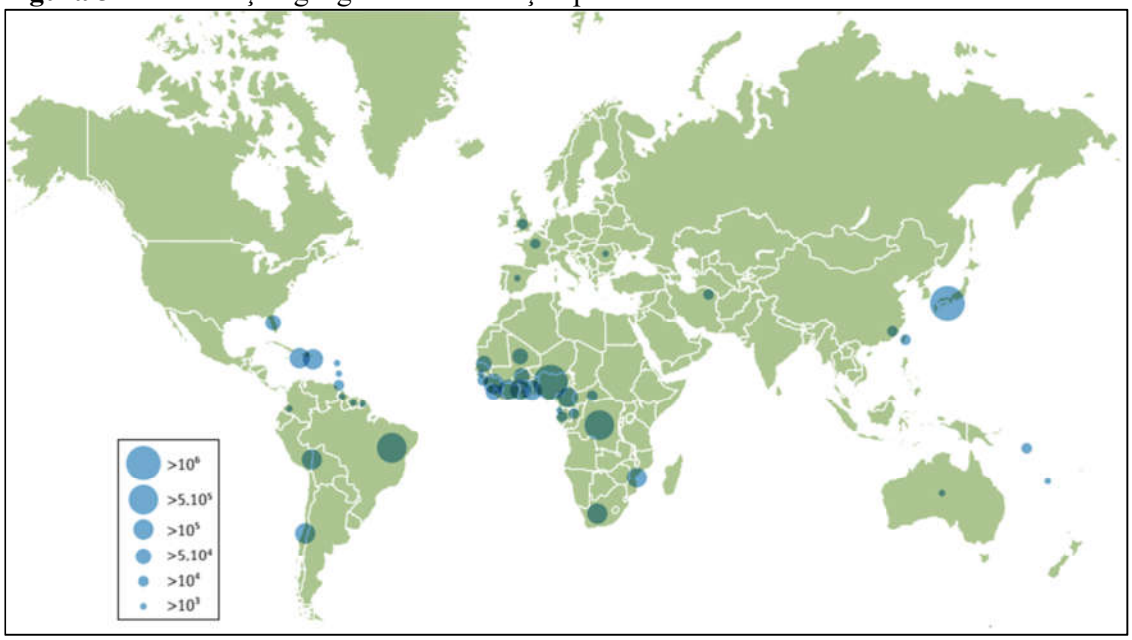
2.4 EPIDEMIOLOGIA

O HTLV é descrito como um vírus de ocorrência endêmica em várias regiões do mundo. O HTLV-1 é mais disseminado em comparação aos demais HTLVs, e tem sido estudado devido à sua maior associação a doenças (até 10% dos infectados desenvolvem alguma doença relacionada ao vírus, e o restante permanece assintomático). As principais doenças associadas ao HTLV-1 são a PET/MAH e a LLcTA, sendo que até 50% dos pacientes com PET/MAH ficam dependentes de cadeira de rodas e a sobrevida média dos pacientes com LLcTA é de 6-8 meses (HLELA et al., 2009).

Apesar de aceitar-se uma prevalência de infecção pelo HTLV-1 de 10 a 20 milhões de pessoas (BRASIL, 2013) (ou 5 a 10 milhões, como sugerido por Gessain e Cassar, 2012), há dificuldades relacionadas à estimativa da real prevalência: poucos estudos de base populacional foram realizados e a maioria dos estudos são realizados sobre grupos restritos, como pacientes sintomáticos, grupos selecionados ou estudos realizados em hemocentros. Por exemplo, estudos feitos em grávidas geralmente incluem mulheres de várias idades, apesar de que o HTLV-1 está mais presente nas faixas etárias superiores. Também, as características de doadores de sangue podem diferir muito entre um país e outro, sendo que em alguns contextos os doadores são parentes de pessoas hospitalizadas, em outros são principalmente pessoas de baixa classe econômica que recebem dinheiro para doar, ou ainda são doadores homens jovens.

A Figura 3, retirada de Bangham et al. (2015), mostra a distribuição geográfica da infecção por HTLV-1, sugerida pela revisão de literatura realizada por Gessain e Cassar (2012). Observa-se que as principais áreas endêmicas no mundo são o sudoeste do Japão, região do Caribe, várias regiões da África, partes da América do Sul e da Melanésia. A origem desse padrão de distribuição típico não é bem conhecida, mas provavelmente se deve ao efeito do fundador em alguns grupos, seguido por uma taxa aumentada de transmissão viral nesses grupos.

Figura 3 - Distribuição geográfica da infecção por HTLV-1.



Fonte: Bangham et al. (2015) apud Gessain e Cassar (2012).

No Japão, têm sido realizados inúmeros estudos relacionados ao HTLV desde os últimos 30 anos devido à alta prevalência de casos de LLcTA e PET/MAH, em especial no sudoeste desse país (TAJIMA, 1990; OSAME et al., 1990). Estima-se que cerca de um milhão de pessoas estejam infectadas pelo HTLV-1 no Japão (SATAKE; YAMAGUCHI; TADOKORO, 2012). A América do Sul, de forma geral, é uma região endêmica para o HTLV-1 e as doenças associadas a ele, sendo que países como Peru, Chile, Colômbia, Guiana Francesa e Brasil apresentam prevalência relativamente elevada, em especial quando se consideram grupos ameríndios isolados ou populações com descendência africana (MACÍA et al., 2016; GESSAIN; CASSAR, 2012; ITA et al., 2014).

No Brasil, têm sido realizados vários estudos relacionados ao HTLV-1, sugerindo um padrão geográfico heterogêneo de prevalência. Muitos desses estudos são baseados em dados de hemocentros, desde 1993 quando a pesquisa por HTLV passou a ser obrigatória sobre potenciais doadores de sangue. Catalan-Soares, Carneiro-Proietti e Proietti (2005), em trabalho realizado a partir de dados de vários hemocentros distribuídos nas 27 principais áreas metropolitanas do Brasil (dados referentes a 1995 até 2000) encontraram prevalência entre 0,04 e 1%, a qual se mostrou maior nas regiões Norte e Nordeste em relação ao Sul, similar aos achados de Carneiro-Proietti et al. (2012), que encontraram maior prevalência nos doadores de Pernambuco em relação a doadores de Minas Gerais e São Paulo. Catalan-Soares, Carneiro-Proietti e Proietti (2005) encontraram que os três estados onde houve maior

soroprevalência via ensaio imunoenzimático foram Maranhão, seguido da Bahia, seguida do Pará.

Mesmo sabendo que doadores de sangue, sendo compostos de pessoas relativamente saudáveis, geralmente não são representativos da população em geral, os bancos de sangue podem ser úteis como termômetro no monitoramento de tendências da infecção por HTLV no decorrer do tempo, mas a prevalência do vírus na população em geral é maior. Além disso, têm sido publicados séries de casos de PET/MAH e LLcTA no país (SILVA et al., 2013; FACCHINETTI et al., 2013; BITTENCOURT et al., 2013; CHAMPS et al., 2010; BITTENCOURT et al., 2009).

Similarmente, no Estado do Pará, não há estudos com base populacional que possam disponibilizar informações fidedignas sobre a real prevalência do HTLV na população urbana. Entre 190 pacientes sofrendo de doenças neurológicas crônicas, atendidas na capital Belém, 15 pacientes mostraram sororreatividade para HTLV-1/HTLV-2, sendo 10 confirmados via Western Blot para infecção por HTLV-1 (MACÊDO et al., 2004). Também, Santos et al. (2009) encontraram, em um grupo de 79 doadores soropositivos para HTLV-1 e 2, no banco de sangue HEMOPA, que 71% destes eram positivos para HTLV-1 e 29% para o HTLV-2, contrastando com a maior proporção HTLV-2/HTLV-1 encontrada por Vallinoto et al. (1998) em pacientes HIV soropositivos e comumente encontrado em várias comunidades indígenas do estado (PAIVA; CASSEB, 2015; ISHAK et al., 2003).

Ainda em 1995, entre índios Kayapó, foi encontrada uma alta prevalência de HTLV-2: acima de 30% (ISHAK et al., 1995). Neste estudo foram analisados soros de 1324 indígenas, via ELISA e confirmação por Western Blot. Na população urbana da capital Belém tem sido descrita maior prevalência de HTLV-2 em relação a outras regiões do país, provavelmente refletindo a contribuição indígena durante o processo de colonização da região amazônica (VALLINOTO et al., 2002; PAIVA; CASSEB, 2015).

Mais recentemente, em 2013, Magno Falcão et al. (2013) encontraram, prevalência de anticorpos anti-HTLV da ordem de 4,7% em 657 indivíduos de duas comunidades formadas aos arredores da usina hidroelétrica de Tucuruí, apoiando estudos que apontam maior prevalência e incidência de doenças transmissíveis que provavelmente decorrem de movimentos migratórios e crescimento populacional desordenado após construção de hidroelétricas (TETTEH; FREMPONG; AWUAH, 2004). Nesse estudo a soroprevalência foi acessada pela técnica de ELISA e os participantes foram voluntários das comunidades, submetidos à coleta de sangue venoso do membro superior.

2.5 DOENÇAS LIGADAS AO HTLV-1

Atualmente várias outras condições são associadas à infecção pelo HTLV-1, além da PET/MAH e da LLcTA, como: uveíte, alterações dermatológicas e reumatológicas, síndrome de Sjögren (insuficiência glandular caracterizada pela diminuição das secreções lacrimais e salivares) e polimiosite (inflamação muscular).

Uveíte, que é a inflamação da camada intermediária do olho, prevalente em regiões endêmicas como o Japão, ocorre quando células T $CD3^+$ infectadas pelo HTLV-1 (células não neoplásicas) se infiltram na região ocular e se multiplicam. As células T $CD4^+$ resultantes da multiplicação das células $CD3^+$ então passam a liberar citocinas inflamatórias (IL-1, IL-6, TNF-α, IFN-γ) que causam inflamação intraocular (KAMOI; MOCHIZUKI, 2012), podendo levar o paciente à cegueira nos casos não tratados. É possível detectar alta carga proviral nessas células infiltrantes, além de proteínas produzidas pelo HTLV-1. A terapia indicada para a condição é o tratamento com corticosteroides, que visa à inibição da resposta inflamatória exagerada.

Entre as alterações dermatológicas associadas ao HTLV-1 temos xerose (ressecamento epitelial patológico), micose superficial e dermatite seborreica, que podem ser o resultado da presença de linfócitos infectados na pele, ou resultado da migração de linfócitos infectados após a ocorrência de lesões. Essas alterações podem servir como indicadoras da presença do vírus, assim são úteis como um alerta clínico para a infecção em indivíduos inicialmente assintomáticos, principalmente em regiões endêmicas para o vírus (DANTAS et al., 2014).

O HTLV-1 também têm sido associado a artropatia inflamatória crônica, relacionada à hiperplasia de células sinoviais e à resposta autoimune desregulada na região articular, onde pode-se observar inflamação e destruição de ossos e articulações em intensidades variadas. Frequentemente são encontradas alterações nos níveis de marcadores de inflamação (proteína C-reativa, ácido siálico), além de anticorpos contra produtos gênicos do HTLV-1 tanto no soro (IgG) quanto no líquido sinovial (IgM), indicando uma resposta imunológica ativa contra estes antígenos virais. Algumas áreas afetadas são joelho, ombro e pulso. No entanto, é difícil fazer a distinção destas características clínicas como causadas pelo próprio HTLV-1 ou relacionadas a uma artrite reumatoide idiopática (NISHIOKA; SUMIDA; HASUNUMA, 1996).

2.5.1 PET/MAH

2.5.1.1 Descrição

A PET/MAH é uma condição neurológica reconhecida como prevalente, inicialmente, em regiões tropicais, como Colômbia, Martinica e sul da Índia (GESSAIN et al., 1985). Caracteriza-se como uma mielopatia crônica, progressiva, com comprometimento espinhal, podendo envolver paraparesia (perda parcial da função motora principalmente dos membros inferiores), paraplegia (perda total da função motora principalmente dos membros inferiores), distúrbios sensoriais e desregulação esfincteriana. Essas manifestações possuem um início insidioso e lento, tornando-se muitas vezes difícil determinar quando exatamente os sinais iniciaram. A principal região acometida é a medula espinhal torácica. Aceita-se que as manifestações resultam de um processo inflamatório de natureza crônica e progressiva, que resulta da ativação de linfócitos e consequente liberação de citocinas no sistema nervoso central, o que causa lesões desmielinizantes e citotóxicas, comprometendo o tecido nervoso (FUZII et al., 2014).

2.5.1.2 Diagnóstico

Após a descrição da PET/MAH na década de 80 (GESSAIN et al., 1985), ainda em 1988 a Organização Mundial da Saúde (OMS) elaborou um guia de auxílio ao diagnóstico da PET/MAH, o qual foi revisado em 1989. De acordo com este guia, o paciente em avaliação precisa atender aos critérios estabelecidos para que o diagnóstico de PET/MAH seja confirmado. Logo, há dois estados possíveis deste paciente: ou com ou sem PET/MAH. Estes critérios são divididos em critério clínico, onde se avalia a presença de manifestações neurológicas (como fraqueza nos membros inferiores, hiperreflexia, sintomas sensoriais disfunções sexuais e da bexiga) e não neurológicas (como síndrome de Sjögren, uveíte e alveolite pulmonar) e critério laboratorial, onde se avalia a presença de antígenos ou anticorpos para o HTLV no sangue ou no fluido cefalorraquidiano, aumento da concentração de proteínas no fluido cefalorraquidiano, ou alterações no número ou na morfologia dos linfócitos.

Mais recentemente, Castro-Costa et al. (2006) propuseram nova implementação destes critérios de classificação, simplificando-os com base nos critérios inicialmente propostos pela OMS e nas práticas correntes descritas na literatura pesquisada, entre artigos científicos relacionados aos achados clínicos da PET/MAH e publicados no período de 1985 a

2002. Em particular, foi dada atenção aos doentes iniciais, nos quais PET/MAH é suspeita, mas nem todos os critérios de diagnóstico são atingidos.

Neste contexto, reunidos na cidade de Belém (Estado do Pará) em 2002 e posteriormente na cidade de Belo Horizonte (Minas Gerais) em 2004, Castro-Costa et al. (2006) elaboraram uma classificação consensual de três grupos: definido (atendem os quesitos de diagnóstico de PET/MAH), provável (com uma apresentação monossintomática) e possível (possuem algumas manifestações, mas outras causas não foram excluídas). No entanto, segundo os próprios autores, esta implementação não substitui os critérios da OMS, que são mais estritos, mas deve ser vista como uma ferramenta complementar de diagnóstico e acompanhamento de pessoas infectadas pelo HTLV-1 e 2 e que apresentam sintomatologia completa ou parcialmente desenvolvida. Abaixo temos o resumo destes três níveis de determinação:

- Definido: o paciente apresenta paraparesia espástica progressiva, com deficiência de marcha perceptível por ele próprio. Sinais/sintomas sensoriais estão presentes ou não. Se sinais/sintomas sensoriais estão presentes, possuem aspecto súbito. Disfunções urinária e esfincteriana estão presentes ou não. Possui anticorpos anti-HTLV-1 no plasma e fluido cefalorraquidiano (com confirmação por Western Blot) e/ou possuem PCR positivo para HTLV-1 no sangue e/ou no fluido cefalorraquidiano. Não possui outras condições que lembrem PET/MAH.
- Provável: o paciente apresenta espasticidade ou hiperreflexia nos membros inferiores, ou sinal de Babinski (com ou sem sinais/sintomas sensoriais de aspecto súbito), ou bexiga neurogênica (confirmada por testes urodinâmicos). Possui anticorpos anti-HTLV-1 no plasma e/ou fluido cefalorraquidiano (com confirmação por Western Blot) e/ou possuem PCR positivo para HTLV-1 no sangue e/ou no fluido cefalorraquidiano. Não possui outras condições que lembrem PET/MAH.
- Possível: o paciente apresenta caracterização clínica completa ou incompleta para PET/MAH. Possui anticorpos anti-HTLV-1 no plasma e/ou fluido cefalorraquidiano (com confirmação por Western Blot) e/ou possui PCR positivo para HTLV-1 no sangue e/ou no fluido cefalorraquidiano. Ao contrário dos demais grupos, desordens que lembrem PET/MAH não foram excluídas.

Há várias condições que lembram PET/MAH (condições confundidoras) e que devem ser excluídas por exame clínico ou laboratorial. O artigo original lista as seguintes: Esclerose múltipla, esclerose lateral amiotrófica, meningite carcinomatosa, paraparesia espástica familiar, mielite transversa, esclerose lateral primária, síndromes paraneoplásicas,

siringomielia, doença de Lyme, deficiência de folato ou de vitamina B12, doença de Behçet, neurosífilis, neurotuberculose, sarcoidose, mielopatia vacuolar ligada ao HIV, compressão da medula espinhal, doenças colágeno-vasculares, mielopatias autoimunes, síndrome de Sjögren, mielopatia hepática, mielopatia fúngica, mielopatia parasítica (larva migrans visceral de *Toxocara canis* e *Ascaris suum*), mielopatia tóxica, fístula arteriovenosa espinhal e mielopatias regionais endêmicas com manifestações clínicas similares (como a esquistossomíase e neurocisticercose).

2.5.1.3 Patogênese

A patogênese da PET/MAH ainda não é bem compreendida (FUZII et al., 2014), mas sabe-se que uma das principais características que desencadeiam esta condição é a infiltração perivascular de linfócitos T $CD4^+$ (frequentemente nas lesões iniciais) e T $CD8^+$ (frequentemente nas lesões tardias) no tecido espinhal, os quais medeiam inflamação através da liberação de citocinas pró-inflamatórias, como IL-1, IL-6, IFN-γ e TNF-α. Em decorrência disso há um processo neurodegenerativo que leva à desmielinização e meningomielite. Frequentemente na PET/MAH são observados astrócitos infectados pelo HTLV-1, bem como o acúmulo de IgG no parênquima do sistema nervoso central, o que é evidência de rompimento da barreira hematoencefálica (OSAME et al., 1986). Além disso, populações clonais de linfócitos infectados com o HTLV-1 encontrados no fluido cefalorraquidiano derivam dos mesmos progenitores que os linfócitos infectados encontrados no sangue periférico (CAVROIS et al., 2000).

Os eventos que levam à desorganização da barreira hematoencefálica e à entrada de linfócitos infectados ainda estão sob investigação. Por exemplo, Afonso et al. (2008) encontraram que células endoteliais da barreira hematoencefálica expressam em sua superfície receptores usados pelo HTLV-1 (GLUT-1, Neuropilina-1 e *Heparan Sulfate Proteoglycans* - HSPG), tanto na medula espinhal de humanos com PET/MAH quanto em indivíduos não infectados. Nesse estudo, a linhagem celular de células endoteliais humanas hCMEC/D3, a qual também expressa esses mesmos receptores, foi infectada pelo vírus e passou a produzir partículas virais viáveis (infectantes). Além disso, os autores mostraram que essa mesma linhagem formava uma monocamada *in vitro* quando infectada pelo HTLV-1, permitindo maior transmigração de linfócitos não infectados pelo vírus, ou de linfócitos infectados.

Recentemente, Curis et al. (2016) encontraram que a proteína Tax recruta o complexo NF-κB para a região promotora do gene *alcam*, o qual codifica a proteína Molécula de Adesão de Leucócitos Ativados (*Activated Leukocyte Cell Adhesion Molecule* ou ALCAM). Dessa forma a Tax aumenta a expressão de ALCAM pelas células infectadas, por meio da via NF-κB. O extravasamento normal de leucócitos ocorre em três etapas: rolagem, quando os leucócitos aderem fracamente ao endotélio; aderência, quando há aumento da adesão por intermédio de moléculas de aderência na superfície do leucócito e do endotélio (como as integrinas); e transmigração através do epitélio (diapedese). A proteína ALCAM é importante na transmigração de monócitos através do tecido endotelial. Os autores encontraram que células T $CD4^+$ infectadas com o HTLV-1 superexpressam ALCAM em sua superfície, o que foi associado a uma maior migração desses linfócitos através de um modelo experimental de barreira hematoencefálica, sugerindo assim uma função no estágio inicial da PET/MAH. Também, o bloqueio dessa expressão diminuiu significativamente a migração das células infectadas.

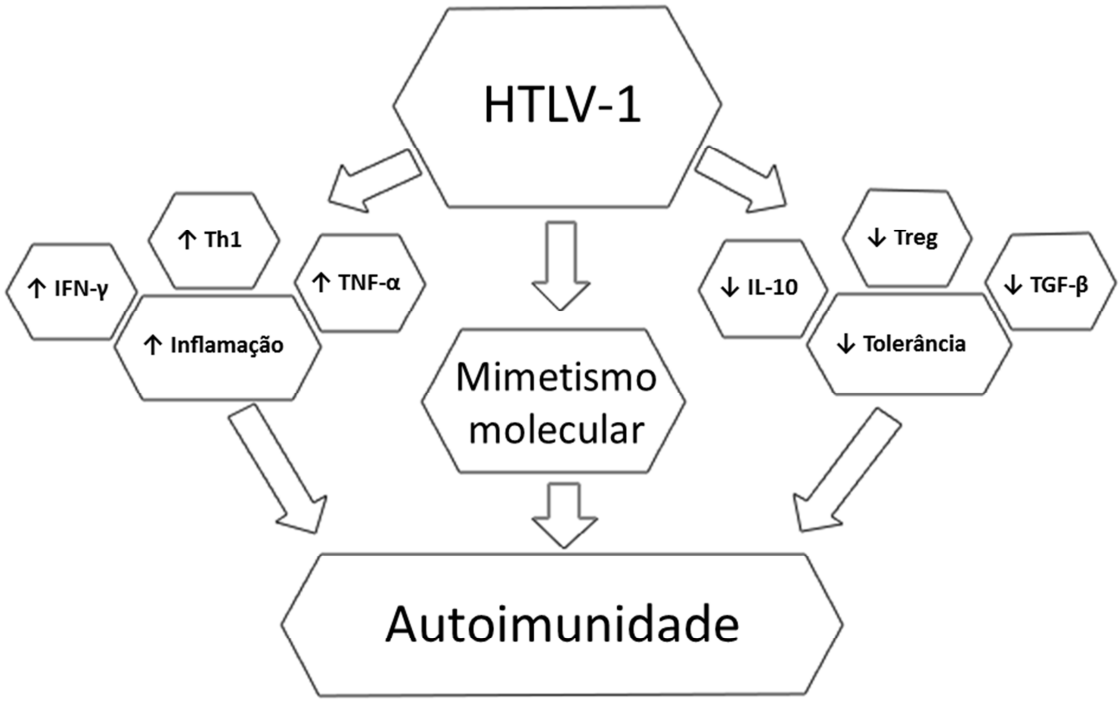
ALCAM é expressa em células Treg de indivíduos não infectados. O fato de linfócitos infectados pelo HTLV-1 expressarem essa proteína corrobora outros achados que mostram que as células T $CD4^+$ podem exibir um fenótipo de células Treg, apresentando em sua superfície moléculas CD25, CD45RO e CD127, específicas deste tipo celular (KRESS; GRASSMANN; FLECKENSTEIN, 2011). Células Treg estão envolvidas em mecanismos de controle da resposta imunológica, secretando citocinas anti-inflamatórias como IL-10 e TGF-β (características de uma resposta Th2), sob influência da citocina IL-4 e sob regulação do fator de transcrição FOXP3. Possuem importante papel nas doenças autoimunes pois sua desregulação aumenta a sensibilidade da resposta imunológica, permitindo uma resposta exacerbada e deletéria. Por outro lado, estudos indicam que há aumento da produção de citocinas pró-inflamatórias favorecendo um fenótipo Th1, em indivíduos com PET/MAH, com o aumento da relação IFN-γ/IL-10 e TNF-α/IL-10 tanto em células T $CD4^+$ quanto em T $CD8^+$ nestes pacientes (ESPÍNDOLA et al., 2015; BRITO-MELO et al., 2007; FURUYA et al., 1999), bem como o aumento da população de células T $CD8^+$ acompanhando o aumento da carga proviral (KUBOTA et al., 2000). Em especial, o aumento da carga proviral resulta da expansão espontânea das células T $CD4^+$ infectadas.

Outro aspecto menos explorado é a possibilidade do mimetismo molecular estar presente na patogênese da PET/MAH. Mimetismo molecular ocorre quando componentes virais compartilham locais antigênicos com componentes normais do hospedeiro. Neste caso, a infecção viral pode levar a uma resposta autoimune, pois anticorpos gerados contra esses

componentes compartilhados acabam reconhecendo o próprio como não-próprio. Por exemplo, García-Vallejo, Domínguez e Tamayo (2005) encontraram que anticorpos obtidos de 11 pacientes com PET/MAH reagiam contra componentes do organismo não infectado pelo HTLV-1: o anticorpo monoclonal NOR-1 (anti proteína p-24 do vírus) reagia contra o núcleo de astrócitos fetais humanos não infectados pelo HTLV-1 e o anticorpo monoclonal LT-4 (anti proteína Tax do vírus) reagia contra neurônios motores da medula espinhal de ratos. Logo, como um potencial mecanismo, assim como células T citotóxicas atacam células T $CD4^+$ infectadas pelo vírus, elas poderiam atacar o tecido saudável sob intermédio de anticorpos e de células T $CD4^+$ que entraram no SNC devido ao rompimento da barreira hematoencefálica.

A Figura 4 resume os possíveis mecanismos relacionados à autoimunidade nos indivíduos infectados pelo HTLV-1. Um desbalanço na homeostase imunológica, com aumento da resposta pro-inflamatória (pela maior expressão de citocinas do perfil Th1 como IFN-γ, TNF-α e IL-6) e diminuição da tolerância (pela menor expressão de citocinas relacionadas ao perfil Th2 como IL-10 e TGF-β) aumentam a sensibilidade da resposta e resultam em dado tecidual, e o mimetismo molecular pode atuar acionando ou exacerbando este dano.

Figura 4 - Mecanismos envolvidos na resposta autoimune observada na infecção pelo HTLV-1.



Fonte: adaptado de Quaresma et al. (2015).

Uma vez estabelecida a PET/MAH, Tax passa a ser expressa em maiores níveis tanto no sangue periférico (TAROKHIAN et al., 2017) quanto no fluido cefalorraquidiano (CARTIER; RAMIREZ, 2005). Essa oncoproteína é um potente ativador de várias vias de transcrição celular e é o mais bem caracterizado antígeno do HTLV-1. Sua interação com proteínas e complexos transcricionais da célula infectada são fatores essenciais para a replicação viral e a progressão das doenças associadas ao HTLV-1. A Tax pode ser encontrada no núcleo, no espaço extracelular e em compartimentos extranucleares: ao redor do centro organizador de microtúbulos, justaposto ao centrossomo, no complexo de Golgi e na região da sinapse virológica.

Sabe-se que a Tax é secretada para o ambiente extracelular pela via comum retículo endoplasmático - complexo de Golgi (ALEFANTIS et al., 2005). Também, têm sido demonstrado que sua secreção (a partir da linhagem de células infectadas MT-2) resulta no encurtamento de neuritos de células nervosas humanas derivadas de neuroblastoma (neurito é a denominação do prolongamento da célula nervosa, pois em neuroblastomas axônios e dendritos não são distinguíveis) (MEDINA et al., 2014). Assim, é prejudicada a função da célula nervosa. Muitas das interações da Tax com proteínas ou complexos celulares dependem de modificações pós-traducionais dela, como fosforilação, ubiquitinação, SUMOilação e acetilação. Essas modificações podem determinar a localização da proteína: a ubiquitinação da Tax está associada a uma localização citoplasmática, enquanto a SUMOilação é um sinal de retenção nuclear.

A via NF-κB possui importante papel na regulação da inflamação e da apoptose, ativando-se sob variados estímulos, como estresse oxidativo, acionamento do receptor do TNF-α (TNFR), do receptor da IL-1 (IL1R), do Toll Like Receptor (TLR) e do receptor de células T (TCR). A acetilação, a fosforilação e a SUMOilação da Tax podem contribuir para a ativação da via NF-κB (JEANG, 2001; CURRER et al., 2012). A família de fatores de transcrição NF-κB interage de forma pleiotrópica, influenciando diversos aspectos da resposta inflamatória e imune, do crescimento e da diferenciação celular.

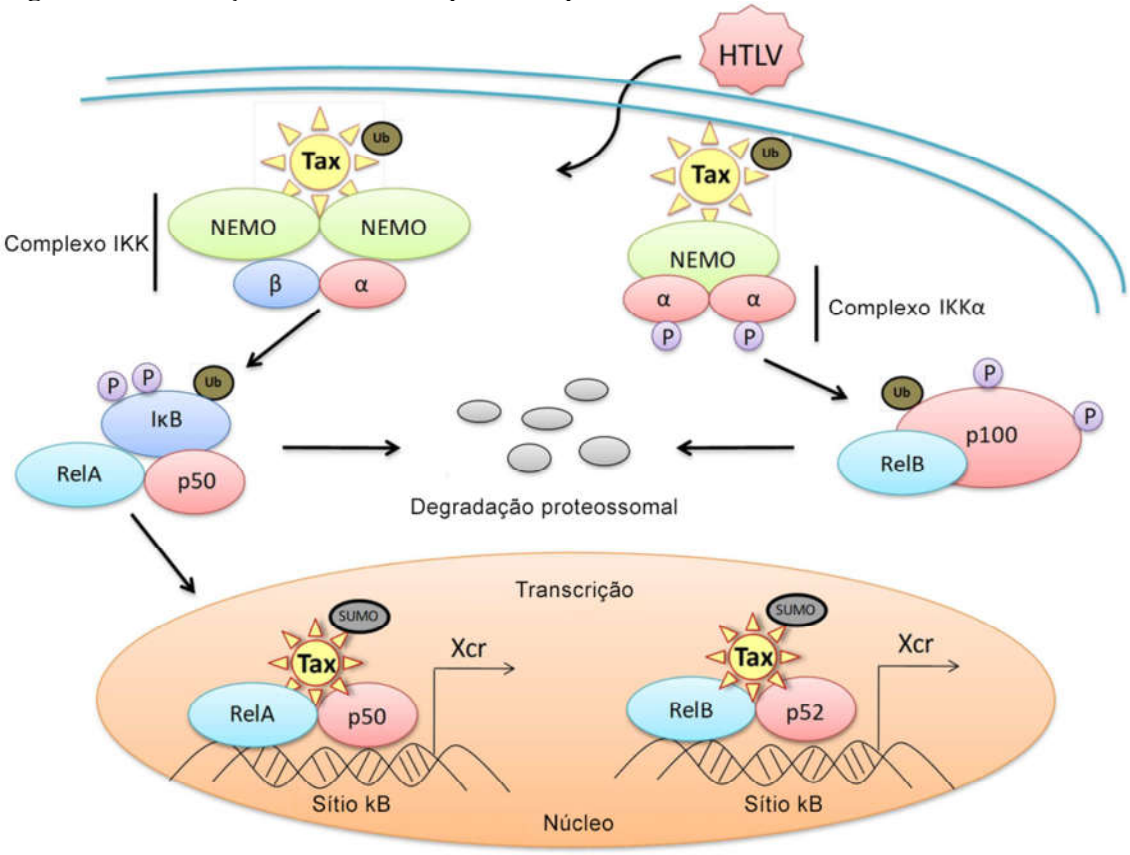
A Tax é capaz de modular a atividade transcricional celular de forma dependente da via NF-κB, o que coaduna com a observação de que as células infectadas pelo HTLV-1 mantém constitutivamente a atividade ligada a esta via: em células T inativas, as proteínas da família NF-κB encontram-se sequestradas no citoplasma por membros da família inibidores de κB (IκB) como IκBα e IκBβ. Durante a resposta imune, de forma transitória, a IκBα é fosforilada e marcada pela ubiquitina para degradação no proteassomo 26S, liberando a NF-κB que então é translocada ao núcleo e age como fator de transcrição dos seus genes-alvo

(como aqueles que codificam o fator de crescimento IL-2 e seu respectivo receptor IL-2Rα). Tax é capaz de estimular a fosforilação e a degradação da IκBα de forma transitória e da IκBβ de forma crônica, desta forma modulando e sustentando a ativação da via NF-κB (MCKINSEY et al., 1996) nas células infectadas.

Duas vias NF-κB são distintas com base em diferenças nos mecanismos de ativação: a via canônica e a via não-canônica. A Tax pode ativar as duas vias, conforme esquematizado na Figura 5: à esquerda está representada a via canônica. O complexo IκB quinase (IKK) é composto das subunidades catalíticas IKKα e IKKβ e da subunidade reguladora IKKγ (também chamada de NEMO). A Tax ubiquitinada interage com o complexo IKK por meio de contato com NEMO, o que resulta na fosforilação da IκB, sua ubiquitinação e posterior degradação no proteassomo. Com a liberação do heterodímero RelA-p50 (membros da família NF-κB), este pode migrar ao núcleo, onde ativará a transcrição dos genes-alvo, inclusive por intermédio da Tax SUMOilada que se encontra retida no núcleo.

Na Figura 5 à direita está representada a via não-canônica. Neste caso, apenas a IKKα e a IKKγ estão presentes no complexo IKK, e a Tax ubiquitinada, por meio de contato com NEMO, promove a fosforilação e ubiquitinação da proteína precursora p100 e seu processamento para p50 que então, formando um dímero com outro membro da família NF-κB (RelB) migra ao núcleo e promove a transcrição dos genes-alvo, inclusive por intermédio da Tax SUMOilada. Ao contrário, as proteínas CYLD e Peptidase 20 Específica da Ubiquitina (*Ubiquitin-Specific Peptidase 20* ou USP20) promovem a deubiquitinação da Tax e assim inibem a interação Tax-complexo IKK e por conseguinte a indução da via NF-κB. Apesar disso, em células transformadas há repressão da expressão de CYLD e USP-20 (WU; ZHANG; SUN, 2011).

Figura 5 - Modulação da via NF-κB pela oncoproteína Tax do HTLV-1.



Fonte: adaptado de Currer et al. (2012).

Muitas outras interações da Tax têm sido descritas (revisado em Currer et al., 2012): o *Serum Response Factor* (SRF) é uma proteína ubíqua de localização nuclear e ativável por Tax. A SRF possui sítios de ligação na região promotora de vários genes, como *c-fos* e *c-jun*, cujos produtos gênicos (proteínas c-Fos e c-Jun) compõem o fator de transcrição *Activator Protein 1* (AP-1), que por sua vez é associado à regulação de outros genes, ligados à proliferação, diferenciação e transformação celular. As citocinas TGF-α, TGF-β e IL-2 estimulam naturalmente a c-Fos, e por conseguinte a proliferação celular via AP-1; a Tax promove a ligação da proteína Proteína de Ligação ao Elemento de Resposta do AMPc (*cAMP-Response Element Binding Protein* ou CREB) à região do Elemento Responsivo à Tax (*Tax Responsive Element* ou TRE-1), que fica na região U3 da LTR viral, desta forma, promovendo a expressão dos genes provirais na célula infectada; a Tax extracelular estimula a liberação das citocinas TNF-α, IL-6 e IL-1β pelas células microgliais, por até 8 horas após a estimulação. Até mesmo a linhagem celular neuronal NT2-N, cujas células compartilham estreitas características com neurônios humanos maduros, são rapidamente induzidas a

produzir TNF-α na presença de Tax (COWAN et al., 1997). A TNF- α é particularmente tóxica aos oligodendrócitos, causando desmielinização e consequente degeneração axonal.

Em resumo, estes estudos que esclarecem as estratégias do HTLV-1 em termos de regulação da maquinaria celular são importantes, pois apresentam novas possibilidades de intervenção terapêutica para o controle das doenças associadas ao vírus.

2.6 ANÁLISE MULTIVARIADA

Várias ferramentas de análise quantitativa têm sido propostas, permitindo o tratamento de problemas cada vez mais complexos nos diversos campos de investigação científica. Este fenômeno é resultado da crescente quantidade de informações geradas na pesquisa básica e aplicada, aliada ao desenvolvimento de *software* e *hardware* que permitem o tratamento adequado dessas informações. Dessa forma, permite-se aos investigadores acessar questões científicas de maneira que seria infactível há algumas décadas. Por outro lado, desordens multifatoriais de diagnóstico relativamente complexo como a PET/MAH, podem se sujeitar a vieses de interpretação subjetiva do avaliador, o qual precisa se basear na apreciação de critérios estabelecidos consensualmente como os critérios da OMS para o diagnóstico da PET/MAH ou os propostos por Castro-Costa et al. (2006) e em sua própria experiência profissional para classificar um novo paciente que chegue ao seu atendimento.

Assim, técnicas multivariadas adaptadas aos tipos de variáveis disponíveis no contexto clínico do diagnóstico da PET/MAH podem ser usadas como classificadoras dos pacientes. Ou seja, é possível utilizar métodos multivariados para acessar o problema da classificação de pacientes de forma mais objetiva e reprodutível. Ademais, a análise multivariada nos permite acessar o quão boas preditoras são as variáveis disponíveis, quanto à predição do grupo no qual os pacientes são melhor classificados de acordo com as suas características. Analisando diversas variáveis clínicas num contexto estatístico multivariado, mais do que acessar seu efeito individual sobre uma dada variável resposta, consegue-se verificar se sua importância está relacionada a um aspecto mais fundamental, subjacente ao efeito individual imediatamente perceptível. Quando se considera, além de variáveis clínicas, níveis de expressão gênica de citocinas que reconhecidamente possuem papel importante nos mecanismos inflamatórios subjacentes ao quadro observado na PET/MAH, pode-se verificar se estas variáveis contribuem em melhorar o poder de classificação e predição do estado do paciente.

Disso, depreende-se que uma questão inicial a abordar em problemas de modelagem e predição é quais preditores usar, dentre vários potenciais preditores disponíveis. Se muitos dos preditores usados na modelagem estatística contribuem pouco com a capacidade de explicação da variável resposta, então o modelo como um todo perde em capacidade de discriminação. De forma similar, se incluímos dois ou mais preditores altamente relacionados na modelagem (em alguns contextos chamados preditores colineares), pode haver perda espúria da capacidade de predição do modelo, levando a conclusões errôneas sobre as relações entre a variável resposta (no caso a classificação de pacientes em um ou mais estados) e as variáveis preditoras consideradas.

2.6.1 Análise de componentes principais

A Análise de Componentes Principais (*Principal Component Analysis* ou PCA) é reconhecida como uma técnica exploratória capaz de sintetizar um conjunto de variáveis em um conjunto menor de variáveis denominadas componentes principais, por meio de transformações ortogonais destas variáveis. As componentes principais, assim, são variáveis sintéticas obtidas a partir do banco de dados original e representam a parte mais importante da variabilidade dos dados iniciais. Ora, é mais simples descrever um banco de dados usando poucas variáveis sintéticas do que observando todas as variáveis originais. Além disso, as componentes principais são ortogonais, isto é, não são correlacionadas entre si, resolvendo dessa forma o problema da multicolinearidade em relações lineares no âmbito da PCA. Logo, a PCA foca em relações lineares entre as variáveis (RENCHER; CHRISTENSEN, 2012).

A técnica se concentra na estrutura de uma única amostra de observações mensuradas em p variáveis. Se há correlações fortes entre algumas variáveis, então é provável que haja uma estrutura bem definida. Se as p variáveis se correlacionam fracamente, então elas são mais ou menos independentes e não há uma estrutura bem definida. O número de componentes principais obtidas sempre é menor ou igual a p, de tal forma que a primeira componente extrai a maior parte possível da estrutura do banco de dados original, e cada componente seguinte extrai também a maior parte possível, sob a condição de que ela seja ortogonal às componentes obtidas nos passos anteriores. Pensando em termos de correlação, se um subconjunto de variáveis emprega alta carga sobre a mesma componente principal, então esse conjunto de variáveis é bem correlacionado entre si. Dessa forma, cada componente extraída dos dados representa variáveis tão correlacionadas quanto possível.

Discute-se o resultado da PCA em termos de escores dos indivíduos e em termos dos coeficientes das variáveis preditoras em cada componente: a Equação (1) ilustra a primeira componente principal, em que $e_1, \dots, e_p$ correspondem aos coeficientes das p variáveis originais na primeira componente e indicam a importância de cada variável nesta componente (quanto maior o valor absoluto do coeficiente, maior a importância da respectiva variável, em uma dada componente). Já $y_{1i}, \dots, y_{pi}$ correspondem aos valores das p variáveis originais para a observação de ordem i. A equação resulta em um escore z que corresponde à relevância da observação de ordem i nesta primeira componente, considerando-se $i = 1, \dots, n$ observações:

$$z_{1i} = e_1 y_{1i} + e_2 y_{2i} + e_3 y_{3i} + \cdots + e_p y_{pi}. \quad (1)$$

Neste trabalho consideramos uma variante da PCA chamada também de PCA mista, proposta inicialmente por Hill e Smith (1976). Ela permite incluir na mesma análise tanto variáveis quantitativas quanto qualitativas, uma vez que a proposição inicial da PCA se aplica apenas às variáveis quantitativas. Para isso, na PCA mista, considera-se um método específico de cálculo de correlação entre uma variável qualitativa e uma quantitativa, desta forma possibilitando incluir as duas naturezas de dados em uma mesma análise.

Um interessante efeito da PCA mista é que, se todas as variáveis independentes são quantitativas contínuas, o resultado corresponde à PCA convencional. Similarmente, se todas as variáveis são categóricas, o resultado corresponde à convencional Análise de Correspondência Múltipla (TENENHAUS; YOUNG, 1985; RENCHER; CHRISTENSEN, 2012), a qual é adequada para lidar com preditores exclusivamente categóricos e se baseia nas medidas de distâncias do qui-quadrado para descrever as relações entre diferentes categorias das variáveis preditoras e entre os diferentes indivíduos (observações).

O pacote R *ade4* (CHESSEL; DUFOUR; THIOULOUSE, 2004; DRAY; DUFOUR et al., 2007) foi proposto inicialmente para ecólogos de comunidades, os quais necessitam lidar com bancos de dados grandes, buscando entender o papel de vários fatores ambientais, geralmente sobre muitas espécies de diferentes locais em um mesmo estudo. Algumas técnicas de análise multivariada que estão disponíveis neste pacote são: Análise de Componentes Principais, Análise de Correspondência (simples e múltipla), Análise de Coordenadas Principais, Análise Entre Classes e Análise Linear Discriminante. Do nosso interesse, o pacote traz a função *dudi.mix* que implementa e generaliza as ideias de Hill e Smith (1976) e Kiers (1991) para a PCA mista.

Nessa implementação da PCA mista as componentes principais são obtidas de forma que maximizem, para as variáveis quantitativas, a soma dos quadrados dos coeficientes de

correlação, para as variáveis qualitativas ordinais, a soma dos quadrados dos coeficientes de correlação múltipla e para as variáveis qualitativas não ordinais, as razões de correlação. Além disso, essa implementação tende a separar clusters de indivíduos um pouco melhor, conforme discutido em Kiers. Antes de proceder à PCA mista propriamente dita, as variáveis são pré-processadas por *dudi.mix* de acordo com sua natureza da seguinte maneira:

- As variáveis numéricas são consideradas como supridas. Elas podem ser medidas em diferentes escalas, uma vez que a matriz de correlações, utilizada aqui, é invariante à escala. Logo, não é necessário padronizar as variáveis para evitar o viés associado a faixas de valores desiguais, o que poderia mascarar a contribuição de cada variável quantitativa (discutido em Rencher e Christensen, 2012, p. 419).
- As variáveis categóricas não ordinais são consideradas na análise através de seus contrastes. Por exemplo, a variável *sexo* é substituída por outras duas variáveis: *sexo.M* que representará a importância da categoria *masculino* em uma dada componente principal e *sexo.F* que representará a importância da categoria *feminino* para esta componente. Então, para uma variável de entrada com *k* níveis, haverá *k* contrastes na saída da PCA. Observe que os contrastes são independentes, pois se a categoria assumida é masculino, esta não pode ser feminino ao mesmo tempo.
- As variáveis categóricas ordinais são também consideradas através de seus contrastes, porém, são contrastes ortogonais polinomiais. Ortogonais porque são independentes entre si. Polinomiais porque representam uma tendência exponencial específica da variável original. Por exemplo, uma variável ordinal *risco* que pode assumir os valores *baixo*, *médio* e *alto* é substituída por outras duas variáveis: *risco.L* que representa a tendência linear (expoente 1) da variável inicial risco e *risco.Q* que representa a tendência quadrática dela. Então, para uma variável ordinal de entrada com *k* níveis, haverá dois contrastes na saída da PCA. É possível considerar outras tendências também, como a cúbica (expoente três), mas esta implementação se limita ao grau dois. Uma desvantagem destes contrastes, porém, é que se considera que as categorias da variável ordinal sejam igualmente espaçadas, o que nas variáveis verdadeiramente ordinais não pode ser garantido. No entanto, é uma abordagem mais adequada do que considerar que as categorias não possuem um ordenamento intrínseco.

2.6.2 Regressão linear e métodos de contração de coeficientes (*shrinkage*)

A Regressão Linear convencional (RENCHER; CHRISTENSEN, 2012) tem por objetivo modelar a relação linear existente entre uma variável resposta contínua *y* (também chamada de variável dependente) e uma (Regressão Linear Simples) ou várias (Regressão Linear Múltipla) variáveis preditoras (também chamadas variáveis independentes), denotadas como $x_1, \dots, x_p$ para p variáveis preditoras. Essa modelagem envolve a estimação de parâmetros a partir do banco de dados que está disponível ao pesquisador. Estes parâmetros são os coeficientes das variáveis preditoras, denotados como $\beta_1, \dots, \beta_p$ para p preditoras. Logo, o modelo geral possui a forma:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \tag{2}$$

onde há $i = 1, \dots, n$ observações. β_0 corresponde ao intercepto, ou seja, o valor *y* assumido se todas as variáveis independentes forem iguais a zero. β_0 pode ser estimado ou não.

O termo ε_1 representa o erro, também chamado resíduo ou ruído, associado à estimativa da variável *y* associada à observação *i*, uma vez que o valor previsto pela equação geralmente difere daquele fornecido no início do processo de modelagem. Assim, $\varepsilon_i = y_i - \hat{y}_i$, onde *y* é o valor observado da variável resposta e $\hat{y}$ é o valor previsto pela equação. Esse erro pode ser entendido como todos os outros fatores que influenciam a variável *y* que não as variáveis preditoras incluídas no modelo. Logo, assume-se que ε é uma variável aleatória não observada. Uma medida sintética do erro associado ao modelo como um todo é a soma dos quadrados dos erros individuais (SSE ou *Sum of Squared Errors*), ou seja, eleva-se cada erro individual ao quadrado e por fim somam-se todos os quadrados:

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 \tag{3}$$

Conseguimos uma forma de representação mais concisa de (2) através da notação matricial. Para isso, consideramos:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \text{ vetor variável resposta,}$$

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \text{ matriz de } p \text{ variáveis preditoras em } n \text{ observações,} \tag{4}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \text{ vetor com os parâmetros desconhecidos,}$$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_i \end{pmatrix}, \text{ vetor com os erros de predição.}$$

Podemos então representar sinteticamente (2) da seguinte forma:

$$y = X\beta + \varepsilon. \tag{5}$$

O método mais utilizado na literatura para estimar o vetor β do modelo de regressão linear (5) é o método dos mínimos quadrados, onde se obtém os coeficientes de regressão de forma a minimizar a soma dos quadrados dos erros (SSE). Na forma matricial isto é alcançado ao multiplicar $\varepsilon^T\varepsilon$, onde ε^T é a matriz transposta de ε. Da Equação (5) observamos que $\varepsilon = y - X\beta$. Logo:

$$\varepsilon^T\varepsilon = (y - X\beta)^T(y - X\beta). \tag{6}$$

Ao igualar (6) a zero e simplificar conforme Reynaldo (1997, p. 5), obtém-se a estimativa do vetor de coeficientes:

$$\beta = (X^TX)^{-1}X^Ty, \tag{7}$$

onde X^T é a matriz transposta de X. Note que a primeira coluna de (4) é composta de uns. Isso é necessário para que (7) estime também o intercepto, além dos coeficientes das variáveis preditoras.

No entanto, essa estimação clássica de coeficientes apresenta alguns problemas em situações específicas. Um deles é a multicolinearidade. Esta é definida como a dependência linear de várias colunas de X, ou seja, preditoras estão correlacionadas de forma que a sua relação com a variável resposta é obscurecida. Além disso, a multicolinearidade aumenta a variabilidade do estimador de mínimos quadrados, resultando em estimativas mais dispersas. Assim, o erro padrão dos coeficientes de regressão pode ser alto, ao ponto de nenhuma variável preditora ser estatisticamente significante (REYNALDO, 1997).

Outro problema está ligado à relação entre n e p: quanto mais parâmetros sendo estimados em relação ao tamanho amostral, mais provavelmente podemos encontrar uma solução perfeita, em que y é completamente predita por X, mas apenas como um artefato resultante do elevado número de variáveis preditoras incluídas no modelo. Denomina-se este fenômeno de sobre-ajuste (*overfitting*), ou seja, o modelo se ajusta muito aos dados, seu desempenho é ótimo quando aplicado sobre os próprios dados iniciais, mas o desempenho é baixo quando aplicado a dados novos, que não os usados para derivar o modelo. Atenção especial deve ser dada quando muitas preditoras são variáveis categóricas, caso em que uma única variável com cinco categorias resultará na estimação de quatro coeficientes, quando a codificamos em quatro variáveis *dummy*.

Neste contexto, Hoerl e Kennard (1970) propuseram que se adicionasse uma penalização k ao estimador de mínimos quadrados, resultando numa variância menor que a do estimador original, porém, a estimação é enviesada, mas os problemas citados acima são amenizados ou sanados. Além disso, ao promover contração dos coeficientes, em algumas situações é melhorada a acurácia preditiva, ou o quão próximo dos valores reais são os valores previstos pelo modelo penalizado, quando aplicado sobre novos dados. A Equação (7) modificada é o estimador RIDGE dado por:

$$\beta_k = (X^T X + kI)^{-1} X^T y,\ k \geq 0, \tag{8}$$

onde I é a matriz identidade. Como se considera o inverso da porção entre parênteses (elevado a -1), aumentando k diminuímos β_k, o qual tende a zero (ou seja, tende ao modelo apenas com o intercepto). De forma análoga, se k tende a zero, β_k tende ao estimador dos mínimos quadrados, Equação (7). Percebe-se que a escolha de k é crítica para a determinação da magnitude do efeito de contração dos coeficientes e redução da variância da estimativa. Há vários métodos para escolha do valor ótimo de k, mas neste trabalho focamos o método de validação cruzada (*cross-validation*), que escolhe esse índice de penalização de forma dependente dos dados (discutido mais adiante).

Apesar de ser um estimador mais estável, RIDGE não realiza seleção de variáveis, uma vez que aplica a contração dos coeficientes de forma contínua, levando todos os coeficientes a zero conjuntamente. Outra proposta de penalização de coeficientes é o estimador Operador de Seleção e Contração Mínimo Absoluto (*Least Absolute Shrinkage and Selection Operator* ou LASSO), de Tibshirani (1996), no qual a estimação dos coeficientes é realizada iterativamente e o processo ao mesmo tempo realiza contração de alguns coeficientes e força outros a zero. Para isso, em LASSO busca-se a minimização da soma dos quadrados dos erros, porém sob a condição de que a soma dos valores absolutos dos coeficientes seja menor ou igual a um parâmetro de penalização t, ou seja:

$$\sum_p |\beta_p| \leq t, \text{ sendo } t \geq 0.$$

Logo, t controla a magnitude da contração que é aplicada. Consideramos t_0 a soma dos valores absolutos dos coeficientes obtidos pela estimação clássica de mínimos quadrados: $t_0 = \sum_p |\beta_p^0|$. Valores $t < t_0$ causarão contração dos coeficientes tendendo ao zero, sendo que alguns serão mais brevemente definidos como zero do que outros. Similarmente, se $t \geq t_0$, a solução equivale ao método clássico de mínimos quadrados. Por ser um processo iterativo, a estimação LASSO costuma ser mais intensiva computacionalmente do que RIDGE, o que é verdade quanto mais preditoras houver no modelo.

Note, porém, que o método LASSO de contração de coeficientes e seleção de variáveis considera β_p parâmetros individualmente. Caso haja variáveis preditoras categóricas, não é conveniente selecionar somente algumas das variáveis *dummy* (variáveis binárias codificando os níveis da categórica), correspondentes a algumas das categorias da variável original, mas sim considerar os níveis dessa variável preditora como um grupo de coeficientes, a serem selecionados conjuntamente. Para resolver este problema, Yuan e Lin (2006) propuseram o método de penalização *Group*-LASSO.

Em *Group*-LASSO, definem-se G grupos de variáveis preditoras, e cada grupo possui uma matriz X_g de preditoras e um vetor de parâmetros β_g. O método combina as vantagens de RIDGE e LASSO: sobre os G grupos relativos aos conjuntos de parâmetros sendo estimados é aplicada a penalização do tipo LASSO, e dentro de cada matriz X_g é aplicada a penalização do tipo RIDGE. De fato, se cada grupo possui um único preditor, a solução equivale a LASSO.

Conceitualmente, em termos de otimização do modelo, podemos definir que o método dos mínimos quadrados busca minimizar a soma dos quadrados dos erros (SSE):

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \sum_{p}\beta_p x_{ip}\right)^2,$$

sendo $\hat{y}$ o valor predito da variável resposta; o método RIDGE busca minimizar a soma penalizada dos quadrados dos erros (*Penalized Sum of Squared Errors* ou PSSE), com base em β_p^2:

$$\sum_{i=1}^{n}\left(y_i - \sum_{p}\beta_p x_{ip}\right)^2 + \lambda\sum_{p}\beta_p^2,$$

o método LASSO também busca minimizar PSSE, mas com outra forma de penalização, baseada em $|\beta_p|$:

$$\sum_{i=1}^{n}\left(y_i - \sum_{p}\beta_p x_{ip}\right)^2 + \lambda\sum_{p}|\beta_p|,$$

e o método Group-LASSO busca minimizar (Friedman; Hastie; Tibshirani, 2010):

$$\sum_{i=1}^{n}\left(y_i - \sum_{g}\beta_g X_g\right)^2 + \lambda\sum_{g}\sqrt{g_n}\,\beta_g,$$

sendo X_g a matriz de variáveis preditoras no grupo g, β_g o vetor de coeficientes no grupo, g_n o número de variáveis no grupo g, e $\sqrt{g_n}$ um fator de compensação para os diferentes tamanhos entre os grupos. O efeito de LASSO em relação a RIDGE se deve ao fato de que $|\beta|$ é muito maior que β^2 para valores baixos de β. Nos três últimos casos, $\lambda \geq 0$ é o parâmetro de penalização. Podemos resumidamente dizer que os métodos de contração visam minimizar a soma dos quadrados dos erros mais uma penalidade aplicada sobre os coeficientes. Quanto maior λ maior a contração dos coeficientes estimados, e quanto menor, mais a solução se aproxima do método dos mínimos quadrados.

2.6.3 Modelagem de variáveis preditoras ordinais e métodos de *shrinkage*

Variáveis categóricas ordinais, como o nome sugere, são aquelas em que apenas a ordem possui um significado prático. Considerando uma variável *risco* que possui os níveis *baixo*, *médio* e *alto*, podemos dizer que *baixo* < *médio* < *alto*, porém, não podemos quantificar a diferença entre dois níveis consecutivos. Mesmo se categorizarmos uma variável numérica, como idade, nos níveis 0-18 anos, 19-24 anos, 25-35 anos, maior que 35 anos, vemos que as categorias são de tamanhos diferentes, com "maior que 35 anos" muito maior que as demais, logo apenas convém considerar o ordenamento intrínseco. Variáveis ordinais são frequentemente vistas nas áreas de ciências sociais e ciências biomédicas, por exemplo na codificação de escalas funcionais, como as preconizadas pela OMS para a Classificação Internacional de Funcionalidades (*International Classification of Functioning, Disability and Health* ou ICF) (World Health Organization, 2001). Por isso, são necessários métodos que melhor se utilizem da natureza dessas variáveis.

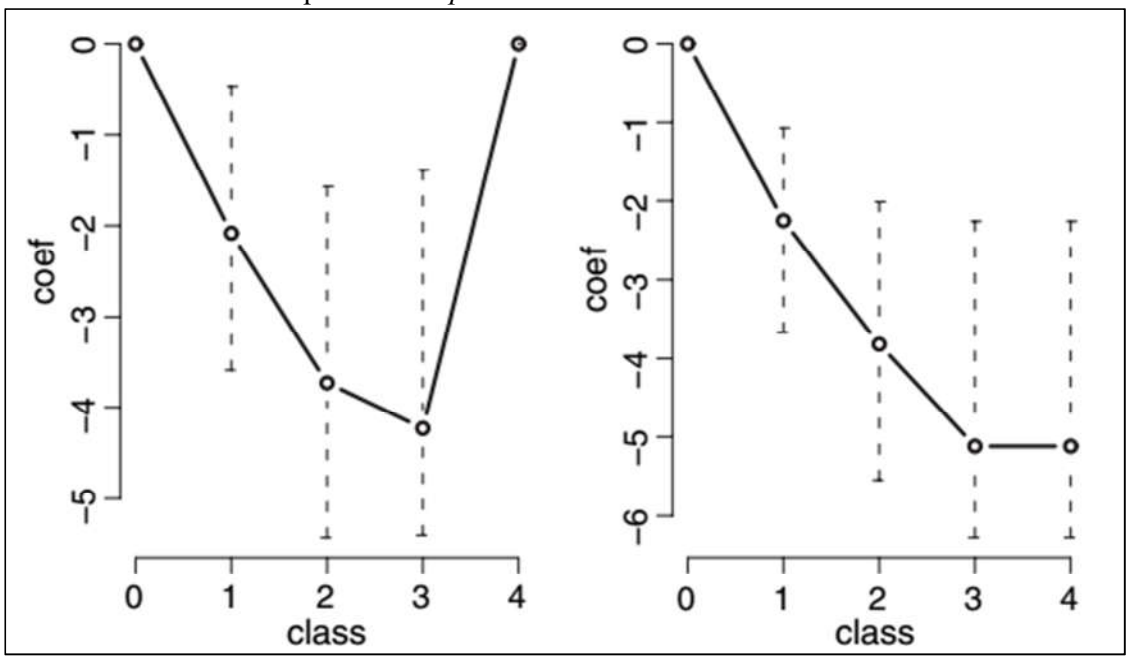
Modelos em que a variável resposta é ordinal têm sido relativamente bem explorados na literatura (revisado em Liu e Agresti, 2005), por exemplo na Regressão Logística Ordinal. Porém, o caso das variáveis ordinais preditoras têm sido negligenciado, apesar de que elas são frequentemente vistas em modelos de regressão. Muitas vezes, as preditoras ordinais são tratadas ou como variáveis puramente categóricas (ignorando a informação sobre ordenamento das categorias), ou como variáveis numéricas discretas ou contínuas (artificialmente, assumindo atributos de mais alto nível que a variável não possui).

Gertheiss e Tutz (2009) e Gertheiss et al. (2011) propuseram uma abordagem diferenciada para tratar de variáveis preditoras ordinais no contexto dos modelos de regressão, utilizando para isso os conceitos relacionados à regressão penalizada, discutidos nas seções

anteriores. Essa abordagem consiste em codificar cada fator ordinal em um grupo de co-variáveis binárias (no mesmo sentido das variáveis *dummy*), com $j = 0, 1, \dots, k$ co-variáveis e o primeiro nível servindo como nível de referência (k é o número de variáveis *dummy* que codificam a variável ordinal). Para fins de seleção de variáveis utiliza-se o algoritmo *Group*-Lasso, que penaliza via parâmetro λ os grupos de coeficientes (os grupos são as variáveis *dummy* do mesmo fator). Para fins apenas de "suavização" de coeficientes de um mesmo fator, é aplicado um λ de forma a obter uma transição mais suave entre os coeficientes, uma vez que se assume que a variável resposta y deve alterar-se suavemente entre duas categorias adjacentes do preditor ordinal. No entanto, a penalização não é aplicada sobre os coeficientes em si, mas sim sobre a diferença entre coeficientes consecutivos, para que se obtenha esse efeito de suavização, como um método RIDGE generalizado. Logo, espera-se que os coeficientes das variáveis preditoras ordinais não sejam muito irregulares.

Além dos efeitos benéficos inerentes à aplicação de métodos de *Shrinkage*, essa forma de penalização sobre diferenças entre coeficientes permite que um coeficiente seja obtido mesmo se a respectiva categoria não é observada nos dados. Isso permite que categorias raras sejam incluídas no modelo e consideradas em previsões futuras. O efeito de suavização devido à natureza ordinal da variável absorve parcialmente a ausência dessas categorias. Por exemplo, a Figura 6 ilustra os coeficientes de uma mesma variável preditora ordinal, ajustados pelo método LASSO clássico (à esquerda) e pelo método de penalização de diferenças entre coeficientes (à direita). O eixo y representa o valor do coeficiente ajustado e no eixo x estão representadas as respectivas cinco categorias da variável ordinal. A categoria quatro não é observada nos dados, por esse motivo o respectivo coeficiente assume o mesmo valor da categoria adjacente três, o que melhora a interpretabilidade do modelo, pois a natureza ordinal da variável fica mais bem representada.

Figura 6 - Coeficientes obtidos pela penalização LASSO padrão (à esquerda) e pela penalização LASSO ordinal. As barras verticais são intervalos de confiança de 90% obtidos por *bootstrap*.



Fonte: adaptado de Gertheiss et al. (2011).

No âmbito da regressão linear convencional (com variável resposta numérica e contínua), o método de suavização de coeficientes proposto pelos autores objetiva minimizar a seguinte soma penalizada dos quadrados dos erros (TUTZ; GERTHEISS, 2014):

$$\frac{(y - X\beta)^T (y - X\beta)}{(2\sigma^2)} + \lambda \sum_{j=1}^{k} \left(\beta_j - \beta_{j-1}\right)^2, \tag{9}$$

onde o termo $(y - X\beta)^T (y - X\beta)$ corresponde ao estimador dos mínimos quadrados como na Equação (6), σ^2 corresponde à variância do erro (resíduo) e $\lambda \sum_{j=1}^{k} \left(\beta_j - \beta_{j-1}\right)^2$ corresponde ao termo de penalização. Para fins de simplificação, consideramos o modelo em que apenas uma variável preditora ordinal está presente, sendo que esta é codificada em k variáveis dummy onde a categoria de ordem zero é o nível de referência. A penalização ocorre sobre coeficientes adjacentes (subtraindo-se β_j e β_{j-1}). O efeito de suavização ocorre pois, conforme λ aumenta, coeficientes correspondentes a categorias adjacentes são forçados a ter valores similares, suavizando a curva de coeficientes, com o efeito extremo de que todos eles sejam contraídos para zero.

Gertheiss e Tutz (2009) compararam este método com outros possivelmente usados com preditores ordinais na regressão linear: estimação RIDGE convencional, regressão linear usando apenas as co-variáveis *dummy* em vez da variável ordinal original (dessa forma

tratando-a como categórica não ordinal) e usando os escores numéricos da variável ordinal (neste caso tratando a mesma como variável numérica). Por meio de simulações e da aplicação sobre um banco de dados real, os autores encontraram que seu método melhorou bastante a acurácia preditiva do modelo, diminuindo a soma dos quadrados dos resíduos tanto quando o modelo é aplicado sobre os mesmos dados que o geraram quanto se aplicados sobre dados novos.

Gertheiss et al. (2011) exploraram estes conceitos sobre a seleção de variáveis preditoras ordinais, como extensão de *Group*-LASSO. Como banco de dados de exemplo, foi empregado um conjunto de dados com 67 variáveis preditoras categóricas ordinais, todas variáveis funcionais do ICF, as quais resultaram na estimação de 332 coeficientes para um $n = 420$. Logo, trata-se de um problema multidimensional. Observou-se que o método proposto resultou em 33 variáveis selecionadas de forma relativamente estável, usando-se o método *bootstrap*, onde as iterações selecionaram com alta probabilidade esses 33 preditores. Naquela oportunidade foi empregada uma extensão das ideias dos autores para o caso em que a variável resposta é binária, no contexto dos modelos lineares generalizados (*Generalized Linear Models* ou GLM), como discutido a seguir.

2.6.4 Penalização do modelo de regressão logística binária

Os modelos lineares generalizados são uma generalização da regressão linear clássica, como forma de unificar vários métodos estatísticos num mesmo arcabouço teórico. Um dos principais métodos neste grupo é a Regressão Logística Binária, onde a variável resposta é binária, possuindo apenas dois valores: 0 (ausência de uma característica) ou 1 (presença da característica), do tipo sim/não, cura/morte, presença/ausência.

O modelo da regressão logística binária possui a seguinte forma:

$$P(y = 1) = \pi = \frac{e^{X\beta}}{1 + e^{X\beta}}, \tag{10}$$

onde modela-se a probabilidade de que a variável resposta y assuma o valor 1 ($P\ (y\ =\ 1)$). Aqui, definimos essa probabilidade como π para facilitar a representação, sendo que π assume um valor de 0 a 1, previsto pelo modelo. e é a base dos logaritmos naturais, β é o vetor estimado dos coeficientes das variáveis preditoras e X é a matriz com os valores das variáveis preditoras em colunas para cada observação nas linhas.

A Equação (10) define uma relação não linear. Se quisermos considerar seu equivalente linear, teremos:

$$ln\left(\frac{\pi}{1-\pi}\right) = X\beta, \tag{11}$$

onde ln é o logaritmo natural e equivale a $\log e$. A Equação (11) define uma equação de regressão linear onde a variável resposta $ln\left(\frac{\pi}{1-\pi}\right)$ é chamada de *logit*. Define-se como chance (ou *odds* em inglês) a probabilidade de um evento ocorrer dividida pela probabilidade de ele não ocorrer, ou seja: $\frac{\pi}{1-\pi}$. Deduz-se então que a variável resposta de (11) é o logaritmo natural da chance associada à variável y. Este logaritmo é resumidamente chamado de *logit*. Mas, por que considerar-se o *logit* como variável resposta do modelo? Uma das razões é que dessa forma pode-se converter uma probabilidade π, que varia sempre de 0 a 1, em uma faixa de valores que pode variar de $(-\infty, +\infty)$, como o conjunto de valores possíveis das variáveis preditoras, no lado direito do sinal de igualdade.

No caso da Regressão Logística Binária, o método de cálculo dos erros deve ser alterado, devido à natureza da variável resposta y, que não é quantitativa como na estimação dos mínimos quadrados. Aqui, precisamos considerar a probabilidade prevista pelo modelo: se as probabilidades previstas (no intervalo de 0 até 1) se distanciam muito dos valores observados da variável resposta (com apenas dois valores possíveis: 0 ou 1), então os erros são elevados, ou o modelo não se ajusta tão bem aos dados. No contexto dos modelos lineares generalizados, chamam-se os erros de *deviance* ou desvios. O cálculo de *deviance* se dá por:

$$\sum_{i=1}^{n} -2\log|\pi_i - s_i|, \tag{12}$$

onde $s_i = 1$ se $y_i = 0$ e $s_i = 0$ se $y_i = 1$.

Similar ao caso da regressão linear, precisamos de um método de estimação dos parâmetros do modelo, ou seja, dos coeficientes β. Aqui, o método usado para estimar os valores dos coeficientes é chamado máxima verossimilhança (ou *maximum likelihood*), onde se busca estimar, iterativamente, os parâmetros de forma a maximizar a probabilidade de obter as frequências observadas em cada uma das duas categorias da variável resposta y. Em outras palavras, busca-se estimar os parâmetros de forma que sua combinação linear minimize os resíduos do modelo preditivo, ou seja, de forma que as probabilidades previstas para a variável y sejam tão próximas do valor real observado da variável resposta quanto possível. Neste ponto, diz-se que o processo de estimação convergiu. Conceitualmente, dizemos que o processo de estimação busca maximizar o log da verossimilhança ($l(\beta)$), o qual possui a seguinte forma (MEIER; GEER; BÜHLMANN, 2008):

$$l(\beta) = \sum_{i=1}^{n} y_i X\beta - \log\left(1 + e^{X\beta}\right),$$

onde $X\beta$ corresponde à Equação (11) e y_i é o valor observado da variável resposta (0 ou 1).

Valendo-se do conceito de modelos lineares generalizados, Gertheiss e Tutz (2009) também propuseram uma extensão de sua penalização baseada em coeficientes adjacentes de uma mesma variável ordinal. Dessa forma, é possível usar o método proposto pelos autores, para o caso em que a variável resposta é binária. Neste caso, o processo de otimização do modelo objetiva maximizar o log penalizado da verossimilhança ($l_p(\beta)$) (TUTZ; GERTHEISS, 2014):

$$l_p(\beta) = l(\beta) - \gamma \sum_{j=1}^{k} \left(\beta_j - \beta_{j-1}\right)^2,$$

onde o termo de penalização de coeficientes adjacentes é idêntico ao termo de penalização da Equação (9).

2.7 MODELOS DE PREDIÇÃO CLÍNICA

Um dos principais objetivos da modelagem estatística é a predição, com base em dados que surgiram após o processo de obtenção do modelo. Por exemplo, se derivamos uma equação de regressão logística a partir de uma amostra onde a variável resposta é um desfecho clínico binário (dicotômico, como morte/cura), essa equação pode ser usada futuramente para prever o risco de morte de certo paciente novo, conhecidos os valores das variáveis preditoras, para este paciente. A Regressão Logística Binária é frequentemente usada neste contexto porque muitos fenômenos clínicos possuem como variável resposta dois estados possíveis. No entanto, algumas recomendações têm sido publicadas com relação à derivação destes modelos para uso no contexto clínico, bem como sua correta avaliação.

2.7.1 Visão geral

No contexto biomédico geralmente usa-se a denominação modelo de predição clínica. Distinguem-se dois tipos: modelos de predição diagnóstica, quando objetiva-se verificar se um paciente possui ou não determinada condição, e modelos de predição prognóstica, quando se sabe que o paciente possui a condição e deseja-se predizer o seu desfecho clínico (REILLY; EVANS, 2006). Ambos os estudos podem incluir uma simples

variável preditora, ou mais comumente, várias variáveis preditoras, buscando identificar aquelas que mais contribuem em explicar a variável resposta. Além disso, reconhece-se a importância de acessar alguns aspectos do modelo obtido, como: população alvo ao qual se aplica, definição objetiva da variável resposta, considerações a respeito do tamanho amostral, validação do modelo e, de importância prática, como o modelo se comporta em relação às medidas de desempenho (discutidas mais adiante).

Bouwmeester et al. (2012) Fizeram uma revisão de literatura sobre os estudos que envolvem predição clínica multivariada, publicados ainda no ano de 2008. De 1204 possíveis artigos identificados na pesquisa de bancos de dados biomédicos, 71 foram selecionados para a revisão. Os autores observaram que estes artigos se encaixavam em basicamente cinco tipos de estudos:

- Pesquisa de preditores: objetivam encontrar algumas, dentre várias variáveis preditoras, que melhor predizem, ou que melhor se associam à variável resposta;
- Desenvolvimento de modelo sem validação externa: objetivam desenvolver um modelo multivariado de predição, com identificação de importantes preditores, podendo incluir algum tipo de validação interna, mas sem validação externa;
- Desenvolvimento de modelo com validação externa: objetivam também desenvolver um modelo multivariado de predição, porém acessam o desempenho do modelo usando um banco de dados externo, independente, o qual não foi utilizado no processo inicial de obtenção do modelo. Por exemplo, dados de outro local (validação geográfica), ou de outro período (validação temporal);
- Estudos de validação externa: objetivam acessar o desempenho de um modelo já proposto, usando um banco de dados externo ao usado no processo de modelagem inicial. Podem eventualmente ajustar ou atualizar um modelo anteriormente estabelecido;
- Estudos de impacto do modelo: objetivam avaliar o impacto ou o efeito de um modelo de predição já definido e em vigor. Podem avaliar o comportamento médico, o desfecho clínico dos pacientes ou o custo-efetividade relacionados ao uso do modelo na prática, comparando-o ao seu não uso.

Algumas recomendações têm sido publicadas sobre o processo de desenvolvimento e avaliação de um modelo preditivo clínico. Steyerberg e Vergouwe (2014) resumiram os procedimentos em sete passos básicos, usando como exemplo ilustrativo a aplicação da

regressão logística binária sobre dados relativos ao desfecho mortalidade dentro de 30 dias após infarto agudo do miocárdio (logo, aqui o desfecho é binário - morte/não-morte):

1. Definição do problema e inspeção dos dados. Requer algum tratamento dos dados antes da elaboração do modelo: correta escolha dos potenciais preditores a serem verificados quanto à associação com a variável resposta, com base na revisão de literatura e/ou orientações de especialistas; definição dos critérios de inclusão dos pacientes no estudo; se alguns preditores possuem poucos dados perdidos, considerar métodos de múltipla imputação, em que estes dados ausentes são estimados. Isto é preferível a usar análise de casos completos, quando todos os pacientes com alguns dados perdidos são excluídos da pesquisa; e definir de forma objetiva a variável resposta de interesse.
2. Codificação dos preditores. Dependendo da natureza dos preditores (se categóricos ou contínuos), várias formas de codificação são possíveis. Por exemplo, variáveis categóricas podem ser codificadas como variáveis *dummy*. Opcionalmente pode-se fundir uma categoria rara (com poucas observações) com outras categorias, por exemplo se há as categorias anterior, posterior e as demais posições possuem uma frequência reduzida, pode-se reunir estas últimas numa categoria denominada "outros". É desencorajado dicotomizar preditores contínuos, devido à perda de informações no processo (ROYSTON; ALTMAN; SAUERBREI, 2006). Apesar de que há perda de informações ao categorizar uma variável contínua em mais de duas categorias, isso pode resultar em um modelo mais "amigável" ao usuário, desde que não haja muita perda da capacidade preditiva.
3. Especificação do modelo. Aqui, há o processo de seleção dos preditores mais apropriados ao modelo e os testes relacionados aos pressupostos do método de modelagem. A modelagem *stepwise* é muito usada na regressão binária clássica, no entanto, se o número de eventos da variável resposta é muito reduzido em relação ao número de coeficientes sendo estimados, as estimativas podem ser instáveis ou pode ocorrer sobre-ajuste do modelo aos dados. Um modelo robusto pode não se ajustar bem aos dados, mas é preferível a um modelo muito ajustado aos dados, mas que possui desempenho reduzido quando aplicado sobre dados externos.
4. Estimação do modelo. É necessário estimar os coeficientes das variáveis preditoras, após estas serem escolhidas. O método mais usado no caso da Regressão Logística Binária é o método de máxima verossimilhança, mas métodos mais recentes, como

os baseados em contração de coeficientes, possuem características importantes, como redução do sobre-ajuste.

5. Desempenho do modelo. Faz-se necessário acessar a qualidade do modelo, com medidas de desempenho frequentemente utilizadas em estudos de predição clínica.
6. Validação do modelo. Aqui, distinguem-se duas formas distintas: validação interna e validação externa. No primeiro caso, acessa-se a estabilidade do modelo em relação à população de onde os dados iniciais foram obtidos. No segundo, avalia-se como o modelo se comporta em populações suficientemente independentes. É considerada uma validação mais "forte" porque permite avaliar como os resultados obtidos no banco de dados de treino são transponíveis para dados externos.
7. Apresentação do Modelo. O formato como o modelo é apresentado ao usuário pode facilitar o seu uso. É possível disponibilizá-lo como uma fórmula, como gráfico em forma de diagrama, ou como calculadoras informatizadas, disponibilizadas *on line*. No futuro, quando os modelos de predição clínica estiverem mais amplamente em uso, seu emprego associado aos registros eletrônicos dos pacientes pode facilitar a tomada de decisões no contexto clínico.

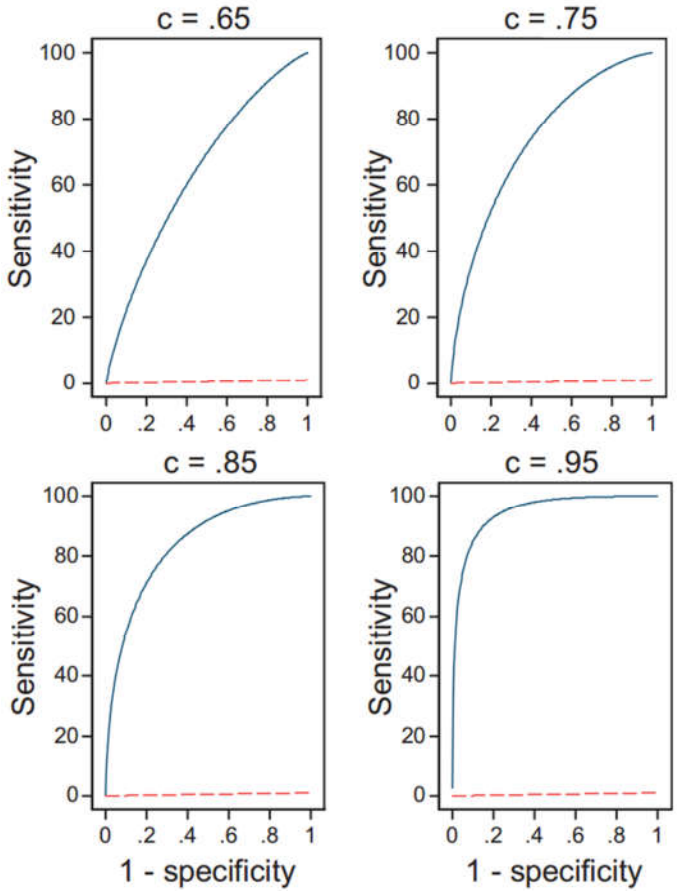
2.7.2 Validação dos modelos de predição clínica

A fase de validação do modelo é uma das mais importantes. Abaixo listamos as principais medidas utilizadas na avaliação dos modelos de predição clínica (STEYERBERG et al., 2010). Elas podem ser obtidas na validação interna ou na externa:

- Calibração: se refere à concordância entre a variável resposta observada e a prevista pelo modelo. Considerando uma variável resposta dicotômica, a calibração varia de 0 a 1, com 1 indicando perfeita concordância.
- Discriminação: se refere à capacidade de distinguir pacientes com o desfecho clínico dos pacientes sem o desfecho clínico (ROYSTON; ALTMAN, 2010). Geralmente se usa a área sob a curva ROC (*Area Under the Receiver Operating Characteristic* ou AUC) como capacidade de discriminação do modelo. A curva ROC é uma apresentação gráfica onde plota-se no eixo y a [sensibilidade] de um teste e no eixo x plota-se [1 - especificidade], formando uma curva que representa diferentes pontos de cortes. A sensibilidade é definida como a proporção de verdadeiros positivos, ou seja, a proporção de pacientes realmente com a condição que foram previstos pelo modelo como tendo a condição. [1 - especificidade] é definida como a proporção de

falsos positivos, ou seja, a proporção de pacientes previstos pelo modelo como tendo a condição, mas que na realidade não a possuíam. A Figura 7 ilustra quatro curvas ROC. Quanto maior a área sob a curva, que também é chamada de estatística *c*, maior é a capacidade de discriminação do modelo. Se AUC=0.5, não há poder de discriminação. Se AUC=1, o poder de discriminação é máximo.

Figura 7 - Curvas ROC hipotéticas. As curvas em azul são as curvas ROC, com área sob a curva ROC (também chamada estatística *c*) no topo de cada gráfico. Se *c* = 0,5, o modelo não possui capacidade de discriminação.



Fonte: adaptado de Royston e Altman (2010).

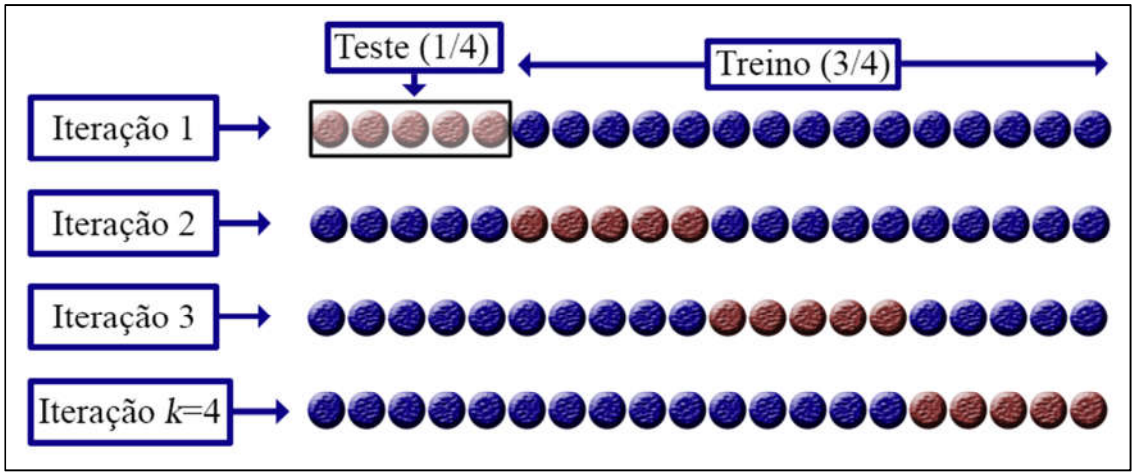
- Sensibilidade: reflete a capacidade de o modelo prever uma condição quando ela está presente. Calcula-se dividindo a quantidade de classificações corretas (a quantidade de pacientes que possuíam a condição e que foram classificados assim pelo modelo, ou seja, o número de verdadeiros positivos), pelo total de pacientes realmente com a condição.
- Especificidade: reflete a capacidade de o modelo prever a ausência da condição quando ela realmente está ausente. Calcula-se dividindo a quantidade de pacientes que não possuíam a condição e que foram classificados pelo modelo como realmente

não possuindo a condição (ou seja, o número de verdadeiros negativos), pelo total de pacientes que realmente não tinham a condição.

- Acurácia: é definida como o quão bem o modelo identifica corretamente ou exclui corretamente uma condição. Soma-se a quantidade de verdadeiros positivos e verdadeiros negativos e divide-se pelo total de predições (n amostral).

Na validação interna (STEYERBERG et al., 2001), avalia-se o modelo em relação à população de onde ele foi obtido, acessando-se sua estabilidade e a qualidade das predições. Duas abordagens muito utilizadas são a validação cruzada (*cross-validation*) e o *bootstrap*. Na validação cruzada, particiona-se o banco de dados, de forma que sejam obtidas k partições de tamanhos aproximadamente iguais. Frequentemente se usa $k = 10$. Neste caso, a primeira partição é tomada como banco de dados de teste e as demais nove partições são tomadas como um banco de dados de treino. O banco de dados de treino é usado para derivar o modelo, e este modelo é aplicado sobre o banco de dados de teste, gerando predições. Em seguida, toma-se a segunda partição como banco de dados de teste e as restantes como banco de dados de treino, e o processo se repete. No fim das 10 iterações, todos os indivíduos participaram do banco de dados de teste e do banco de dados de treino, e pode-se avaliar as predições resultantes deste processo, em função das combinações realizadas. A Figura 8 ilustra o método quando $k = 4$.

Figura 8 - Diagrama de *4-fold cross-validation*. O diagrama mostra quatro iterações, onde sucessivamente cada 1/4 das observações é usado como banco de dados de teste e o restante 3/4 é usado como banco de dados de treino.



Fonte: produzido pelo autor.

Já no *bootstrap*, a partir do banco de dados completo com n observações, obtêm-se várias amostras também de tamanho n, retirando aleatoriamente observações do banco de dados inicial, com reposição (ou seja, quando uma observação é escolhida, ela é

imediatamente reposta ao conjunto). O resultado é que cada amostra pode possuir observações repetidas do banco de dados inicial, e sobre cada amostra executa-se a análise, e por fim avalia-se o comportamento médio ou mediano da análise aplicada repetidas vezes. Tal processo de reamostragem com reposição permite uma aleatorização dos resultados, baseada no banco de dados inicial. Pode-se, por exemplo, estimar os coeficientes de regressão, de forma não-paramétrica, através do *bootstrap*. Métodos não paramétricos são usados caso não se tenham informações sobre a distribuição da população de onde a amostra original foi obtida. Neste caso, tomamos os percentis de 5% e 95% dos valores observados em 1000 repetições do experimento, como intervalo de confiança não paramétrico de 90% de um coeficiente (HENDERSON, 2005; SAUERBREI; ROYSTON, 2007).

2.7.3 Penalização como estratégia de redução do otimismo

Um dos principais objetivos de um modelo (diagnóstico ou prognóstico) de predição clínica é que ele mostre medidas de desempenho similares quando aplicado sobre outra população (diferente daquela que foi usada para derivar o modelo). Diz-se que o modelo é otimista se ele se ajusta muito aos dados de treino (dados usados para derivá-lo) e se mostra fraco quando aplicado sobre dados externos de teste (MOONS et al., 2004). Neste caso, a acurácia preditiva do modelo é prejudicada. Isso é mais provável de ocorrer se o banco de dados de treino é pequeno, ou se um grande número de preditores candidatos é usado, em relação ao número de eventos da variável resposta binária (como no caso de doenças raras), uma vez que o estimador clássico de máxima verossimilhança se atém demais às tendências dos dados de treino. Se este é o caso, o risco do evento tende a ser subestimado em pacientes com baixo risco e tende a ser superestimado nos pacientes com alto risco, ao ponto de o modelo separar completamente os dois grupos de pacientes.

Frequentemente considera-se o número de Eventos Por Variável (EPV) útil em quantificar as informações disponíveis nos dados em relação à complexidade do modelo. Costuma-se considerar um mínimo EPV=10 como regra de ouro no contexto da Análise de Regressão Binária clássica, antes de proceder-se à seleção de variáveis. Isso implica uma quantidade substancial de observações necessárias para proceder à análise, apesar de que têm sido demonstrado que técnicas modernas de modelagem, como *Random Forest*, *Suport Vector Machine* e Redes Neurais, demandam ainda mais dados que a regressão baseada em máxima verossimilhança, com um EPV muito maior para alcançar resultados estáveis quanto à capacidade de discriminação (AUC) e otimismo (PLOEG; AUSTIN; STEYERBERG, 2014).

Têm-se demonstrado que em cenários de baixo EPV, os métodos de *shrinkage* são preferíveis aos métodos de regressão não penalizada (STEYERBERG et al., 2000; STEYERBERG; EIJKEMANS; HABBEMA, 2001). Um efeito benéfico de incluir-se um critério de penalização sobre as estimativas dos coeficientes é que o modelo torna-se menos otimista em cenários de baixo EPV, ao custo de inserir-se um viés nas estimativas dos parâmetros. Pavlou et al. (2016) revisaram vários métodos de *shrinkage* e aplicaram estes métodos sobre um banco de dados com um n grande, mas baixo EPV, onde o evento é a ocorrência de câncer de pênis. Neste estudo, as medidas de calibração, discriminação e o desvio (erro) preditivo médio foram comparados quando mensurados sobre um banco de dados externo (simulado) em dois cenários de baixo EPV: com três e com cinco eventos por variável. Neste estudo, os autores encontraram que as medidas de desempenho das regressões RIDGE, LASSO e LASSO Bayesiano foram superiores sobre a estimação convencional de máxima verossimilhança, e sobre outros métodos de penalização, assim como os desvios de predição foram menores.

3 OBJETIVOS

3.1 GERAL

Propor modelo de predição diagnóstica de PET/MAH para pacientes infectados com HTLV-1, a partir da análise multivariada de características clínicas e dos níveis de expressão gênica.

3.2 ESPECÍFICOS

- Acessar a importância de variáveis utilizadas no contexto clínico de diagnóstico da PET/MAH (variáveis de força muscular, espasticidade, marcha, equilíbrio, função da bexiga, etc.), bem como níveis de expressão gênica (de IFN-γ, IL-4 e IL-10), na determinação do estado do paciente HTLV-1 positivo, para isso utilizando-se da PCA mista e da Regressão Logística Penalizada;
- Derivar um modelo de predição clínica diagnóstica, que permita prever o estado do paciente em com ou sem PET/MAH, a partir das variáveis preditoras mais importantes encontradas no passo anterior, para isso utilizando-se da Regressão Logística Penalizada;
- Avaliar o desempenho do modelo de predição clínica obtido, com base em medidas de validação interna mensuradas no contexto da validação cruzada (*10-fold cross-validation*).

4 MATERIAL E MÉTODOS

4.1 BANCO DE DADOS

Para a viabilização desta pesquisa, nos foi gentilmente cedido o banco de dados utilizado na tese de doutoramento de George Alberto da Silva Dias (DIAS, 2014; DIAS et al., 2016a). O banco de dados é composto de 63 pacientes HTLV-1 positivos, maiores de 18 anos, os quais foram selecionados por conveniência dentre pacientes apresentados ao acompanhamento no Laboratório de Clínica e Epidemiologia de Doenças Endêmicas, do Núcleo de Medicina Tropical da Universidade Federal do Pará, no período de agosto de 2010 a agosto de 2014.

Este estudo herda as características do trabalho original, possuindo natureza observacional e analítica, com delineamento transversal. O diagnóstico da infecção por HTLV-1 ocorreu primeiro por detecção de anticorpos anti-HTLV via ELISA e depois por confirmação molecular e tipagem das amostras soropositivas via PCR. O critério de classificação foi o da OMS, onde os indivíduos são diagnosticados com ou sem PET/MAH. Dos 63 pacientes disponíveis, 23 eram PET/MAH e 40 classificados como sem a doença, e 49 eram do sexo feminino. A Tabela 1 nos dá uma visão geral das variáveis disponíveis.

Tabela 1 - Visão geral das variáveis disponíveis no banco de dados de pacientes com HTLV-1 atendidos no Laboratório de Clínica e Epidemiologia das Doenças Endêmicas, da UFPA, em Belém/PA, no período de agosto de 2010 a agosto de 2014.

Grupo de Variáveis	Variáveis no Grupo	Natureza	Descrição	Referências
Condição	Condição.	Dicotômica	O paciente foi diagnosticado com PET/MAH ou não.	Castro-Costa et al. (2006)
Força	Força proximal/ distal do membro inferior direito/ esquerdo.	Ordinal	Escore inteiro de força variando de 0 (sem contração) a 5 (força normal).	Paternostro-Sluga et al. (2008)
Tônus muscular	Tônus dos músculos adutor direito/ esquerdo, quadríceps femoral direito/esquerdo e tríceps sural direito/ esquerdo.	Ordinal	Escore variando de 0 (sem aumento do tônus) a 4 (tônus rígido em flexão ou extensão).	Naghdi et al. (2008)
Equilíbrio e mobilidade	Escore codificado de equilíbrio e	Ordinal	Escore inteiro variando de 0 (não aplicável ao	Tinetti (1986)

Grupo de Variáveis	Variáveis no Grupo	Natureza	Descrição	Referências
	mobilidade de Tinetti.		paciente) a 3 (alto risco de queda).	
Auxílio na marcha	Grau de auxílio na marcha.	Ordinal	Escore inteiro variando de 0 (deambula sozinho) a 5 (necessita de cadeira de rodas).	-
Urodinâmica	Função da bexiga.	Ordinal	Escore de função da bexiga variando de 0 a 4.	-
Expressão gênica	Níveis de expressão gênica das citocinas IFN-γ, IL-4 e IL-10.	Numérica	Mensuradas pela quantificação da expressão de RNA de células linfomononucleares periféricas. Os níveis de expressão são relativos aos genes constitutivos GAPDH e β-actina.	-

Fonte: produzido pelo autor.

Como se pode observar, a maioria das variáveis é medida na escala ordinal. O escore de força assume uma das categorias: 0 (ausência de contração), 1 (contração muscular visível ou palpável sem movimentação), 2 (movimento ativo com eliminação da gravidade), 3 (movimento ativo contra a gravidade), 4 (movimento ativo contra a gravidade e resistência) e 5 (força normal). O escore de tônus muscular assume uma das categorias: 0 (nenhum aumento do tônus muscular), 1 (leve aumento do tônus), 1+ (moderado aumento do tônus), 2 (aumento do tônus durante a maior parte do movimento), 3 (aumento considerável do tônus com dificuldade de movimentação passiva) e 4 (apresenta rigidez na flexão ou extensão). O codificado escore de equilíbrio e mobilidade assume uma das categorias: 0 (não aplicável ao paciente), 1 (baixo risco de queda, ou escore de Tinetti de 25 a 28 pontos), 2 (risco médio de queda, ou escore de Tinetti de 19 a 24 pontos), e 3 (alto risco de queda, ou escore de Tinetti abaixo de 19 pontos). O escore grau de auxílio na marcha assume uma das categorias: 0 (deambula sozinho), 1 (necessita da ajuda de terceiros), 2 (necessita de bengala ou muleta unilateral), 3 (necessita de muleta bilateral), 4 (necessita do auxílio de andador), 5 (depende de cadeira de rodas).

A escala de força muscular é chamada de escala do Conselho de Pesquisa Médica (*Medical Research Council* ou escala MRC), e se baseia nos estudos de Paternostro-Sluga et al. (2008). A escala de tônus é uma modificação da escala de Ashworth, segundo Bohannon e

Smith (1987). O aumento de tônus muscular é frequentemente chamado de espasticidade. O escore de equilíbrio e mobilidade de Tinetti avalia o equilíbrio e as anormalidades da marcha, segundo Tinetti (1986).

A expressão gênica foi mensurada pela quantificação de RNA para as citocinas IFN-γ, IL-4 e IL-10 nas células mononucleares periféricas, por meio de PCR em tempo real (RT-PCR). O nível de expressão das citocinas foi relativa aos genes constitutivos GAPDH e β-actina: determinou-se o *cycle threshold* (CT) para cada um dos genes de interesse, e este foi normalizado em relação ao CT dos genes constitutivos, resultando em um valor $\Delta CT = CT_{gene} - CT_{constitutivo}$. O valor de expressão gênica considerado no trabalho original foi $2^{-\Delta CT}$. Aqui, foi necessário realizar a transformação logarítmica (FENG et al., 2014) dos níveis de expressão gênica baseados em $2^{-\Delta CT}$, para que fosse atingida a normalidade univariada e para aliviar-se *outliers*.

Por se tratar de dados secundários, dispensa-se a necessidade de pré-aprovação no Comitê de Ética em Pesquisa Envolvendo Seres Humanos local, nos termos da Resolução 466/2012 do Conselho Nacional de Saúde brasileiro.

4.2 ANÁLISE ESTATÍSTICA

O *software* R (R Core Team, 2016) foi usado para as análises estatísticas. Esta é uma ferramenta de análise estatística e um ambiente computacional, disponibilizada publicamente e mantida por um time de contribuidores de vários países, sob a filosofia de *software Open Source* (MORIN; URBAN; SLIZ, 2012). Dada a natureza mista do banco de dados (composto de variáveis de várias naturezas - ordinal, dicotômica e numérica), foi necessário utilizar métodos estatísticos que considerem adequadamente estas características. Com este objetivo, outros pacotes do R, além dos disponibilizados na instalação padrão, foram necessários.

A primeira etapa da pesquisa aplicou métodos de descoberta de padrões (PCA e Análise de Conglomerados Hierárquica), ou seja, métodos não supervisionados (pois não consideram uma variável resposta), para comparar esses padrões com a classificação realizada pelos profissionais clínicos, que usaram os critérios qualitativos já estabelecidos pela OMS, os quais classificam os pacientes em dois estados possíveis: com ou sem PET/MAH. A segunda etapa objetivou desenvolver um modelo de predição clínica diagnóstica, de forma a predizer uma probabilidade de o paciente ser diagnosticado com PET/MAH, conhecidas as preditoras mais importantes.

4.2.1 Análise não-supervisionada

O pacote *ade4* (DRAY; DUFOUR et al., 2007; CHESSEL; DUFOUR; THIOULOUSE, 2004) foi usado para realizar a PCA mista, por meio da função *dudi.mix*. Todas as variáveis do banco de dados (Tabela 1), exceto "Condição", foram incluídas numa primeira PCA. Aquelas preditoras que possuíam um coeficiente de correlação maior que 0,6 com qualquer uma das três primeiras componentes principais foram selecionadas para uma segunda PCA. Ou seja, a primeira PCA objetivou redução de itens. Considerando o resultado da segunda PCA (realizada sobre um banco de dados reduzido), os escores dos indivíduos nas três primeiras componentes foram usados como variáveis preditoras na Análise de Conglomerados Hierárquica, para estudar a classificação dos pacientes com base neste outro método de análise não supervisionada. Assim, considerou-se na análise de conglomerados a maior parte da informação do banco de dados reduzido, diminuindo o ruído.

4.2.2 Análise supervisionada

Para a segunda parte da pesquisa, o pacote *ordPens* (GERTHEISS, 2015) foi usado na análise de regressão logística penalizada, para seleção das variáveis preditoras mais associadas à discriminação dos pacientes com/sem PET/MAH e derivação do modelo de predição clínica diagnóstica. A função *ordSelect* realiza a seleção de preditores aplicando uma derivação de *Group*-LASSO, e suaviza os coeficientes de um mesmo preditor ordinal pela penalização das diferenças entre coeficientes adjacentes, como já discutido (GERTHEISS et al., 2011). A função *ordSmooth* apenas penaliza os coeficientes das variáveis preditoras, aplicando esta mesma suavização dos coeficientes quando a preditora é ordinal (GERTHEISS; TUTZ, 2009), mas sem selecionar as variáveis independentes.

Para encontrar o valor ideal do parâmetro de penalização λ (lambda), foi empregado o método de validação cruzada (*10-fold cross-validation*) (TUTZ; GERTHEISS, 2014): dividiu-se o banco de dados em 10 partes aproximadamente iguais, a primeira parte foi considerada o banco de dados de teste e as demais formaram o banco de dados de treino. A partir do banco de treino foi obtido um modelo (ou por meio de *ordSelect* ou por meio de *ordSmooth*) e aplicou-se esse modelo sobre o banco de teste, resultando em valores observados da variável resposta de teste e em probabilidades associadas à variável resposta de teste (probabilidades previstas pelo modelo de treino, variando de 0 até 1). Este mesmo

processo se repetiu por 10 vezes, até que todas as observações tivessem composto o banco de dados de teste em algum momento. Sobre todos estes valores finais observados da variável resposta, e sobre todas as respectivas probabilidades previstas pelos modelos de treino, aplicou-se a Equação (12) para obter o erro geral de predição (*deviance*), denominada aqui de desvio da validação cruzada.

Esse mesmo processo descrito acima foi repetido várias vezes, com diferentes valores de penalização λ. Escolheu-se como valor λ ideal aquele que resulta no menor desvio da validação cruzada. Conforme notado por Roberts e Nowak (2014), no entanto, pode existir uma instabilidade associada à atribuição das observações a cada partição da validação cruzada. Dessa forma, se houver diferentes atribuições de partições à amostra, os valores definidos para λ podem diferir. Os autores sugeriram executar a validação cruzada várias vezes, por exemplo 50 vezes, obtendo 50 valores de λ e considerar o valor que corresponda a um determinado percentil, por exemplo 50% ou 95%. Aqui, se denomina a abordagem de validação cruzada repetida, e considera-se o percentil de 50% como um valor mediano de λ e o percentil de 95% como um valor mais extremo de λ, resultando em um modelo com menor sobre-ajuste (Roberts e Nowak sugerem que 95% seja um valor apropriado na maioria dos cenários).

Outros métodos de escolha do valor ideal de λ existem. Porém, o método baseado em validação cruzada tem sido usado com mais frequência, devido ao fato de avaliar o desvio obtido sobre dados não usados na derivação do modelo (banco de dados de teste), apesar de que todas as observações foram obtidas de um mesmo contexto original.

Para a seleção de variáveis usando *ordSelect*, primeiro foi escolhido um valor λ ideal como descrito nos parágrafos anteriores, em seguida foram gerados intervalos de confiança não-paramétricos para os coeficientes, por meio de *bootstrap* (TUTZ; GERTHEISS, 2014): da amostra inicial de tamanho n, retiraram-se aleatoriamente 1000 amostras com reposição, cada uma também de tamanho n. Sobre cada uma das 1000 amostras foi empregada *ordSelect*, gerando 1000 estimativas de cada um dos coeficientes. A partir dessas estimativas foram gerados os intervalos de confiança de 90% com base nos percentis de 5% e 95%. Se todos os intervalos de confiança dos coeficientes de uma variável incluíram convincentemente o zero, então a respectiva variável foi excluída. Se a variável é categórica (ordinal ou não ordinal), ela foi excluída apenas se todos os seus coeficientes *dummy* possuíam intervalos de confiança que incluíam convincentemente o zero. Tanto o *bootstrap* quanto os intervalos de confiança de percentis foram gerados pelo pacote *boot* (CANTY; RIPLEY, 2016).

Para a obtenção do modelo final penalizado, usando *ordSmooth*, foram empregados os valores de λ estimados por validação cruzada e por validação cruzada repetida. As variáveis preditoras consideradas foram as obtidas no processo de seleção acima descrito. Para fins de interpretação dos modelos penalizados, foram gerados intervalos de confiança não paramétricos de 90% de percentis, nos mesmos moldes da seleção de variáveis (por *bootstrap*).

Procedeu-se, então, à validação interna dos modelos finais. As medidas de desempenho foram estimadas por validação cruzada de 10 vezes, para cada modelo: acurácia preditiva, sensibilidade, especificidade e Área sob a Curva ROC (AUC), a qual reflete a capacidade de discriminação do modelo. Foram comparados os desvios associados a cada equação, bem como a sua calibração (pela diferença entre médias das probabilidades previstas para cada classe da variável resposta binária: com/sem PET/MAH). Por fim, simularam-se todos os cenários em que as variáveis preditoras ordinais poderiam ser empregadas na prática, pela permutação dos níveis de cada variável, em cada um dos modelos finais.

5 RESULTADOS

5.1 ANÁLISE DE COMPONENTES PRINCIPAIS MISTA

A PCA mista foi obtida usando a função *dudi.mix*, pacote *ade4*, aplicada sobre 63 pacientes e 17 variáveis preditoras, com uma relação observações/variáveis de 3,7. Foram geradas 30 componentes principais. A Tabela 2 exibe os autovalores e as proporções de variância explicada pelas 14 primeiras componentes principais (que representam 95% da variação do banco de dados), quando todas as variáveis preditoras são consideradas (Tabela 1, exceto a variável resposta "Condição"). Observa-se que as duas primeiras componentes principais representam quase metade da variação dos dados (48%) e as seis primeiras componentes representam 75% da variação.

Tabela 2 - Autovalores e proporção de variância explicada pelas componentes principais do banco de dados completo.

Eixo	Inércia	Inércia cumulativa	Proporção	Proporção cumulativa
1	9,56	9,56	0,32	0,32
2	4,62	14,18	0,16	0,48
3	3,27	17,45	0,11	0,59
4	1,94	19,39	0,07	0,66
5	1,53	20,92	0,05	0,71
6	1,28	22,20	0,04	0,75
7	1,13	23,33	0,04	0,79
8	0,95	24,28	0,03	0,82
9	0,87	25,15	0,03	0,85
10	0,76	25,91	0,03	0,88
11	0,69	26,61	0,02	0,90
12	0,52	27,12	0,02	0,92
13	0,50	27,62	0,02	0,93
14	0,43	28,05	0,01	0,95

Como a variação total do banco de dados encontra-se mais ou menos dispersa nas componentes, ou seja, a grande quantidade de variáveis preditoras em relação ao número de observações aumentou o ruído da análise, executou-se uma nova PCA mista, agora apenas com as variáveis mais representativas das três primeiras componentes principais da análise inicial: selecionaram-se aquelas variáveis com coeficiente de correlação maior que 0,6 em

qualquer uma das três primeiras componentes. A Tabela 3 exibe os coeficientes de correlação entre cada variável preditora e as três primeiras componentes principais. Em negrito são exibidas as correlações maiores que 0,6. Observa-se que sete variáveis são removidas por esse critério: sexo, força distal dos membros inferiores (direito e esquerdo), escore de equilíbrio de Tinetti, e as três variáveis de expressão gênica. Dessa forma a relação observações/variáveis passa a ser 6,3.

Tabela 3 - Coeficientes de correlação entre todas as variáveis preditoras do banco de dados completo e as três primeiras componente principais da PCA mista.

Variável	CP1	CP2	CP3
Sexo	0,03	0,01	0,03
Força Proximal mmii D	**0,80**	0,05	**0,67**
Força Proximal mmii E	**0,77**	0,09	**0,64**
Força Distal mmii D	0,58	0,32	0,26
Força Distal mmii E	0,52	0,5	0,29
Bexiga	**0,61**	0,31	0,01
Marcha	**0,62**	0,11	0,26
Escore de Tinetti	0,56	0,08	0,11
Adutor D do Quadril	**0,79**	**0,63**	0,04
Adutor E do Quadril	**0,81**	**0,7**	0,11
Quadríceps Femoral D	**0,82**	**0,72**	0,08
Quadríceps Femoral E	**0,82**	0,32	0,23
Tríceps Sural D	**0,83**	0,38	0,18
Tríceps Sural E	**0,90**	0,10	0,26
IFN-γ	0,02	0,02	0,06
IL-4	0,07	0,21	0
IL-10	0,01	0,05	0,04

Mmii: membros inferiores. D: direito. E: esquerdo. CP: Componente Principal.

A Tabela 4 exibe os autovalores e as proporções de variação explicada pelas 9 primeiras componentes principais obtidas a partir do banco de dados reduzido, as quais representam 95% da variação total. As quatro primeiras componentes principais explicam 79% da variância do banco de dados. Verifica-se que houve diminuição da dispersão pela redução do número de variáveis preditoras.

Tabela 4 - Autovalores e proporção de variância explicada pelas componentes principais do banco de dados reduzido.

Eixo	Inércia	Inércia cumulativa	Proporção	Proporção cumulativa
1	7,91	7,91	0,40	0,40
2	3,89	11,80	0,20	0,60
3	2,39	14,19	0,12	0,72
4	1,43	15,62	0,07	0,79
5	0,86	16,48	0,04	0,84
6	0,78	17,26	0,04	0,88
7	0,56	17,82	0,03	0,91
8	0,46	18,29	0,02	0,93
9	0,42	18,71	0,02	0,95

A Tabela 5 exibe os coeficientes de correlação das 10 variáveis preditoras selecionadas neste segundo contexto, sobre as três primeiras componentes principais. Verifica-se que a componente mais importante, que corresponde a 40% da variância total, possui correlação mais forte com as variáveis relacionadas à força muscular e as relacionadas ao tônus muscular, em especial o tríceps esquerdo. Verifica-se também que a segunda componente representa especialmente as variáveis relacionadas ao tônus muscular, enquanto que a terceira componente representa principalmente as variáveis relacionadas à força muscular proximal.

Tabela 5 - Coeficientes de correlação entre as variáveis preditoras que permaneceram no banco de dados reduzido e as três primeiras componente principais da PCA mista.

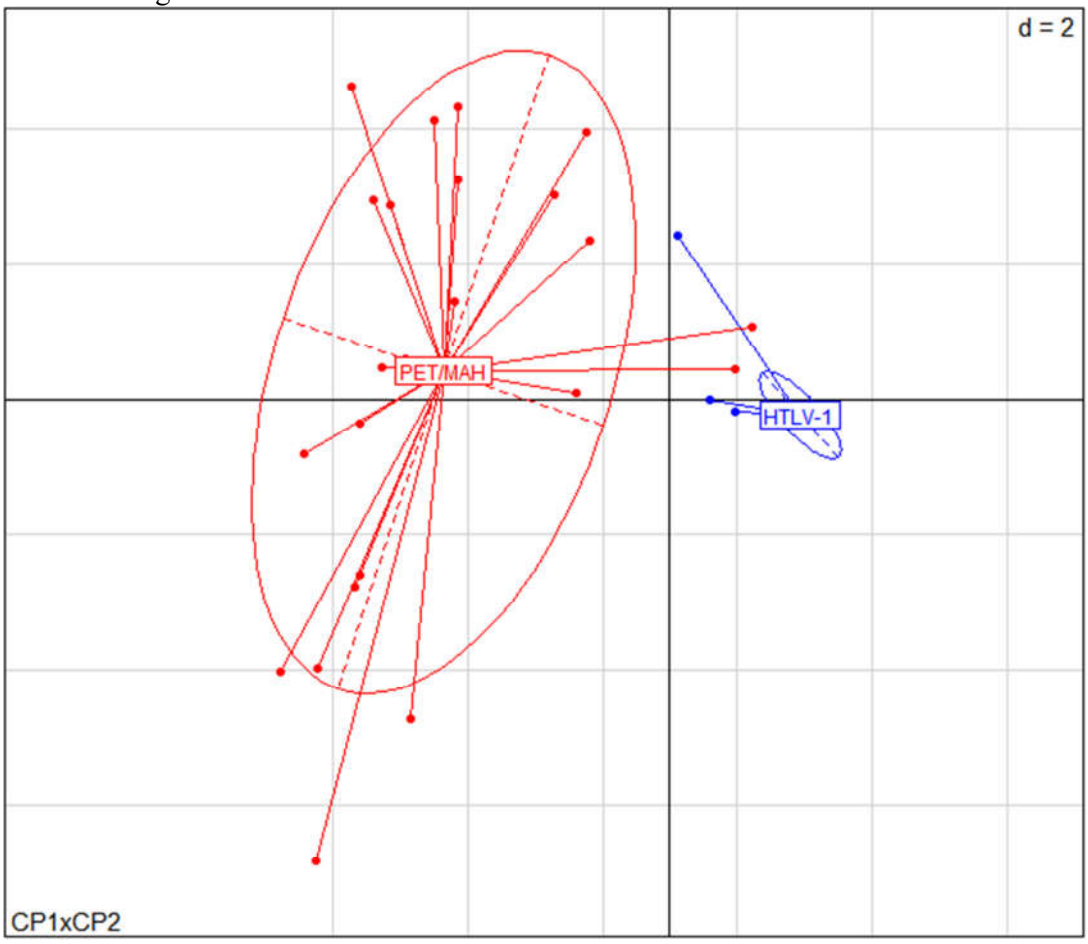
Variável	CP1	CP2	CP3
Força Proximal mmii D	0,80	0,01	0,81
Força Proximal mmii E	0,75	0,01	0,88
Bexiga	0,58	0,17	0,05
Marcha	0,64	0,02	0,32
Adutor D do Quadril	0,80	0,68	0,03
Adutor E do Quadril	0,82	0,82	0,00
Quadríceps Femoral D	0,89	0,79	0,06
Quadríceps Femoral E	0,88	0,55	0,04
Tríceps Sural D	0,83	0,56	0,04
Tríceps Sural E	0,92	0,28	0,15

Mmii: membros inferiores. D: direito. E: esquerdo. CP: Componente Principal.

Tomando as coordenadas dos indivíduos em cada uma das componentes principais, podemos estudar como as observações se comportam em relação às componentes. A Figura 9

exibe a disposição dos indivíduos nas duas primeiras componentes principais. O eixo x representa a primeira componente e o eixo y, a segunda. Cada ponto do gráfico representa um indivíduo, e estes foram identificados de acordo com sua condição (com ou sem PET/MAH, variável que não entrou na computação da PCA), por meio da função *s.class*. Além de identificar os indivíduos por grupos, essa função cria uma elipse por grupo, que permite observar o baricentro de cada conjunto no gráfico. Percebe-se que a primeira componente, que representa 40% da variabilidade total, representa um gradiente de pacientes: indivíduos com escores maiores nesta componente são principalmente sem PET/MAH, enquanto que indivíduos com escores menores são pacientes com PET/MAH. É possível observar também que os dois grupos se distinguem bem na primeira componente, e que o grupo sem PET/MAH é bem mais homogêneo, provavelmente porque possui menos variáveis preditoras alteradas. No entanto, percebem-se alguns indivíduos, próximo ao centro do gráfico, que são intermediários entre os dois grupos, ou seja, provavelmente possuem características tanto de um grupo quanto de outro.

Figura 9 - Coordenadas dos indivíduos nas duas primeiras componentes principais. CP: Componente Principal. PET/MAH: pacientes infectados pelo HTLV-1 e diagnosticados com a doença. HTLV-1: Pacientes infectados pelo HTLV-1 e não diagnosticados com PET/MAH.

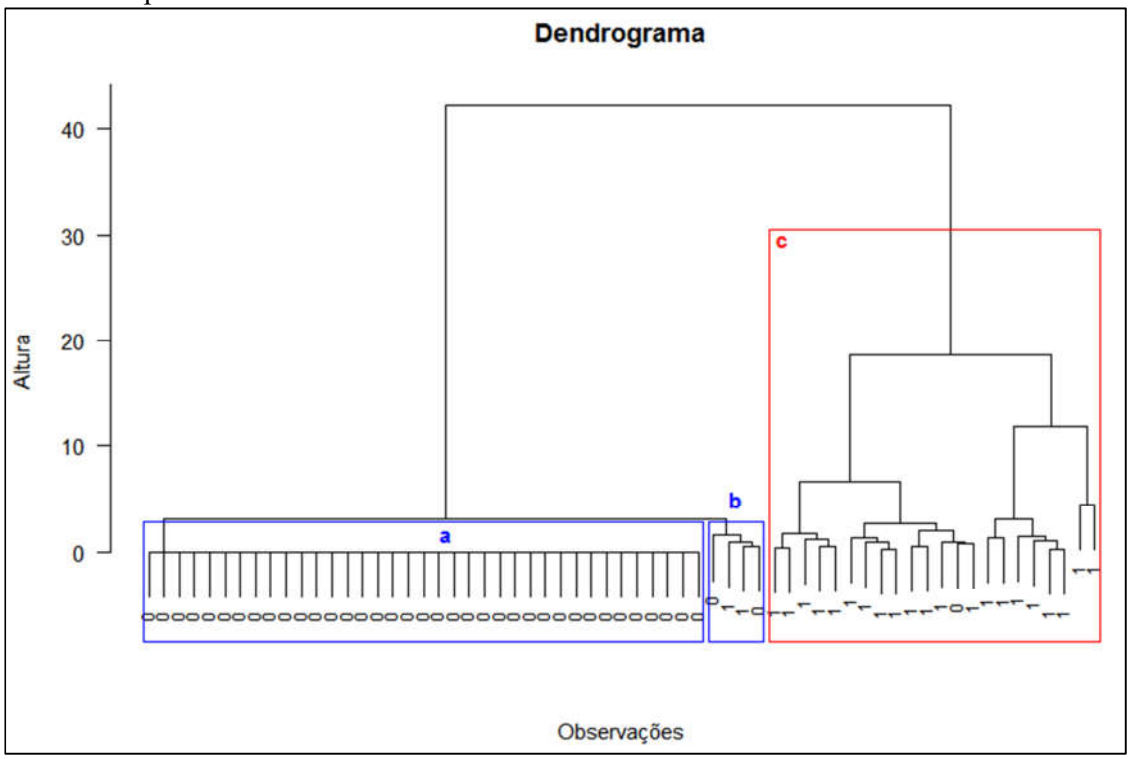


5.2 ANÁLISE DE CONGLOMERADOS HIERÁRQUICA

Para esmiuçar a disposição dos pacientes observada na Figura 9, foi realizada uma análise de conglomerados hierárquica com base nos resultados obtidos da PCA do banco de dados reduzido. Considerando os escores dos indivíduos em cada uma das três primeiras componentes principais, as quais representam grande parte (72%) da variação do banco de dados reduzido, obteve-se uma matriz de distâncias entre as observações de acordo com o perfil multivariado dos pacientes nessas três componentes. A partir dessa matriz de distâncias, foi derivado um dendrograma usando a função *hclust*, com o método de aglomeração *ward.D*. A Figura 10 mostra o dendrograma resultante. Observa-se que dois conjuntos de pacientes são bem definidos: o conjunto em azul forma os *clusters a* e *b* e o conjunto em vermelho forma o

cluster c. O *cluster c* é composto de indivíduos com PET/MAH (identificados como um) e apenas um indivíduo sem PET/MAH (identificado como zero), o *cluster a* é composto de indivíduos sem PET/MAH, e o *cluster b* possui quatro pacientes, sendo dois indivíduos com e dois indivíduos sem a condição. Nota-se também que os pacientes do *cluster b* estão mais próximos dos indivíduos sem PET/MAH do que dos indivíduos com.

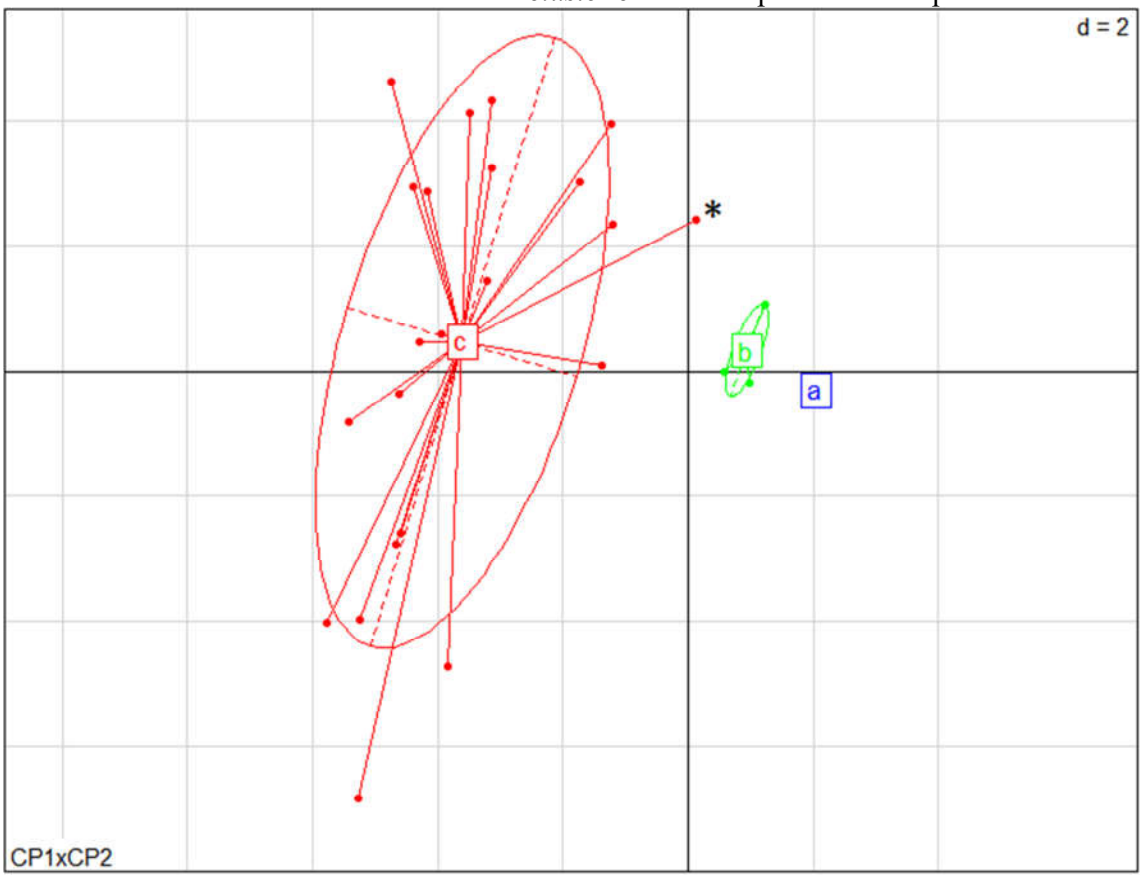
Figura 10 - Dendrograma obtido pelo método de aglomeração hierárquica de Ward. Os pacientes diagnosticados com PET/MAH são identificados como um e os pacientes sem PET/MAH são identificados como zero.



Para identificar estes três grupos de pacientes no gráfico de indivíduos da PCA reduzida, foi também usada a função *s.class*, como no passo anterior. A Figura 11 exibe as coordenadas dos indivíduos nas duas primeiras componentes da PCA derivada do banco de dados reduzido. Os pacientes estão identificados de acordo com os três *clusters* nos quais foram classificados pela análise de conglomerados hierárquica. Os baricentros dos grupos indicam esses *clusters*. Observa-se que os pacientes estão bem separados nos três grupos, e que os grupos *a* e *b* estão mais próximos e o grupo *c* mais afastado dos demais, similar ao demostrado no dendrograma. Observa-se também que o grupo *a* é bem mais homogêneo, onde os pacientes provavelmente não possuem variáveis clínicas alteradas. O paciente sem PET/MAH classificado no *cluster c* da Figura 10 encontra-se evidenciado com um asterisco.

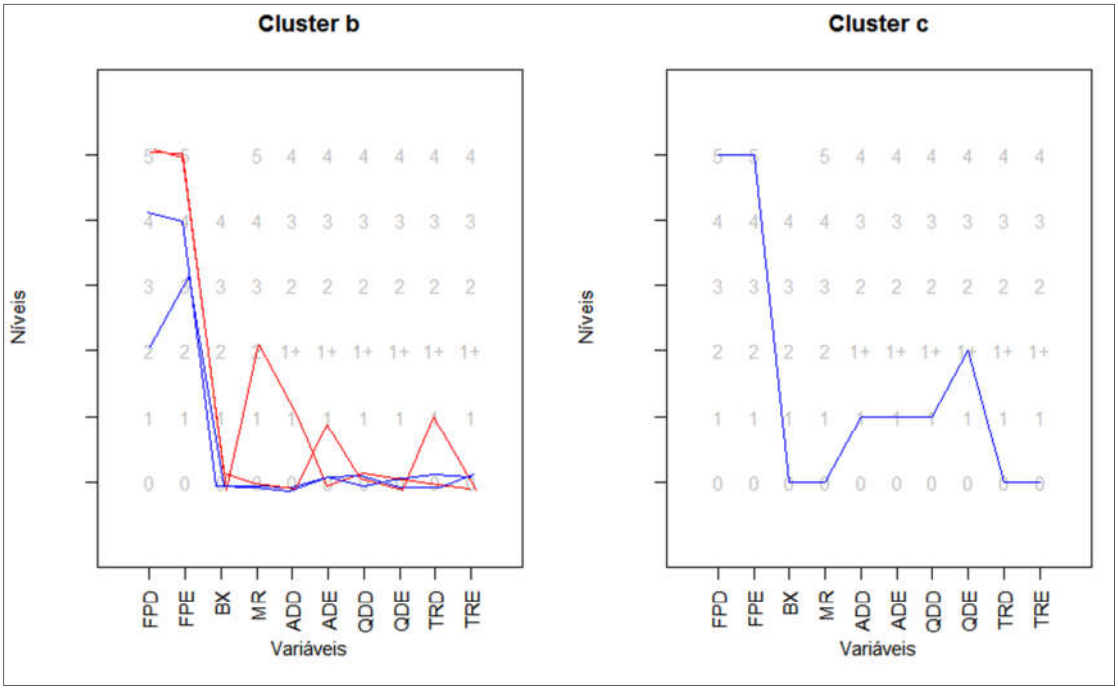
Percebe-se que ele realmente se encontra mais próximo dos pacientes com PET/MAH de acordo com o resultado da PCA.

Figura 11 - Coordenadas dos indivíduos nas duas primeiras componentes principais da PCA do banco de dados reduzido. Os indivíduos estão identificados de acordo com os *clusters* encontrados na Figura 10. O asterisco indica o paciente sem diagnóstico de PET/MAH e classificado no *cluster c*. CP: Componente Principal.



A Figura 12 exibe, à esquerda, o perfil das variáveis ordinais dos quatro pacientes classificados no *cluster b* e, à direita, o perfil do único paciente classificado pela Análise de Conglomerados junto ao *cluster c*, mas não diagnosticado com PET/MAH. À esquerda, observa-se que dois pacientes foram diagnosticados com PET/MAH (linhas em vermelho) e dois não (linhas em azul). Destes, os pacientes sem PET/MAH possuíam alteração de força, mas não alteração de tônus nos três músculos considerados (adutor do quadril, quadríceps femoral e tríceps sural), nem alteração da função da bexiga ou de marcha. Enquanto que os dois com PET/MAH possuíam escores normais de força (escore 5), um deles possuíam alteração de marcha e o outro possuíam alteração de tônus no músculo adutor esquerdo e no tríceps sural direito. À direita, observa-se que o paciente sem PET/MAH possuía alteração de tônus exceto no músculo tríceps sural.

Figura 12 - Perfil multivariado dos quatro pacientes classificados no *cluster b* (gráfico à esquerda) pela Análise de Conglomerados, e do único paciente sem diagnóstico de PET/MAH, mas classificado no *cluster c* (gráfico à direita). Cada linha representa um paciente. Linhas azuis: pacientes sem diagnóstico de PET/MAH. Linhas vermelhas: pacientes com diagnóstico de PET/MAH. Variáveis mostradas: força proximal do membro inferior direito (FPD) e esquerdo (FPE), função da bexiga (BX), grau de alteração da marcha (MR), tônus do músculo adutor direito (ADD) e do esquerdo (ADE) do quadril, tônus do quadríceps femoral direito (QDD) e do esquerdo (QDE), tônus do tríceps sural direito (TRD) e do esquerdo (TRE).



5.3 MODELO DE PREDIÇÃO CLÍNICA DIAGNÓSTICA

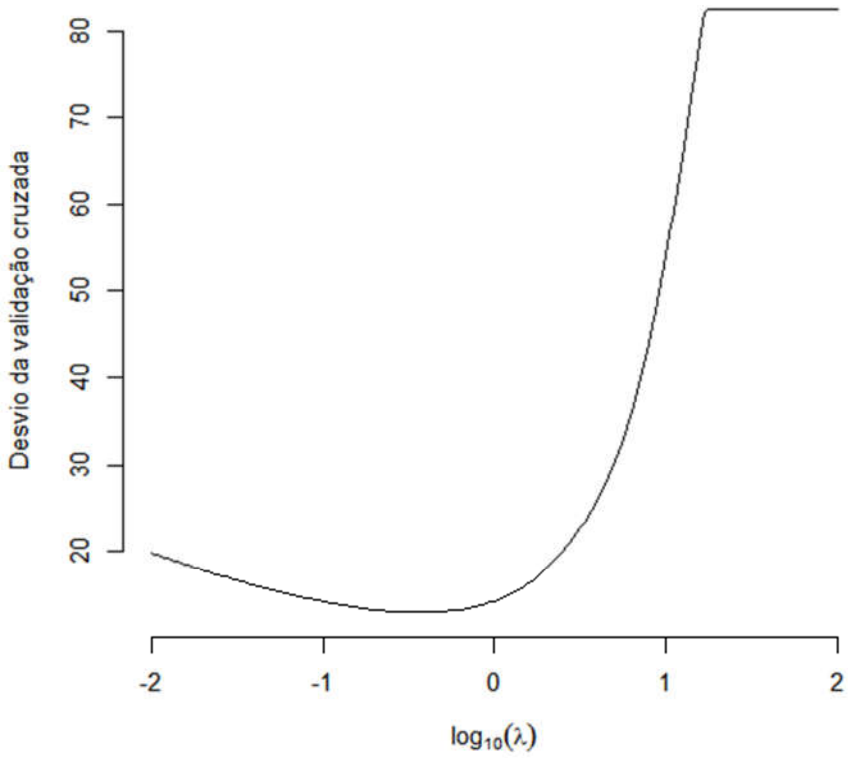
5.3.1 Seleção de variáveis preditoras por penalização de coeficientes

Geralmente é desejável obter um modelo parcimonioso, com menos variáveis preditoras e sem muita perda da acurácia preditiva. Usando a função *ordSelect* do pacote *ordPens*, foi realizada a seleção de variáveis pela contração de coeficientes por meio de penalização *Group*-LASSO: sobre os preditores ordinais foi penalizada a diferença entre coeficientes *dummy* adjacentes, e sobre os preditores categóricos não ordinais e sobre as variáveis numéricas foi aplicada a penalização padrão, segundo Gertheiss et al. (2011). A variável resposta foi a condição do paciente (com ou sem PET/MAH) conforme diagnóstico definido no estudo original, logo, a técnica usada foi uma modificação da regressão logística

binária. O paciente indicado com um asterisco na Figura 11 foi considerado uma observação atípica e excluído da modelagem.

Para escolha do parâmetro de penalização λ, o qual representa a intensidade da contração a ser aplicada sobre os coeficientes estimados, foi usado o método de *10-fold cross-validation* (TUTZ; GERTHEISS, 2014), já comentado. A Figura 13 exibe o desvio de validação cruzada como uma função de $\log_{10} \lambda$. Os valores de λ estão em log porque assim permite-se melhor representação de uma ampla faixa de valores de λ. Verificou-se que o valor de λ onde o desvio de validação cruzada é minimizado é $0{,}35$ (ou $10^{-0{,}46}$). O método validação cruzada repetida resultou nos valores de $\lambda = 0,32$ para o percentil de 50% e de $\lambda = 3,37$ para o percentil de 95%. A penalização de 3,37 foi mais agressiva, contraindo todos os coeficientes para zero (dados não mostrados), por isso, considerou-se como lambda ideal o obtido por validação cruzada simples, cujo valor foi bem próximo do percentil de 50% da validação cruzada repetida.

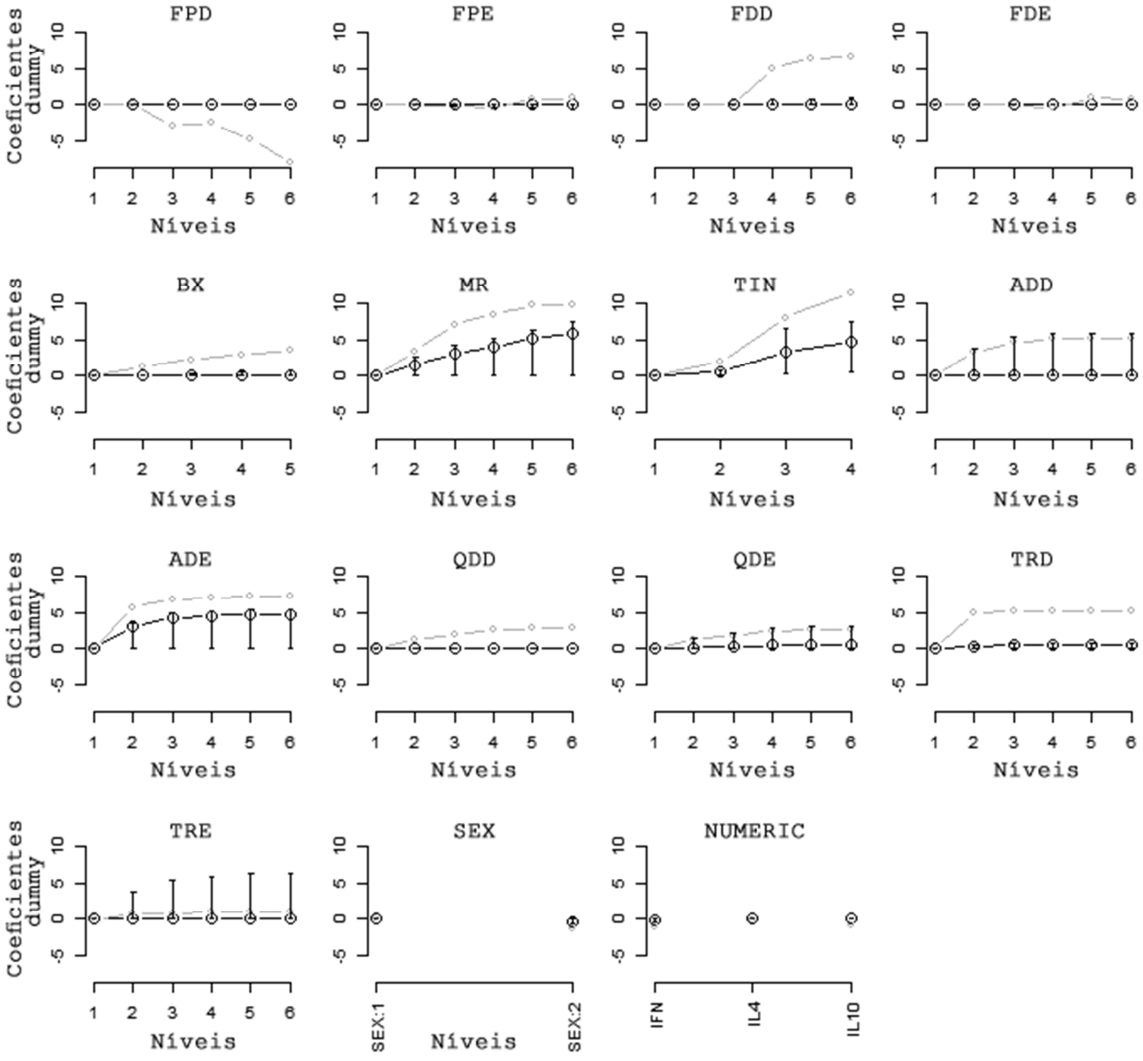
Figura 13 - Desvio da validação cruzada como função de $\boldsymbol{\log_{10} \lambda}$, para seleção de variáveis.



Escolhido o valor ideal de lambda para a seleção das variáveis, procedeu-se à seleção propriamente dita. Foram feitas 1000 replicações do banco de dados original por meio de *bootstrap*. Sobre cada réplica um modelo foi obtido. Foram gerados intervalos de confiança (não paramétricos) de 90% para os coeficientes com base em percentis, pelo pacote *boot*, com

base nas 1000 réplicas. A Figura 14 exibe os coeficientes obtidos pela penalização (em preto), os coeficientes sem penalização (em cinza) e os intervalos de confiança 90% baseados em percentis (barras verticais). Observa-se que as variáveis cujos intervalos de confiança dos coeficientes não incluem (convincentemente) o zero são a marcha (MR), o escore codificado de equilíbrio e mobilidade de Tinetti (TIN), o escore de tônus dos músculos adutores direito (ADD) e esquerdo (ADE) e o escore do tríceps sural esquerdo (TRE). As demais variáveis ordinais, a variável sexo e as variáveis de expressão gênica foram então eliminadas do modelo preditivo.

Figura 14 - Coeficientes da regressão e respectivos intervalos de confiança de 90% com base em percentis. Em preto estão os coeficientes penalizados usando $\boldsymbol{\lambda = 0,35}$. Em cinza estão os coeficientes sem penalização com $\boldsymbol{\lambda = 0}$ (exibidos para referência), e as barras verticais indicam o intervalo de confiança obtido por *bootstrap* com 1000 réplicas. Variáveis preditoras: força proximal (FPD) e força distal (FDD) do membro inferior direito e esquerdo (FPE, FDE), função da bexiga (BX), grau de alteração da marcha (MR), escala de Tinetti codificada (TIN), tônus do músculo adutor direito (ADD) e do esquerdo (ADE) do quadril, tônus do quadríceps femoral direito (QDD) e do esquerdo (QDE), tônus do tríceps sural direito (TRD) e do esquerdo (TRE), níveis de expressão gênica de IFN-γ, IL-4 e IL-10 (NUMERIC) e sexo do paciente (SEX), sendo 1 masculino e 2 feminino.

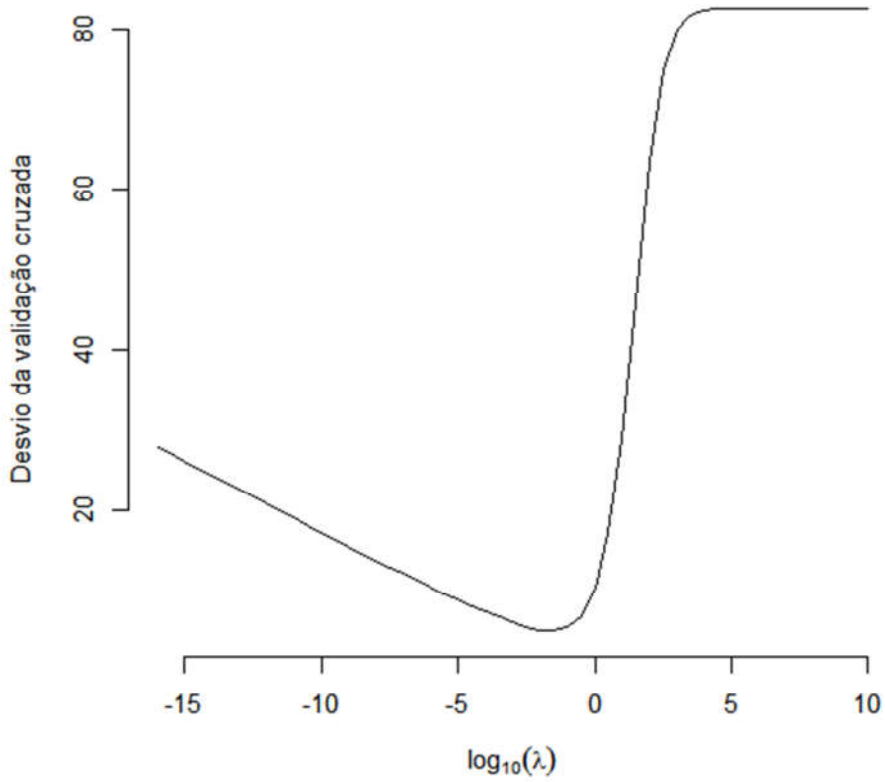


5.3.2 Penalização das variáveis preditoras selecionadas

Usando a função *ordSmooth* do pacote *ordPens*, foi realizada a penalização das variáveis preditoras ordinais escolhidas no passo anterior: sobre os preditores ordinais foi

penalizada a diferença entre coeficientes *dummy* adjacentes, segundo Gertheiss e Tutz (2009). A variável resposta foi a condição do paciente (com ou sem PET/MAH). Aqui também o paciente indicado com um asterisco na Figura 11 foi considerado uma observação atípica e eliminado. Para escolha do parâmetro de penalização λ, utilizou-se dois métodos: validação cruzada simples (*10-fold cross-validation*) (TUTZ; GERTHEISS, 2014) e validação cruzada repetida. A Figura 15 exibe o desvio da validação cruzada simples como uma função de $\log_{10}\lambda$. Neste caso, o valor de λ onde o desvio de validação cruzada é minimizado foi $0{,}032$ (ou $10^{-1{,}5}$). No caso da validação cruzada repetida, os valores encontrados foram $\lambda = 0{,}032$ para o percentil de 50% e $\lambda = 0,1$ para o percentil de 95%.

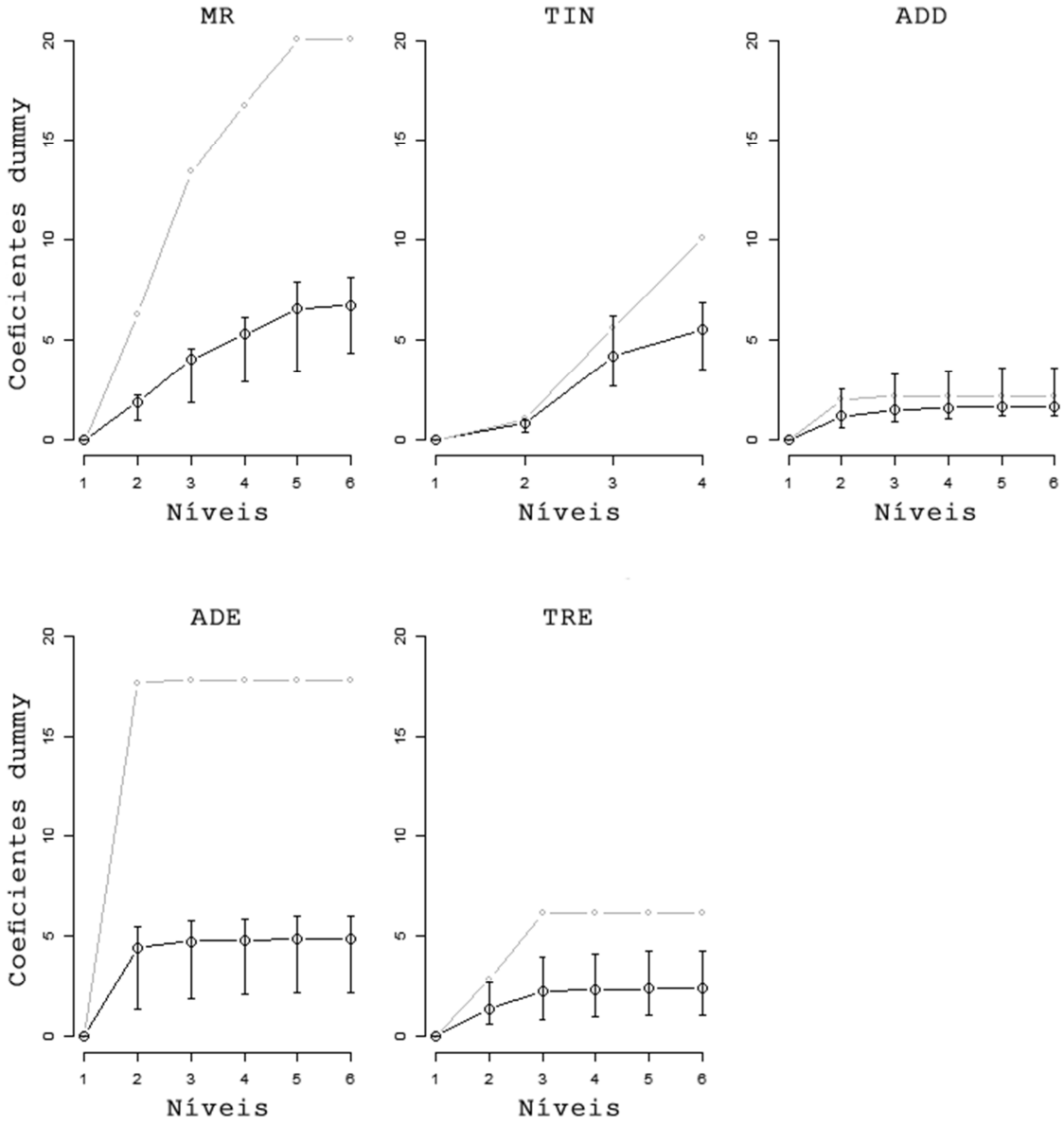
Figura 15 - Desvio da validação cruzada como função de $\log_{10}\lambda$, para suavização de coeficientes.



Escolhido o valor ideal de lambda, procedeu-se à obtenção do modelo final. Foram feitas 1000 replicações do banco de dados original por meio de *bootstrap*, como no passo anterior, gerando-se intervalos de confiança (não paramétricos) de 90% para os coeficientes com base em percentis. A Figura 16 exibe os coeficientes obtidos pela penalização $\lambda = 0{,}032$ (em preto), os coeficientes sem penalização (em cinza) e os intervalos de confiança de 90% baseados em percentis (barras verticais). Observa-se que a variável alteração da marcha possui maior importância sobre a determinação do estado do paciente, mesmo os coeficientes *dummy* tendo sido bastante contraídos pelo processo de penalização, como se pode observar

pelas diferenças entre as estimativas em preto e as estimativas em cinza. Similarmente, as variáveis tônus do músculo adutor esquerdo do quadril e a codificação da escala de equilíbrio de Tinetti foram importantes na determinação da variável resposta. De forma geral, observa-se que as cinco variáveis escolhidas possuem comportamento aproximadamente monotônico, em que quanto maior o escore, mais provável é que o paciente possua a condição (PET/MAH=1).

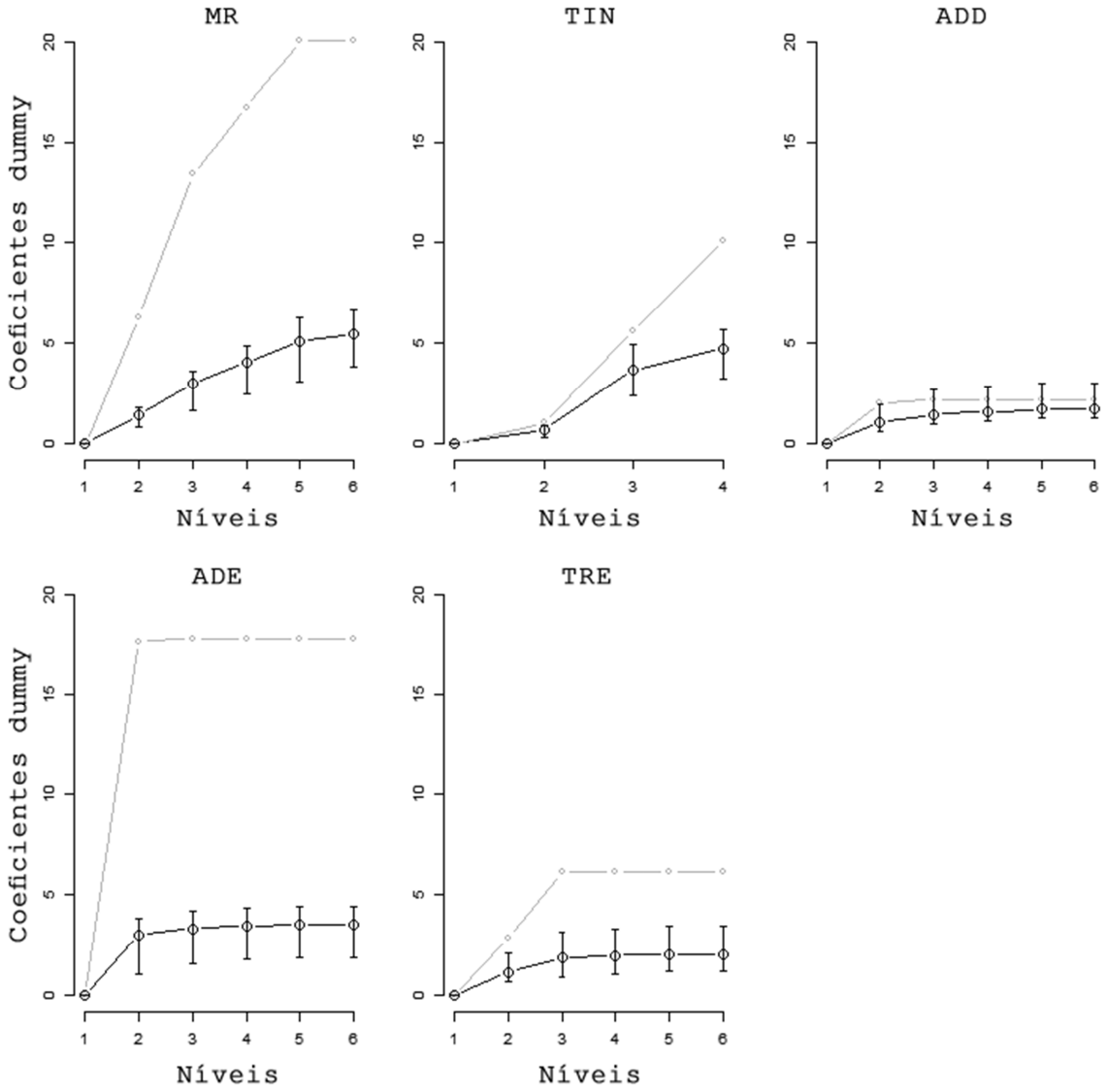
Figura 16 - Coeficientes da regressão e respectivos intervalos de confiança de 90% com base em percentis, obtidos pelo critério de penalização $\boldsymbol{\lambda = 0,032}$. Em preto estão os coeficientes penalizados. Em cinza estão os coeficientes sem penalização com $\boldsymbol{\lambda = 10^{-5}}$ (exibidos para referência), e as barras verticais indicam o intervalo de confiança obtido por *bootstrap* com 1000 réplicas. Variáveis preditoras: grau de alteração da marcha (MR), escala de Tinetti codificada (TIN), tônus do músculo adutor direito (ADD) e do esquerdo (ADE) do quadril, e tônus do tríceps sural esquerdo (TRE).



Já a Figura 17 exibe as estimativas dos coeficientes nas mesmas condições, mas quando o parâmetro de penalização $\lambda = 0{,}1$ é empregado. Percebe-se que os coeficientes são um pouco mais contraídos, e que a variabilidade das suas estimativas é diminuída (barras verticais que correspondem ao intervalo de confiança de percentis).

Os dois modelos finais e exemplos de aplicação estão disponibilizados no Apêndice A.

Figura 17 - Coeficientes da regressão e respectivos intervalos de confiança de 90% com base em percentis, obtidos pelo critério de penalização $\boldsymbol{\lambda = 0,1}$. Em preto estão os coeficientes penalizados. Em cinza estão os coeficientes sem penalização com $\boldsymbol{\lambda = 10^{-5}}$ (exibidos para referência), e as barras verticais indicam o intervalo de confiança obtido por *bootstrap* com 1000 réplicas. Variáveis preditoras: grau de alteração da marcha (MR), escala de Tinetti codificada (TIN), tônus do músculo adutor direito (ADD) e do esquerdo (ADE) do quadril, e tônus do tríceps sural esquerdo (TRE).

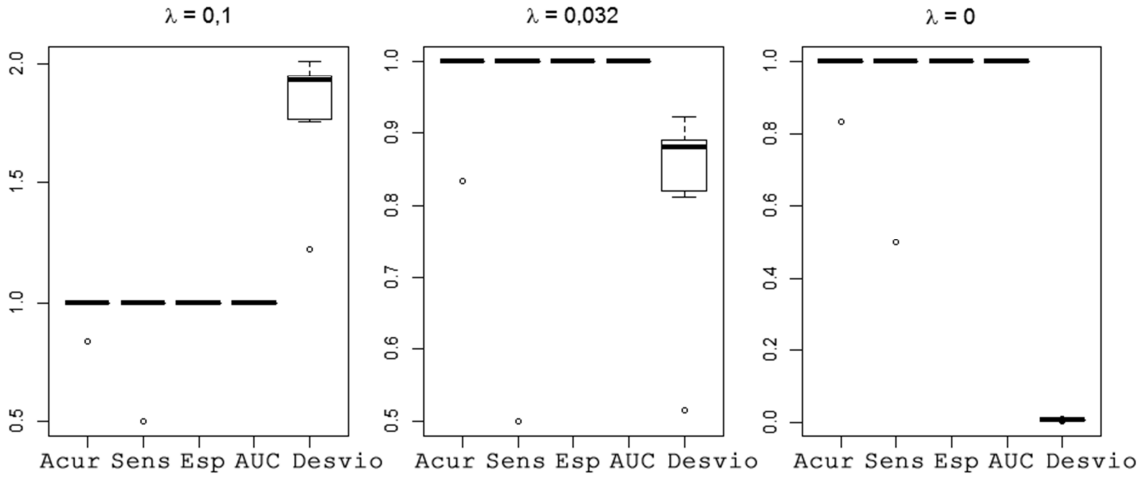


5.3.3 Avaliação dos modelos preditivos

Obtiveram-se dois modelos que diferiram quanto ao parâmetro de penalização empregado: $0{,}032$ ou $0{,}1$. A validação interna de cada modelo foi realizada por meio de validação cruzada (*10-fold cross-validation*): o banco de dados foi dividido em 10 partes aproximadamente iguais, e na primeira iteração considerou-se a décima parte como banco de dados de teste e o restante considerou-se como banco de dados de treino. Do banco de treino obteve-se o modelo usando a função *ordSmooth* e este modelo foi usado para prever as classes do banco de dados de teste. 10 iterações foram realizadas e, assim, foram obtidos 10 conjuntos de predições. Essas predições foram comparadas com as classes observadas e foram computadas as medidas de desempenho.

A Figura 18 mostra os resultados obtidos para a acurácia, sensibilidade, especificidade, área sob a curva ROC (AUC) e os desvios calculados pela Equação (12) quando $\lambda = 0{,}1$ é usado (à esquerda), quando $\lambda = 0{,}032$ (no centro) e quando $\lambda = 0$ é usado (à direita). Percebe-se que os desvios aumentam quanto maior a penalização empregada. Pelo fato do erro ser zero quando não se usa a penalização ($\lambda = 0$), depreende-se que neste cenário há sobre-ajuste. Porém, pode-se desconfiar de sobre-ajuste nos três casos devido ao ótimo desempenho das medidas utilizadas, quando aplicadas neste cenário de validação interna. No entanto, uma avaliação mais criteriosa seria possível em um contexto de validação externa, quando o modelo é avaliado sobre uma amostra independente de pacientes.

Figura 18 - Medidas de desempenho e erro obtidos por validação cruzada, quando $\boldsymbol{\lambda = 0{,}1}$ (à esquerda), $\boldsymbol{\lambda = 0{,}032}$ (no centro) ou quando $\boldsymbol{\lambda = 0}$ (à direita). Para facilitar a representação, o erro é plotado no mesmo gráfico, porém, este não é restringido no intervalo de 0 a 1. Acur: acurácia. Sens: sensibilidade. Esp: especificidade. AUC: *Area Under the ROC Curve*.



A calibração do modelo preditivo reflete a concordância entre a variável resposta binária observada e as previsões do modelo. A Figura 19 resume a calibração do modelo

quando não há penalização (primeira coluna, $\lambda = 0$) e quando $\lambda = 0{,}1$ ou $\lambda = 0{,}032$, se a calibração é obtida usando o próprio modelo ajustado (primeira coluna) ou se a calibração é obtida por meio de validação cruzada (segunda coluna). Aqui, a inclinação é definida como a diferença entre médias das probabilidades previstas para os dois níveis da variável resposta binária. As probabilidades assim entendidas como risco de o paciente possuir PET/MAH. Se a inclinação é um, o modelo prevê perfeita separação entre os dois níveis da variável resposta. Observa-se que no último gráfico à esquerda (modelo não penalizado) ocorre esta perfeita separação. Nos demais casos, as inclinações são altas, acima de 0,9, e similares entre si.

A Figura 20 mostra as probabilidades previstas pelo modelo ajustado, quando aplicado sobre os pacientes do banco de dados original. Como se esperaria no caso de sobre-ajuste do modelo, o risco associado a cada paciente tende a ser subestimado se o paciente possui baixo risco e tende a ser superestimado se o paciente possui alto risco de possuir PET/MAH, ao ponto de haver completa separação dos dois grupos de pacientes, como é o caso quando $\lambda = 0$ (Figura 20 à direita), o que não é desejável do ponto de vista da interpretabilidade dos resultados. Também se observa que há relaxamento dessa tendência conforme o valor de λ aumenta.

Como as variáveis preditoras que permaneceram no modelo são variáveis categóricas ordinais, com um número limitado de níveis, o número de combinações entre níveis dessas variáveis preditoras também é limitado. Logo, é possível predizer todos os cenários em que os modelos seriam usados na prática, simulando as probabilidades previstas ao permutar todos os níveis das variáveis preditoras usando os modelos penalizados de regressão logística. Para o problema abordado, há 5.184 possíveis combinações das cinco variáveis preditoras que permaneceram no modelo da Figura 16.

A Figura 21 exibe o resultado da simulação quando $\lambda = 0{,}1$ (à esquerda), $\lambda = 0{,}032$ (no centro) e $\lambda = 0$ (à direita). Percebem-se padrões similares nas três situações, enquanto na penalização mais forte a linha mediana dos gráficos em caixa se desloca um pouco mais ao centro do gráfico. Das 5.184 possibilidades, 91,3% preveem que o paciente é PET/MAH quando $\lambda = 0{,}1$, 90,5% preveem que o paciente é PET/MAH quanto $\lambda = 0{,}032$ e 90,2% preveem que o paciente é PET/MAH quando $\lambda = 0$.

No Apêndice B estão disponibilidados os códigos R utilizados nas análises.

Figura 19 - Calibração do modelo sem (primeira coluna) e com (segunda coluna) validação cruzada, com e sem penalização dos coeficientes do modelo preditivo. O eixo $\boldsymbol{x}$ indica a condição (0 sem PET/MAH e 1 com PET/MAH) e o eixo $\boldsymbol{y}$ indica a probabilidade prevista pelo modelo preditivo em cada cenário. CV: *10-fold cross-validation*.

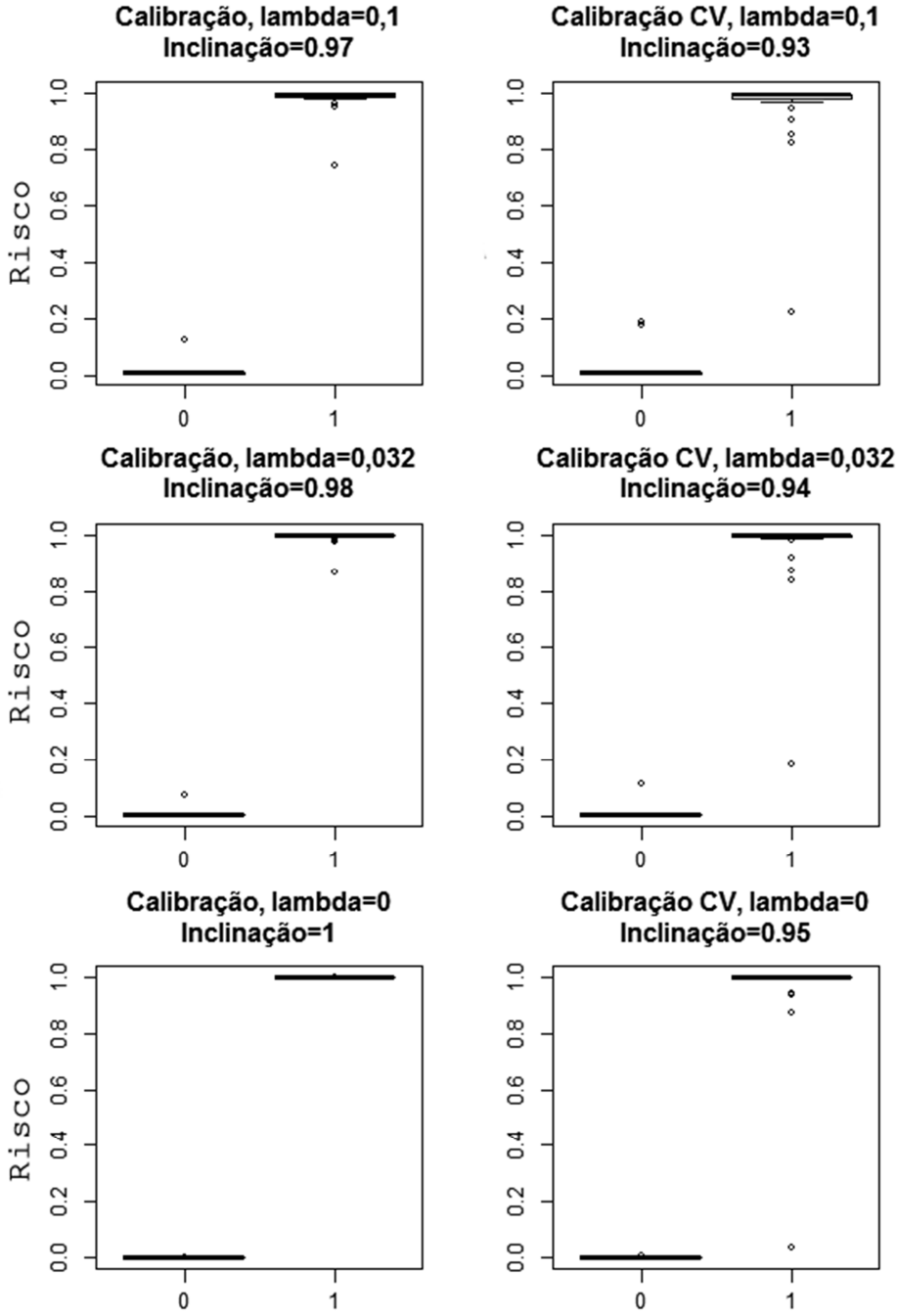


Figura 20 - Probabilidades associadas às observações do banco de dados original. O eixo $\boldsymbol{x}$ mostra os diferentes valores de lambda ($\mathbf{0,1}$, $\mathbf{0,032}$ e $\mathbf{0}$, respectivamente) e o eixo $\boldsymbol{y}$ indica a probabilidade prevista pelo modelo preditivo em cada cenário.

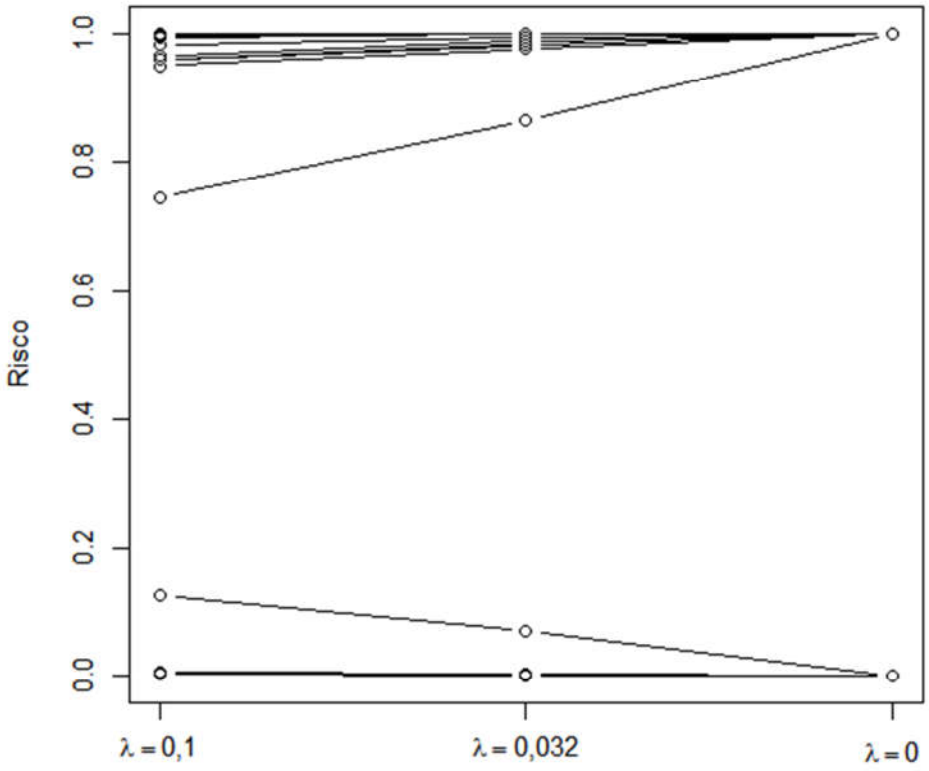
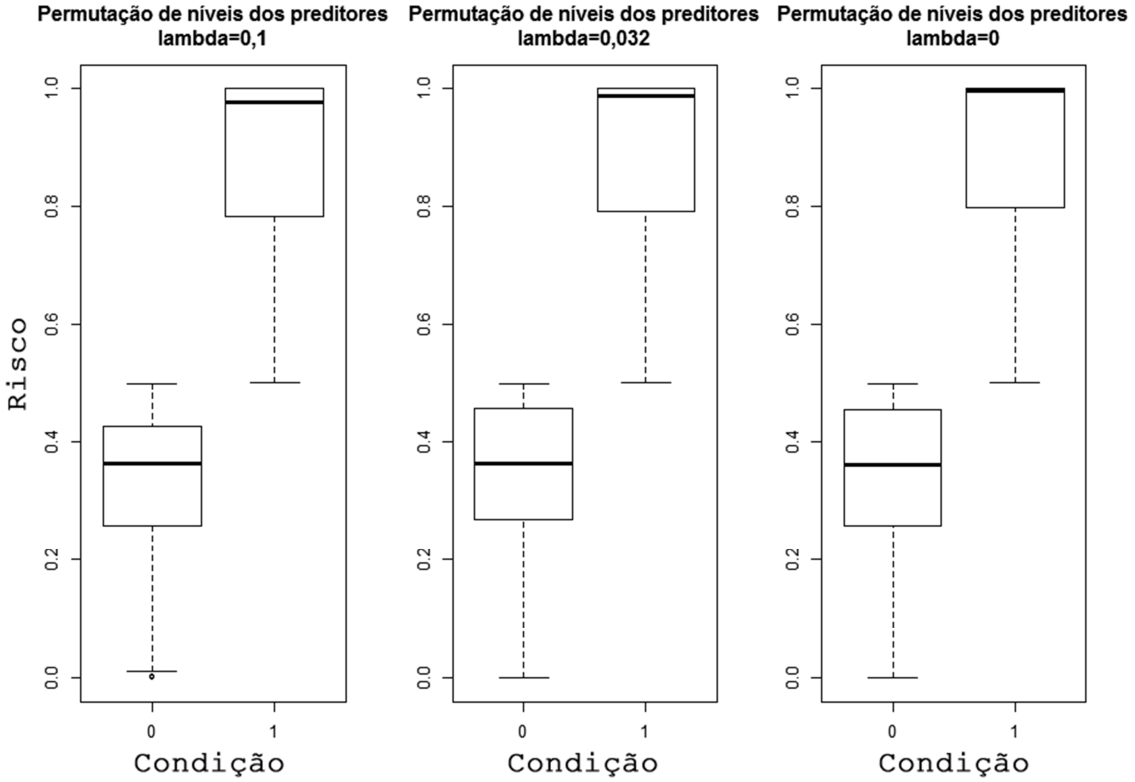


Figura 21 - Simulação das probabilidades associadas à predição usando o modelo preditivo diagnóstico. Todos os níveis das variáveis preditoras são permutados. O eixo $\boldsymbol{x}$ indica a condição (0 sem PET/MAH e 1 com PET/MAH) e o eixo $\boldsymbol{y}$ indica a probabilidade prevista pelo modelo preditivo em cada cenário.



6 DISCUSSÃO

A Paraparesia Espástica Tropical/Mielopatia Associada ao HTLV-1 (PET/MAH) é uma condição relativamente rara na população em geral, mas altamente debilitante, pois até 50% dos acometidos podem depender de cadeira de rodas (HLELA et al., 2009). Uma atenção especial é necessária, sobretudo nas regiões endêmicas como o Brasil, onde a infecção por HTLV-1 entre doadores de sangue pode chegar a cerca de 1% (CATALAN-SOARES; CARNEIRO-PROIETTI; PROIETTI, 2005; CARNEIRO-PROIETTI et al., 2012). Sobretudo porque não há ainda intervenções terapêuticas efetivas a disponibilizar aos pacientes com PET/MAH (WILLEMS et al., 2017; YAMANO; SATO, 2012).

Neste estudo investigou-se a classificação de pacientes infectados com o HTLV-1 em grupos de acordo com suas características, concretizadas em variáveis funcionais clínicas, variáveis de expressão gênica e variável demográfica (sexo). Numa primeira etapa, foram empregados métodos estatísticos exploratórios e de descobrimento de padrões para encontrar uma classificação em grupos de pacientes, e comparar estes padrões com a classificação clínica definida pelos avaliadores com base nos critérios oficiais definidos pela OMS para o diagnóstico da PET/MAH. Na segunda etapa, selecionaram-se variáveis e obtiveram-se dois possíveis modelos com base em regressão logística penalizada.

Na análise não supervisionada, verificou-se que é possível classificar os pacientes em grupos, basicamente três, de acordo com o perfil multivariado. Já era esperado que os pacientes se agrupassem em pelo menos dois conjuntos distintos, sabendo-se que havia no banco de dados tanto indivíduos com o diagnóstico de PET/MAH quanto sem. O mais notável, porém, foi a definição de um terceiro grupo de pacientes localizados no intermédio dos outros dois (Figura 11). Quando os pacientes são separados em indivíduos com e sem PET/MAH (como definido pelos profissionais clínicos) os dois grupos não são bem separados, e alguns indivíduos confundem a classificação no centro da Figura 9. No entanto, ao usar uma classificação em três grupos pelos clusters encontrados na Análise de Conglomerados (Figura 10), os indivíduos são melhor distintos. O *cluster b*, intermediário, é composto de indivíduos classificados tanto com quanto sem o diagnóstico positivo.

Castro-Costa et al. (2006) já notavam que alguns pacientes não atendiam os critérios de diagnóstico da OMS, ou porque a doença estava em estágio inicial ou porque o contexto local não provia os meios diagnósticos suficientes. Dessa forma, sugeriram adicionalmente a classificação em níveis de certeza, como condição definida (certamente há PET/MAH), condição provável (com apresentação monossintomática e exclusão de outras desordens

similares à PET/MAH) ou condição possível (com apresentação sintomática completa ou incompleta, mas sem a exclusão de outras causas similares à PET/MAH).

Essa caracterização inicial que não configura PET/MAH definida já está documentada. Umeki et al. (2009) seguiram por 10 anos três crianças jamaicanas infectadas verticalmente pelo HTLV-1. Observaram que uma das crianças apresentou expansão de dois clones de células infectadas, aumentando a carga proviral em mais de 40 vezes, e apresentou hiperreflexia apesar de não configurar o diagnóstico da PET/MAH. Estudos urodinâmicos que avaliam manifestações como a bexiga neurogênica mostram que os sintomas urinários são frequentes nos indivíduos com HTLV-1. Em Andrade et al. (2013), encontrou-se que 42,4% de 118 infectados e sem diagnóstico de PET/MAH apresentaram sintomas urinários, o que foi associado com pior qualidade de vida desses indivíduos. Também, fisioterapia combinada (terapia comportamental, cinesioterapia e eletroterapia) tem sido utilizada com sucesso no alívio desses sintomas, resultando em melhora dos parâmetros urodinâmicos e qualidade de vida, quanto mais cedo ela for empregada (ANDRADE et al., 2016).

Biswas et al. (2009) encontraram maior chance de desenvolvimento de fraqueza nas pernas, problemas de marcha, alteração no senso de vibração, sinal de Babinski e incontinência urinária em indivíduos infectados pelo HTLV-1 e HTLV-2 quando comparados com indivíduos não infectados, em uma grande coorte (153 HTLV-1 positivos, 388 HTLV-2 positivos e 810 indivíduos não infectados, num seguimento de 15 anos), excluídos casos confirmados de PET/MAH. Conclusões similares foram obtidas por Poetker et al. (2011) sobre 71 indivíduos com HTLV-1 (sem PET/MAH) comparados com 71 indivíduos não infectados, na Bahia, e também na Bahia, por Tanajura et al. (2015), que avaliaram 251 indivíduos sem manifestações neurológicas e HTLV-1 positivos num estudo de seguimento de 8 anos, dos quais cinco desenvolveram PET/MAH definida e 78 (31%) desenvolveram PET/MAH provável. Além disso, lesões da substância branca do cérebro podem ser encontradas por ressonância magnética em níveis similares em pacientes com PET/MAH comparados com pacientes infectados, mas sem PET/MAH (MORGAN et al., 2007).

Embora alguns desses sintomas tenham um caráter isolado e não configurem PET/MAH conforme os critérios estabelecidos na literatura, eles são consistentes com os principais achados da doença. Parafraseando Biswas et al. (2009), a PET/MAH pode ser a "ponta do *iceberg*" de um amplo aspecto de manifestações neurológicas estáveis associadas à infecção pelo HTLV. Reforça essa ideia o fato de vários estudos registrarem que a idade dos pacientes com PET/MAH era estatisticamente superior à idade dos pacientes HTLV-1

positivos (ROSADO et al., 2017; TANAJURA et al., 2015; ISHIHARA et al., 2015; FURTADO et al., 2012).

Outro aspecto abordado no presente estudo foi a obtenção de um modelo preditivo clínico diagnóstico para predição do risco de ocorrência da PET/MAH, em indivíduos infectados pelo vírus. Tanto os critérios da OMS quanto a classificação proposta por Castro-Costa et al. (2006) são dotados de certa complexidade e sujeitos à subjetividade da interpretação do profissional clínico avaliador. Uma forma de mitigação seria submeter os pacientes a mais de uma avaliação independente, porém há custos associados quanto maior o número de avaliadores. Outra solução seria submeter o paciente a um modelo estatístico de predição que derive um risco com base em um número "médio" de outras avaliações. Dessa forma, esse risco poderia dar suporte à decisão diagnóstica do avaliador.

Esta é a primeira iniciativa de derivar um modelo de predição clínica diagnóstica (BOUWMEESTER et al., 2012; REILLY; EVANS, 2006) de PET/MAH com base em variáveis funcionais que refletem o acometimento neurológico dos indivíduos infectados pelo HTLV-1. Ishihara et al. (2015) propuseram um modelo preditivo diagnóstico de PET/MAH usando a regressão logística binária clássica (não penalizada). Três variáveis preditoras foram consideradas: nível da proteína *Secreted Protein Acidic and Rich in Cysteine* (SPARC), nível da proteína *Vascular Cell Adhesion Molecule 1* (VCAM-1) e a carga proviral do HTLV-1. Foi utilizado um banco de dados de treino de 71 indivíduos (34 com PET/MAH, 11 eventos por variável) e um banco de dados de teste independente com 34 indivíduos (16 com PET/MAH), resultando em um poder de discriminação (mensurado por AUC) de 0,897, comparado com um AUC=0,756 do modelo que possui um único preditor (carga proviral do HTLV-1). Além disso, o modelo de três variáveis alcançou uma sensibilidade de 86% e especificidade de 81,8%. No entanto, essas variáveis preditoras necessitam de aparato laboratorial para serem mensuradas, um fator limitante para o uso do modelo em alguns contextos de poucos recursos.

Também, há limitações relacionadas ao uso da regressão logística clássica quanto menor o tamanho amostral. Na presente pesquisa, foi empregado o método de regressão penalizada conforme proposto por Gertheiss e Tutz (2009), em que as diferenças entre coeficientes *dummy* adjacentes das variáveis ordinais são contraídas de forma a suavizar a transição entre os diversos níveis da variável preditora ordinal, e as variáveis categóricas e numéricas são penalizadas da forma convencional (por *Group*-LASSO ou RIDGE) (GERTHEISS et al., 2011; TUTZ; GERTHEISS, 2014). A intensidade da penalização aplicada é crítica para a derivação do modelo, e se reflete num parâmetro de penalização λ

(lambda). Aqui, foram considerados dois métodos de obtenção de λ: minimização do desvio (erro) de predição obtido por validação cruzada de 10 vezes (empregado por Tutz e Gertheiss, 2014) e o mesmo procedimento, mas repetido por 50 vezes com atribuição aleatória das observações às partições da validação cruzada em cada uma das 50 repetições (sugerido por Roberts e Nowak, 2014 e denominado aqui de validação cruzada repetida). Além de diminuir o viés associado à atribuição das observações às partições da validação cruzada, a validação cruzada repetida provê percentis para a escolha de λ, permitindo escolher um percentil superior se é desejada uma maior penalização ou um percentil de 50% se é desejada uma penalização moderada.

Outros métodos de modelagem estão disponíveis ao pesquisador, quando a variável resposta é binária, como *Random Forests*, Redes Neurais e Árvores de Classificação e Regressão (CART ou *Classification and Regression Trees*) (MAROCO et al., 2011). No entanto, Ploeg, Austin e Steyerberg (2014) encontraram que essas técnicas relativamente recentes demandam uma amostra maior que a regressão logística clássica para alcançar estimativas estáveis de AUC e alcançar menor otimismo quando aplicadas sobre bancos de dados simulados, necessitando de cerca de 10 vezes mais observações por variável que o método clássico. Logo, elas seriam apropriadas para bancos de dados grandes. Mesmo a regressão logística, neste contexto de simulações, alcançou estabilidade de AUC com cerca de 20 a 50 observações por variável.

Ao contrário, os métodos de *shrinkage* possibilitam amenizar problemas relacionados ao baixo número de eventos por variável em bancos de dados pequenos (STEYERBERG et al., 2000; HOERL; KENNARD, 1970), particularmente o otimismo ou sobre-ajuste (MOONS et al., 2004). Isso é interessante em várias aplicações biomédicas, onde a condição em análise é pouco frequente na população como no caso da PET/MAH, ou quando o número de variáveis preditoras excede o número de observações disponíveis. Apesar disso, as medidas de desempenho (acurácia, sensibilidade, especificidade e AUC) obtidas por validação cruzada (Figura 18) foram ótimas, permitindo desconfiar de sobre-ajuste, mesmo quando o maior lambda (0,1) é empregado. Logo, novas avaliações a partir de um banco de dados externo são necessárias.

Foi realizada apenas a validação interna (STEYERBERG et al., 2010) dos modelos propostos, devido à indisponibilidade de um banco de dados externo para avaliação do desempenho, o que é necessário para aplicação prática (BLEEKER et al., 2003). Foi priorizada a avaliação interna por validação cruzada de 10 vezes, pois a técnica tem sido demonstrada robusta na correção de otimismo em amostras pequenas em comparação com o

não uso de técnicas de replicação, apesar de mostrar maior variabilidade quando comparada com a validação cruzada repetida na avaliação de AUC (SMITH et al., 2014; ROBERTS; NOWAK, 2014). Outra potencial limitação está relacionada à estimação de coeficientes por métodos de *shrinkage* onde se introduz um viés, restringindo a derivação de um p-valor ao qual grande parte dos pesquisadores das ciências biomédicas está acostumada a interpretar (BREIMAN, 2001). No entanto, a importância prática da estatística baseada em modelos de população pode ser questionada (KATTAN, 2011).

Tanto na PCA mista quanto na regressão logística penalizada as variáveis contínuas de expressão gênica foram eliminadas durante o processo de seleção de variáveis. Isso significa que os níveis de expressão das citocinas IFN-γ, IL4 e IL-10 são compatíveis entre os indivíduos, o que vai de encontro a outros estudos que indicam um aumento de citocinas pró-inflamatórias e diminuição de citocinas anti-inflamatórias nos pacientes com PET/MAH (ESPíNDOLA et al., 2015; FUZII et al., 2014; FURUYA et al., 1999). Logo, é possível que esta seja uma característica restrita à amostra em análise. Também é possível que, considerando os pressupostos assumidos pelos métodos usados no presente trabalho, a importância das variáveis de expressão gênica tenha sido obscurecida pela importância das variáveis clínicas ordinais. Outra possibilidade é que os níveis de expressão mensurados sejam compatíveis entre indivíduos com e sem PET/MAH utilizados neste estudo, mas alterados em relação aos indivíduos não infectados, os quais não foram objeto da presente investigação.

Outro fato interessante a considerar é que a seleção de variáveis realizada por regressão logística penalizada selecionou menos variáveis (cinco, conforme a Figura 16) do que a seleção de variáveis realizada pelo critério de correlação maior que 0,6 entre as variáveis e as três primeiras componentes principais da PCA mista (Tabela 5). Isso provavelmente reflete os diferentes pressupostos que cada método usa para lidar com as variáveis preditoras do banco de dados, mas também ao critério de seleção de variáveis escolhido na PCA mista. Este critério é frequentemente usado, mas a escolha da correlação crítica é um tanto quanto arbitrária, de modo que se fosse considerada uma correlação mínima maior que 0,6, menos variáveis seriam selecionadas na PCA mista. Neste sentido, a seleção de variáveis pelo método de regressão aqui considerado é menos arbitrária. Todas as variáveis constantes no modelo logístico final, exceto o escore de Tinetti, foram selecionadas na análise de componentes principais mista.

Por fim, uma vantagem da predição com base em variáveis preditoras funcionais de natureza ordinal é que todos os valores de probabilidade possíveis podem ser previstos pela

permutação de níveis dos fatores. Conforme observado na simulação da Figura 21, as probabilidades previstas pelos dois modelos finais são interpretáveis, distinguindo-se razoavelmente bem entre as previsões possíveis de PET/MAH (probabilidade maior que 0,5) e sem PET/MAH (probabilidade menor ou igual a 0,5).

7 CONCLUSÕES E RECOMENDAÇÕES

É possível classificar os indivíduos infectados pelo HTLV-1 em três grupos: um grupo composto de pacientes com PET/MAH definida, um grupo de pacientes sem PET/MAH, e um grupo intermediário, que comporta tanto pacientes com quanto sem a doença. Esses pacientes intermediários possuem características tanto de um grupo quanto de outro e devem ser vistos com a necessária atenção, pois se distinguem suficientemente dos pacientes assintomáticos.

As variáveis mais importantes segundo a análise de componentes principais mista foram: força proximal do membro inferior direito e do esquerdo, escore de função da bexiga, grau de auxílio na marcha, tônus do músculo adutor direito e do esquerdo, tônus do quadríceps femoral direito e do esquerdo e tônus do tríceps sural direito e do esquerdo. As variáveis mais importantes segundo a análise de regressão logística binária penalizada foram: grau de auxílio na marcha, escore codificado de equilíbrio e mobilidade de Tinetti, tônus do músculo adutor direito e do esquerdo e tônus do tríceps sural esquerdo. Todas as variáveis selecionadas pela regressão penalizada foram também selecionadas pela PCA mista, com exceção do escore codificado de Tinetti.

Foram derivados dois modelos de regressão logística binária penalizada: um com parâmetro de penalização 0,032 e outro com parâmetro de penalização 0,1. Logo, é possível derivar modelos de predição clínica diagnóstica adequados para a natureza ordinal das variáveis preditoras funcionais, que aperfeiçoem a classificação dos pacientes em com ou sem PET/MAH e que auxiliem os profissionais clínicos na tomada de decisão. Também, a literatura mostra que os métodos de regressão baseados em penalização de coeficientes (métodos de *shrinkage*) são adequados para os bancos de dados pequenos, os quais geralmente estão disponíveis nas doenças de frequência rara na população em geral, como a PET/MAH.

O modelo final com parâmetro de penalização 0,032 é dado por:

$$P(y = 1) = \frac{e^x}{1+e^x}, \text{ onde}$$

$x =$ -6,73 + 1,9(MR:1) + 4,01(MR:2) + 5,27(MR:3) + 6,53(MR:4) + 6,71(MR:5) + 0,81(TIN:1) + 4,17(TIN:2) + 5,54(TIN:3) + 1,17(ADD:1) + 1,52(ADD:1+) + 1,6(ADD:2) + 1,68(ADD:3) + 1,68(ADD:4) + 4,42(ADE:1) + 4,73(ADE:1+) + 4,81(ADE:2) + 4,84(ADE:3) + 4,84(ADE:4) + 1,33(TRE:1) + 2,26(TRE:1+) + 2,33(TRE:2) + 2,36(TRE:3) + 2,36(TRE:4),

sendo: MR: grau de auxílio na marcha; TIN: escala codificada de Tinetti; ADD e ADE: tônus do músculo adutor direito e do esquerdo, respectivamente; TRE: tônus do tríceps sural esquerdo; e: base dos logaritmos naturais; $P\,(y = 1)$: probabilidade de o paciente ser PET/MAH.

O modelo final com parâmetro de penalização 0,1 possui o mesmo formato, porém, o valor de x é dado por:

$x =$ -5,6 + 1,44(MR:1) + 3(MR:2) + 4,05(MR:3) + 5,10(MR:4) + 5,45(MR:5) + 0,70(TIN:1) + 3,67(TIN:2) + 4,72(TIN:3) + 1,06(ADD:1) + 1,45(ADD:1+) + 1,58(ADD:2) + 1,69(ADD:3) + 1,69(ADD:4) + 3(ADE:1) + 3,32(ADE:1+) + 3,45(ADE:2) + 3,52(ADE:3) + 3,52(ADE:4) + 1,16(TRE:1) + 1,87(TRE:1+) + 1,98(TRE:2) + 2,05(TRE:3) + 2,05(TRE:4).

Estes modelos finais e exemplos de aplicação estão disponíveis no Apêndice A. No Apêndice B estão disponibilidados os códigos R usados nas análises.

Em relação à avaliação dos modelos propostos neste trabalho, estes mostraram medidas de desempenho ótimas, mesmo quando é empregado o parâmetro de penalização mais extremo, podendo-se desconfiar de sobreajuste. Porém, é necessário proceder à sua validação externa para uma avaliação mais criteriosa, ou propor outros modelos seguindo os mesmos preceitos metodológicos em um banco de dados com mais eventos por variável.

REFERÊNCIAS

AFONSO, P. V. et al. Alteration of blood-brain barrier integrity by retroviral infection. **PLoS pathogens**, v. 4, p. 1–11, nov. 2008. ISSN 1553-7374.

ALEFANTIS, T. et al. Secretion of the human T cell leukemia virus type I transactivator protein tax. **The Journal of biological chemistry**, v. 280, p. 17353–17362, abr. 2005. ISSN 0021-9258.

ANDRADE, R. et al. Association between urinary symptoms and quality of life in HTLV-1 infected subjects without myelopathy. **International braz j urol: official journal of the Brazilian Society of Urology**, v. 39, p. 861–866, 2013. ISSN 1677-6119.

ANDRADE, R. C. P. et al. Effects of physiotherapy in the treatment of neurogenic bladder in patients infected with Human T-Lymphotropic Virus 1. **Urology**, v. 89, p. 33–38, mar. 2016. ISSN 1527-9995.

BAI, X. T.; NICOT, C. Overview on HTLV-1 p12, p8, p30, p13: accomplices in persistent infection and viral pathogenesis. **Frontiers in microbiology**, v. 3, p. 400, 2012. ISSN 1664-302X.

BANGHAM, C. R. M. et al. HTLV-1-associated myelopathy/tropical spastic paraparesis. **Nature reviews - Disease primers**, v. 1, p. 15012, Jun 2015. ISSN 2056-676X.

BISWAS, H. H. et al. Neurologic abnormalities in HTLV-I- and HTLV-II-infected individuals without overt myelopathy. **Neurology**, v. 73, p. 781–789, set. 2009. ISSN 1526-632X.

BITTENCOURT, A. L. et al. Adult t-cell leukemia/lymphoma (ATL) presenting in the skin: clinical, histological and immunohistochemical features of 52 cases. **Acta Oncol**, v. 48, n. 4, p. 598–604, 2009. Disponível em: <http://dx.doi.org/10.1080/02841860802657235>.

BITTENCOURT, A. L. et al. Analysis of cutaneous lymphomas in a medical center in Bahia, Brazil. **American journal of clinical pathology**, v. 140, p. 348–354, Sep 2013. ISSN 1943-7722.

BLEEKER, S. E. et al. External validation is necessary in prediction research: a clinical example. **Journal of clinical epidemiology**, v. 56, p. 826–832, Sep 2003. ISSN 0895-4356.

BOHANNON, R. W.; SMITH, M. B. Interrater reliability of a modified Ashworth scale of muscle spasticity. **Physical therapy**, v. 67, p. 206–207, fev. 1987. ISSN 0031-9023.

BOUWMEESTER, W. et al. Reporting and methods in clinical prediction research: a systematic review. **PLoS medicine**, v. 9, p. 1–12, 2012. ISSN 1549-1676.

BREIMAN, L. Statistical modeling: The two cultures. Statististial Science, **The Institute of Mathematical Statistics**, v. 16, n. 3, p. 199–231, 08 2001. Disponível em: <http://dx.doi.org/10.1214/ss/1009213726>.

BRITO-MELO, G. E. A. et al. IL-10 produced by CD4+ and CD8+ T cells emerge as a putative immunoregulatory mechanism to counterbalance the monocyte-derived TNF-alpha and guarantee asymptomatic clinical status during chronic HTLV-I infection. **Clinical and experimental immunology**, v. 147, p. 35–44, jan. 2007. ISSN 0009-9104.

CALATTINI, S. et al. Discovery of a new human T-cell lymphotropic virus (HTLV-3) in central africa. **Retrovirology**, v. 2, p. 30, 2005. Disponível em: <http://dx.doi.org/10.1186/1742-4690-2-30>.

CANTY, A.; RIPLEY, B. **boot: Bootstrap Functions (Originally by Angelo Canty for S)**. [S.l.], 2016. R package version 1.3-18. Disponível em: <https: //CRAN.R-project.org/package=boot>.

CARNEIRO-PROIETTI, A. B. F. et al. Human T-lymphotropic virus type 1 and type 2 seroprevalence, incidence, and residual transfusion risk among blood donors in Brazil during 2007-2009. **AIDS Res Hum Retroviruses**, v. 28, n. 10, p. 1265–1272, Oct 2012. Disponível em: <http://dx.doi.org/10.1089/AID.2011.0143>.

CARTIER, L.; RAMIREZ, E. Presence of HTLV-I Tax protein in cerebrospinal fluid from HAM/TSP patients. **Archives of virology**, v. 150, p. 743–753, abr. 2005. ISSN 0304-8608.

CASTRO-COSTA, C. M. D. et al. Proposal for diagnostic criteria of tropical spastic paraparesis/HTLV-I-associated myelopathy (TSP/HAM). **AIDS research and human retroviruses**, v. 22, p. 931–935, Oct 2006. ISSN 0889-2229.

CATALAN-SOARES, B.; CARNEIRO-PROIETTI, A. B. d. F.; PROIETTI, F. A. Heterogeneous geographic distribution of human T-cell lymphotropic viruses I and II (HTLV-I/II): serological screening prevalence rates in blood donors from large urban areas in Brazil. **Cadernos de Saude Publica**, v. 21, n. 3, p. 926–931, 2005.

CAVROIS, M. et al. Common human T cell leukemia virus type 1 (HTLV-1) integration sites in cerebrospinal fluid and blood lymphocytes of patients with HTLV-1-associated myelopathy/tropical spastic paraparesis indicate that HTLV-1 crosses the blood-brain barrier via clonal HTLV-1-infected cells. **The Journal of infectious diseases**, v. 182, p. 1044–1050, out. 2000. ISSN 0022-1899.

CHAMPS, A. P. S. et al. [HTLV-1 associated myelopathy: clinical and epidemiological profile in a 10-year case series study]. **Rev Soc Bras Med Trop**, v. 43, n. 6, p. 668–672, 2010.

CHESSEL, D.; DUFOUR, A. B.; THIOULOUSE, J. The ade4 package - i : One-table methods. **R news**, v. 4, n. 1, p. 5–10, 2004.

COWAN, E. P. et al. Induction of tumor necrosis factor alpha in human neuronal cells by extracellular human T-cell lymphotropic virus type 1 Tax. **Journal of virology**, v. 71, p. 6982–6989, set. 1997. ISSN 0022-538X.

CURIS, C. et al. Human T-Lymphotropic Virus Type 1-induced overexpression of Activated Leukocyte Cell Adhesion Molecule (ALCAM) facilitates trafficking of infected lymphocytes

through the Blood-Brain Barrier. **Journal of virology**, v. 90, p. 7303–7312, ago. 2016. ISSN 1098-5514.

CURRER, R. et al. HTLV tax: a fascinating multifunctional co-regulator of viral and cellular pathways. **Frontiers in microbiology**, v. 3, p. 406, 2012. ISSN 1664-302X.

DANTAS, L. et al. Dermatological manifestations of individuals infected with human T cell lymphotropic virus type I (HTLV-I). **International journal of dermatology**, v. 53, p. 1098–1102, Sep 2014. ISSN 1365-4632.

DIAS, G. A. da S. Padrão de resposta imunológica periférica em pacientes infectados pelo HTLV-1 e sua correlação com as manifestações neurológicas funcionais nos indivíduos com PET/MAH. Tese (Doutorado) — Universidade Federal do Pará, 2014.

DIAS, G. A. da S. et al. Correlation between clinical symptoms and peripheral immune response in HAM/TSP. **Microbial pathogenesis**, v. 92, p. 72–75, Mar 2016. ISSN 1096-1208.

DIAS, G. A. S. et al. Neurological manifestations in individuals with HTLV-1-associated myelopathy/tropical spastic paraparesis in the Amazon. **Spinal cord**, v. 54, p. 154–157, Feb 2016. ISSN 1476-5624.

DRAY, S.; DUFOUR, A.-B. et al. The ade4 package: implementing the duality diagram for ecologists. **Journal of statistical software**, v. 22, n. 4, p. 1–20, 2007.

ESPíNDOLA, O. M. et al. High IFN-γ/IL-10 expression ratio and increased frequency of persistent human T-cell lymphotropic virus type 1-infected clones are associated with human T-cell lymphotropic virus type 1-associated myelopathy/tropical spastic paraparesis development. **Intervirology**, v. 58, p. 106–114, 2015. ISSN 1423-0100.

FACCHINETTI, L. D. et al. Falls in patients with HTLV-i-associated myelopathy/tropical spastic paraparesis (HAM/TSP). **Spinal Cord**, v. 51, n. 3, p. 222–225, Mar 2013. Disponível em: <http://dx.doi.org/10.1038/sc.2012.134>.

FENG, C. et al. Log-transformation and its implications for data analysis. **Shanghai Arch Psychiatry**, v. 26, n. 2, p. 105–109, Apr 2014. Disponível em: <http://dx.doi.org/10.3969/j.issn.1002-0829.2014.02.009>.

FERRER, J. F. et al. Further studies on the antigenic properties and distribution of the putative bovine leukemia virus. **Bibl Haematol**, n. 40, p. 59–66, 1975.

FILIPPONE, C. et al. A severe bite from a nonhuman primate is a major risk factor for HTLV-1 infection in hunters from Central Africa. **Clinical infectious diseases: an official publication of the Infectious Diseases Society of America**, v. 60, p. 1667–1676, Jun 2015. ISSN 1537-6591.

Friedman, J.; Hastie, T.; Tibshirani, R. A note on the group lasso and a sparse group lasso. [S.l.], 2010. Disponível em: <https://arxiv.org/pdf/1001.0736v1.pdf>.

FURTADO, M. d. S. B. S. et al. Monitoring the HTLV-1 proviral load in the peripheral blood of asymptomatic carriers and patients with HTLV-associated myelopathy/tropical spastic paraparesis from a Brazilian cohort: ROC curve analysis to establish the threshold for risk disease. **Journal of medical virology**, v. 84, p. 664–671, abr. 2012. ISSN 1096-9071.

FURUKAWA, Y. et al. Existence of escape mutant in HTLV-i tax during the development of adult T-cell leukemia. **Blood, American Society of Hematology**, v. 97, n. 4, p. 987–993, 2001. ISSN 0006-4971. Disponível em: <http://www.bloodjournal.org/content/97/4/987>.

FURUYA, T. et al. Elevated levels of interleukin-12 and interferon-gamma in patients with human T lymphotropic virus type i-associated myelopathy. **Journal of neuroimmunology**, v. 95, p. 185–189, mar. 1999. ISSN 0165-5728.

FUZII, H. T. et al. Immunopathogenesis of HTLV-1-assoaciated myelopathy/tropical spastic paraparesis (HAM/TSP). **Life Sci**, v. 104, n. 1-2, p. 9–14, May 2014. Disponível em: <http://dx.doi.org/10.1016/j.lfs.2014.03.025>.

GALLO, R. C. et al. Isolation and tissue distribution of type-C virus and viral components from a gibbon ape (Hylobates lar) with lymphocytic leukemia. **Virology**, v. 84, n. 2, p. 359–373, Feb 1978.

GALLO, R. C. et al. Human T-cell leukemia-lymphoma virus (HTLV) is in T but not B lymphocytes from a patient with cutaneous T-cell lymphoma. **Proc Natl Acad Sci USA**, v. 79, n. 18, p. 5680–5683, Sep 1982.

GARCíA-VALLEJO, F.; DOMíNGUEZ, M. C.; TAMAYO, O. Autoimmunity and molecular mimicry in tropical spastic paraparesis/human T-lymphotropic virus-associated myelopathy. **Brazilian journal of medical and biological research = Revista brasileira de pesquisas medicas e biologicas**, v. 38, p. 241–250, fev. 2005. ISSN 0100-879X.

GERTHEISS, J. **ordPens: Selection and/or Smoothing of Ordinal Predictors**. [S.l.], 2015. R package version 0.3-1. Disponível em: <https://CRAN.R-project.org/package=ordPens>.

GERTHEISS, J. et al. Selection of ordinally scaled independent variables with applications to international classification of functioning core sets. **Journal of the Royal Statistical Society: Series C (Applied Statistics)**, Blackwell Publishing Ltd, v. 60, n. 3, p. 377–395, 2011. ISSN 1467-9876. Disponível em: <http://dx.doi.org/10.1111/j.1467-9876.2010.00753.x>.

GERTHEISS, J.; TUTZ, G. Penalized regression with ordinal predictors. **International Statistical Review**, Blackwell Publishing Ltd, v. 77, n. 3, p. 345–365, 2009. ISSN 1751-5823. Disponível em: <http://dx.doi.org/10.1111/j.1751-5823.2009.00088.x>.

GESSAIN, A. et al. Antibodies to human T-lymphotropic virus type-I in patients with tropical spastic paraparesis. **Lancet**, v. 2, n. 8452, p. 407–410, Aug 1985.

GESSAIN, A.; CASSAR, O. Epidemiological aspects and world distribution of HTLV-1 infection. **Front Microbiol**, v. 3, p. 388, 2012. Disponível em: <http: //dx.doi.org/10.3389/fmicb.2012.00388>.

GIAM, C.-Z.; SEMMES, O. J. HTLV-1 infection and Adult T-Cell Leukemia/Lymphoma-a tale of two proteins: Tax and HBZ. **Viruses**, v. 8, jun. 2016. ISSN 1999-4915.

HENDERSON, A. R. The bootstrap: a technique for data-driven statistics. using computer-intensive analyses to explore experimental data. **Clinica chimica acta; international journal of clinical chemistry**, v. 359, p. 1–26, Sep 2005. ISSN 0009-8981.

HILL, M. O.; SMITH, A. J. E. Principal component analysis of taxonomic data with multi-state discrete characters. **Taxon, International Association for Plant Taxonomy (IAPT)**, v. 25, n. 2/3, p. 249–255, 1976. Disponível em: <http://www.jstor.org/stable/1219449>.

HLELA, C. et al. The prevalence of human T-cell lymphotropic virus type 1 in the general population is unknown. **AIDS Rev**, v. 11, n. 4, p. 205–214, 2009.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems. **Technometrics**, Taylor & Francis Group, v. 12, n. 1, p. 55–67, 1970.

HOSHINO, H. Cellular factors involved in HTLV-1 entry and pathogenicity. **Frontiers in microbiology**, v. 3, p. 222, 2012. ISSN 1664-302X.

ISHAK, R. et al. Identification of human T cell lymphotropic virus type iia infection in the Kayapo, an indigenous population of Brazil. **AIDS Res Hum Retroviruses**, v. 11, n. 7, p. 813–821, Jul 1995.

ISHAK, R. et al. Epidemiological aspects of retrovirus (HTLV) infection among indian populations in the Amazon region of Brazil. **Cad Saude Publica**, v. 19, n. 4, p. 901–914, 2003.

ISHIHARA, M. et al. A plasma diagnostic model of human T-cell leukemia virus-1 associated myelopathy. **Annals of clinical and translational neurology**, v. 2, p. 231–240, mar. 2015.

ITA, F. et al. Human T-lymphotropic virus type 1 infection is frequent in rural communities of the southern Andes of Peru. **Int J Infect Dis**, v. 19, p. 46–52, Feb 2014. Disponível em: <http://dx.doi.org/10.1016/j.ijid.2013.10.005>.

JEANG, K. T. Functional activities of the human T-cell leukemia virus type i Tax oncoprotein: cellular signaling through NF-kappa B. **Cytokine & growth factor reviews**, v. 12, p. 207–217, 2001. ISSN 1359-6101.

JONES, K. S. et al. Cell-free HTLV-1 infects dendritic cells leading to transmission and transformation of CD4(+) T cells. **Nature medicine**, v. 14, p. 429–436, abr. 2008. ISSN 1546-170X.

KALYANARAMAN, V. S. et al. A new subtype of human T-cell leukemia virus (HTLV-ii) associated with a T-cell variant of hairy cell leukemia. **Science**, v. 218, n. 4572, p. 571–573, Nov 1982.

KAMOI, K.; MOCHIZUKI, M. HTLV-1 uveitis. **Frontiers in microbiology**, v. 3, p. 270, 2012. ISSN 1664-302X.

KAPLAN, J. E. et al. Male-to-female transmission of human T-cell lymphotropic virus types I and II: association with viral load. the Retrovirus Epidemiology Donor Study Group. **J Acquir Immune Defic Syndr Hum Retrovirol**, v. 12, n. 2, p. 193–201, Jun 1996.

KATTAN, M. W. Doc, what are my chances? a conversation about prognostic uncertainty. **European urology**, v. 59, p. 224, Feb 2011. ISSN 1873-7560.

KHABBAZ, R. F. et al. Seroprevalence of HTLV-1 and HTLV-2 among intravenous drug users and persons in clinics for sexually transmitted diseases. **N Engl J Med**, v. 326, n. 6, p. 375–380, Feb 1992. Disponível em: <http://dx.doi.org/10.1056/NEJM199202063260604>.

KIERS, H. A. L. Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. **Psychometrika**, v. 56, n. 2, p. 197–212, 1991. ISSN 1860-0980. Disponível em: <http://dx.doi.org/10.1007/BF02294458>.

KINOSADA, H. et al. HTLV-1 bZIP factor enhances T-cell proliferation by impeding the suppressive signaling of co-inhibitory receptors. **PLoS pathogens**, v. 13, p. 1006120, jan. 2017. ISSN 1553-7374.

KRESS, A. K.; GRASSMANN, R.; FLECKENSTEIN, B. Cell surface markers in HTLV-1 pathogenesis. **Viruses**, v. 3, n. 8, p. 1439–1459, Aug 2011. Disponível em: <http://dx.doi.org/10.3390/v3081439>.

KUBOTA, R. et al. HTLV-I specific IFN-gamma+ CD8+ lymphocytes correlate with the proviral load in peripheral blood of infected individuals. **Journal of neuroimmunology**, v. 102, p. 208–215, jan. 2000. ISSN 0165-5728.

LAIRMORE, M. D. et al. Molecular determinants of human T-lymphotropic virus type 1 transmission and spread. **Viruses**, v. 3, p. 1131–1165, jul. 2011. ISSN 1999-4915.

LEVINE, P. H. et al. Human T-cell lymphotropic virus type I and adult T-cell leukemia/lymphoma outside Japan and the Caribbean basin. **Yale J Biol Med**, v. 61, n. 3, p. 215–222, 1988.

LEVINE, P. H. et al. Human T-cell leukemia virus-i and hematologic malignancies in Panama. **Cancer**, v. 63, n. 11, p. 2186–2191, Jun 1989.

LIU, I.; AGRESTI, A. The analysis of ordered categorical data: An overview and a survey of recent developments. **Test**, Springer, v. 14, n. 1, p. 1–73, 2005.

MACíA, C. et al. Seroprevalence of human T-lymphotropic virus in blood bank donors at Fundación Valle del Lili, Cali, Colombia, 2008-2014. **Biomedica: revista del Instituto Nacional de Salud**, v. 36, p. 108–115, Feb 2016. ISSN 0120-4157.

MACêDO, O. et al. Human T-cell lymphotropic virus types I and II infections in a cohort of patients with neurological disorders in Belém, Pará, Brazil. **Rev Inst Med Trop Sao Paulo**, v. 46, n. 1, p. 13–17, 2004.

Magno Falcão, L. F. et al. Environmental impact and seroepidemiology of HTLV in two communities in the eastern Brazilian amazon. **J Med Virol**, v. 85, n. 9, p. 1585–1590, Sep 2013. Disponível em: <http://dx.doi.org/10.1002/jmv.23620>.

MAHIEUX, R.; GESSAIN, A. HTLV-3/STLV-3 and HTLV-4 viruses: discovery, epidemiology, serology and molecular aspects. **Viruses**, v. 3, n. 7, p. 1074–1090, Jul 2011. Disponível em: <http://dx.doi.org/10.3390/v3071074>.

MANEL, N. et al. The ubiquitous glucose transporter GLUT-1 is a receptor for HTLV. **Cell**, v. 115, p. 449–459, nov. 2003. ISSN 0092-8674.

MARANO, G. et al. Human T-lymphotropic virus and transfusion safety: does one size fit all? **Transfusion**, v. 56, p. 249–260, jan. 2016. ISSN 1537-2995.

MAROCO, J. et al. Data mining methods in the prediction of dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. **BMC research notes**, v. 4, p. 299, Aug 2011. ISSN 1756-0500.

MCKINSEY, T. A. et al. Inactivation of IkappaBbeta by the tax protein of human T-cell leukemia virus type 1: a potential mechanism for constitutive induction of NF-kappaB. **Molecular and cellular biology**, v. 16, p. 2083–2090, maio 1996. ISSN 0270-7306.

MEDINA, F. et al. Tax posttranslational modifications and interaction with calreticulin in MT-2 cells and human peripheral blood mononuclear cells of human T cell lymphotropic virus type-I-associated myelopathy/tropical spastic paraparesis patients. **AIDS research and human retroviruses**, v. 30, p. 370–379, abr. 2014. ISSN 1931-8405.

MEIER, L.; GEER, S. V. D.; BÜHLMANN, P. The group lasso for logistic regression. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 70, n. 1, p. 53–71, 2008.

BRASIL (Ed.). Guia de manejo clínico da infecção pelo HTLV, v. 1. [S.l.]: Ministério da Saúde. Secretaria de Vigilância em Saúde. Programa Nacional de DST e AIDS, 2013.

MIYAZATO, P. et al. Transcriptional and epigenetic regulatory mechanisms affecting HTLV-1 provirus. **Viruses**, v. 8, Jun 2016. ISSN 1999-4915.

MOONS, K. G. M. et al. Penalized maximum likelihood estimation to directly adjust diagnostic and prognostic prediction models for overoptimism: a clinical example. **Journal of clinical epidemiology**, v. 57, p. 1262–1270, Dec 2004. ISSN 0895-4356.

MORGAN, D. J. et al. Brain magnetic resonance imaging white matter lesions are frequent in HTLV-I carriers and do not discriminate from HAM/TSP. **AIDS research and human retroviruses**, v. 23, p. 1499–1504, dez. 2007. ISSN 0889-2229.

MORIN, A.; URBAN, J.; SLIZ, P. A quick guide to software licensing for the scientist-programmer. **PLoS computational biology**, v. 8, p. 1002598, 2012. ISSN 1553-7358.

NAGHDI, S. et al. Interrater reliability of the Modified Modified Ashworth Scale (MMAS) for patients with wrist flexor muscle spasticity. **Physiotherapy theory and practice**, v. 24, p. 372–379, 2008. ISSN 1532-5040.

NISHIOKA, K.; SUMIDA, T.; HASUNUMA, T. Human T lymphotropic virus type I in arthropathy and autoimmune disorders. **Arthritis and rheumatism**, v. 39, p. 1410–1418, Aug 1996. ISSN 0004-3591.

OSAME, M. et al. Nationwide survey of HTLV-I-associated myelopathy in Japan: association with blood transfusion. **Ann Neurol**, v. 28, n. 1, p. 50–56, Jul 1990. Disponível em: <http://dx.doi.org/10.1002/ana.410280110>.

OSAME, M. et al. HTLV-I associated myelopathy, a new clinical entity. **Lancet**, v. 1, n. 8488, p. 1031–1032, May 1986.

PAIVA, A.; CASSEB, J. Origin and prevalence of human T-lymphotropic virus type 1 (HTLV-1) and type 2 (HTLV-2) among indigenous populations in the Americas. **Rev Inst Med Trop Sao Paulo**, v. 57, n. 1, p. 1–13, 2015. Disponível em: <http://dx.doi.org/10.1590/S0036-46652015000100001>.

PATERNOSTRO-SLUGA, T. et al. Reliability and validity of the Medical Research Council (MRC) scale and a modified scale for testing muscle strength in patients with radial palsy. Journal of rehabilitation medicine, v. 40, p. 665–671, ago. 2008. ISSN 1651-2081.

PAVLOU, M. et al. Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. **Statistics in medicine**, v. 35, p. 1159–1177, Mar 2016. ISSN 1097-0258.

PERCHER, F. et al. Mother-to-child transmission of HTLV-1 epidemiological aspects, mechanisms and determinants of mother-to-child transmission. **Viruses**, v. 8, n. 2, 2016. Disponível em: <http://dx.doi.org/10.3390/v8020040>.

PISE-MASISON, C. A. et al. Co-dependence of HTLV-1 p12 and p8 functions in virus persistence. **PLoS pathogens**, v. 10, p. 1004454, nov. 2014. ISSN 1553-7374.

PLOEG, T. van der; AUSTIN, P. C.; STEYERBERG, E. W. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. **BMC medical research methodology**, v. 14, p. 137, Dec 2014. ISSN 1471-2288.

POETKER, S. K. W. et al. Clinical manifestations in individuals with recent diagnosis of HTLV type i infection. **Journal of clinical virology: the official publication of the Pan American Society for Clinical Virology**, v. 51, p. 54–58, maio 2011. ISSN 1873-5967.

POIESZ, B. J. et al. Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. **Proc Natl Acad Sci USA**, v. 77, n. 12, p. 7415–7419, Dec 1980.

POIESZ, B. J. et al. Isolation of a new type C retrovirus (HTLV) in primary uncultured cells of a patient with sézary T-cell leukaemia. **Nature**, v. 294, n. 5838, p. 268–271, Nov 1981.

QUARESMA, J. A. S. et al. HTLV-1, immune response and autoimmunity. **Viruses**, v. 8, Dec 2015. ISSN 1999-4915.

R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2016. Disponível em: <https://www.R-project.org/>.

REILLY, B. M.; EVANS, A. T. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. **Annals of internal medicine**, v. 144, p. 201–209, Feb 2006. ISSN 1539-3704.

RENCHER, A. C.; CHRISTENSEN, W. F. Methods of Multivariate Analysis. JOHN WILEY & SONS INC, 2012. ISBN 0470178965. Disponível em: <http: //www.ebook.de/de/product/18600379/alvin_c_rencher_william_f_christensen_ methods_of_multivariate_analysis.html>.

REYNALDO, C. Regressão "Ridge": um método alternativo para o mal condicionamento da matrix das regressoras. Dissertação (Mestrado), 1997. Disponível em: <http: //www.bibliotecadigital.unicamp.br/document/?code=000120534>.

ROBERTS, S.; NOWAK, G. Stabilizing the lasso against cross-validation variability. **Computational Statistics & Data Analysis**, Elsevier, v. 70, p. 198–211, 2014.

ROSADO, J. et al. The FAS-670 AA genotype is associated with high proviral load in peruvian HAM/TSP patients. **Journal of medical virology**, v. 89, p. 726–731, abr. 2017. ISSN 1096-9071.

ROYSTON, P.; ALTMAN, D. G. Visualizing and assessing discrimination in the logistic regression model. **Statistics in medicine**, v. 29, p. 2508–2520, out. 2010. ISSN 1097-0258.

ROYSTON, P.; ALTMAN, D. G.; SAUERBREI, W. Dichotomizing continuous predictors in multiple regression: a bad idea. **Statistics in medicine**, v. 25, p. 127–141, jan. 2006. ISSN 0277-6715.

SANTOS, E. L. d. et al. Molecular characterization of HTLV-1/2 among blood donors in Belém, state of Pará: first description of HTLV-2b subtype in the Amazon region. **Rev Soc Bras Med Trop**, v. 42, n. 3, p. 271–276, 2009.

SATAKE, M.; YAMAGUCHI, K.; TADOKORO, K. Current prevalence of HTLV-1 in japan as determined by screening of blood donors**. J Med Virol**, v. 84, n. 2, p. 327–335, Feb 2012. Disponível em: <http://dx.doi.org/10.1002/jmv.23181>.

SAUERBREI, W.; ROYSTON, P. Modelling to extract more information from clinical trials data: On some roles for the bootstrap. **Statistics in medicine**, v. 26, p. 4989–5001, Nov 2007. ISSN 0277-6715.

SEIKI, M. et al. Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. **Proceedings of the National Academy of Sciences of the United States of America**, v. 80, p. 3618–3622, Jun 1983. ISSN 0027-8424.

SILVA, J. L. S. da et al. Clustering of HTLV-1 associated myelopathy/tropical spastic paraparesis (HAM/TSP) and infective dermatitis associated with HTLV-1 (IDH) in Salvador, Bahia, Brazil. **J Clin Virol**, v. 58, n. 2, p. 482–485, Oct 2013. Disponível em: <http://dx.doi.org/10.1016/j.jcv.2013.07.012>.

SMITH, G. C. S. et al. Correcting for optimistic prediction in small data sets. **American journal of epidemiology**, v. 180, p. 318–324, Aug 2014. ISSN 1476-6256.

STEYERBERG, E. W. et al. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. **Statistics in medicine**, v. 19, p. 1059–1079, abr. 2000. ISSN 0277-6715.

STEYERBERG, E. W.; EIJKEMANS, M. J. C.; HABBEMA, J. D. F. Application of shrinkage techniques in logistic regression analysis: A case study. **Statistica Neerlandica**, Blackwell Publishers Ltd, v. 55, n. 1, p. 76–88, 2001. ISSN 1467-9574. Disponível em: <http://dx.doi.org/10.1111/1467-9574.00157>.

STEYERBERG, E. W. et al. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. **Journal of clinical epidemiology**, v. 54, p. 774–781, Aug 2001. ISSN 0895-4356.

STEYERBERG, E. W.; VERGOUWE, Y. Towards better clinical prediction models: seven steps for development and an abcd for validation. **European heart journal**, v. 35, p. 1925–1931, Aug 2014. ISSN 1522-9645.

STEYERBERG, E. W. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. **Epidemiology (Cambridge)**, v. 21, p. 128–138, Jan 2010. ISSN 1531-5487.

TAJIMA, K. The 4th nation-wide study of adult T-cell leukemia/lymphoma (ATL) in Japan: estimates of risk of ATL and its geographical and clinical features. the T- and B-cell Malignancy Study Group. **Int J Cancer**, v. 45, n. 2, p. 237–243, Feb 1990.

TANAJURA, D. et al. Neurological manifestations in Human T-Cell Lymphotropic Virus Type 1 (HTLV-1)-infected individuals without HTLV-1-Associated Myelopathy/Tropical Spastic Paraparesis: A longitudinal cohort study. **Clinical infectious diseases: an official publication of the Infectious Diseases Society of America**, v. 61, p. 49–56, jul. 2015. ISSN 1537-6591.

TAROKHIAN, H. et al. The effect of HTLV-1 virulence factors (HBZ, Tax, proviral load), HLA class I and plasma neopterin on manifestation of HTLV-1 associated myelopathy tropical spastic paraparesis. **Virus research**, v. 228, p. 1–6, jan. 2017. ISSN 1872-7492.

TENENHAUS, M.; YOUNG, F. W. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. **Psychometrika**, v. 50, n. 1, p. 91–119, 1985. ISSN 1860-0980. Disponível em: <http://dx.doi.org/10.1007/BF02294151>.

TETTEH, I. K.; FREMPONG, E.; AWUAH, E. An analysis of the environmental health impact of the Barekese Dam in Kumasi, Ghana. **J Environ Manage**, v. 72, n. 3, p. 189–194, Sep 2004. Disponível em: <http://dx.doi.org/10.1016/j.jenvman.2004.04.012>.

THé, G. de; BOMFORD, R. An HTLV-I vaccine: why, how, for whom? **AIDS research and human retroviruses**, v. 9, p. 381–386, May 1993. ISSN 0889-2229.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 267–288, 1996.

TINETTI, M. E. Performance-oriented assessment of mobility problems in elderly patients. **Journal of the American Geriatrics Society**, v. 34, p. 119–126, fev. 1986. ISSN 0002-8614.

TUTZ, G.; GERTHEISS, J. Rating scales as predictors–the old question of scale level and some answers. **Psychometrika**, v. 79, p. 357–376, Jul 2014. ISSN 1860-0980.

UCHIYAMA, T. et al. Adult T-cell leukemia: clinical and hematologic features of 16 cases. **Blood**, v. 50, n. 3, p. 481–492, Sep 1977.

UMEKI, K. et al. Proviral loads and clonal expansion of HTLV-1-infected cells following vertical transmission: a 10-year follow-up of children in Jamaica. **Intervirology**, v. 52, p. 115–122, 2009. ISSN 1423-0100.

VALLINOTO, A. C. et al. Serological evidence of HTLV-I and HTLV-II coinfections in HIV-1 positive patients in Belém, state of Pará, Brazil. **Mem Inst Oswaldo Cruz**, v. 93, n. 3, p. 407–409, 1998.

VALLINOTO, A. C. R. et al. Molecular epidemiology of human T-lymphotropic virus type II infection in Amerindian and urban populations of the Amazon region of Brazil. **Hum Biol**, v. 74, n. 5, p. 633–644, Oct 2002.

WANG, Z. et al. Histone posttranslational modifications of CD4+ T cell in autoimmune diseases. **International journal of molecular sciences**, v. 17, Sep 2016. ISSN 1422-0067.

WILLEMS, L. et al. Reducing the global burden of HTLV-1 infection: An agenda for research and action. **Antiviral Research**, v. 137, p. 41–48, 2017. ISSN 0166-3542. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0166354216306258>.

WOLFE, N. D. et al. Emergence of unique primate T-lymphotropic viruses among central African bushmeat hunters. **Proc Natl Acad Sci USA**, v. 102, n. 22, p. 7994–7999, May 2005. Disponível em: <http://dx.doi.org/10.1073/pnas.0501734102>.

World Health Organization. International Classification of Functioning, Disability and Health: ICF. [S.l.]: World Health Organization, 2001.

WU, X.; ZHANG, M.; SUN, S.-C. Mutual regulation between deubiquitinase CYLD and retroviral oncoprotein Tax. **Cell & bioscience**, v. 1, p. 27, ago. 2011. ISSN 2045-3701.

XIE, J. et al. The prevalence of human T-lymphotropic virus infection among blood donors in southeast China, 2004-2013. **PLoS Negl Trop Dis**, v. 9, n. 4, p. 0003685, Apr 2015. Disponível em: <http://dx.doi.org/10.1371/journal.pntd.0003685>.

YAMANO, Y.; SATO, T. Clinical pathophysiology of human T-lymphotropic virus-type 1-associated myelopathy/tropical spastic paraparesis. **Front Microbiol**, v. 3, p. 389, 2012. Disponível em: <http://dx.doi.org/10.3389/fmicb.2012.00389>.

YOSHIDA, M.; MIYOSHI, I.; HINUMA, Y. Isolation and characterization of retrovirus from cell lines of human adult T-cell leukemia and its implication in the disease. **Proc Natl Acad Sci USA**, v. 79, n. 6, p. 2031–2035, Mar 1982.

YUAN, M.; LIN, Y. Model selection and estimation in regression with grouped variables. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 68, n. 1, p. 49–67, 2006.

APÊNDICE

APÊNDICE A – MODELOS FINAIS DE REGRESSÃO LOGÍSTICA BINÁRIA PENALIZADA E EXEMPLOS DE APLICAÇÃO

O modelo de regressão logística binária penalizada, obtido com critério de penalização $\lambda = 0{,}032$, é dado por:

$$P(y = 1) = \frac{e^x}{1+e^x}, \text{ onde}$$

$x =$ -6,73 + 1,9(MR:1) + 4,01(MR:2) + 5,27(MR:3) + 6,53(MR:4) + 6,71(MR:5) + 0,81(TIN:1) + 4,17(TIN:2) + 5,54(TIN:3) + 1,17(ADD:1) + 1,52(ADD:1+) + 1,6(ADD:2) + 1,68(ADD:3) + 1,68(ADD:4) + 4,42(ADE:1) + 4,73(ADE:1+) + 4,81(ADE:2) + 4,84(ADE:3) + 4,84(ADE:4) + 1,33(TRE:1) + 2,26(TRE:1+) + 2,33(TRE:2) + 2,36(TRE:3) + 2,36(TRE:4), sendo: MR: grau de alteração na marcha; TIN: escala codificada de Tinetti; ADD e ADE tônus do músculo adutor direito e esquerdo, respectivamente; TRE: tônus do tríceps sural esquerdo; e: base dos logaritmos naturais; $P\,(y = 1)$: probabilidade de o paciente ser PET/MAH.

O código após cada variável indica o nível daquela variável ordinal. Logo, se o paciente possui escore 3 de marcha então o código será MR:3. Se um paciente possui escore 1 de marcha, escore codificado de Tinetti igual a zero, tônus do músculo adutor direito e do esquerdo iguais a 1+ e tônus do tríceps sural esquerdo igual a 1, então teremos:

$x =$ -6,73 + 1,9(MR:1) + 1,52(ADD:1+) + 4,73(ADE:1+) + 1,33(TRE:1) = **2,75**,

Logo o paciente tem 94% de probabilidade de ser PET/MAH:

$$P(y = 1) = \frac{e^{2,7\ 5}}{1 + e^{2,7\ 5}} = 0{,}94.$$

Considerando o modelo de regressão logística binária penalizada com critério de penalização $\lambda = 0{,}1$, o valor de x seria dado por:

$x =$ -5,6 + 1,44(MR:1) + 3(MR:2) + 4,05(MR:3) + 5,10(MR:4) + 5,45(MR:5) + 0,70(TIN:1) + 3,67(TIN:2) + 4,72(TIN:3) + 1,06(ADD:1) + 1,45(ADD:1+) + 1,58(ADD:2) + 1,69(ADD:3) + 1,69(ADD:4) + 3(ADE:1) + 3,32(ADE:1+) + 3,45(ADE:2) + 3,52(ADE:3) + 3,52(ADE:4) + 1,16(TRE:1) + 1,87(TRE:1+) + 1,98(TRE:2) + 2,05(TRE:3) + 2,05(TRE:4).

Neste caso, a probabilidade prevista para o paciente de exemplo seria 85,5%

APÊNDICE B – CÓDIGOS R UTILIZADOS NAS ANÁLISES

```
#Aqui são disponibilizados primeiro os scripts R usados nas análises,
#depois são apresentadas as funções customizadas usadas nestes scripts. Por
#exemplo, a função h$plotfactor() que plota um fator. Foi utilizada a
#versão 3.3.2 do R.

#==================ANÁLISE NÃO-SUPERVISIONADA==================

#PCA mista sobre o banco de dados completo (com todas as variáveis
#preditoras). db é o banco de dados usado na análise. end.dudi.cr são
#os coeficientes de correlação das variáveis com as componentes:
end.dudi <- ade4::dudi.mix(db, scannf=F, nf=5)

#PCA mista sobre o banco de dados reduzido:
end.dudi0 <- ade4::dudi.mix(db[,c('fpid','fpie','bex','mar','add','ade',
      'qdd','qde','trd','tre')], scannf=F, nf=5)

#Gráfico dos indivíduos (primeira e segunda componentes do banco de
#dados reduzido):
png('pca_red.png', width=641, height=552, res=90)
ade4::s.class(end.dudi0$li, xax=1, yax=2, fac=db$class,
      col=c('blue','red'), label=c('HTLV-1','PET/MAH'), clabel=0.8,
      sub='CP1xCP2')
dev.off()

#Dendrograma a partir da PCA sobre o banco de dados reduzido:
#método de aglomeração hierárquica:
end.clus <- hclust(dist(end.dudi0$l1[,1:3]), method='ward.D')

#gera imagem png:
png('dendro.png', width=822, height=552, res=90)
plot(end.clus, labels=paste(db$class,rownames(db),sep='.'),
      cex=0.8, sub='', xlab='Observações', ylab='Altura',
      main='Dendrograma', las=1)
#identifica os 3 clusters (a, b e c):
rect.hclust(end.clus, h=3, which=1:2, border='blue')
rect.hclust(end.clus, h=20, which=2, border='red')
text(x=c(20.5,39.5,42.5), y=c(1.5,5.0,29.5), labels=c('a','b','c'), font=2,
      col=c('blue','blue','red'))
dev.off()

#Obtém, como um vetor, os clusters a, b e c, gerados pelo gráfico anterior:
end.clus.3 <- cutree(end.clus, h=3)
end.clus.3[!(end.clus.3%in%1:2)] <- 3
end.clus.3 <- factor(end.clus.3, labels=c('b','a','c'))

#plota os 3 clusters no gráfico de indivíduos da PCA reduzida:
png('pca_cluster.png', width=724, height=552, res=90)
ade4::s.class(end.dudi0$li, xax=1, yax=2, fac=end.clus.3,
      col=c('green','blue','red'),
      label=c('b','a','c'), clabel=1.1, sub='CP1xCP2')
dev.off()

#identifica os indivíduos do cluster b, bem como o indivíduo zero
#do cluster c:
png('factor_clus.png', width=883, height=552, res=90)
par(mfcol=c(1,2))
h$plotfactor(db[rownames(db)%in%c('G1','G22','G46','G61'),
      c('fpid','fpie','bex','mar','add','ade','qdd','qde','trd',
      'tre')], col=c('red','blue','red','blue'),
```

```
      main='Cluster b', cnames=c('FPD','FPE','BX','MR','ADD','ADE','QDD',
      'QDE','TRD','TRE'), rdm=.15)

h$plotfactor(db[rownames(db)%in%c('G25'),
      c('fpid','fpie','bex','mar','add','ade','qdd','qde','trd',
      'tre')], col=c('blue'),
      main='Cluster c', cnames=c('FPD','FPE','BX','MR','ADD','ADE','QDD',
      'QDE','TRD','TRE'), rdm=0)
dev.off()
shell('factor_clus.png')

#==================ANÁLISE SUPERVISIONADA==================

#Para a análise supervisionada (regressão logística binária penalizada),
#o banco de dados inicial (db) precisou ser recodificado, pois o pacote
#ordPens reconhece as variáveis ordinais codificadas como números inteiros.
#Por exemplo, se uma variável ordinal inicialmente possuía os níveis
#0, 1, 1+, 2, 3 e 4 (como os escores de tônus muscular), esta variável
#precisou ser codificada para os escores 1, 2, 3, 4, 5 e 6, pois é
#esta codificação requerida pelo pacote ordPens para as variáveis
#categóricas (mais informações na documentação do pacote). Este tratamento
#inicial resultou num banco de dados end.db1 com os seguintes componentes,
#como requerido pelo pacote: end.db1$x (preditores ordinais),
#end.db1$y (variável resposta), end.db1$u (preditores categóricos
#não ordinais) e end.db1$z (preditores numéricos).

#Escolha do lambda para ordselect sobre o banco de dados completo. Aqui,
#o lambda ideal encontrado foi 0.35. Método de validação cruzada:
png('lamb_select.png', width=552, height=552, res=90)
with(end.db1, {
      h$chooselambda(x=x,y=y,u=u,z=z,select=T,
            elambda=seq(from=2, to=-2, by=-0.02),
            ylab='Desvio da validação cruzada')
})
dev.off()

#Mesma coisa, mas usando a abordagem repeated cross-validation
#Aqui, os lambda baseados em percentis de 50 e 95% foram: 0.316 (50%) e
#3.37 (95%)
with(end.db1, {
      h$repeatedlambda(x=x,y=y,u=u,z=z,select=T,R=50,
            elambda=seq(from=1.5, to=-3, by=-0.08))
})

#Seleção de variáveis usando ordSelect. Sobre o banco de dados completo:
png('ordselect.png', width=800, height=850, res=110)
with(end.db1, {
      h$penIC(x=x,y=y,u=u,z=z,lambda=c(0,0.35),select=T,p=90,R=1000,
      ncol=4,xlab='Níveis',ylab='Coeficientes dummy')
})
dev.off()

#obtenção do banco de dados numérico reduzido:
end.db2 <- end.db1
end.db2$x <- end.db2$x[,c('MR','TIN','ADD','ADE','TRE')]
end.db2$u <- NULL
end.db2$z <- NULL

#Escolha do lambda para ordSmooth sobre o banco de dados reduzido:
png('lamb_smooth.png', width=552, height=552, res=90)
with(end.db2, {
```

```
      h$chooselambda(x=x,y=y,select=F,onlyreturn=T,seed=NULL,
            elambda=seq(from=10, to=-16, by=-0.5),
            ylab='Desvio da validação cruzada')
})
dev.off()

#Escolha do lambda para ordSmooth por repeated cross-validation sobre
#o banco de dados reduzido. Valores encontrados: 0,032 (percentil 50%)
#e 0,1 (percentil 95%):
with(end.db2, {
      h$repeatedlambda(x=x,y=y,select=F,R=100,
            elambda=seq(from=10, to=-16, by=-0.5))
})

#Obtenção do modelo final por suavização de coeficientes via
#ordSmooth, lambda 0.032:
png('ordsmooth.png', width=800, height=800, res=90)
with(end.db2, {
      h$penIC(x=x,y=y,lambda=c(1e-5,0.032),select=F,p=90,R=1000,
      ncol=3,xlab='Níveis',ylab='Coeficientes dummy')
})
dev.off()

#Obtenção do modelo final por suavização de coeficientes via
#ordSmooth, lambda 0.1. Sobre o banco de dados reduzido:
png('ordsmooth1.png', width=800, height=800, res=90)
with(end.db2, {
      h$penIC(x=x,y=y,lambda=c(1e-5,0.1),select=F,p=90,R=1000,
      ncol=3,xlab='Níveis',ylab='Coeficientes dummy')
})
dev.off()

#Obtenção dos modelos finais com os três lambdas encontrados: 0.1, 0.032
#e o não penalizado (10e-5):
end.model <- with(end.db2,
      ordPens::ordSmooth(y=y,x=x,lambda=c(0.1,0.032,10e-5),model='logit')
)
x <- apply(end.db2$x,1,function(x){paste(names(end.db2$x),x,sep=':')})
sum(end.model$coefficients[c('intercept',x[,1]),])
paste(names(end.db2$x))

#Plotagem das medidas de desempenho dos modelos finais pela
#função ordSmooth:
png('performance.png',width=977,height=537,res=90)
png('performance1.png',width=1079,height=488,res=140)
par(mfcol=c(1,3))
with(end.db2,{
      h$performance(x=x,y=y,lambda=0.1,K=10, pressq=F,
            title=expression(paste(lambda,' = 0,1')),
            labels=c('Acur','Sens','Espec','AUC','Desvios'))
})
with(end.db2,{
      h$performance(x=x,y=y,lambda=0.032,K=10, pressq=F,
            title=expression(paste(lambda,' = 0,032')),
            labels=c('Acur','Sens','Espec','AUC','Desvios'))
})
with(end.db2,{
      h$performance(x=x,y=y,lambda=10e-5,K=10, pressq=F,
            title=expression(paste(lambda,' = 0')),
            labels=c('Acur','Sens','Espec','AUC','Desvios'))
})
```

```
dev.off()

#Plotagem da calibração dos modelos finais obtidos por ordSmooth:
png('calibration.png',width=700,height=900,res=90)
png('calibration.png',width=600,height=1000,res=120)
par(mfcol=c(3,2))
with(end.db2,{
      h$calibration(x=x,y=y,lambda=0.1,cv=F,title='Calibração,
lambda=0,1',slabel='Inclinação',xlab='Condição',ylab='Risco')
})
with(end.db2,{
      h$calibration(x=x,y=y,lambda=0.032,cv=F,title='Calibração,
lambda=0,032',slabel='Inclinação',xlab='Condição',ylab='Risco')
})
with(end.db2,{
      h$calibration(x=x,y=y,lambda=10e-5,cv=F,title='Calibração,
lambda=0',slabel='Inclinação',xlab='Condição',ylab='Risco')
})
with(end.db2,{
      h$calibration(x=x,y=y,lambda=0.1,cv=T,title='Calibração CV,
lambda=0,1',slabel='Inclinação',xlab='Condição',ylab='Risco')
})
with(end.db2,{
      h$calibration(x=x,y=y,lambda=0.032,cv=T,title='Calibração CV,
lambda=0,032',slabel='Inclinação',xlab='Condição',ylab='Risco')
})
with(end.db2,{
      h$calibration(x=x,y=y,lambda=10e-5,cv=T,title='Calibração CV,
lambda=0',slabel='Inclinação',xlab='Condição',ylab='Risco')
})
dev.off()

#Simulação de todos os valores possíveis de probabilidade dos modelos
#finais, pela permutação de níveis dos preditores ordinais:
png('sim_predict.png',width=925,height=652,res=120)
par(mfcol=c(1,3))
x <- with(end.db2,{
      h$perm.smo(y=y,x=x,lambda=0.1,xlab='Condição',ylab='Risco',ask=F,main
=cb('Permutação de níveis dos preditores\nlambda=0,1'))
      h$perm.smo(y=y,x=x,lambda=0.032,xlab='Condição',ylab='Risco',ask=F,ma
in=cb('Permutação de níveis dos preditores\nlambda=0,032'))
      h$perm.smo(y=y,x=x,lambda=10e-
5,return=T,xlab='Condição',ylab='Risco',ask=F,main=cb('Permutação de níveis
dos preditores\nlambda=0'))
})
dev.off()

#Plota as probabilidades obtidas pela simulação acima:
png('prob_intern.png',width=552,height=552,res=90)
with(end.model,{
      plot(1,xlim=c(1,3),ylim=0:1,type='n',xaxt='n',xlab='',ylab='Risco')
      for(i in 1:nrow(fitted)){
            lines(x=1:2,y=fitted[i,1:2],type='b')
            lines(x=2:3,y=fitted[i,2:3],type='b')
      }
      axis(side=1,at=1:3,labels=c(
            expression(lambda=='0,1'),
            expression(lambda=='0,032'),
            expression(lambda=='0')))
})
dev.off()
```

```
#=======FUNÇÕES COSTUMIZADAS CRIADAS PARA FACILITAR OS PROCEDIMENTOS=======

h <- list()

#Obtém o resíduo de uma análise de classificação binária, dadas uma
#variável resposta y e uma probabilidade de pertencimento à classe
#"caso" de y (essa probabilidade variando de 0 a 1). No contexto da
#regressão logística, o Residual Deviance é equivalente à soma dos
#quadrados dos erros obtidos da regressão linear (OLS). Quanto menor
#melhor o ajuste preditivo do modelo aos dados.
#y: variável respota binária (pode ser vetor numérico ou fator de dois
#níveis).
#p: probabilidade associada a y, predita por um modelo variando em 0 a 1.
#resdev: retornar o Residual Deviance? Se TRUE, será retornada a soma dos
#quadrados dos resíduos obtidos de cada observação. Se false, será
#retornado um vetor de tamanho length(y) contendo cada valor de
#deviance individual das observações (sem elevar ao quadrado).
h$deviance <- function(y, p, resdev = TRUE){
      #checagens iniciais
      if(!is.numeric(p)){
            stop('p is not numeric')
      }else if(any(is.na(y))){
            stop('y should not have NAs')
      }else if(any(p>1)){
            stop('p has values > 1')
      }else if(any(p<0)){
            stop('p has values < 0')
      }

      #codifica a resposta adequadamente como numérico binário (0 ou 1)
      y_ <- y
      y_ <- factor(y_)
      levels(y_) <- c(0,1)
      y_ <- as.numeric(as.character(y_))

      #complemento yc de y (se y=0, yc=1; se y=1, yc=0)
      yc <- ifelse(y_==1,0,1)

      a <- abs(p - yc)

      #deviance ao quadrado
      d2 <- -2*log(a)

      if(resdev){
            return(sum(d2))
      }else{
            return(sqrt(d2))
      }
}

#Plota boxplots das probabilidades previstas para a variável resposta y.
#O desvio (slope), ou seja, a diferença entre as médias das
#probabilidades das classes 1 e 0, nos dá a idéia de calibração do
#modelo, ou seja, quanto mais próximo de 45° essa diferença (e mais
#próxima de 1), melhor a calibração.
#y: variável resposta binária (pode ser vetor numérico ou fator de dois
#níveis).
#p: probabilidade associada a y, predita por um modelo, variando em 0 a 1.
#xlab, ylab: labels do gráfico
#slope label
```

```
h$plotprob <- function(y, p, main=NULL, xlab='Response', ylab='Predicted
risk', slabel='Slope'){
      #checagens iniciais
      if(!is.numeric(p)){
            stop('p is not numeric')
      }else if(any(is.na(y))){
            stop('y should not have NAs')
      }else if(any(p>1)){
            stop('p has values > 1')
      }else if(any(p<0)){
            stop('p has values < 0')
      }

      #codifica a resposta adequadamente como numérico binário (0 ou 1)
      y_ <- y
      y_ <- factor(y_)
      levels(y_) <- c(0,1)

      mean0 <- mean(p[y_==0])
      mean1 <- mean(p[y_==1])

      main_ <- paste(cb(slabel,'='), round(mean1-mean0,2), sep='')

      if(!is.null(main)){
            main <- paste(main, main_, sep='\n')
      }else{
            main <- main_
      }

      plot(p ~ y_, main=main, xlab=xlab, ylab=ylab)

#     points(x=c(0,1),y=c(mean0,mean1),type=16)
}

#Plota um gráfico que ajuda na escolha do lambda (critério de penalização)
#ideal (conforme Tutz, 2014, essa escolha pode ser por cross-validation).
#Há um lambda ideal para seleção de variáveis (ordPens::ordSelect) e
#outro ideal para suavização dos coeficientes (ordPens::ordSmooth). O
#ideal é um lambda de que minimize a soma dos quadrados dos
#desvios obtidos por cross-validation.
#x, y, u, z: Argumentos a passar a ordPens::ordSelect ou
#ordPens::ordSmooth, formatados corretamente (x, u e z data.frames apenas
#com colunas numéricas, y vetor numérico resposta com valores 0 ou 1).
#Consultar a ajuda para mais detalhes. x e y são obrigatórios.
#select: se TRUE, a modelagem será feita por ordPens::ordSelect. Se
#FALSE, por ordPens::ordSmooth
#elambda: expoentes dos valores de lambda a considerar. O eixo x do
#gráfico é representado na base 10, pois dessa forma se permite valores de
#lambda numa escala exponencial. Devem ser valores decrescentes. Sugestão:
#lambda=seq(from=2,to=-2,by=-0.1), onde o lambda real será 10e2,
#10e1.9, 10e1.8, ...
#lastRow: Se TRUE, há uma linha a mais (a última de x, y, u e z) que
#existe apenas para indicar o nível máximo das variáveis ordinais, e não
#entra na computação. Mais detalhes na ajuda do pacote ordPens.
#K: número de round de cross-validation. Se 10, será considerado um
#10-fold cross-validation.
#ylab: y label do gráfico
#seed: fixa o seed do gerador de números aleatórios (torna os resultados
#replicáveis). Passe NULL para não fixar o seed.
#onlyreturn: se TRUE, apenas será retornado o lambda ideal, sem
#plotar nada.
```

```
h$chooselambda <- function(x, y, u=NULL, z=NULL, select=FALSE, elambda,
lastRow=TRUE, K=10, ylab='Cross-validation deviance', seed=123456,
onlyreturn=FALSE){
      if(is.null(elambda)){
            stop('elambda is a must')
      }
      if(!is.numeric(y)){
            stop('y should be numeric: 0 or 1')
      }

      require(ordPens)

      if(select){
            fun <- ordSelect
      }else{
            fun <- ordSmooth
      }

      if(!is.null(seed)){
            if(!is.numeric(seed)){
                  stop('seed is not numeric')
            }else{
                  set.seed(seed) #torna os resultados replicáveis
            }
      }else{
            set.seed(NULL)
      }

      #n? considerar a última linha?
      if(lastRow){
            n <- nrow(x) - 1
            lastRow <- nrow(x)
      }else{
            n <- nrow(x)
            lastRow <- NULL
      }

      #espalha aleatoriamente os índices
      i <- sample(1:n, n)

      lambda <- 10^(elambda) #critérios de penalização

      #cuts do banco de dados para K-fold cross-validation
      folds <- cut(1:n,breaks=K,labels=FALSE)

      #matriz devianceXlambda (deviance é o resíduo de cada observação
      #no contexto do GLM)
      deviance <- matrix(nrow=0, ncol=length(lambda), byrow=TRUE)
      colnames(deviance) <- lambda

      #K loops de cross-validation
      for(j in 1:K){
            i_ <- which(folds==j,arr.ind=T) #indices de teste de i
            itest <- i[i_] #índices de teste do banco de dados
            itrain <- i[-i_] #índices de treino do banco de dados

            if(!onlyreturn){
                  message(paste('Actual cross-validation round:', j, 'de',
            K))
                  flush.console()
            }
```

```
      x_train <- x[c(itrain,lastRow),] #variáveis ordinais de treino
      y_train <- y[c(itrain,lastRow)] #variável resposta de treino
      if(!is.null(u)){ #u dado?
            u_train <- u[c(itrain,lastRow),] #variáveis categóricas
      #de treino
      }else{
            u_train <- NULL
      }
      if(!is.null(z)){ #z dado?
            z_train <- z[c(itrain,lastRow),] #variáveis numéricas de
            #treino
      }else{
            z_train <- NULL
      }

      x_test <- x[itest,] #variáveis ordinais de teste
      y_test <- y[itest] #variável resposta de teste
      if(!is.null(u)){ #u dado?
            u_test <- u[itest,] #variáveis categóricas de teste
      }else{
            u_test <- NULL
      }
      if(!is.null(z)){ #z dado?
            z_test <- z[itest,] #variáveis numéricas de teste
      }else{
            z_test <- NULL
      }

      #obtém um modelo do banco de dados de treino
      suppressMessages({
            train <- fun(x=x_train,y=y_train,u=u_train,z=z_train,
            lambda=lambda,model='logit')
      })

      #aplica o modelo sobre o banco de dados de teste. Retorna
      #uma matriz
      #probabilidadesXlambda
      test <- predict(train,newx=x_test,newu=u_test,
      newz=z_test,type='response')

      #transforma a matriz probabilidadeXlambda em
      #devianceXlambda (deviance: resíduo
      #individual, de cada observação)
      d <- apply(test, 2, function(x){
                  return(h$deviance(y=y_test, p=x, resdev=FALSE))
            })

      #une d a deviance
      deviance <- rbind(deviance, d)
}

#obtém o residual deviance (resíduo total por lambda) da amostra
#de teste completa
#o residual deviance é igual à soma dos quadrados dos resíduos
#das observações
resdev <- apply(deviance, 2, function(x){sum(x^2)})

lambda10 <- log10(lambda)

if(!onlyreturn){
```

```
            #plota o residual deviance em função do log10(lambda)
            plot(resdev ~ lambda10, type='l',
            xlab=expression('log'[10](lambda)),
                  ylab=ylab, frame.plot=FALSE)
      }

      min_lambda <- lambda10[which(resdev==(min(resdev)[1]))]

      if(!onlyreturn){
            message(paste('O valor mínimo é:', min_lambda, 'que
            corresponde a um lambda de', 10^min_lambda))
      }else{
            return(10^min_lambda)
      }
}

#Roda h$chooselambda R vezes, buscando evitar que haja viés associado à
#escolhe de indivíduos a
#compor cada partição, como poderia ocorrer com a validação cruzada normal.
#Retorna uma lista com os elementos: lambda - vetor com os lambda
#resultantes de cada round; percentiles - percentis de
#5%, 50% e 95% dos lambda. Segundo Roberts2014.
#R: número de repetições
#os demais argumentos são passados a h$chooselambda em cada round
h$repeatedlambda <- function(x, y, u=NULL, z=NULL, R, select=FALSE,
elambda, lastRow=TRUE, K=10){
      out <- list(lambda=c(), percentiles=c())

      for(i in 1:R){
            message(cb('Actual round R of repeated cross-validation: ',i,'
            of ',R))
            flush.console()
            out$lambda <- c(out$lambda,

      h$chooselambda(x=x,y=y,u=u,z=z,select=select,elambda=elambda,lastRow=
      lastRow,K=K,seed=NULL,onlyreturn=T)
            )
      }

      out$percentiles <- quantile(out$lambda, c(.05,.5,.95))

      return(out)
}

#Plota intervalos de confiança (percentil de p%, como calculado pelo pacote
#boot e gerado via bootstrap) para os coeficientes das variáveis preditoras
consideradas em ordPens::ordSelect ou ordPens::ordSmooth.
#No mesmo gráfico, são plotados os coeficientes não penalizados, em cor
#cinza (referentes a lambda[1]).
#x, y, u, z: Argumentos a passar a ordPens::ordSelect ou
#ordPens::ordSmooth, formatados corretamente (x, u e z data.frames apenas
#com colunas numéricas, y vetor numérico resposta com valores 0 ou 1).
#Consultar a ajuda para mais detalhes. x e y são obrigatórios.
#lambda: dois valores de lambda, o primeiro deve ser 0 ou um valor bem
#pequeno, equivalente a não penalizar os coeficientes. O segundo é o valor
#ótimo de penalização obtido, por exemplo, por #h$chooselambda. Esse é o
#valor de penalização considerado na modelagem em cada replicação R.
#select: se TRUE, a modelagem será feita por ordPens::ordSelect, se FALSE,
#por ordPens::ordSmooth
#lastRow: Se TRUE, há uma linha a mais (a última de x, y, u e z) que existe
#apenas para indicar o nível máximo das variáveis ordinais, e não entra na
```

```
#computação. Mais detalhes na ajuda do pacote ordPens.
#p: gerar intervalos de confiança de p% baseados em percentil, como
#calculado pelo pacote boot.
#ncol: quantas colunas considerar no dispositivo gráfico?
#R: núm. de replicações a considerar no bootstrap
#seed: fixa o seed do gerador de números aleatórios (torna os resultados
#replicáveis). Passe NULL
#para não fixar o seed.
#xlab: x label de cada gráfico
#ylab: y label de cada gráfico
h$penIC <- function(x, y, u=NULL, z=NULL, lambda, select=FALSE,
lastRow=TRUE, p=95, ncol=4, R=10, seed=123456, xlab='level', ylab='Dummy
coefficient'){
      require(ordPens)
      require(Hmisc)
      require(boot)

      if(length(lambda)!=2){
            stop('lambda should have length 2')
      }
      if(!is.numeric(y)){
            stop('y should be numeric: 0 or 1')
      }

      if(select){
            fun <- ordSelect
      }else{
            fun <- ordSmooth
      }

      if(!select & lambda[1]==0){
            warning('If lambda[1] is zero, ordSmooth can fail. You can set
            a small value, as 1e-5.')
      }

      if(!is.null(seed)){
            if(!is.numeric(seed)){
                  stop('seed is not numeric')
            }else{
                  set.seed(seed) #torna os resultados replicáveis
            }
      }else{
            set.seed(NULL)
      }

      #quantos gráficos serão plotados? (um para cada variável ordinal,
      #um pra cada variável categórica
      # e um para todas as contínuas.
      ngraph <- ncol(x)
      if(!is.null(u)){
            ngraph <- ngraph + ncol(u)
      }
      if(!is.null(z)){
            ngraph <- ngraph + 1
      }
      if((ngraph %% 2) > 0){ #se número ímpar, soma com 1
            ngraph <- ngraph + 1
      }

      if((ngraph%%ncol)!=0){
            stop(paste('Revise o num. de colunas do dispositivo gráfico:',
```

```
            ngraph, 'células para', ncol, 'colunas'))
      }

      #n? considerar a última linha?
      if(lastRow){
            n <- nrow(x) - 1
            lastRow <- nrow(x)
      }else{
            n <- nrow(x)
            lastRow <- NULL
      }

      #dados (x, y, u, z). Se a última linha é especial, é criado lr
      #apenas com ela
      data <- list(); data$x <- x; data$y <- y; data$u <- u; data$z <- z
      if(!is.null(lastRow)){ #a última linha contém apenas o último
      #nível das variáveis ordinais?
            lr <- list() #data apenas com a última linha
            lr$x <- data$x[lastRow,,drop=F]; data$x <- data$x[-lastRow,,
            drop=F] #x
            lr$y <- data$y[lastRow]; data$y <- data$y[-lastRow] #y
            if(!is.null(data$u)){
                  lr$u <- data$u[lastRow,,drop=F]; data$u <- data$u[-
lastRow,,drop=F] #u
            }
            if(!is.null(data$z)){
                  lr$z <- data$z[lastRow,,drop=F]; data$z <- data$z[-
lastRow,,drop=F] #z
            }
      }else{
            lr <- NULL
      }


      #mensagens de validação
      message('Variáveis sendo consideradas:')
      message(paste(c(names(x), names(u), names(z), 'y'), collapse=' '))

      mod0 <- fun(x=x,y=y,u=u,z=z,lambda=lambda[1],model='logit') #modelo
      #não penalizado

      #mantém um track da sequência de boot
      i_boot_track <<- 0

      #função que gera a estatística em cada réplica do bootstrap
      stat <- function(data, y0, u0, z0, indices, lr, lambda){
            x_ <- data[indices,,drop=F] #variáveis ordinais do bootstrap
            if(!is.null(lr)){ #une a última linha
                  x_ <- rbind(x_, lr$x)
            }
            y_ <- y0[indices] #variável resposta do bootstrap
            if(!is.null(lr)){ #une a última linha
                  y_ <- c(y_, lr$y)
            }
            if(!is.null(u0)){ #variáveis categóricas do bootstrap
            #(podem não existir)
                  u_ <- u0[indices,,drop=F]
                  if(!is.null(lr$u)){ #une a última linha
                        u_ <- rbind(u_, lr$u)
                  }
            }else{
```

```
            u_ <- NULL
      }
      if(!is.null(z0)){ #variáveis numéricas do bootstrap (podem
      #não existir)
            z_ <- z0[indices,,drop=F]
            if(!is.null(lr$z)){ #une a última linha
                  z_ <- rbind(z_, lr$z)
            }
      }else{
            z_ <- NULL
      }

      #consegue o modelo penalizado
      suppressMessages({
            mod <- fun(x=x_,y=y_,u=u_,z=z_,lambda=lambda,
            model='logit')
      })

      #andamento do boot
      message(paste('Round atual:', i_boot_track, 'de', R))
      flush.console()
      i_boot_track <<- i_boot_track + 1

      #vetor com os coeficientes relativos ao lambda
      return(mod$coefficients[,1])
}

#bootstrap
boot <- boot(data=data$x, y0=data$y, u0=data$u, z0=data$z,
statistic=stat, R=R, lr=lr, lambda=lambda[2])

#constrói a matrix de resultados (coeficientes, lower CI e upper CI)
ci <- matrix(nrow=0, ncol=3)
colnames(ci) <- c('coef','lower','upper')

if(p<0 | p>100){
      stop('p should be between 0 and 100')
}
message('Considerando IC de ', p, '% do bootstrap.')

#preenche o objeto ci, uma variável por vez
for(i in 1:ncol(boot$t)){
      ci.boot <- boot.ci(boot, type='perc', index=i, conf=p/100)
      ci <- rbind(ci, c(ci.boot$t0, ci.boot$percent[,4:5]))
}
rownames(ci) <- names(boot$t0)

layout(matrix(1:ngraph, ncol=ncol, byrow=T))

#define ylim. O mínimo será o menor valor entre os coeficientes
#não penalizados e o lower CI. O máximo será o maior valor entre
#os coeficientes não penalizados e o upper CI. (exceto o intercepto)
ymin <- min(c(ci[-1,'lower'], mod0$coefficients[-1,1]))
ymax <- max(c(ci[-1,'upper'], mod0$coefficients[-1,1]))
ylim <- c(ymin,ymax)

xlev <- mod0$xlevels #vetor com o máximo de níveis das
#variáveis ordinais

i_ini <- 2 #ci[i,] onde iniciam os coeficientes das
#variáveis ordinais
```

```
#plota as variáveis ordinais
for(i in 1:length(xlev)){
      name <- toupper(names(xlev)[i]) #nome da variável ordinal

      levs <- xlev[i] #total de níveis da variável ordinal atual

      j0 <- i_ini #índice (linha) inicial da variável ordinal em ci
      j1 <- i_ini + levs - 1 #índice (linha) final da variável
      #ordinal em ci

      coef <- ci[j0:j1,1] #coeficientes reais dummy da variável
      #índice i
      lower <- ci[j0:j1,2] #limite inferior do intervalo de confiança
      upper <- ci[j0:j1,3] #limite superior do intervalo de confiança
      coef0 <- mod0$coefficients[j0:j1,1] #coeficientes não
      #penalizados (lambda[1])

      #plota os coeficientes obtidos
      plot(1:levs,coef,type='b',frame.plot=F,cex=1.5,lwd=1,xlab=xlab,
            ylab=ylab,ylim=ylim,main=name,xaxt='n')

      #eixo
      axis(side=1,at=1:levs,labels=1:levs)

      #plota os coeficientes não penalizados (mod0)
      lines(1:levs,coef0,type='b',pch=1.5,col='gray')

      #plota o intervalo de confiança (percentil) obtido do bootstrap

      errbar(1:levs,coef,yplus=upper,yminus=lower,add=T,cex=0,
      ylim=ylim,cap=0.02)

      i_ini <- i_ini + levs #recicla i_ini
}

if(!is.null(u)){
      ulev <- mod0$ulevels #vetor com o máximo de níveis das
      #variáveis categóricas

      #plota as variáveis categóricas
      for(i in 1:length(ulev)){
            name <- toupper(names(ulev)[i]) #nome da variável

            levs <- ulev[i] #total de níveis da variável atual

            j0 <- i_ini #índice (linha) inicial da variável em ci
            j1 <- i_ini + levs - 1 #índice (linha) final da variável
            #em ci

            coef <- ci[j0:j1,1] #coeficientes reais dummy da variável
            #índice i
            lower <- ci[j0:j1,2] #limite inferior do intervalo de
            #confiança
            upper <- ci[j0:j1,3] #limite superior do intervalo de
            #confiança
            coef0 <- mod0$coefficients[j0:j1,1] #coeficientes não
            #penalizados (lambda[1])

            #plota os coeficientes obtidos
```

```
                plot(1:levs,coef,type='p',frame.plot=F,cex=1.5,lwd=1,
                xlab=xlab,ylab=ylab,ylim=ylim,xlim=c(.8,levs+.2),
                main=name,xaxt='n')

                #eixo
                axis(side=1,at=1:levs,labels=rownames(ci)[j0:j1],las=3)

                #plota os coeficientes não penalizados (mod0)
                points(1:levs,coef0,pch=1.5,col='gray')

                #plota o intervalo de confiança (percentil) obtido
                #do bootstrap

                errbar(1:levs,coef,yplus=upper,yminus=lower,add=T,cex=0,
                ylim=ylim)

                i_ini <- i_ini + levs #recicla i_ini
          }
    }

    if(!is.null(z)){
          znum <- mod0$zcovars #vetor com o número de variáveis contínuas

          j0 <- i_ini #índice inicial das variáveis numéricas em ci
          j1 <- nrow(ci) #índice final das variáveis numéricas em ci

          name <- toupper(rownames(ci)[j0:j1]) #nomes das variáveis
          #contínuas

          coef <- ci[j0:j1,1] #coeficientes reais das variáveis
          lower <- ci[j0:j1,2] #limite inferior do intervalo de confiança
          upper <- ci[j0:j1,3] #limite superior do intervalo de confiança
          coef0 <- mod0$coefficients[j0:j1,1] #coeficientes não
          #penalizados (lambda[1])

          #plota os coeficientes obtidos

    plot(1:znum,coef,type='p',frame.plot=F,cex=1.5,lwd=1,xlab='var',

    ylab='Coefficientes',ylim=ylim,xlim=c(.8,znum+.2),main='NUMERIC',
    xaxt='n')

          #eixo
          axis(side=1,at=1:znum,labels=name,las=3)

          #plota os coeficientes não penalizados (mod0)
          points(1:znum,coef0,pch=1.5,col='gray')

          #plota o intervalo de confiança (percentil) obtido do bootstrap

          errbar(1:znum,coef,yplus=upper,yminus=lower,add=T,cex=0,
          ylim=ylim)
    }

    include_zero <- ifelse(ci[,'lower']<=0 & ci[,'upper']>=0, 1, 0)
    View(cbind(ci,include_zero))
}

#Plota medidas de desempenho com base nos resultados de K-fold cross
#validation: acurácia, sensibilidade, especificidade, AUC e Press'Q, além
#da soma dos quadrados dos resíduos(deviance) do modelo de treino
```

```
#quando aplicado sobre os mesmos dados de treino.
#x, y, u, z: Argumentos a passar a ordPens::ordSelect ou
#ordPens::ordSmooth, formatados corretamente
#(x, u e z data.frames apenas com colunas numéricas, y vetor
#numérico resposta com valores 0 ou 1).
#Consultar a ajuda para mais detalhes. x e y são obrigatórios.
#lambda: valor único de lambda. Valor ótimo de penalização obtido, por
#exemplo, por h$chooselambda. Esse é o valor de penalização considerado na
#modelagem em cada replicação K.
#select: se TRUE, a modelagem será feita por ordPens::ordSelect, se FALSE,
#por ordPens::ordSmooth
#lastRow: Se TRUE, há uma linha a mais (a última de x, y, u e z) que existe
#apenas para indicar o nível máximo das variáveis ordinais, e não entra na
#computação. Mais detalhes na ajuda do pacote ordPens.
#K: núm. de rounds a considerar no cross-validation
#seed: fixa o seed do gerador de números aleatórios (torna os resultados
#replicáveis). Passe NULL para não fixar o seed.
#pressq: Se TRUE, será plotada também a estatística PressQ
#labels: nomes a aparecer como labels do eixo x do gráfico;
#title: título opcional do gráfico
h$performance <- function(x, y, u=NULL, z=NULL, lambda, K, select=FALSE,
lastRow=TRUE, seed=123456, pressq=TRUE,
labels=c('accuracy','sensibility','specificity','AUC','deviance'),title='')
{
      if(length(lambda)!=1){
            stop('lambda length should be 1')
      }else if(length(labels)!=5){
            stop('labels should have length 5')
      }

      require(ordPens)
      require(DescTools)
      require(caTools)
      if(!is.numeric(y)){
            stop('y should be numeric: 0 or 1')
      }

      if(select){
            fun <- ordSelect
      }else{
            fun <- ordSmooth
      }

      if(!is.null(seed)){
            if(!is.numeric(seed)){
                  stop('seed is not numeric')
            }else{
                  set.seed(seed) #torna os resultados replicáveis
            }
      }

      #n? considerar a última linha?
      if(lastRow){
            n <- nrow(x) - 1
            lastRow <- nrow(x)
      }else{
            n <- nrow(x)
            lastRow <- NULL
      }

      #espalha aleatoriamente os índices (randomly shuffle the data)
```

```
i <- sample(1:n, n)

#cuts do banco de dados para cross-validation
folds <- cut(1:n, breaks=K, labels=FALSE)

#matriz cvXdesempenho
performance <- matrix(nrow=K, ncol=6, byrow=TRUE)
colnames(performance) <- c('accuracy','sensibility','especificity',
'AUC','pressQ','deviance')

#K loops de cross-validation
for(j in 1:K){
      i_ <- which(folds==j,arr.ind=T) #indices de teste de i
      itest <- i[i_] #índices de teste do banco de dados
      itrain <- i[-i_] #índices de treino do banco de dados

      message(paste('Actual cross-validation round:', j, 'de', K))
      flush.console()

      x_train <- x[c(itrain,lastRow),] #variáveis ordinais de treino
      y_train <- y[c(itrain,lastRow)] #variável resposta de treino
      if(!is.null(u)){ #u dado?
            u_train <- u[c(itrain,lastRow),] #variáveis categóricas
            #de treino
      }else{
            u_train <- NULL
      }
      if(!is.null(z)){ #z dado?
            z_train <- z[c(itrain,lastRow),] #variáveis numéricas de
            #treino
      }else{
            z_train <- NULL
      }

      x_test <- x[itest,] #variáveis ordinais de teste
      y_test <- y[itest] #variável resposta de teste
      if(!is.null(u)){ #u dado?
            u_test <- u[itest,] #variáveis categóricas de teste
      }else{
            u_test <- NULL
      }
      if(!is.null(z)){ #z dado?
            z_test <- z[itest,] #variáveis numéricas de teste
      }else{
            z_test <- NULL
      }

      #obtém um modelo do banco de dados de treino
      suppressMessages({
            train <- fun(x=x_train,y=y_train,u=u_train,z=z_train,
            lambda=lambda,model='logit')
      })

      #adiante o banco de dados de treino não precisa ter a linha 64

      x_train <- x[itrain,] #variáveis ordinais de treino
      y_train <- y[itrain] #variável resposta de treino
      if(!is.null(u_train)){
            u_train <- u[itrain,] #variáveis categóricas de treino
      }
      if(!is.null(z_train)){
```

```
            z_train <- z[itrain,] #variáveis numéricas de treino
        }

        #aplica o modelo sobre o banco de dados de teste. predict
        #retorna uma matriz de probabilidades em linhas de acordo com o
        #lambda em colunas
        prob <- predict(train,newx=x_test,newu=u_test,
        newz=z_test,type='response')[,1]
        pred <- ifelse(prob>0.5, 1, 0) #predição para classes 0 ou 1

        #confusion matrix e estatísticas associadas
        conf <- Conf(pred, y_test, pos='1')

        #porcentagem de casos corretamente classificados
        accuracy <- conf$acc
        #habilidade de predizer a condição quando ela está presente
        sensibility <- conf$byclass['sens',1]
        #habilidade de predizer a ausência da condição quando ela não
        #está presente
        specificity <- conf$byclass['spec',1]
        #poder discriminante do classificador (AUC da probab. de teste
        #sobre a classificação real de teste)
        AUC <- colAUC(prob, y_test)
        #compara a performance do classificador com pura chance,
        #através de uma distribuição de qui-quadrado com 1 grau de
        #liberdade. Se Q>=3.84, há diferença ao nível de 5%
        #(Maroco, 2011)
        pressQ <- (n - conf$table[1,1]*2)^2 / (n * (2-1))
        #soma dos quadrados do resíduo (deviance do modelo de treino)
        desdev <- h$deviance(y=y_train, p=train$fitted[,1])

        #preenche deviance
        performance[j,] <- c(accuracy, sensibility, specificity, AUC,
        pressQ, desdev)
    }

    if(pressq){
        boxplot(performance[,5], xlab='PressQ',main=title)
        abline(h=3.84, col='blue')
        dev.new()
    }

    boxplot(performance[,-5],names=labels,main=title)

    View(round(performance,2))
}

#Exibe calibração do modelo com base em K-fold cross-validation. São
#plotados dois boxplots, um para cada nível (0 ou 1) da variável resposta
#y, e as probabilidades previstas para cada caso.
#Slope é a diferença entre médias entre as respostas 0 e 1.
#x, y, u, z: Argumentos a passar a ordPens::ordSelect ou
#ordPens::ordSmooth, formatados corretamente
#(x, u e z data.frames apenas com colunas numéricas, y vetor numérico
#resposta com valores 0 ou 1).
#Consultar a ajuda para mais detalhes. x e y são obrigatórios.
#lambda: valor único de lambda. Valor ótimo de penalização obtido, por
#exemplo, por h$chooselambda. Esse é o valor de penalização considerado na
#modelagem em cada replicação K.
#cv: Se FALSE, será obtido o gráfico de calibração a partir dos dados
#originais fornecidos (sem cross validation). Se TRUE, a calibração será
```

```
#obtida por meio de cross-validation.
#select: se TRUE, a modelagem será feita por ordPens::ordSelect, se FALSE,
#por ordPens::ordSmooth
#lastRow: Se TRUE, há uma linha a mais (a última de x, y, u e z) que existe
#apenas para indicar o nível máximo das variáveis ordinais, e não entra na
#computação. Mais detalhes na ajuda do pacote ordPens.
#K: núm. de rounds de cross validation
#seed: fixa o seed do gerador de números aleatórios (torna os resultados
#replicáveis). Passe NULL para não fixar o seed.
#xlab, ylab: labels x e y do gráfico.
#title: título do gráfico. Se 'default', será ifelse(cv==TRUE,'Cross-
#validated calibration','Model calibration')
#slabel: label do slope
h$calibration <- function(x, y, u=NULL, z=NULL, lambda, K=10, cv=TRUE,
select=FALSE, lastRow=TRUE, seed=123456, xlab='Response', ylab='Risk',
title='default', slabel='Slope'){
      if(length(lambda)!=1){
            stop('lambda length should be 1')
      }

      require(ordPens)
      require(DescTools)
      require(caTools)

      if(!is.numeric(y)){
            stop('y should be numeric: 0 or 1')
      }

      if(select){
            fun <- ordSelect
      }else{
            fun <- ordSmooth
      }

      main <- if(title=='default'){
            ifelse(cv==FALSE,'Model calibration','Cross-validated
            calibration')
      }else{
            title
      }

      if(!cv){ #deseja a calibração do modelo original, sem cross-
      #validation
            mod0 <- fun(x=x,y=y,u=u,z=z,lambda=lambda,model='logit')

            if(lastRow){
                  y_ <- y[-length(y)]
            }else{
                  Y_ <- y
            }

            h$plotprob(y_, mod0$fitted[,1], main=main, xlab=xlab,
            ylab=ylab, slabel=slabel)
            return()
      }

      if(!is.null(seed)){
            if(!is.numeric(seed)){
                  stop('seed is not numeric')
            }else{
                  set.seed(seed) #torna os resultados replicáveis
```

```
      }
}

#n? considerar a última linha?
if(lastRow){
      n <- nrow(x) - 1
      lastRow <- nrow(x)
}else{
      n <- nrow(x)
      lastRow <- NULL
}

#espalha aleatoriamente os índices (randomly shuffle the data)
i <- sample(1:n, n)

#cuts do banco de dados para cross-validation
folds <- cut(1:n, breaks=K, labels=FALSE)

#matriz [observations,c('y','prob')]
out <- matrix(nrow=0, ncol=2, byrow=FALSE)
colnames(out) <- c('y','prob')

#K loops de cross-validation
for(j in 1:K){
      i_ <- which(folds==j,arr.ind=T) #indices de teste de i
      itest <- i[i_] #índices de teste do banco de dados
      itrain <- i[-i_] #índices de treino do banco de dados

      message(paste('Actual cross-validation round:', j, 'de', K))

      x_train <- x[c(itrain,lastRow),] #variáveis ordinais de treino
      y_train <- y[c(itrain,lastRow)] #variável resposta de treino
      if(!is.null(u)){ #u dado?
            u_train <- u[c(itrain,lastRow),] #variáveis categóricas
            #de treino
      }else{
            u_train <- NULL
      }
      if(!is.null(z)){ #z dado?
            z_train <- z[c(itrain,lastRow),] #variáveis numéricas de
            #treino
      }else{
            z_train <- NULL
      }

      x_test <- x[itest,] #variáveis ordinais de teste
      y_test <- y[itest] #variável resposta de teste
      if(!is.null(u)){ #u dado?
            u_test <- u[itest,] #variáveis categóricas de teste
      }else{
            u_test <- NULL
      }
      if(!is.null(z)){ #z dado?
            z_test <- z[itest,] #variáveis numéricas de teste
      }else{
            z_test <- NULL
      }

      #obtém um modelo do banco de dados de treino
      suppressMessages({
            train <- fun(x=x_train,y=y_train,u=u_train,z=z_train,
```

```
                  lambda=lambda,model='logit')
            })

            #adiante o banco de dados de treino não precisa ter a
            #linha final

            x_train <- x[itrain,] #variáveis ordinais de treino
            y_train <- y[itrain] #variável resposta de treino
            if(!is.null(u_train)){
                  u_train <- u[itrain,] #variáveis categóricas de treino
            }
            if(!is.null(z_train)){
                  z_train <- z[itrain,] #variáveis numéricas de treino
            }

            #aplica o modelo sobre o banco de dados de teste. predict
            #retorna uma matriz de probabilidades em linhas de acordo com o
            #lambda em colunas
            prob <- predict(train,newx=x_test,newu=u_test,newz=z_test,
            type='response')[,1]

            out <- rbind(out, matrix(c(y_test,prob), ncol=2, byrow=F))
      }

      h$plotprob(out[,'y'], out[,'prob'], main=main, xlab=xlab,
      ylab=ylab, slabel=slabel)

      View(out)
}

#Dado um data.frame composto apenas de fatores, retorna um data frame com
#as mesmas colunas e níveis dos fatores originais, mas as linhas são todas
#as possíveis combinações (permutações) dos níveis dos fatores. Sempre
#mostra para o usuário o número de permutações possíveis e pergunta se
#deseja prosseguir.
#df: data.frame composto apenas de fatores
#all.fact: se TRUE, espera-se que todas as colunas sejam fatores. Se FALSE,
#espera-se que todos os fatores sejam codificados como inteiros (neste caso
#os níveis serão considerados como o resultado da função unique)
#return: data.frame
#ask: se TRUE, o usuário é perguntado se deseja prosseguir, dado o num de
#permutações que serão realizadas.
h$permute.df <- function(df, all.fact=TRUE, ask=TRUE){
      if(all.fact){
            if(any(unlist(lapply(df,
                  function(x){return(!is.factor(x))})))){
                  stop('df should have only factors')
            }
      }else{
            if(any(unlist(lapply(df,
                  function(x){return(!is.numeric(x))})))){
                  stop('df should have only integers')
            }
      }

      #mostra quantas linhas (permutações) serão obtidas e pede confirmação
      if(all.fact){
            perm <- prod(unlist(lapply(df,nlevels))) #número de linhas
            #necessárias
      }else{
            #número de linhas necessárias
```

```
            perm <- prod(unlist(lapply(df,function(x){length(1:max(x))})))
      }

      if(ask){
            cont <- readline(paste('Há',perm,'permutações entre os níveis
            dos fatores.\nPressione 0 para parar ou 1 para continuar:\n'))
            if(cont != 1){
                  stop('Parado pelo usuário')
            }
      }

      df0 <- df[0,] #data frame com as mesmas colunas e sem linhas

      if(all.fact){
            df1 <- expand.grid(lapply(df,levels))

            for(i in 1:ncol(df0)){
                  if(is.ordered(df0[,i])){
                        df1[,i] <- as.ordered(df1[,i])
                  }
            }
      }else{
            df1 <- expand.grid(lapply(df,function(x){1:max(x)}))
      }
      return(df1)
}

#plota num gráfico estilo calibração as probabilidades de todas as
#possíveis combinações dos níveis de variáveis ordinais e categóricas
#não ordinais. No eixo X estão os grupos previstos (0 ou 1) e no eixo
#Y as respectivas probabilidades. O modelo é obtido por
#ordPens::ordSmooth considerando o valor lambda fornecido, a
#variável resposta y, o banco de dados x composto apenas de variáveis
#ordinais (no formato aceito pelo pacote ordPens) e o banco de dados
#u composto apenas de variáveis categóricas (no mesmo formato). Em
#seguida esse banco de dados é substituído por todas as possíveis
#combinações dos níveis das variáveis supridas em x e u e sobre esse
#banco de dados "permutado" é obtida a probabilidade prevista com base
#no modelo citado. Em seguida esses resultados são plotados.
#y: variável resposta no formato aceito pelo pacote ordPens
#x: data.frame com variáveis ordinais no formato aceito pelo pacote ordPens
#u: data.frame com variáveis categóricas no formato aceito pelo pacote
#ordPens
#lambda: valor de penalização a considerar
#strip: se TRUE, será plotado via stripchart. Se FALSE, via boxplot;
#return: se TRUE, será retornado um data.frame com todas as possíveis
#permutações dos níveis dos fatores e o vetor de probabilidades previstas
#para cada linha, como uma coluna 'prob'.
#xlab,ylab main: labels dos eixos e títulos dos gráficos. Se
#main='default', será usado título padrão.
#ask: se TRUE, o usuário é perguntado se deseja prosseguir, dado o num de
#permutações que serão realizadas.
h$perm.smo <- function(y, x, u=NULL, lambda, strip=FALSE, return=FALSE,
xlab='Response', ylab='Predicted risk', main='default', ask=TRUE){
      require(ordPens)

      mod0 <- ordSmooth(y=y, x=x, u=u, lambda=lambda) #obtém o modelo

      #número de variáveis categóricas e ordinais
      nx <- ncol(x)
      nu <- if(is.null(u)){
```

```
      0
    }else{
      ncol(u)
    }

    #data.frame unindo x e u
    df <- if(is.null(u)){
      df <- x
    }else{
      df <- data.frame(x,u)
    }

    df <- h$permute.df(df, all.fact=F, ask=ask) #banco de dados permutado

    x_ <- df[,1:nx] #df permutado apenas com variáveis ordinais
    u_ <- if(is.null(u)){
      NULL
    }else{
      df[,(nx+1):ncol(df),drop=F]
    }

    #probabilidade prevista do df permutado
    prob <- predict(mod0,newx=x_,newu=u_,type='response')[,1]

    #devido a penalização, pode ser que as probabilidades
    #estejam além dos limites 0-1:
    prob <- ifelse(prob>1,1,prob)
    prob <- ifelse(prob<0,0,prob)

    resp <- ifelse(prob>0.5,1,0) #variável resposta

    tot <- length(resp) #total permutações
    posit <- length(resp[resp==1]) #=1
    negat <- length(resp[resp==0]) #=0

    message(cb('Total de ', tot, ' permutações, ', posit , '==1 e ',
      negat, '==0.'))

    main <- if(main=='default'){
      paste('Risk predicted by permuting all\n',
            'levels of factor predictors.\n',
            'Lambda:', lambda)
    }else{
      cb(main)
    }

    if(strip){
      stripchart(prob ~ resp, main=main, xlab=xlab, ylab=ylab)
    }else{
      boxplot(prob ~ resp, main=main, xlab=xlab, ylab=ylab)
    }


    if(return){
      return(data.frame(df,prob=prob))
    }
}

#Plota o perfil de fatores de um banco de dados composto apenas de fatores.
#rdm: dado o valor exato de um ponto, ele será desviado aleatoriamente até
#no máximo 'rdm' unidades de seu ponto original. Defina 0 para não desviar
```

```
#aleatoriamente os pontos.
#fac: fator opcional que informa permite colorir os indivíduos em grupos
#col: cores das linhas.
#cnames: nomes das variaveis a constar no eixo X. Se NULL, constará
#names(df)
h$plotfactor <- function(df, main=NULL, rdm=0.3, fac=NULL, col=NULL,
cnames=NULL){
      if(!is.data.frame(df)){
            stop('data.frame deve ser suprido')
      }else if(any(as.numeric(lapply(df, is.factor))==0)){
            stop('todas as colunas de df precisam ser fatores')
      }

      nc <- ncol(df) #total de colunas

      mx <- max(as.numeric(lapply(df, nlevels))) #maior nível encontrado

      #plot em branco
      plot.default(0, 0, type='n', xlim=c(0,nc+1), ylim=c(0,mx+1),
            xlab='Variáveis', ylab='Níveis', xaxt='n', yaxt='n',
            main=main)

      #labels do eixo X
      if(is.null(cnames)){
            labs <- names(df)
      }else{
            if(length(cnames)!=length(names(df))){
                  stop('cnames have wrong length')
            }
            labs <- cnames
      }

      #eixos
      axis(side=1, at=1:nc, labels=cnames, las=3)
      axis(side=2, at=1:mx, labels=F)

      #insere os níveis em gray
      for(i in 1:nc){
            vari <- df[,i] #variável

            lev <- levels(vari) # níveis da variável

            text(x=rep(i,nlevels(vari)), y=1:nlevels(vari),
                  labels=levels(vari), col='gray', cex=1)
      }

      #df codificado como numérico
      dfnum <- df
      dfnum <- lapply(df, function(x){
            y <- x
            levels(y) <- 1:nlevels(y)
            y <- as.numeric(x)
            return(y)
      })
      dfnum <- as.data.frame(dfnum)

      #define possíveis cores
      colr <- 'black'
      if(!is.null(fac)){
            if(!is.factor(fac)){
                  stop('fac should be a factor')
```

```
            }else if(length(fac) != nrow(df)){
                  stop('fac and df should have compatible size')
            }

            if(is.null(col)){
                  colr <- rainbow(nlevels(fac))[fac]
            }else{
                  if(nlevels(fac)!=length(col)){
                        stop('nlevels(fac) and length(col) should be
                        equal')
                  }
                  colr <- col[fac]
            }
      }else if(!is.null(col)){
            if(length(col)==1){
                  colr <- rep(col, nrow(df))
            }else{
                  colr <- col
            }
      }

      #insere os indivíduos
      for(i in 1:nrow(df)){
            rw <- dfnum[i,] #row atual (numérico)

            #x e y exatos
            y <- as.numeric(as.vector(rw))
            x <- 1:nc

            #x e y inexatos
            x <- x + (rdm * runif(n=length(x), min=-1, max=1))
            y <- y + (rdm * runif(n=length(y), min=-1, max=1))

            lines(x=x, y=y, col=ifelse(length(colr)==1,colr,colr[i]))
      }
}
```