**Seleção de covariáveis
para modelos de sobrevivência
via verossimilhança penalizada**

Jony Arrais Pinto Júnior

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO TÍTULO
DE
MESTRE EM CIÊNCIAS

Programa: Estatística

Orientador: Prof. Dr. Antonio Carlos Pedroso de Lima

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro do CNPq

São Paulo, fevereiro de 2009

**Seleção de covariáveis
para modelos de sobrevivência
via verossimilhança penalizada**

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e defendida por Jony Arrais Pinto Júnior
e aprovada pela Comissão Julgadora.

Banca Examinadora:

- Prof. Dr. Antonio Carlos Pedroso de Lima (orientador) - IME-USP.
- Prof. Dr. Heleno Bolfarine - IME-USP.
- Prof. Dr. Edwin Moises Ortega - ESALQ-USP.

À minha mãe.

Agradecimentos

Ao Antonio Carlos, pela orientação, apoio, paciência e incentivo.

À Gisela, pela ajuda e apoio, que foram fundamentais para o desenvolvimento deste trabalho.

Aos meus queridos amigos Patrícia, Márcio, Marcelo, Artemir, Éverton, Fernanda, Álvaro, Nubia, Victor, Cibele, Mari e Juliana que me deram suporte para enfrentar todos os desafios dessa trajetória.

Às minhas queridas amigas de Fortaleza que me incentivaram desde o início e ainda hoje me perguntam quando eu voltarei para ficar definitivamente entre elas, Eugenia, Débora, Cibele, Aline, Ticiane, Neila e Fernandinha.

Aos professores Maurício, Manoel e Luis, que me conduziram por esse caminho de forma brilhante.

À minha família, que mesmo de longe foi parte fundamental neste processo.

Aos meus pais que me ensinaram a valorizar minhas conquistas e são os principais responsáveis por essa.

Resumo

A seleção de variáveis é uma importante fase para a construção de um modelo parcimonioso. Entretanto, as técnicas mais populares de seleção de variáveis, como, por exemplo, a seleção do melhor subconjunto de variáveis e o método *stepwise*, ignoram erros estocásticos inerentes à fase de seleção das variáveis. Neste trabalho, foram estudados procedimentos alternativos aos métodos mais populares para o modelo de riscos proporcionais de Cox e o modelo de Cox com fragilidade gama. Os métodos alternativos são baseados em verossimilhança penalizada e diferem dos métodos usuais de seleção de variáveis, pois têm como objetivo excluir do modelo variáveis não significantes estimando seus coeficientes como zero. O estimador resultante possui propriedades desejáveis com escolhas apropriadas de funções de penalidade e do parâmetro de suavização λ . A avaliação desses métodos foi realizada por meio de simulação e uma aplicação a um conjunto de dados reais foi considerada.

Palavras-chave: seleção de variáveis, verossimilhança penalizada, funções de penalidade.

Abstract

Variable selection is an important step when setting a parsimonious model. However, the most popular variable selection techniques, such as the best subset variable selection and the stepwise method, do not take into account inherent stochastic errors in the variable selection step. This work presents a study of alternative procedures to more popular methods for the Cox proportional hazards model and the frailty model. The alternative methods are based on penalized likelihood and differ from the usual variable selection methods, since their objective is to exclude from the model non significant variables, estimating their coefficient as zero. The resulting estimator has nice properties with appropriate choices of penalty functions and the tuning parameter λ . The assessment of these methods was studied through simulations, and an application to a real data set was considered.

Keywords: variable selection, penalized likelihood, penalty functions.

Sumário

1	Introdução	1
2	Seleção de Variáveis via Verossimilhança Penalizada	5
2.1	Log-verossimilhança penalizada	7
2.2	Funções de penalidade	9
3	Modelos de Riscos Proporcionais e Fragilidade Gama	27
3.1	Modelo de Riscos Proporcionais de Cox	29
3.1.1	Inclusão da penalidade na verossimilhança parcial de Cox	33
3.1.2	Propriedade <i>oracle</i> para o modelo de riscos proporcionais de Cox	35
3.2	Modelo de Cox com Fragilidade Gama	39
3.2.1	Inclusão da penalidade na verossimilhança do modelo de Cox com fragilidade gama	45
3.2.2	Propriedade <i>oracle</i> para o modelo de Cox com fragilidade gama	47
3.3	Aproximação para as funções de penalidade	48

3.4	Estimação do parâmetro λ por validação cruzada	50
4	Estudo de Simulação e Aplicação	53
4.1	Estudo de simulação	53
4.1.1	Simulação para o modelo riscos proporcionais de Cox	55
4.1.2	Simulação para o modelo de Cox com fragilidade gama	61
4.2	Aplicação a dados reais	71
5	Considerações Finais e Estudos Futuros	75
A	Aspectos Computacionais	77

Lista de Tabelas

3.1	Tempos de sobrevivência, em dias, dos pacientes dos grupos 1 e 2.	34
4.1	MERM estimado e número médio de coeficientes nulos para o modelo de riscos proporcionais de Cox.	56
4.2	SD e SD_{dma} dos coeficientes não nulos para o modelo de riscos proporcionais de Cox. .	58
4.3	MERM estimado e número médio de coeficientes nulos para o modelo de Cox com fragilidade gama.	64
4.4	SD e SD_{dma} dos coeficientes não nulos para o modelo de Cox com fragilidade gama. .	67
4.5	Coeficientes e Erros padrões estimados.	73

Lista de Figuras

2.1	Log-verossimilhança de uma $N(\mu, 4)$	8
2.2	Penalidade entropia na log-verossimilhança de uma $N(\mu, 4)$	11
2.3	Funções de penalidade.	14
2.4	Gráfico da $\partial QP/\partial\beta$ com penalidade HT.	16
2.5	Gráfico da função QP com penalidade HT.	18
2.6	Gráfico da $\partial QP/\partial\beta$ com penalidade L_1	19
2.7	Gráfico da função QP com penalidade L_1	21
2.8	Gráfico da $\partial QP/\partial\beta$ com penalidade SCAD.	22
2.9	Gráfico da função QP com penalidade SCAD.	24
2.10	Estimadores baseados na verossimilhança penalizada.	25
3.1	Log-verossimilhança parcial penalizada.	35
3.2	Funções de penalidades e suas aproximações quadráticas.	49

4.1	Boxplot do Modelo de riscos proporcionais de Cox – 0% de censura.	59
4.2	Boxplot do Modelo de riscos proporcionais de Cox – 30% de censura.	60
4.3	Boxplot do Modelo de riscos proporcionais de Cox – 60% de censura.	61
4.4	Boxplot do Modelo de Cox com fragilidade gama – 0% de censura e correlação 0,2. . .	68
4.5	Boxplot do Modelo de Cox com fragilidade gama – 0% de censura e correlação 0,5. . .	69
4.6	Boxplot do Modelo de Cox com fragilidade gama – 30% de censura e correlação 0,2. .	70
4.7	Boxplot do Modelo de Cox com fragilidade gama – 30% de censura e correlação 0,5. .	71

Capítulo 1

Introdução

A seleção de variáveis é uma importante fase de qualquer análise estatística, quando se objetiva construir um modelo parcimonioso. Em análise de sobrevivência é coletado, em geral, um conjunto grande de covariáveis, com o objetivo de identificar fatores de risco para o evento de interesse ou para reduzir um possível viés na modelagem. Entretanto, modelos que englobam muitas covariáveis podem apresentar problemas computacionais como, por exemplo, instabilidade numérica no processo de maximização da verossimilhança, levando a estimativas imprecisas. Isto pode ocorrer quando existe multicolineariedade ou quando o número de observações não censuradas não é muito maior que o número de covariáveis. Portanto, é de interesse estudar técnicas que selecionem covariáveis de forma criteriosa.

Em modelos lineares clássicos, vários métodos podem ser utilizados na presença de multicolineariedade como, por exemplo, a regressão sobre cristas (Oishi, 1983). Nessa abordagem é verificada a existência de soluções distintas da de mínimos quadrados, que talvez sejam mais estáveis a pequenas perturbações nos dados. Embora viesadas, essas soluções representam estimativas mais estáveis que as convencionais não viesadas. Segundo Verweij e Van Houwelingen (1994), um método equivalente ao de regressão sobre cristas, no caso de modelos baseados em verossimilhanças, é a estimação dos coeficientes por verossimilhança penalizada.

Quanto ao problema da multicolineariedade, Huang e Harrington (2002) sugerem em seu tra-

balho duas possíveis abordagens: (i) modificar as equações de estimação para diminuir o efeito da multicolineariedade; (ii) reduzir o número de covariáveis utilizando uma técnica de seleção. Existem muitas técnicas de seleção de variáveis em modelos de regressão linear. Algumas delas, tais como *stepwise* e a seleção do melhor subconjunto de variáveis, foram estendidas para os modelos de análise de sobrevivência. Nesse contexto, estas técnicas objetivam selecionar um subconjunto de covariáveis significantes sobre as quais a função de taxa de falha depende. Apesar de úteis na prática, pois se encontram implementadas na maioria dos programas estatísticos, esses métodos ignoram erros estocásticos inerentes à fase de seleção das variáveis. Vale ressaltar que as propriedades dos estimadores resultantes dessas técnicas de seleção são amplamente desconhecidas (Fan e Li, 2002).

Com base nas informações anteriores, surge o interesse em estudar a verossimilhança parcial penalizada como um método de seleção de variáveis e as propriedades dos estimadores resultantes. Neste trabalho, será estudada uma classe de procedimentos de seleção de variáveis por meio de verossimilhanças penalizadas não côncavas. As propriedades dos estimadores resultantes para o modelo de riscos proporcionais de Cox e o modelo de Cox com fragilidade gama serão consideradas. Os procedimentos abordados neste trabalho foram propostos por Fan e Li (2001) e possuem como característica principal o fato de estimar parâmetros e selecionar covariáveis simultaneamente. Os métodos em questão diferem dos métodos usuais de seleção de covariáveis, pois têm como objetivo excluir do modelo variáveis não significantes estimando seus coeficientes como zero.

Nesse contexto, a função de penalidade tem importante papel na seleção das variáveis. Sendo assim, é aconselhável escolher funções de penalidade com características que garantam estimadores com propriedades interessantes. Dentre as características, podem-se citar, como em Fan e Li (2001), **não viés** (associada à superpenalização do estimador), **esparsividade** (relacionada à estimação de coeficientes como zero) e **continuidade** (relacionada à predição do modelo). Serão consideradas as penalidades HT, SCAD e a família L_q . Os estimadores resultantes da log-verossimilhança parcial penalizada, com alguma destas funções de penalidade, estimam os efeitos de variáveis sem importância como zero, e os coeficientes restantes são estimados como se fossem conhecidos, a priori, os coeficientes nulos (propriedade *oracle*).

As funções de penalidade, conforme será visto posteriormente, dependem de um parâmetro de suavização λ . Na prática, no entanto, não é fácil escolher esse valor apropriadamente, e algumas metodologias têm sido propostas na literatura (Fan e Li, 2001). Neste trabalho, considera-se uma metodologia para a escolha do parâmetro de suavização das penalidades baseada no valor preditivo do modelo, conforme Verweij e Van Houwelingen (1994). Em modelos lineares, o valor preditivo pode ser medido pela soma de quadrados de predição, que é uma medida de quão bem os valores ajustados podem prever as variáveis respostas observadas por meio de um modelo proposto. Em modelos baseados em verossimilhança parcial, o mesmo pode ser feito usando-se uma log-verossimilhança parcial de validação cruzada (Verweij e Van Houwelingen, 1993).

A seguir, apresenta-se a organização deste trabalho. No Capítulo 2, faz-se um breve histórico dos aspectos que motivaram a utilização de verossimilhanças penalizadas e do seu uso para selecionar variáveis. Introduz-se também a verossimilhança penalizada e as funções de penalidade que serão utilizadas para a composição dos métodos de seleção. Ainda nesse capítulo, será discutido o método de seleção de variáveis via verossimilhança penalizada para um exemplo envolvendo mínimos quadrados. No Capítulo 3, serão abordados os modelos de riscos proporcionais de Cox e modelo de Cox com fragilidade. Serão discutidas também a inclusão da penalidade na log-verossimilhança parcial e a propriedade *oracle* em cada modelo. No Capítulo 4, serão comparados os resultados das simulações para os modelos abordados no Capítulo 3 com o Critério de Informação *Akaike* (AIC) e com o modelo ótimo, que será definido posteriormente. Também serão considerados a aplicação dos métodos baseados em verossimilhança penalizada e o AIC a um conjunto de dados reais. No Capítulo 5, serão apresentados as considerações finais e possíveis estudos futuros. Por fim será apresentado um apêndice contendo aspectos computacionais.

Capítulo 2

Seleção de Variáveis via Verossimilhança Penalizada

Uma das grandes vantagens de trabalhar com a verossimilhança penalizada é que se selecionam e se estimam simultaneamente os coeficientes das variáveis explicativas. Dessa forma, os métodos apresentados neste trabalho são diferentes das abordagens tradicionais de seleção de variáveis, pois excluem variáveis do modelo, fazendo com que as estimativas de seus coeficientes sejam zero.

Algumas técnicas de seleção de variáveis, muito utilizadas no contexto de modelos de regressão linear, foram estendidas para problemas de análise de sobrevivência, por exemplo, o método *stepwise* e a seleção do melhor subconjunto de variáveis. Neste capítulo, apresenta-se o método de seleção de variáveis via verossimilhança penalizada para um exemplo simples de mínimos quadrados, com o objetivo de preparar o leitor para a extensão a ser feita no capítulo seguinte para modelos de análise de sobrevivência. Além disso, faz-se uma breve revisão bibliográfica, discute-se o conceito de verossimilhança penalizada e apresenta-se algumas funções de penalidade.

Os estimadores de máxima verossimilhança penalizada (EMVPE) surgiram em problemas de estimação não paramétrica de curvas ou funções, tornando-se uma alternativa quando os estimadores de máxima verossimilhança não podiam ser aplicados ou não conduziam a resultados satisfatórios. O uso de estimativas de máxima verossimilhança penalizada para a estimação de densidades foi sugerido por Good e Gaskins (1971). O EMVPE possui dois objetivos na estimação de curvas: (i) maximizar a probabilidade da amostra observada, medida pela log-verossimilhança usual; (ii) evitar curvas que

exibam grandes picos.

Métodos de estimação de densidades não paramétricos podem ser utilizados para estimar uma densidade f . Supondo que seja uma função suave, isto é, para um número especificado n de derivadas, tem-se $\int [f^{(n)}(t)]^2 dt < \infty$ (Simonoff, 2008). Nesse contexto, o estimador de máxima verossimilhança não paramétrico (EMVNP) de f não é um estimador adequado da densidade, pois envolve um conjunto de picos formado pelos dados observados. A fim de suavizar esses picos, pode-se utilizar a abordagem de Good e Gaskins (1971), que consiste em modificar a log-verossimilhança original, incorporando-se um termo que suaviza os picos do EMVNP de $f(\cdot)$. Define-se, então, a log-verossimilhança penalizada $lp(f)$ por

$$lp(f) = l(f) - \lambda \Phi(f),$$

em que $l(f)$ é a log-verossimilhança original de f , $\Phi(f)$ é uma função não negativa convenientemente escolhida, cuja magnitude decresce à medida que f se torna mais suave, e λ é um parâmetro de suavização que permite controlar a quantidade que é subtraída de $l(f)$, tornando $lp(f)$ mais suave. A forma de $\Phi(f)$ será objeto de um estudo mais detalhado nas seções seguintes. Uma discussão detalhada sobre verossimilhança penalizada pode ser vista em Silverman (1986).

Tibshirani (1996), usando verossimilhança penalizada para selecionar variáveis, propôs o método LASSO (*least absolute shrinkage and selection operator*) para modelos de regressão. No ano seguinte, Tibshirani (1997) estendeu o LASSO para o contexto de dados censurados, em que o autor apresentou um método de seleção para o modelo de riscos proporcionais de Cox. Fan e Li (2001) propuseram uma família de novos métodos de seleção de variáveis baseados em verossimilhança penalizada. Esses métodos são compostos pelas funções de penalidade HT, SCAD e a família L_q . Vale ressaltar que o LASSO é um membro da família L_q , quando $q = 1$. A abordagem foi aplicada em regressão linear, regressão linear robusta e modelos lineares generalizados. Posteriormente, Fan e Li (2002) estenderam esses novos métodos para o modelo de riscos proporcionais de Cox, incluindo o modelo com efeitos aleatórios. Na seção a seguir, faz-se uma apresentação mais formal da log-verossimilhança penalizada.

2.1 Log-verossimilhança penalizada

Conforme mencionado anteriormente, o parâmetro de suavização permite controlar os picos da verossimilhança penalizada e, assim, pode influenciar no valor do estimador de máxima verossimilhança. Nesta seção, considera-se o uso dessa metodologia na seleção de variáveis. A idéia é modificar a verossimilhança original, com o intuito de forçar valores nulos para os coeficientes de variáveis independentes cujo efeito não é estatisticamente significativo.

A fim de apresentar o procedimento geral, considere que $\{(\mathbf{x}_i, Y_i), i = 1, \dots, n\}$ é uma amostra aleatória em que Y_i é a variável resposta que depende de um vetor d -dimensional de covariáveis $\mathbf{x}_i = (x_{1i}, x_{2i}, \dots, x_{di})$. Assuma que $Y_i|\mathbf{x}_i$ tem uma densidade $f_i(y_i; \mathbf{x}_i^T \boldsymbol{\beta})$, $i = 1, \dots, n$, em que $\boldsymbol{\beta}$ é um vetor d -dimensional de parâmetros desconhecidos. Seja $l_i = \log f_i$ a log-verossimilhança condicional associada ao indivíduo i . A contribuição desse indivíduo na log-verossimilhança penalizada é dada por

$$lp_i(\boldsymbol{\beta}) = l_i(y_i; \mathbf{x}_i^T \boldsymbol{\beta}) - \sum_{j=1}^d \lambda_j p_j(|\beta_j|), \quad i = 1, \dots, n, \quad (2.1)$$

em que $p_j(\cdot)$ são funções de penalidade não negativas e λ_j são parâmetros de suavização.

A idéia do procedimento é deslocar a função de log-verossimilhança pelo fator $\sum_{j=1}^d \lambda_j p_j(|\beta_j|)$ em direção ao valor zero, de modo que os coeficientes desprezíveis, isto é, coeficientes associados a covariáveis com efeito desprezível na variável resposta sejam estimados como zero. Os coeficientes que sejam diferentes de zero são as variáveis selecionadas para a composição do modelo final. Observe que a penalidade depende dos valores de β_j e, assim, é possível deslocar o máximo da função de log-verossimilhança. No caso em que a influência da covariável sobre a variável resposta for desprezível, a verossimilhança penalizada objetiva fazer com que $\hat{\beta}_j = 0$.

Com o intuito de fixar idéias e analisar o efeito da penalização sem que se deseje fazer qualquer tipo de inferência, considere uma situação simples de uma amostra de tamanho 1 com distribuição $N(\mu, \sigma^2)$. Admita que a amostra fornece $x = 3$. Os gráficos da log-verossimilhança original, bem como

da log-verossimilhança penalizada usando a função de penalidade HT (a ser apresentada adiante), são apresentados na Figura 2.1, considerando-se $\lambda = 2$. Note que a log-verossimilhança do indivíduo tem o seu máximo deslocado quando é penalizada, fazendo com que o máximo da log-verossimilhança penalizada ocorra em $\hat{\mu} = 0$. Na log-verossimilhança usual, observe que o máximo ocorre em $\hat{\mu} = 3$.

Figura 2.1: Log-verossimilhança de uma $N(\mu, 4)$.

É importante ressaltar que em (2.1) as funções de penalidade $p_j(\cdot)$ e os parâmetros de suavização λ_j não são necessariamente os mesmos para todo j . Se existirem variáveis com poder interpretativo importante para o pesquisador, seus coeficientes podem não ser penalizados e suas significâncias estatísticas avaliadas da forma usual.

Para simplificar a notação, assuma que a mesma penalização será aplicada a todos os compo-

nentes de β , o que implica que a função de penalidade da log-verossimilhança do indivíduo i será da forma $\lambda p(|\beta_j|)$, $j = 1, \dots, d$, que será denotada por $p_\lambda(|\beta_j|)$. Observe que a função $p_\lambda(|\beta_j|)$ está definida estabelecendo-se uma dependência da função de penalidade com o parâmetro de suavização λ . Logo, a log-verossimilhança penalizada constituída por todos os elementos amostrais, como discutida em Fan e Li (2002), será expressa por

$$lp(\beta) = \sum_{i=1}^n l_i(y_i; \mathbf{x}_i^T \beta) - n \sum_{j=1}^d p_\lambda(|\beta_j|). \quad (2.2)$$

Extensões para situações nas quais existam diferentes funções de penalidades para os componentes de β não envolvem dificuldades extras, apenas a notação torna-se bastante onerosa.

2.2 Funções de penalidade

Existem na literatura algumas funções ou famílias de funções de penalidade bastante estudadas que têm mostrado bons resultados na seleção de variáveis. As funções de penalidade influenciam, em geral, propriedades dos estimadores de máxima verossimilhança penalizados. Segundo Fan e Li (2001), uma função de penalidade adequada deve resultar em um estimador com as seguintes propriedades:

- **Não viesado:** O estimador resultante deve ser o mais não viesado possível, não havendo uma superpenalização quando a influência da covariável na variável resposta é importante, evitando, assim, um viés desnecessário na modelagem.
- **Esparsividade:** Com uma escolha apropriada do parâmetro de suavização λ , a função de penalidade deve favorecer a estimação de componentes nulos do verdadeiro valor de parâmetros como zero com probabilidade tendendo à 1 conforme se aumenta o tamanho da amostra, mantendo-se a consistência para os componentes não nulos.
- **Continuidade:** O estimador resultante é contínuo para diferentes valores amostrais, a fim de evitar variação desnecessária na predição do modelo.

A propriedade de esparsividade está relacionada com o conceito de limitantes (*thresholdings*). Uma função $h(x)$ na qual os valores de $|x|$ menores que um valor especificado possuem imagens iguais a zero é uma função limitante (Antoniadis, 1997). Note que é interessante que os estimadores possuam essa característica, pois quando alguns coeficientes de β são iguais a zero, é possível excluir as correspondentes variáveis do modelo. Uma função limitante pode ser aquela em que todos os outros pontos que não possuam imagens iguais a zero atribuam o seu próprio valor como imagem. Neste caso, a função é limitante severa e é definida por

$$h(x) = \begin{cases} x, & \text{se } |x| > \lambda, \\ 0, & \text{caso contrário.} \end{cases}$$

No caso em que os pontos com imagens diferentes de zero possuam imagens iguais ao valor de x acrescido de uma constante, a função é chamada limitante suave, definida por

$$h(x) = \begin{cases} x - \lambda, & \text{se } |x| > \lambda, \\ 0, & \text{se } |x| \leq \lambda, \\ x + \lambda, & \text{se } |x| < -\lambda. \end{cases}$$

Algumas funções penalidade que têm sido bastante utilizadas na literatura serão descritas a seguir. Uma observação importante é que muitos critérios populares de seleção de variáveis são casos especiais da verossimilhança penalizada (2.2).

Considere inicialmente a função de penalidade denominada “entropia”, dada por $p_\lambda(|\beta|) = 0.5\lambda^2 I(|\beta| \neq 0)$, em que $I(\cdot)$ é a função indicadora. Observe que esta função penaliza igualmente a verossimilhança para valores de β diferentes de zero. Para uma amostra aleatória observada de n elementos, a forma geral da verossimilhança penalizada com a penalidade entropia pode ser reescrita como

$$\text{LP}(\beta) = \sum_{i=1}^n l_i(y_i; \mathbf{x}_i^T \beta) - 0.5n\lambda^2 |M|, \quad (2.3)$$

em que $|M| = \sum_{j=1}^d I(|\beta_j| \neq 0)$. Note que a dimensão do modelo é igual a $|M|$, o número de coeficientes não zeros no modelo.

Com escolhas apropriadas de λ , pode-se derivar de (2.3) alguns dos métodos usuais de seleção de variáveis. Por exemplo, o critério de Akaike (AIC) corresponde à escolha de $\lambda = \sqrt{2/n}$ e o critério Bayesiano de Schwarz (BIC) é obtido com $\lambda = \sqrt{(\log n)/n}$, embora esses critérios sejam motivados por princípios diferentes (Hunter e Li, 2005).

Figura 2.2: Penalidade entropia na log-verossimilhança de uma $N(\mu, 4)$.

A Figura 2.2 refere-se ao exemplo da $N(\mu, 4)$, citado anteriormente, com a penalidade entropia. O Gráfico 2.2.a apresenta as funções de log-verossimilhanças usual e penalizada, no caso em que o valor da covariável observada é $x = 3$ e $\lambda = 2$. Note que a log-verossimilhança penalizada é

descontínua em zero. No ponto zero, o valor da log-verossimilhança penalizada coincide com o valor da log-verossimilhança não-penalizada. Neste caso, observe que o máximo da log-verossimilhança penalizada ocorre em $\hat{\mu} = 0$, enquanto na verossimilhança usual o máximo ocorre em $\hat{\mu} = 3$. No Gráfico 2.2.b encontram-se as funções de log-verossimilhanças usual e penalizada quando $x = 5$ e $\lambda = 2$. Neste caso, observe que os máximos das funções de log-verossimilhanças penalizada e usual coincidem em $\hat{\mu} = 5$. Novamente foram ilustradas situações com o intuito de fixar idéias e analisar o efeito da função de penalidade na log-verossimilhança penalizada, sem que se deseje fazer qualquer tipo de inferência.

Ainda serão consideradas neste trabalho as seguintes funções de penalidade ou famílias de funções de penalidade:

- ***Hard thresholding (HT)*** .

A função penalidade *hard thresholding*, que será denotada em todo o trabalho por HT, foi proposta por Fan (1997) e melhorada por Antoniadis (1997). Esta penalidade tem sido bastante explorada na literatura e é definida por

$$p_{\lambda}(|\beta|) = \lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda). \quad (2.4)$$

Pela expressão (2.4), verifica-se que a penalização é constante para valores de $|\beta|$ maiores que λ , ou seja, a função de penalidade HT não superpenaliza valores grandes de $|\beta|$. Isso significa que o estimador resultante é não viesado.

- ***L_q*** .

A família de funções de penalidade L_q é dada por

$$p_{\lambda}(|\beta|) = \lambda |\beta|^q. \quad (2.5)$$

Essa família tem sido muito utilizada quando se considera a estimação por meio do método de mínimos quadrados penalizados. Um fato relevante na família L_q consiste na incapacidade de

manter os estimadores resultantes não viesados para parâmetros com valores absolutos grandes, pois há uma penalização excessiva. Quando $q = 2$, a penalidade conduz a uma regressão sobre cristas. Quando $q = 1$, a penalidade gera o processo LASSO, proposto por Tibshirani (1996).

- ***Smoothly Clipped Absolute Deviation (SCAD)***.

Fan (1997) propôs uma função de penalidade diferenciável e contínua denominada *smoothly clipped absolute deviation*, definida por

$$p'_\lambda(\beta) = \lambda \left\{ I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \right\} \quad \text{para } a > 2 \text{ e } \beta > 0, \quad (2.6)$$

em que $p'_\lambda(\beta)$ denota a derivada de $p_\lambda(\cdot)$ em relação a β e $(a\lambda - \beta)_+ = (a\lambda - \beta)I(a\lambda > \beta)$. Esta função, que foi motivada por estudos em ondaletas (Antoniadis, 1997), não superpenaliza valores grandes de $|\beta|$. Como sugerido em Fan e Li (2001), será utilizado o valor $a = 3,7$ para a aplicação da função de penalidade SCAD em todo o trabalho.

A Figura 2.3 mostra o comportamento de algumas funções de penalidade citadas anteriormente. Os gráficos das penalidades L_1 e L_2 mostram que, à medida que β cresce, o valor de $p_\lambda(|\beta|)$ também cresce, ou seja, há uma superpenalização dos parâmetros. O mesmo não se observa nos gráficos das funções de penalidade HT e SCAD, já que, a partir de um determinado valor essas funções ficam constantes. Para a construção dos gráficos utilizou-se $\lambda = 2,2$. Note que as funções de penalidade apresentadas são simétricas em torno do ponto zero e convexas no intervalo $(0, \infty)$.

As funções de penalidade HT, L_1 e SCAD serão discutidas num exemplo simples a seguir. Considere uma situação em que se utilizou o método de mínimos quadrados para estimação dos parâmetros. Denote por Z a variável resposta e suponha uma amostra de tamanho $n = 1$, ou seja, $Z = \beta + \epsilon$, em que β é o parâmetro de interesse e ϵ é o erro. Sem a inclusão da penalidade, deseja-se encontrar $\hat{\beta}$ que minimize $Q(\beta) = \frac{1}{2}(z - \beta)^2$ ou, equivalentemente, maximize $Q(\beta) = -\frac{1}{2}(z - \beta)^2$. O método de mínimos quadrados penalizado consiste em minimizar

Figura 2.3: Funções de penalidade.

$$QP(\beta) = \frac{1}{2}(z - \beta)^2 + p_\lambda(|\beta|). \quad (2.7)$$

Analisando primeiramente o problema de mínimos quadrados sem a penalidade, é claro que o estimador será da forma $\hat{\beta} = z$. Ao incluir a penalidade, a derivada primeira de (2.7) em relação a β é

$$\begin{aligned} \frac{\partial QP}{\partial \beta} &= -(z - \beta) + \text{sgn}(\beta)p'_\lambda(|\beta|) \\ &= \text{sgn}(\beta)[|\beta| + p'_\lambda(|\beta|)] - z, \end{aligned}$$

em que $\text{sgn}(\cdot)$ é uma função dicotômica que assume os valores 1, se β for positivo, e -1 , se β for negativo. Pela expressão acima, pode-se perceber que, se $p'_\lambda(|\beta|) = 0$ para valores grandes de $|\beta|$, o estimador resultante é igual a z quando $|z|$ for suficientemente grande. Por outro lado, na estrutura deste exemplo, valores grandes para $|\beta|$ fazem com que $|z|$ seja grande com alta probabilidade. Logo, como nessa situação $p'_\lambda(|\beta|) = 0$, o estimador de β será o estimador de mínimos quadrados usual ($\beta = z$), que é não viciado. Conclui-se que neste modelo, a condição $p'_\lambda(|\beta|) = 0$ para β grande é uma condição suficiente para garantir que o estimador resultante seja não viciado.

Para este exemplo serão considerados alguns casos particulares para $p_\lambda(|\beta|)$. Considere a função penalidade HT definida em (2.4). A derivada primeira de (2.7) é da forma

$$\begin{aligned} \frac{\partial \text{QP}}{\partial \beta} &= -(z - \beta) - 2\text{sgn}(\beta)(|\beta| - \lambda)I(|\beta| < \lambda) \\ &= \beta - 2\text{sgn}(\beta)(|\beta| - \lambda)I(|\beta| < \lambda) - z. \end{aligned}$$

A derivada pode ser reescrita em função do sinal de β e, para $\lambda > 0$, tem-se que

$$\frac{\partial \text{QP}}{\partial \beta} = \begin{cases} -\beta + 2\lambda - z, & \text{se } 0 < \beta < \lambda, \\ -\beta - 2\lambda - z, & \text{se } -\lambda < \beta < 0, \\ \beta - z, & \text{se } \beta \geq \lambda \text{ ou } \beta \leq -\lambda. \end{cases}$$

Para auxiliar na obtenção do estimador de mínimos quadrados penalizado, com penalidade HT, representou-se, por meio de gráficos, a derivada primeira da função QP, que está baseada nos possíveis valores de λ e Z . A Figura 2.4, neste caso, torna-se mais um instrumento útil na busca dos estimadores, pois $\partial \text{QP} / \partial \beta$ é descontínua no zero.

- Caso I: $z < -2\lambda$.

Pela Figura 2.4, verifica-se que o máximo da função QP deve ocorrer para $\beta < 0$. Assim, o mínimo pode ocorrer em $-\hat{\beta} - 2\lambda - z = 0$ ou $\hat{\beta} - z = 0$. Note que as derivadas segundas são $\partial^2 \text{QP} / \partial \beta^2 < 0$ e $\partial^2 \text{QP} / \partial \beta^2 > 0$, respectivamente. Isso implica que o ponto de mínimo ocorre em $\hat{\beta} - z = 0$, logo:

Figura 2.4: Gráfico da $\partial QP/\partial\beta$ com penalidade HT.

$$\frac{\partial QP}{\partial\beta} = 0 \quad \Leftrightarrow \quad \hat{\beta} = zI(z < -2\lambda) .$$

- Caso II: $-\lambda < z < \lambda$.

Observando-se a figura (2.4), verifica-se que este é o caso em que $\hat{\beta} = 0$. Os gráficos que representam esta situação são os casos (II.1), quando $0 < z < \lambda$ e (II.2), em que $-\lambda < z < 0$.

- Caso III: $z > 2\lambda$.

Este é o caso em que $\hat{\beta} > 0$. O mínimo pode ocorrer em $-\hat{\beta} + 2\lambda - z = 0$ ou $\hat{\beta} - z = 0$. Note que as derivadas segundas são $\partial^2 QP/\partial\beta^2 < 0$ e $\partial^2 QP/\partial\beta^2 > 0$, respectivamente. Isso implica que o ponto de mínimo ocorre em $\hat{\beta} - z = 0$, logo:

$$\frac{\partial \text{QP}}{\partial \beta} = 0 \Leftrightarrow \hat{\beta} = zI(z > 2\lambda).$$

- Caso IV: $-2\lambda < z < -\lambda$.

Este caso está representado pelo gráfico (IV) na Figura 2.4. Note que existe mais de um ponto crítico, ou seja, existe mais de um possível candidato a mínimo ou máximo, local ou global. O mínimo de $\partial \text{QP}/\partial \beta$, função descontínua, é obtido comparando-se o valor de $\text{QP}(\beta)$ nos pontos de mínimos obtidos nos intervalos abertos, excluindo-se os pontos de descontinuidade e comparando-os com o valor de QP nos pontos de descontinuidade. Note que z é negativo e assim $\beta < 0$.

No intervalo $-\lambda < \beta < 0$, $\partial \text{QP}/\partial \beta = 0 \Leftrightarrow -\hat{\beta} - 2\lambda - z = 0$. Note que $\partial^2 \text{QP}/\partial \beta^2 < 0$; logo, trata-se de ponto de máximo.

No intervalo $\beta \leq -\lambda$, $\partial \text{QP}/\partial \beta = 0 \Leftrightarrow \hat{\beta} - z = 0$. Note que $\partial^2 \text{QP}/\partial \beta^2 > 0$; logo, trata-se de ponto de mínimo e $\hat{\beta} = z$. Neste caso foi utilizado $\lambda = 2$ e $z = -3$. O valor de $\text{QP}(-3) = 4$.

Considerando o ponto de descontinuidade zero, tem-se que $\text{QP}(0) = 4,5$.

Neste caso, $\text{QP}(-3) < \text{QP}(0)$ e o estimador será $\hat{\beta} = z = -3$. Como pode ser observado na Figura 2.5.

- Caso V: $\lambda < z < 2\lambda$.

Novamente, neste caso, tem-se mais de um ponto crítico. O gráfico (V) ilustra o comportamento de $\partial \text{QP}/\partial \beta$, na Figura 2.4 e a função QP , na Figura 2.5. Note que z é positivo e, assim, $\beta > 0$.

Considerando o intervalo $0 < \beta < \lambda$, $\partial \text{QP}/\partial \beta = 0 \Leftrightarrow -\hat{\beta} + 2\lambda - z = 0$. Note que $\partial^2 \text{QP}/\partial \beta^2 < 0$. Assim, o $\hat{\beta}$ resultante é ponto de máximo.

No intervalo $\beta > \lambda$, $\partial \text{QP}/\partial \beta = 0 \Leftrightarrow \hat{\beta} - z = 0$. Note que $\partial^2 \text{QP}/\partial \beta^2 > 0$; logo, trata-se de ponto de mínimo e $\hat{\beta} = z$. Neste caso utilizou-se $\lambda = 2$ e $z = 2,5$. O valor de $\text{QP}(2,5) = 4$.

Considerando o ponto de descontinuidade zero, tem-se que $\text{QP}(0) = 3,125$.

Assim, tem-se que $\hat{\beta} = 0$. Esse fato pode ser observado graficamente na Figura 2.5.

Figura 2.5: Gráfico da função QP com penalidade HT.

Dos casos I, II, III, IV e V, tem-se que o estimador de mínimos quadrados penalizados pode ser descrito na forma

$$\begin{aligned} \hat{\beta} &= z, & \text{se } |z| \geq 2\lambda, \\ \hat{\beta} &= z \text{ ou } \hat{\beta} = 0, & \text{se } \lambda < |z| < 2\lambda, \\ \hat{\beta} &= 0, & \text{se } |z| < \lambda. \end{aligned}$$

Observe que quando $\lambda < |z| < 2\lambda$, o estimador pode assumir dois valores. Neste caso em que a verossimilhança é simples, é possível certificar-se de qual dos dois possíveis valores é o ponto de

mínimo global graficamente, pois no caso (V), no qual $\hat{\beta} = 0$, o ponto $\hat{\beta} = z = 2,5$ é ponto de mínimo.

Uma condição necessária e suficiente para que a função de penalidade possua a propriedade de esparsividade é que ela seja irregular na origem, ou seja, implica que a mesma possua uma derivada descontínua no ponto zero. Verifica-se que se $p'_\lambda(0+) > 0$, o estimator resultante é automaticamente uma função limitante; neste caso, uma função limitante severa que estima coeficientes pequenos como zero. Sendo assim, a penalidade HT produz estimadores esparsos e não viesados.

Note que, considerando-se a penalidade L_1 definida em (2.5), quando $q = 1$, a derivada primeira de (2.7) para o exemplo de mínimos quadrados penalizado será

$$\begin{aligned}\frac{\partial \text{QP}}{\partial \beta} &= -(z - \beta) + \text{sgn}(\beta)\lambda \\ &= \beta + \text{sgn}(\beta)\lambda - z.\end{aligned}$$

Reescrevendo em função do sinal de β , tem-se que

$$\frac{\partial \text{QP}}{\partial \beta} = \begin{cases} \beta + \lambda - z, & \text{se } \beta \geq 0, \\ \beta - \lambda - z, & \text{se } \beta < 0, \end{cases}$$

Figura 2.6: Gráfico da $\partial \text{QP} / \partial \beta$ com penalidade L_1 .

O auxílio do gráfico da $\partial \text{QP} / \partial \beta$ será utilizado para a obtenção do estimador de mínimos quadrados penalizado com penalidade L_1 . Os possíveis casos foram representados em função dos valores de λ e Z na Figura 2.6. A seguir, supomos $\lambda > 0$.

- Caso I: $\lambda - z < 0$ e $-\lambda - z < 0$.

Neste caso, tem-se que $z > \lambda$ e $z > -\lambda$ simultaneamente e, assim, tem-se que $z > \lambda$. Pela Figura 2.6, verifica-se que o máximo da função QP deve ocorrer para $\beta \geq 0$, pois, lembrando que $\lambda > 0$, tem-se

$$\frac{\partial \text{QP}}{\partial \beta} = 0 \Leftrightarrow \hat{\beta} = z - \lambda \Leftrightarrow \hat{\beta} = \text{sgn}(z)(|z| - \lambda)I(z > \lambda).$$

- Caso II: $\lambda - z > 0$ e $-\lambda - z > 0$.

Neste caso, tem-se que $z < \lambda$ e $z < -\lambda$, o que implica em $z < -\lambda$. Observando-se a Figura 2.6, verifica-se que o máximo deve ocorrer para $\beta < 0$ e o estimador é da forma

$$\frac{\partial \text{QP}}{\partial \beta} = 0 \Leftrightarrow \hat{\beta} = \lambda + z \Leftrightarrow \hat{\beta} = \lambda - |z| \Leftrightarrow \hat{\beta} = \text{sgn}(z)(|z| - \lambda)I(-z > \lambda).$$

- Caso III: $\lambda - z > 0$ e $-\lambda - z < 0$.

Neste caso, tem-se que $z < \lambda$ e $z > -\lambda$, ou seja, $-\lambda < z < \lambda$. Este é o caso em que $\hat{\beta} = 0$, como se pode verificar na Figura 2.6.

- Caso IV: $\lambda - z < 0$ e $-\lambda - z > 0$.

Neste caso, tem-se que $z < -\lambda$ e $z > \lambda$ simultaneamente, de modo que não existe z que satisfaça as inequações.

A Figura 2.7 mostra o comportamento da função QP em cada um dos casos citados anteriormente. Utilizaram-se $\lambda = 2$ e $z = 4$, $z = -4$ e $z = 1,6$, para ilustrar os casos I, II e III. No caso I, note que $\hat{\beta} = 2$, pois $\hat{\beta} = \text{sgn}(z)(|z| - \lambda) = (|4| - 2) = 2$, no caso II, $\hat{\beta} = \text{sgn}(z)(| - 4| - 2) = -2$ e, finalmente, no caso III, em que $-\lambda < z < \lambda$, tem-se o mínimo de QP em $\hat{\beta} = 0$.

Figura 2.7: Gráfico da função QP com penalidade L_1 .

Combinando-se os casos I, II, III e IV, tem-se que o estimador de mínimos quadrados penalizados pode ser escrito na forma $\hat{\beta} = \text{sgn}(z)(|z| - \lambda)_+$, em que $(|z| - \lambda)_+ = (|z| - \lambda)I(|z| > \lambda)$, que é uma função limitante suave. Se considerada toda a família L_q , observe que satisfaz $p'_\lambda(0+) > 0$, quando $q \leq 1$, ou seja, nesse caso, os estimadores são funções limitantes. Outro ponto importante é que as soluções são contínuas para $q \geq 1$ (Fan e Li, 2001). Assim, L_1 parece ser a penalidade mais interessante da família L_q , pois possui duas das três propriedades desejadas em uma boa função de penalidade (esparsividade e continuidade).

Considere a função de penalidade SCAD definida em (2.6) no exemplo de mínimos quadrados penalizado. A derivada primeira de (2.7) é

$$\begin{aligned} \frac{\partial QP}{\partial \beta} &= -(z - \beta) + \lambda I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{a - 1} I(\beta > \lambda) \\ &= \beta + \lambda I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{a - 1} I(\beta > \lambda) - z, \end{aligned}$$

que pode ser reescrita como

$$\frac{\partial \text{QP}}{\partial \beta} = \begin{cases} \beta + \lambda - z, & \text{se } 0 < \beta \leq \lambda, \\ \beta + \frac{a\lambda - \beta}{a-1} - z, & \text{se } \lambda < \beta < a\lambda, \\ \beta - z, & \text{se } \beta \geq a\lambda. \end{cases}$$

Figura 2.8: Gráfico da $\partial \text{QP} / \partial \beta$ com penalidade SCAD.

Novamente será feita uma partição que cubra todos os possíveis valores de λ e Z para obtenção do estimador de mínimos quadrados penalizado com penalidade SCAD. A Figura 2.8 ilustra os possíveis casos.

- Caso I: $\lambda - z \geq 0$.

Observe que a função $\partial\text{QP}/\partial\beta$ é formada por três expressões. Para $0 < \beta \leq \lambda$, o gráfico é uma reta com intercepto igual a $\lambda - z$ que, neste caso, é positivo. Para $\lambda < \beta < a\lambda$, o gráfico também será uma reta. Observe que o coeficiente angular é igual a $(a-2)/(a-1)$ e, como $a > 2$, trata-se de uma reta crescente. Por fim, quando $\beta \geq \lambda$, o gráfico novamente será uma reta crescente. Pela Figura 2.8, obtém-se que $\hat{\beta} = 0$.

- Caso II: $\lambda - z < 0$ e $2\lambda - z \geq 0$.

Analogamente ao caso I, será feito o estudo da função $\partial\text{QP}/\partial\beta$. Considerando-se $\lambda - z < 0$ e $2\lambda - z \geq 0$, tem-se que $z > \lambda$ e $z \leq 2\lambda$, respectivamente. Assim, tem-se que $\lambda < z \leq 2\lambda$. Pela Figura 2.8, verifica-se que o máximo da função QP deve ocorrer para $0 < \beta \leq \lambda$ e, assim,

$$\frac{\partial\text{QP}}{\partial\beta} = 0 \quad \Leftrightarrow \quad \hat{\beta} = z - \lambda \quad \Leftrightarrow \quad \hat{\beta} = \text{sgn}(z)(|z| - \lambda).$$

- Caso III: $2\lambda - z < 0$ e $a\lambda - z \geq 0$.

Tem-se que, nesta situação, $z > 2\lambda$ e $z \leq a\lambda$ e, desta forma, tem-se que $2\lambda < z \leq a\lambda$. Pela Figura 2.8, verifica-se que o máximo da função QP deve ocorrer para $\lambda < \beta < a\lambda$, de tal forma que

$$\begin{aligned} \frac{\partial\text{QP}}{\partial\beta} = 0 & \quad \Leftrightarrow \quad \hat{\beta} + \frac{a\lambda - \hat{\beta}}{a-1} - z = 0 & \quad \Leftrightarrow \\ \hat{\beta} = \frac{(a-1)z - a\lambda}{a-2} & \quad \Leftrightarrow \quad \hat{\beta} = \frac{(a-1)z - \text{sgn}(z)a\lambda}{a-2}. \end{aligned}$$

- Caso IV: $a\lambda - z < 0$.

Por fim, é considerado o caso em que $a\lambda - z < 0$, ou seja, $z > a\lambda$. Pela Figura 2.8, verifica-se que o máximo da função QP deve ocorrer para $\beta \in (a\lambda, +\infty)$ e, assim,

$$\frac{\partial\text{QP}}{\partial\beta} = 0 \quad \Leftrightarrow \quad \hat{\beta} - z = 0 \quad \Leftrightarrow \quad \hat{\beta} = z.$$

Figura 2.9: Gráfico da função QP com penalidade SCAD.

A Figura 2.9 mostra o comportamento da função QP nos quatro casos citados anteriormente. Utilizou-se $a = 3$, $\lambda = 2$ e $z = 1, 2$, $z = 3, 8$, $z = 5, 1$ e $z = 7, 2$ para ilustrar os casos I, II, III e IV, respectivamente. É importante ressaltar que, satisfeita a condição do caso I, o máximo ocorre em $\hat{\beta} = 0$.

Observe que a penalidade SCAD está definida para $\beta > 0$. Para incluir o caso $\beta < 0$, utiliza-se $p'_\lambda(\beta)$ para obter a verossimilhança penalizada. Dos casos I, II, III e IV, considerando-se que os cálculos para $\beta < 0$ são similares, o estimador resultante fica dado por:

$$\hat{\beta} = \begin{cases} \operatorname{sgn}(z)(|z| - \lambda)_+, & \text{quando } |z| \leq 2\lambda, \\ \{(a-1)z - \operatorname{sgn}(z)a\lambda\}/(a-2), & \text{quando } 2\lambda < |z| \leq a\lambda, \\ z, & \text{quando } |z| > a\lambda. \end{cases}$$

Esta solução foi obtida por Fan (1997), no cenário de ondaletas. Note que o estimador é uma função limitante, pois valores os de $|z| < \lambda$ são estimados como zero. Esta função possui as três propriedades desejadas.

Figura 2.10: Estimadores baseados na verossimilhança penalizada.

Como citado anteriormente, os estimadores resultantes das penalidades L_1 , HT e SCAD são funções limitantes. A Figura 2.10 mostra o comportamento dos estimadores obtidos com base nas funções de penalidade para valores de z . As linhas tracejadas representam os estimadores não penalizados para β .

Os gráficos apresentados na Figura 2.10 são úteis para verificar, no exemplo de mínimos quadrados penalizados, as propriedades desejadas para as funções de penalidade, conforme discutido no começo desta seção. A propriedade de esparsividade está presente nas penalidades HT, L_1 e SCAD, pois para valores de $|z|$ menores que λ , os coeficientes são estimados como zero. A penalidade L_1 , contudo, apresenta a desvantagem de o estimador resultante estar deslocado por uma constante λ .

A penalidade L_1 é a única das três a não possuir a propriedade de não viés. Nos gráficos das penalidades HT e SCAD, fica clara a propriedade de não viés, pois a partir de um determinado valor (que é função de λ), a função do estimador penalizado é $\hat{\beta} = z$, coincidindo com o estimador não penalizado. Como pode ser verificado na Figura 2.10, o comportamento do gráfico com a penalidade SCAD diferencia-se dos demais por possuir um número maior de faixas de penalizações. A propriedade de continuidade é encontrada nas funções de penalidade SCAD e L_1 . Para ilustração dos gráficos das funções limitantes utilizou-se $a = 3, 7$ e $\lambda = 2, 2$.

Considere a seguinte partição de $\beta_0 = (\beta_{10}, \dots, \beta_{d0})^T = (\beta_{10}^T, \beta_{20}^T)$. Assuma sem perda de generalidade que os s primeiros termos de β_0 , agregados em β_{10} , são os componentes diferentes de 0. Os demais coeficientes, compondo β_{20} , são iguais a 0. É claro que, na aplicação prática, não se conhece a dimensão de β_{10} e nem a dimensão de β_{20} .

O estudo das propriedades assintóticas dos estimadores de máxima verossimilhança penalizada recaem em avaliar as condições para cada função de penalidade, necessárias para que o estimador resultante apresente algumas propriedades desejadas.

Uma das propriedades desejadas é conhecida como propriedade *oracle*. Em linhas gerais, a propriedade *oracle* consiste em garantir que para algumas funções de penalidade, com escolhas apropriadas de λ , o estimador resultante da verossimilhança penalizada possua a propriedade de esparsividade e que o vetor de parâmetros β_{10} possua normalidade assintótica.

A matriz de variância e covariâncias de $\hat{\beta}$ será obtida como sugerida em Grey (1992), isto é, considere I a matrix de informação de *Fisher* usual do modelo de riscos proporcionais de Cox, e

$$H = I - \begin{pmatrix} 0 & 0 \\ 0 & p'' \end{pmatrix}$$

em que p é a função de penalidade utilizada. Sendo assim, a estimativa da matriz de variância e covariância será dada por

$$V = H^{-1} I H^{-1}.$$

Capítulo 3

Modelos de Riscos Proporcionais e Fragilidade Gama

O modelo de riscos proporcionais de Cox (Cox, 1972) é um modelo semiparamétrico bastante utilizado em problemas de análise de sobrevivência. A popularidade deste modelo deve-se à flexibilidade causada pela presença de um componente não paramétrico. O modelo foi generalizado com a inclusão de efeitos aleatórios, dando origem ao modelo de Cox com fragilidade.

Existe na literatura um grande número de trabalhos que consideram esses modelos [ver, por exemplo Murphy e Van Der Vaart (2000) e Therneau et al. (2003)], uma vez que métodos de estimação e propriedades dos estimadores resultantes já são amplamente conhecidos (Kalbfleisch e Prentice, 2002). A seleção de covariáveis em modelos de riscos proporcionais, no entanto, é um problema que começou a ser considerado mais recentemente, e os métodos já propostos não são ainda muito difundidos. O objetivo principal deste capítulo é discutir a inclusão da penalidade na função de verossimilhança dos modelos considerados neste capítulo.

Serão apresentados inicialmente a definição do modelo de riscos proporcionais de Cox e do modelo de Cox com fragilidade gama. Considere uma amostra de n indivíduos que são acompanhados no intervalo de tempo $[0, \tau]$. A variável de interesse é $\{T_i : i = 1, \dots, n\}$ e representa o tempo até a ocorrência de um determinado evento de interesse no indivíduo i . Na terminologia de análise de sobrevivência, esse evento de interesse é também chamado de falha. No entanto, na prática, vários motivos podem impedir a observação do valor exato de T_i . Por exemplo, perda de acompanhamento

do indivíduo por qualquer motivo que não seja o de interesse, ou mesmo aqueles indivíduos que chegam ao final do período de observação sem a ocorrência da falha. Nota-se que esses estudos têm como característica observações incompletas ou parciais. Nessas situações, a única informação que temos é que o tempo para a ocorrência do evento é maior que um valor C_i , denominado tempo de censura.

A observação censurada não deve ser desconsiderada, pois fornece informação parcial sobre o tempo de falha. Quando se omitem dados censurados, ocorrem conclusões viciadas. De acordo com a estratégia adotada para determinar o fim do acompanhamento dos indivíduos, os mecanismos de censura podem ser classificados como: (i) censura do tipo I, (ii) censura do tipo II e (iii) censura aleatória, a mais comum em experimentos não controlados envolvendo seres humanos. Para maiores informações, ver Klein e Moeschberger (1997).

Neste trabalho considera-se a censura à direita, quando a variável resposta efetiva é representada por

$$Z_i = \min\{T_i, C_i\} \quad \text{e} \quad \delta_i = I(T_i \leq C_i) \quad (3.1)$$

em que δ_i é o indicador de falha. Note que os dados podem ser resumidos ao par de variáveis (Z_i, δ_i) , ou seja, $(C_i, 0)$ se uma censura foi observada e $(T_i, 1)$ se foi uma falha. Considere ainda que, para cada indivíduo i , observa-se o vetor de covariáveis $\mathbf{x}_i$, de dimensão $d \times 1$.

A sequência de variáveis aleatórias independentes e não negativas $T_i, i = 1, \dots, n$, com função de distribuição $F_i(\cdot)$, tem função de taxa de falha (ou função de risco) definida por

$$h(t|\mathbf{x}_i) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(T_i \in [t, t + \Delta t) | T_i \geq t, \mathbf{x}_i).$$

Suponha que foi observado $\{(\mathbf{x}_i, Z_i, \delta_i) : i = 1, \dots, n\}$, que constitui uma amostra aleatória independente de uma certa população. Denote por $S(\cdot|\mathbf{x})$ a função de sobrevivência, ou seja, $S(t|\mathbf{x}_i) = P(T_i \geq t|\mathbf{x}_i)$. A verossimilhança dos dados, como desenvolvida em Lawless (2003), pode ser escrita como

$$L = \prod_{\bar{c}} f(Z_i|\mathbf{x}_i) \prod_c S(Z_i|\mathbf{x}_i) = \prod_{\bar{c}} h(Z_i|\mathbf{x}_i) \prod_i^n S(Z_i|\mathbf{x}_i), \quad (3.2)$$

em que $\bar{c}$ e c indicam os dados não censurados e censurados, respectivamente, e $f(t|\mathbf{x})$ é a função densidade condicionada a $\mathbf{x}$. A obtenção de estimadores de parâmetros de interesse com base em (3.2) será discutida separadamente para o caso do modelo de riscos proporcionais de Cox e para o modelo de Cox com fragilidade gama a seguir.

3.1 Modelo de Riscos Proporcionais de Cox

A classe composta pelos modelos de riscos proporcionais é a mais popular em análise de sobrevivência. Proposto por Cox (1972), o modelo de riscos proporcionais de Cox é o mais comumente empregado para modelar a função de risco de um indivíduo com covariáveis. O motivo de se incluírem covariáveis nesse modelo pode estar no interesse de comparar grupos ou mesmo descobrir o efeito que determinadas covariáveis exercem sobre a função de taxa de falha. Como o modelo sob apreço é um modelo de regressão, ele também é conhecido na literatura como modelo de regressão de Cox.

O modelo de riscos proporcionais de Cox é definido por

$$h(t|\mathbf{x}) = h_0(t)g(\mathbf{x}^T\boldsymbol{\beta}), \quad (3.3)$$

em que $g(\cdot)$ é uma função não negativa que deve ser especificada, de modo que $g(0) = 1$, $\boldsymbol{\beta}$ é o vetor de parâmetros associado às covariáveis de dimensão $d \times 1$ e $h_0(t)$ é uma função não especificada. Observe que $h(t|\mathbf{x})$ é o produto de duas quantidades: $h_0(t)$, que é a parte não paramétrica, e $g(\mathbf{x}^T\boldsymbol{\beta})$, que é a parte paramétrica do modelo. Aqui será considerado $g(\mathbf{x}^T\boldsymbol{\beta}) = \exp(\beta_1x_1 + \dots + \beta_dx_d) = \exp(\mathbf{x}^T\boldsymbol{\beta})$, de forma que a função de taxa de falha (3.3) fica dada por

$$h(t|\mathbf{x}) = h_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta}). \quad (3.4)$$

Nos modelos paramétricos, é especificada a forma funcional da função de taxa $h_0(t)$, mas

no caso do modelo de regressão de Cox ela é arbitrária (porém não negativa). A forma como a componente não paramétrica é incluída faz com que este modelo possua modelos paramétricos como casos particulares, como, por exemplo, o modelo de regressão de Weibull. Observe que $h_0(t)$ pode ser interpretado como o risco de falha de indivíduos com todas covariáveis iguais a zero.

Note que o intercepto, presente nos modelos paramétricos de análise de sobrevivência, não se inclui em (3.4). Isso ocorre por causa da parte não-paramétrica no modelo que “absorve” essa constante. Observe ainda que a razão das taxas de falha de dois indivíduos diferentes é constante no tempo, ou seja, a razão das funções de taxa de falha para os indivíduos i e l é expressa por

$$\frac{h_i(t)}{h_l(t)} = \frac{h_0(t)\exp(\mathbf{x}'_i\boldsymbol{\beta})}{h_0(t)\exp(\mathbf{x}'_l\boldsymbol{\beta})} = \frac{\exp(\mathbf{x}'_i\boldsymbol{\beta})}{\exp(\mathbf{x}'_l\boldsymbol{\beta})} = \exp\{(\mathbf{x}_i - \mathbf{x}_l)'\boldsymbol{\beta}\},$$

que é independente do tempo. Por exemplo, se um indivíduo no início do estudo tem um risco de remissão igual a três vezes o risco de outro paciente, essa razão será a mesma para todo o período de acompanhamento. Essa propriedade é conhecida como proporcionalidade dos riscos e é a única suposição feita quando se utiliza o modelo de Cox. Assim, para o uso do modelo de regressão de Cox, é necessário que a suposição de proporcionalidade das taxas de falhas seja válida. Este modelo tem sido considerado em situações mais gerais, incluindo casos em que $h_0(\cdot)$ é diferente em estratos distintos da população e quando as covariáveis $\mathbf{x}_i$ são dependentes do tempo.

No modelo de regressão de Cox, assume-se ainda que T_i , o tempo de falha do indivíduo i , e C_i , o tempo de censura do indivíduo i , são condicionalmente independentes dado $\mathbf{x}_i$, e que o mecanismo de censura é não informativo (Klein e Moeschberger, 1997).

O interesse maior quando se utiliza o modelo de riscos proporcionais de Cox é a estimação dos parâmetros $\boldsymbol{\beta}$. Para isso, considere a seguinte ordenação dos N tempos de falhas observados, $t_1^0 < \dots < t_N^0$, indicados por $t_{(1)}^0, \dots, t_{(N)}^0$, e os vetores de covariáveis associados aos tempos de falha $\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(N)}$. Seja $R_j = \{i : Z_i < t_j^0\}$ o conjunto de indivíduos que estão em risco em t_j^0 , isto é, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j^0 . Considere ainda $H(t) = \int_0^t h(u)du$, a função de taxa de falha acumulada que, como o próprio nome

faz antever, fornece a taxa de falha acumulada.

Com base nas quantidades acima, pode-se escrever a verossimilhança dada em (3.2) para o modelo de riscos proporcionais como

$$L = \prod_{i=1}^N h_0(Z_{(i)}) \exp(\mathbf{x}_{(i)}^T \boldsymbol{\beta}) \prod_{i=1}^n \exp\{-H_0(Z_i) \exp(\mathbf{x}_i^T \boldsymbol{\beta})\}, \quad (3.5)$$

em que $H_0(\cdot)$ é a função de taxa de falha padrão acumulada. A log-verossimilhança de (3.5) será dada por

$$l(\boldsymbol{\beta}) = \sum_{i=1}^N [\log\{h_0(Z_{(i)})\} + \mathbf{x}_{(i)}^T \boldsymbol{\beta}] - \sum_{i=1}^n \{H_0(Z_i) \exp(\mathbf{x}_i^T \boldsymbol{\beta})\}. \quad (3.6)$$

Note que, se a função de taxa de falha for da forma $h_0(t, \boldsymbol{\theta})$, ou seja, uma função paramétrica, os estimadores para $(\boldsymbol{\theta}, \boldsymbol{\beta})$ podem ser obtidos pelo método de máxima verossimilhança usual.

Quando não é imposta uma forma paramétrica para $h_0(\cdot)$, o método de máxima verossimilhança usual não pode ser utilizado para obter os estimadores de $\boldsymbol{\beta}$. Cox (1975) propôs e formalizou o uso da máxima verossimilhança parcial para estimar os parâmetros do modelo de riscos proporcionais de Cox. Por esse enfoque, não é necessário estimar $h_0(t)$ [ou $H_0(t)$], e estimativas para $\boldsymbol{\beta}$ são facilmente encontradas. Quando se deseja estimar $S(t|\mathbf{x})$, torna-se necessário estimar $h_0(t)$. Isso é feito considerando-se a estimação de $H_0(t)$, que aqui será abordada utilizando-se a proposta de Breslow (1972). A idéia é estimar a função $H_0(t)$ por uma função escada com saltos em t_j^0 , ou seja, nos tempos de falha distintos. Essa função escada é da forma

$$H_0(Z_i) = \sum_{j=1}^N h_j I(i \in R_j), \quad (3.7)$$

em que $R_j = \{i : Z_i < t_j^0\}$. Quando se reescreve a log-verossimilhança (3.6) em função de (3.7),

tem-se que

$$l(\boldsymbol{\beta}) = \sum_{j=1}^N \{\log(h_j) + \mathbf{x}_j^T \boldsymbol{\beta}\} - \sum_{i=1}^n \left\{ \sum_{j=1}^N h_j I(i \in R_j) \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right\}. \quad (3.8)$$

Existem algumas maneiras de se obter a verossimilhança e, conseqüentemente, os estimadores de $\boldsymbol{\beta}$ no modelo de regressão de Cox e, neste trabalho, será utilizada a mesma abordagem de Fan e Li (2002). Para um dado j , $j = 1, \dots, N$, considere (3.8) como função de h_j e $\boldsymbol{\beta}$ uma constante. A derivada primeira de (3.8) com relação a h_j fica dada por

$$\begin{aligned} \frac{\partial l}{\partial h_j} &= \frac{1}{h_j} - \sum_{i=1}^n I(i \in R_j) \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \\ &= \frac{1}{h_j} - \sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta}). \end{aligned}$$

Fazendo $\partial l / \partial h_j = 0$, obtém-se

$$\hat{h}_j = \left\{ \sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right\}^{-1}. \quad (3.9)$$

Substituindo-se h_j por $\hat{h}_j$ em (3.8), tem-se que

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{j=1}^N \left[\log \left(\frac{1}{\sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right) + \mathbf{x}_{(j)}^T \boldsymbol{\beta} \right] \\ &\quad - \sum_{i=1}^n \left\{ \sum_{j=1}^N \frac{1}{\sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta})} I(i \in R_j) \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right\} \\ &= \sum_{j=1}^N \left[\mathbf{x}_{(j)}^T \boldsymbol{\beta} - \log \left\{ \sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right\} \right] - N. \end{aligned}$$

Como N é uma constante, ela pode ser retirada sem alterações dos resultados e obtém-se a log-verossimilhança parcial abaixo

$$l(\boldsymbol{\beta}) = \sum_{j=1}^N \left[\mathbf{x}_{(j)}^T \boldsymbol{\beta} - \log \left\{ \sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right\} \right]. \quad (3.10)$$

Os estimadores de máxima verossimilhança parcial são obtidos maximizando-se (3.10) em relação a $\boldsymbol{\beta}$. Propriedades assintóticas dos estimadores resultantes foram estabelecidas utilizando-se a teoria de processos de contagem (Andersen e Gill, 1982).

3.1.1 Inclusão da penalidade na verossimilhança parcial de Cox

Será considerada agora a situação em que se deseja selecionar variáveis explicativas por meio da utilização da verossimilhança penalizada. A idéia é incluir a penalidade na verossimilhança (3.5). Da mesma forma com que se fez anteriormente, pode-se utilizar o estimador de Breslow para estimar $H_0(t)$ obtendo-se a log-verossimilhança penalizada

$$lp(\boldsymbol{\beta}) = \sum_{j=1}^N \{\log(h_j) + \mathbf{x}_j^T \boldsymbol{\beta}\} - \sum_{i=1}^n \left\{ \sum_{j=1}^N h_j I(i \in R_j) \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right\} - n \sum_{i=1}^d p_\lambda(|\beta_j|). \quad (3.11)$$

Para a derivação da log-verossimilhança parcial penalizada, os mesmos passos da log-verossimilhança parcial podem ser seguidos. Note que a diferença entre (3.11) e (3.8) reside na inclusão da penalidade, mas o termo que penaliza a log-verossimilhança não depende de h_j , cujo estimador é o mesmo dado em (3.9). Deste modo, a log-verossimilhança parcial penalizada é da forma

$$lp(\boldsymbol{\beta}) = \sum_{j=1}^N \left[\mathbf{x}_{(j)}^T \boldsymbol{\beta} - \log \left\{ \sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right\} \right] - n \sum_{i=1}^d p_\lambda(|\beta_j|). \quad (3.12)$$

É importante ressaltar que a forma da log-verossimilhança parcial penalizada é constituída pela verossimilhança parcial em (3.10) subtraída pela penalidade. Assim, a verossimilhança parcial

proposta por Cox (1975) coincide com (3.12) quando $p_\lambda(\cdot)$ for nula. Considere o seguinte exemplo para compreender o comportamento da log-verossimilhança parcial quando penalizada.

Exemplo: Considere os seguintes tempos de sobrevivência dados na Tabela 3.1. Observe que a covariável é o grupo a que o paciente pertence. Os tempos que estiverem indicados com o sinal + são referentes às censuras.

Grupos	Tempos de sobrevivência
Grupo 1	9, 11, 20+, 22
Grupo 2	12, 13+, 17, 24+, 30+

Tabela 3.1: Tempos de sobrevivência, em dias, dos pacientes dos grupos 1 e 2.

A Figura 3.1 mostra o comportamento da verossimilhança parcial penalizada, com as penalidades HT, SCAD, L_1 e L_2 . Note que em todos os casos o máximo da verossimilhança parcial penalizada ocorre em $\hat{\beta} = 0$, enquanto na verossimilhança parcial usual o máximo ocorre próximo ao valor -1 . Para este exemplo, o nível mínimo de significância encontrado para o teste usual para $H_0 : \beta = 0$ foi 0,3. Para a construção dos gráficos, foi utilizado $\lambda = 2$. Nesse caso, com a estimação via verossimilhança penalizada, os coeficientes são estimados como zero por não terem influência na variável resposta.

As estimativas para β são obtidas via maximização da verossimilhança parcial penalizada (3.12). O algoritmo de *Newton-Raphson* pode ser aplicado para a maximização de (3.12). Na literatura, são fornecidos alguns algoritmos alternativos. Por exemplo, Fan e Li (2002) propõe um algoritmo de *Newton-Raphson* modificado. No algoritmo proposto, as funções de penalidade são aproximadas localmente por funções quadráticas. Essa modificação deve-se ao fato de algumas penalidades serem irregulares na origem, de tal forma que podem não possuir derivada segunda para alguns pontos. Segundo os autores, para bons valores iniciais, o algoritmo comporta-se de maneira satisfatória, fornecendo estimativas em poucas iterações. Outra proposta sugerida por Hunter e Li (2005) é fazer uma conexão entre a aproximação quadrática local, sugerida por Fan e Li (2002) e os algoritmos MM, discutidos por Heiser (1995) e Lange et al. (2000).

Figura 3.1: Log-verossimilhança parcial penalizada.

3.1.2 Propriedade *oracle* para o modelo de riscos proporcionais de Cox

Nesta seção serão apresentados os teoremas para estabelecer a propriedade *oracle* do estimador de máxima verossimilhança parcial penalizado de β . Como citado no capítulo anterior, essa é uma das propriedades desejadas para o estimador resultante. A seguir serão discutidas as condições necessárias para que alguns componentes de $\hat{\beta}$ sejam iguais a zero quando as covariáveis associadas não exercem influência no tempo de sobrevivência e os coeficientes restantes sejam estimados como se o submodelo correto fosse conhecido a priori.

Considere, assim como feito no Capítulo 2, a seguinte partição de $\beta_0 = (\beta_{10}, \dots, \beta_{d0})^T = (\beta_{10}^T, \beta_{20}^T)^T$, assumindo que os s primeiros termos são diferentes de zero e são os componentes de β_1 ,

e que os $(d - s)$ restantes pertencem a $\beta_{20} = \mathbf{0}$.

Defina as quantidades a_n e b_n como

$$a_n = \max_{1 \leq j \leq d} \{p'_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\} \quad \text{e} \quad b_n = \max_{1 \leq j \leq d} \{|p''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}. \quad (3.13)$$

Para verificar a propriedade *oracle* é necessário mostrar que o estimador resultante é $\sqrt{n}$ -consistente, que estima $\hat{\beta}_2$ como zero e que para $n^{-1/2}\lambda_n \rightarrow \infty$, $\hat{\beta}_1$ é assintoticamente normal com matriz de variâncias e covariâncias $\mathbf{I}_1^{-1}$, em que $\mathbf{I}$ é a matriz informação de Fisher para β_1 calculada sob a restrição $\beta_2 = \mathbf{0}$. Se ficar demonstrado que o estimador da verossimilhança penalizada satisfaz essas condições, isso implicará que ele estima β_{10} tão bem como se $\beta_{20} = \mathbf{0}$ fosse conhecido.

Para definir o estimador da verossimilhança parcial penalizada, lembre que a definição da log-verossimilhança parcial de Cox é dada por $l(\beta) = \sum_{j=1}^N [\mathbf{x}_{(j)}^T \beta - \log\{\sum_{i \in R_j} \exp(\mathbf{x}_i^T \beta)\}]$ e que a log-verossimilhança parcial penalizada é $lp(\beta) = l(\beta) - n \sum_{j=1}^d p_\lambda(|\beta_j|)$. Assume-se que a função de penalidade é aplicada para todos os componentes de β .

O teorema a seguir mostra a existência de um estimador baseado na verossimilhança parcial penalizada que é $\sqrt{n}$ -consistente, ou seja, que converge com a razão $O_P(n^{-1/2} + a_n)$, em que a_n é dado por (3.13). O teorema mostra também que a razão de convergência do estimador da verossimilhança parcial penalizada depende dos parâmetros de suavização.

Teorema 3.1 *Assuma que $\{(\mathbf{x}_i, T_i, C_i) : i = 1, \dots, n\}$ são independentes e identicamente distribuídos. T e C são condicionalmente independentes dado $\mathbf{x}$. Satisfeitas as condições de regularidade apresentadas em Fan e Li (2002), se $b_n \rightarrow 0$, então existe um maximizador local $\hat{\beta}$ de $lp(\beta)$ tal que $\|\hat{\beta} - \beta_0\| = O_P(n^{-1/2} + a_n)$, em que a_n e b_n são dados em (3.13).*

Para a prova deste teorema, ver Fan e Li (2001). Note que, com uma escolha apropriada de λ_n , o Teorema 3.1 garante a existência de um estimador para a verossimilhança parcial penalizada que é $\sqrt{n}$ -consistente e tão desejado quanto um estimador que converge com razão $a_n = O(n^{-1/2})$.

Dado que existe um estimador $\sqrt{n}$ -consistente, o passo seguinte para estabelecer a propriedade

oracle é mostrar que este estimador possui a propriedade de esparsividade, ou seja, ele estima β_2 como zero.

Lema 3.1 *Assuma que*

$$\liminf_{n \rightarrow \infty} \liminf_{\theta \rightarrow 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0, \quad (3.14)$$

$\lambda_n \rightarrow 0$, $\sqrt{n}\lambda_n \rightarrow \infty$ e as condições do Teorema 3.1 sejam atendidas. Então, com probabilidade tendendo a 1, para um dado β_1 satisfazendo $\|\beta_1 - \beta_{10}\| = O_P(n^{-1/2})$ e uma constante C ,

$$l_p = \left\{ \begin{pmatrix} \beta_1 \\ \mathbf{0} \end{pmatrix} \right\} = \max_{\|\beta_2\| \leq Cn^{-1/2}} \left\{ \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \right\}.$$

Denote por

$$\Sigma = \text{diag}\{p''_{\lambda_n}(|\beta_{10}|), \dots, p''_{\lambda_n}(|\beta_{s0}|)\}, \quad (3.15)$$

e

$$\mathbf{b} = (p'_{\lambda_n}(|\beta_{10}|)\text{sgn}(\beta_{10}), \dots, p'_{\lambda_n}(|\beta_{s0}|)\text{sgn}(\beta_{s0}))^T, \quad (3.16)$$

em que $\text{sgn}(\beta)$ é a função que retorna o sinal de β e s o número de componentes de β_{10} .

O segundo teorema estabelece a propriedade *oracle* para os estimadores resultantes da verossimilhança parcial penalizada.

Teorema 3.2 (Propriedade *Oracle*) *Assuma que a função de penalidade $p_{\lambda_n}(|\theta|)$ satisfaz a condição (3.14). Se $\lambda_n \rightarrow 0$, $\sqrt{n}\lambda_n \rightarrow \infty$ e $a_n = O(n^{-1/2})$, quando $n \rightarrow \infty$, então, sob as condições do Teorema 3.1, com probabilidade tendendo a 1, o maximizador local $\hat{\beta} = (\hat{\beta}_1^T, \hat{\beta}_2^T)^T$ no Teorema*

3.1 deve satisfazer:

(i) *Esparsividade*

$$\hat{\beta}_2 = \mathbf{0};$$

$$[-0.25cm]$$

(ii) *Normalidade assintótica*

$$\sqrt{n}(\mathbf{I}_1(\beta_{10}) + \Sigma)\{\hat{\beta}_1 - \beta_{10} + (\mathbf{I}_1(\beta_{10}) + \Sigma)^{-1} \mathbf{b}\} \rightarrow N\{\mathbf{0}, \mathbf{I}_1(\beta_{10})\}$$

em que $\mathbf{I}_1(\beta_{10}) = \mathbf{I}_1(\beta_{10}, \mathbf{0})$ é a matriz de informação de Fisher, conhecendo $\beta_{20} = \mathbf{0}$, de dimensão $s \times s$.

Como consequência, a matriz de variância e covariância assintótica de $\hat{\beta}_1$ é

$$\frac{1}{n}\{\mathbf{I}_1(\beta_{10}) + \Sigma\}^{-1}\mathbf{I}_1(\beta_{10})\{\mathbf{I}_1(\beta_{10}) + \Sigma\}^{-1},$$

que é aproximadamente igual a $(1/n)\mathbf{I}_1^{-1}(\beta_{10})$, quando o parâmetro de suavização λ_n tende a zero.

A seguir, detalha-se as condições necessárias para que os estimadores resultantes da verossimilhança parcial penalizada de algumas funções de penalidade possuam a propriedade *oracle*.

Primeiramente será considerada a função de penalidade SCAD. Se $\lambda_n \rightarrow 0$, a função de penalidade SCAD, como pode ser visto em (2.8), é nula, o que implica que $b_n = 0$; logo, do Teorema 3.1, tem-se que existe um estimador que converge com taxa $OP(n^{-1/2} + a_n)$. Ao passo que, se $\lambda_n \rightarrow 0$, isso implica que $a_n = 0$. Sendo assim, existe um estimador $\sqrt{n}$ -consistente. Considerando ainda que $\lambda_n \rightarrow 0$, tem-se que $\Sigma = \mathbf{0}$ e $\mathbf{b} = \mathbf{0}$. Pelo Teorema 3.2 e Lema 3.1, quando $\sqrt{n}\lambda_n \rightarrow \infty$, o estimador deve satisfazer a propriedade de esparsividade e

$$\sqrt{n}\{\hat{\beta}_1 - \beta_{10}\} \xrightarrow{D} N\{\mathbf{0}, \mathbf{I}_1^{-1}(\beta_{10})\},$$

ou seja, é garantida a existência de um estimador $\sqrt{n}$ -consistente, com propriedade de esparsividade e

que $\hat{\beta}_1$ é assintoticamente normal. Logo, para a função de penalidade SCAD, verifica-se a propriedade *oracle* quando $\lambda_n \rightarrow 0$ e $\sqrt{n}\lambda_n \rightarrow \infty$.

Para a função de penalidade HT observa-se que $a_n = 0$ se $\lambda_n \rightarrow 0$, pelo mesmo motivo da penalidade SCAD. Sendo assim, o estimador resultante é $\sqrt{n}$ -consistente, ou seja, converge com taxa $O_P(n^{-1/2})$. Quando a condição $\sqrt{n}\lambda_n \rightarrow \infty$ é satisfeita, o estimador resultante, assim como na penalidade SCAD, possuirá a propriedade de esparsividade e $\sqrt{n}\{\hat{\beta}_1 - \beta_{10}\} \xrightarrow{D} N\{\mathbf{0}, I_1^{-1}(\beta_{10})\}$, e, portanto, executará tão bem quanto se o estimador de β_1 conhecesse $\beta_2 = \mathbf{0}$.

Porém, para a penalidade L_1 , tem-se que $a_n = \lambda_n$. Sendo assim, para que o estimador resultante seja $\sqrt{n}$ -consistente, é necessário que $\lambda_n = O(n^{-1/2})$. Por outro lado, o Teorema 3.2 requer que $\sqrt{n}\lambda_n \rightarrow \infty$ para que o estimador possua a propriedade *oracle*. Note que essas duas condições não podem ser satisfeitas ao mesmo tempo. Nesse caso, a penalidade L_1 não possui a propriedade *oracle*.

3.2 Modelo de Cox com Fragilidade Gama

O modelo de riscos proporcionais de Cox apresentado anteriormente considera a suposição de que os tempos de sobrevivência de indivíduos distintos são independentes. Embora seja razoável fazer essa suposição para a grande maioria dos estudos, existem casos em que ela pode ser violada. Por exemplo, quando a variável de interesse é o tempo de sobrevivência em indivíduos de uma mesma família ou mesmo em animais de uma mesma ninhada. Esses exemplos caracterizam situações em que a suposição de independência dos tempos pode não ser válida. Nesses casos, é razoável supor que existam agrupamentos de indivíduos e, portanto, uma associação entre os tempos para observações de um mesmo grupo. Ou ainda casos de variáveis aleatórias não observadas, como fatores ambientais e genéticos, ou informações que por algum motivo não foram consideradas no planejamento e que podem violar a suposição de independência dos tempos.

Para considerar essa suposta associação entre tempos de sobrevivência, o modelo que tem ganhado destaque é o modelo de Cox com fragilidade (Clayton e Cuzick, 1985). Nesse modelo, um efeito aleatório, denominado fragilidade, é introduzido na função de taxa de falha $h(t)$. Neste trabalho será considerado o caso em que a fragilidade atua de forma multiplicativa em $h(t)$. Sendo assim, o

modelo de fragilidade tem a função de taxa de falha para o j -ésimo sujeito no i -ésimo grupo definida por

$$h_{ij}(t|\mathbf{x}_{ij}, u_i) = h_0(t)u_i \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}), \quad i = 1, \dots, n \quad \text{e} \quad j = 1, \dots, J_i, \quad (3.17)$$

em que $h_0(\cdot)$ é a função de taxa de falha padrão, $\boldsymbol{\beta}$ é o vetor de parâmetros associado às covariáveis de dimensão $d \times 1$, $\mathbf{x}_{ij}$ é o vetor de covariáveis para o indivíduo j do grupo i de dimensão $d \times 1$ e u_i são variáveis de fragilidade supostas positivas, independentes e identicamente distribuídas, com distribuição conhecida. Observe que está sendo considerado n grupos de estudo, sendo que cada grupo é composto por J_i observações, $i = 1, \dots, n$.

A fragilidade u_i é comum a todas observações de um mesmo grupo J_i para todo i , $i = 1, \dots, n$. Supôs-se que, condicional a um dado valor da variável de fragilidade u_i , todas as observações daquele grupo são independentes.

Como a fragilidade não pode ser negativa, dado que ela entra na função de risco de forma multiplicativa e esta, por sua vez, é necessariamente positiva, algumas distribuições ganharam grande popularidade, como, por exemplo, a log-normal e a gama, sendo que esta última é a distribuição utilizada para modelar a fragilidade neste trabalho. Esta escolha deve-se ao tratamento algébrico conveniente.

Assume-se que a média da fragilidade gama é 1, o que garante a identificabilidade do modelo. A necessidade de impor esta relação entre os parâmetros da gama para que o modelo seja identificável é devida ao fato de ser possível dividir e multiplicar u_i e h_0 pela mesma constante e continuar com um modelo gama com um novo parâmetro de escala, apesar do parâmetro de forma não se alterar. Portanto, faz-se necessária a suposição de $E(u_i) = 1$, ou seja, de que a distribuição da fragilidade seja $\text{gama}(\alpha, \alpha)$, isto é, a densidade da fragilidade é dada por

$$g(u) = \frac{\alpha^\alpha u^{\alpha-1} \exp(-\alpha u)}{\Gamma(\alpha)}.$$

Dadas as definições acima e considerando o conjunto de pseudo-dados $\{u_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\delta}_{ij}\}$ (note que u_i não é observado) $i = 1, \dots, n$, tem-se que a forma geral da verossimilhança dada em (3.2)

para o modelo de fragilidade gama é

$$\prod_{i=1}^n \prod_{j=1}^{J_i} [\{h(z_{ij}|\mathbf{x}_{ij}, u_i)\}^{\delta_{ij}} S(z_{ij}|\mathbf{x}_{ij}, u_i)] \prod_{i=1}^n g(u_i). \quad (3.18)$$

Para a obtenção da verossimilhança dos dados observados, integra-se a verossimilhança (3.18) em relação a $u_1, \dots, u_n$, a partir da qual encontra-se a seguinte função de verossimilhança

$$L = \exp \left\{ \boldsymbol{\beta}^T \left(\sum_{i=1}^n \sum_{j=1}^{J_i} \delta_{ij} \mathbf{x}_{ij} \right) \right\} \prod_{i=1}^n \frac{\alpha^\alpha \prod_{j=1}^{J_i} \{h_0(z_{ij})\}^{\delta_{ij}}}{\Gamma(\theta) \{ \sum_{j=1}^{J_i} H_0(z_{ij}) \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \alpha) \}^{A_i + \alpha}}, \quad (3.19)$$

em que $A_i = \sum_{j=1}^{J_i} \delta_{ij}$ é o número de falhas no grupo i , e $H_0(t)$ é a função de taxa de falha padrão acumulada. A log-verossimilhança de (3.19) será dada por

$$\begin{aligned} l(\boldsymbol{\beta}) = & \sum_{i=1}^n \left\{ \sum_{j=1}^{J_i} \delta_{ij} \log h(z_{ij}) - \left[(A_i + \alpha) \log \left\{ \sum_{j=1}^{J_i} H_0(z_{ij}) \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta} + \alpha) \right\} \right] \right\} \\ & + \sum_{i=1}^n \left\{ \boldsymbol{\beta}^T \left(\sum_{j=1}^{J_i} \delta_{ij} \mathbf{x}_{ij} \right) + \alpha \log \alpha - \log \Gamma(\alpha) \right\}. \end{aligned} \quad (3.20)$$

Será considerado o método da verossimilhança perfilada para eliminar o parâmetro de perturbação $h(\cdot)$. Para estimar $H_0(t)$, novamente será utilizada a proposta de Breslow (1972), ou seja, $H_0(\cdot)$ é da forma

$$H_0(Z_i) = \sum_{j=1}^N \psi_j I(z_j \leq z), \quad (3.21)$$

em que ψ_j são constantes a serem estimadas, $z_1, \dots, z_N$ são os tempos de falhas observados. Daqui se tem que $h_0(z_{ij}) = \sum_{l=1}^N \psi_l I(z_l \leq z_{ij}) - \sum_{l=1}^N \psi_l I(z_l < z_{ij})$.

Sustituindo (3.21) em (3.20), tem-se que

$$\begin{aligned}
l(\boldsymbol{\beta}) = & \sum_{i=1}^n \left\{ \sum_{j=1}^{J_i} \delta_{ij} \left[\log \left(\sum_{l=1}^N \psi_l I(z_l \leq z_{ij}) - \sum_{l=1}^N \psi_l I(z_l < z_{ij}) \right) + \mathbf{x}_{ij}^T \boldsymbol{\beta} \right] \right. \\
& - \left. \left[(A_i + \alpha) \log \left\{ \sum_{j=1}^{J_i} \sum_{l=1}^N \psi_l I(z_l \leq z_{ij}) \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}) + \alpha \right\} \right] \right\} \\
& + \sum_{i=1}^n \left\{ \boldsymbol{\beta}^T \left(\sum_{j=1}^{J_i} \delta_{ij} \mathbf{x}_{ij} \right) + \alpha \log \alpha - \log \Gamma(\alpha) \right\}. \tag{3.22}
\end{aligned}$$

Para um dado l , $l = 1, \dots, N$, considere (3.22) como função de ψ_l e $\boldsymbol{\beta}$. A derivada de (3.22) em relação a ψ_l é dada por

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \psi_l} = \frac{1}{\psi_l} - \sum_{i=1}^n \frac{(A_i + \alpha) \sum_{j=1}^{J_i} I(z_l \leq z_{ij}) \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta})}{\sum_{k=1}^N \psi_k \sum_{j=1}^{J_i} I(z_k \leq z_{ij}) \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}) + \alpha}$$

e a raiz da função escore satisfaz a seguinte equação

$$\psi_l^{-1} = \sum_{i=1}^n \frac{(A_i + \alpha) \sum_{j=1}^{J_i} I(z_l \leq z_{ij}) \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta})}{\sum_{k=1}^N \psi_k \sum_{j=1}^{J_i} I(z_k \leq z_{ij}) \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}) + \alpha}. \tag{3.23}$$

Note que a solução de ψ_l não pode ser expressa em função de quantidades conhecidas; logo, a maximização da verossimilhança perfilada não pode ser feita diretamente por um algoritmo *Newton-Raphson*. Sendo assim, para a estimação dos parâmetros de interesse, pode ser usado o algoritmo descrito a seguir.

Algoritmo Fan e Li (2002):

Passo 1: As estimativas iniciais de $\boldsymbol{\beta}$, $\psi_1, \dots, \psi_N$ e α podem ser obtidas facilmente. A de $\boldsymbol{\beta}$ pode ser obtida de uma regressão de Cox usual (desconsiderando uma possível dependência entre os tempos de cada grupo), $h_1, \dots, h_N$, obtidos de (3.9), servem como valores iniciais para $\psi_1, \dots, \psi_N$ e, como estimativa inicial para α , pode ser utilizado um valor qualquer maior que zero;

Passo 2: Atualizar ψ_l , $l = 1, \dots, N$ por (3.23);

Passo 3: Maximizar a log-verossimilhança definida em (3.22) em relação a (α, β) ;

Passo 4: Iterar entre os passos 2 e 3 até a obtenção da convergência.

No passo 3 pode ser utilizado o algoritmo *Newton-Raphson* para a maximização da log-verossimilhança. Segundo os autores, caso seja usado, devem-se tomar certas precauções com relação aos valores de α , pois o algoritmo envolve as derivadas primeira e segunda da função gama, sendo que esta pode não existir para determinados valores de α . Pode-se evitar o problema seguindo a sugestão de Nielsen et al. (1992), que sugerem construir uma grade de possíveis valores para α e maximizar sobre essa grade. Repare que a dificuldade para estimar os parâmetros no modelo de Cox com fragilidade gama é maior que no modelo de riscos proporcionais de Cox, uma vez que tem-se que trabalhar com a verossimilhança completa.

O algoritmo acima, trata-se de uma abordagem alternativa para o algoritmo EM frequentemente utilizado em modelos de fragilidade. O algoritmo EM, recomendado para a obtenção de estimativas de máxima verossimilhança em conjuntos com observações incompletas, surge naturalmente neste contexto, pois as fragilidades são variáveis não observadas. Dois algoritmos EM comumente utilizados em modelos de fragilidade são os propostos por Nielsen et al. (1992) e Klein (1992). A diferença principal entre os dois algoritmos reside no fato de que um usa uma implementação completa do algoritmo EM (Klein, 1992), enquanto o outro usa uma procura unidimensional para a estimação dos parâmetros (Nielsen et al., 1992). Para maiores detalhes desses algoritmos, ver Chicarino (1999).

Os algoritmos de Nielsen et al. (1992), Klein (1992) e Fan e Li (2002) podem apresentar os seguintes complicadores: algoritmos relativamente lentos e podem ocorrer problemas de convergência para grandes amostras. Recentemente uma nova proposta para estimação de β no modelo de fragilidade foi apresentada (Therneau et al., 2003). Nessa proposta, (3.17) é reescrita como

$$h_{ij}(t|\mathbf{x}_{ij}, w_i) = h_0(t)\exp(\mathbf{x}_{ij}^T\beta + w_i), \quad i = 1, \dots, n \quad \text{e} \quad j = 1, \dots, J_i, \quad (3.24)$$

em que u_i no modelo (3.17) corresponde a $\exp\{w_i\}$. Nesse modelo, assume-se que os w_i 's são uma

amostra independente de alguma distribuição com média zero e variância η , de modo que, quando $\eta = 0$, tem-se o modelo de riscos proporcionais de Cox.

Além dos algoritmos citados anteriormente, outra forma de abordar o modelo de fragilidade é por meio de verossimilhança penalizada. Therneau e Grambsch (2000) exploraram a conexão entre regressão penalizada e modelos de fragilidade. Neste trabalho eles notam que

$$l(\beta) = l_{\text{COX}}(\beta, u) - g(u, \eta), \quad (3.25)$$

em que $l(\beta)$ é a log-verossimilhança apresentada em (3.22), $l_{\text{COX}}(\cdot)$ é a log-verossimilhança parcial usual, como definida em (3.10), com $\log(u)$ considerado como um *offset*, isto é, $\log(u)$ é um termo adicionado ao preditor linear durante o ajuste com coeficiente conhecido e $g(\cdot)$ é a função de penalidade, ou seja, o logaritmo da função de verossimilhança do modelo de fragilidade gama, apresentado em (3.22), pode ser escrito como uma verossimilhança penalizada. Como a distribuição da fragilidade é gama com média igual a 1 e variância $\eta = 1/\alpha$, o logaritmo da função densidade de probabilidade de $u = \exp(w)$ pode ser escrito como

$$\begin{aligned} \log[f(u)] &= (\alpha - 1)\log u - \alpha u + \alpha \log \alpha - \log \Gamma(\alpha) \\ &= (\alpha - 1)w - \alpha e^w + \alpha \log \alpha - \log \Gamma(\alpha) = \log[g(w)], \end{aligned}$$

e verifica-se facilmente que o logaritmo da densidade de w é $(w - \exp\{w\})/\eta$ acrescido de uma função de α . Conhecida a forma da função de penalidade, pode-se reescrever (3.25) como

$$l(\beta) = l_{\text{COX}}(\beta, \mathbf{u}, \text{dados}) - \frac{1}{\eta} \sum_{i=1}^n [w_i - \exp\{w_i\}]. \quad (3.26)$$

Lema 3.2 *A solução do modelo de verossimilhança parcial penalizado com função de penalidade $g(u, \eta) = 1/\eta \sum_{i=1}^n [w_j - \exp(w_j)]$ coincide com a solução do algoritmo EM para um valor de η fixo e que utilize fragilidade gama.*

Essa descoberta foi obtida por acidente em alguns testes iniciais, mas posteriormente sua prova foi estabelecida (Therneau e Grambsch, 2000). Para a obtenção das estimativas, considerando-se a perspectiva acima, é necessário inicializar as fragilidades como parâmetros fixos iguais a 1 no primeiro passo da otimização da função de verossimilhança parcial. No passo seguinte, as fragilidades são avaliadas como médias condicionais, dadas suas observações, de modo similar ao que é feito no algoritmo EM. Esse processo é repetido até a convergência ser obtida.

Segundo alguns trabalhos na literatura, por exemplo Hougaard (2000), essa abordagem se comporta bem para o modelo de fragilidade gama. Neste trabalho, será considerada a maximização da log-verossimilhança penalizada do modelo de fragilidade apresentada na seção a seguir.

3.2.1 Inclusão da penalidade na verossimilhança do modelo de Cox com fragilidade gama

Como visto na seção anterior, a estimativa de β no modelo de Cox com fragilidade, pode ser obtida através da maximização de uma verossimilhança penalizada. Para selecionar também variáveis explicativas por meio da utilização da verossimilhança penalizada é necessário incluir uma segunda função de penalidade para a obtenção de $\hat{\beta}$. Ao tratar do modelo de riscos proporcionais de Cox, a inclusão da penalidade na verossimilhança parcial não modifica os passos para obtenção das estimativas de h_j , $j = 1, \dots, n$, já que a penalidade não é função de h_j . Sendo assim, a log-verossimilhança parcial penalizada resume-se à log-verossimilhança parcial acrescida da função de penalidade.

No caso do modelo de fragilidade, a situação é análoga, ou seja, os problemas apresentados para se encontrarem as estimativas para ψ_l , $l = 1, \dots, n$ são os mesmos. Para lidar com essa situação, pode-se considerar o algoritmo de Fan e Li (2002), apresentado na sessão (3.2), modificando-se o passo 3, maximizando-se a log-verossimilhança penalizada

$$\begin{aligned}
lp(\boldsymbol{\beta}) = & \sum_{i=1}^n \left\{ \sum_{j=1}^{J_i} \delta_{ij} \left[\log \left(\sum_{l=1}^N \psi_l I(z_l \leq z_{ij}) - \sum_{l=1}^N \psi_l I(z_l < z_{ij}) \right) + \mathbf{x}_{ij}^T \boldsymbol{\beta} \right] \right. \\
& - \left. \left[(A_i + \alpha) \log \left\{ \sum_{j=1}^{J_i} \sum_{l=1}^N \psi_l I(z_l \leq z_{ij}) \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}) + \alpha \right\} \right] \right\} \\
& + \sum_{i=1}^n \left\{ \boldsymbol{\beta}^T \left(\sum_{j=1}^{J_i} \delta_{ij} \mathbf{x}_{ij} \right) + \alpha \log \alpha - \log \Gamma(\alpha) \right\} - n \sum_{i=1}^d p_\lambda(|\beta_j|),
\end{aligned} \tag{3.27}$$

em relação a α e $\boldsymbol{\beta}$.

Note que a diferença entre (3.22) e (3.27) é a inclusão da penalidade $n \sum_{i=1}^d p_\lambda(|\beta_j|)$. O algoritmo procede exatamente como descrito na seção (3.2). Como dito na seção anterior, a log-verossimilhança do modelo de Cox com fragilidade gama pode ser escrita como uma log-verossimilhança penalizada, segundo a abordagem de Therneau e Grambsch (2000). A inclusão da função de penalidade faz com que a log-verossimilhança seja penalizada por duas funções. Sendo assim é possível reescrever (3.27) como

$$l_{pp}(\boldsymbol{\beta}) = l(\boldsymbol{\beta}, u) - \frac{1}{\eta} \sum_{i=1}^n (w_i - \exp\{w_i\}) - n \sum_{i=1}^d p_\lambda(|\beta_j|), \tag{3.28}$$

Assim como na verossimilhança parcial penalizada, com a escolha apropriada da função de penalidade $p_\lambda(\cdot)$ e do parâmetro de suavização λ , a maximização de (3.28) deve produzir coeficientes estimados como zero para efeitos não significantes, enquanto para os outros coeficientes se obtêm estimativas adequadas. Esse fato é discutido a seguir.

3.2.2 Propriedade *oracle* para o modelo de Cox com fragilidade gama

Nesta seção, assim como discutido para o modelo de riscos proporcionais de Cox, serão apresentados as condições e os teoremas necessários para que o estimador resultante do modelo de fragilidade gama possua a propriedade *oracle*. Dado que as estimativas são assintoticamente normais (para maiores detalhes ver Fan e Li (2002)), é possível estabelecer a propriedade *oracle* para o estimador da verossimilhança penalizada do modelo de fragilidade gama.

Seguindo os mesmos passos do modelo de riscos proporcionais de Cox, primeiramente será discutida a existência de um estimador $\sqrt{n}$ -consistente. Denote $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}^T)^T$ e $\boldsymbol{\theta}_0 = (\alpha_0, \boldsymbol{\beta}_0^T)^T$, o verdadeiro valor de $\boldsymbol{\theta}$.

Teorema 3.3 *Assuma que $(\mathbf{x}_{ij}, T_{ij}, C_{ij}), j = 1, \dots, J$ é uma amostra aleatória independente para $i = 1, \dots, n$, e dado u_i , $(\mathbf{x}_{ij}, T_{ij}, C_{ij}), j = 1, \dots, J$, são distribuídos independentemente de acordo com (3.17). T_{ij} e C_{ij} , dado $\mathbf{x}_i$, são condicionalmente independentes e u_i são independentes e identicamente distribuídos com distribuição gama. Se $b_n \rightarrow 0$ e as condições de regularidade apresentadas em Fan e Li (2002) são válidas, então existe um maximizador local $\hat{\boldsymbol{\theta}}$ de $lp(\boldsymbol{\theta})$ tal que $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_P(n^{-1/2} + a_n)$, em que a_n e b_n são dados por (3.13).*

Para estabelecer a propriedade *oracle*, denote $\hat{\boldsymbol{\theta}}_1 = (\hat{\alpha}, \hat{\boldsymbol{\beta}}_1^T)^T$, $\boldsymbol{\theta}_{10} = (\alpha_0, \boldsymbol{\beta}_{10}^T)^T$, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ e $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$ e defina

$$\boldsymbol{\Sigma}_1 = \text{diag}(0, \boldsymbol{\Sigma}), \quad (3.29)$$

em que $\boldsymbol{\Sigma}$ é definido em (3.15) e

$$\mathbf{b}_1 = (0, \mathbf{b}^T)^T, \quad (3.30)$$

em que $\mathbf{b}$ é dado por (3.16).

Teorema 3.4 *Assuma que a penalidade $p_{\lambda_n}(|\theta|)$ satisfaça a condição (3.14). Se $\lambda_n \rightarrow 0$, $\sqrt{n}\lambda_n \rightarrow \infty$, $a_n = O(n^{-1/2})$ e sob as condições de regularidade do Teorema 3.3, com probabilidade tendendo a 1, o maximizador local $\sqrt{n}$ -consistente $\hat{\theta} = (\hat{\alpha}, \hat{\beta}_1^T, \hat{\beta}_2^T)^T$ dado pelo Teorema 3.3 satisfaz:*

(i) *Esparsividade*

$$\hat{\beta}_2 = \mathbf{0};$$

(ii) *(Normalidade assintótica)*

$$\sqrt{n}(\tilde{I}_1(\theta_{10}) + \Sigma_1)\{\hat{\theta}_1 - \theta_{10} + (\tilde{I}_1(\theta_{10}) + \Sigma_1)^{-1} \mathbf{b}_1\} \rightarrow N\{\mathbf{0}, \tilde{I}_1(\theta_{10})\}$$

em que $\tilde{I}_1(\theta_{10})$ é a submatriz de $\tilde{I}_1(\theta_{10}, \mathbf{0})$ de dimensão $(s+1) \times (s+1)$.

De acordo com os Teoremas 3.3 e 3.4 é possível estabelecer a propriedade *oracle* para os estimadores resultantes da verossimilhança penalizada do modelo de Cox com fragilidade gama, assim como feito para os modelo de riscos proporcionais de Cox, com uma escolha apropriada de λ_n . Para maiores detalhes desses teoremas ver Fan e Li (2002).

A seguir serão discutidos uma aproximação para a função de penalidade e a estimação do parâmetro de suavização λ .

3.3 Aproximação para as funções de penalidade

Neste capítulo foram apresentados alguns algoritmos para a obtenção das estimativas dos parâmetros nos modelos de riscos proporcionais de Cox e modelo de Cox com fragilidade gama. Entretanto, como já citado anteriormente, algumas funções de penalidade são irregulares na origem, como, por exemplo, a HT, L_1 e SCAD. Portanto, essas funções podem não possuir derivada segunda para alguns pontos e cuidados especiais são necessários antes da aplicação de um algoritmo de *Newton-Raphson*. Fan e Li (2001) sugerem que as funções de penalidade sejam aproximadas localmente por funções quadráticas. Para isso, suponha um valor inicial β_0 que seja próximo ao valor que maximiza a log-verossimilhança penalizada. Se β_{j0} é próximo do valor zero, então $\hat{\beta}_j = 0$. Caso contrário, elas podem ser aproximadas

por funções quadráticas como

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\text{sgn}(\beta_j) \approx \{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}\beta_j,$$

quando $\beta_j \neq 0$. Em outras palavras,

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2}\{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}(\beta_j^2 - \beta_{j0}^2), \quad (3.31)$$

para $\beta_j \approx \beta_{j0}$.

A aproximação reduz significativamente o custo computacional (Fan e Li, 2001). Na Figura 3.2 são apresentadas algumas funções de penalidade (linha preenchida) e suas respectivas aproximações quadráticas (linhas tracejadas). Nos gráficos, utilizaram-se os valores 0,9 e 1,3 para as duas aproximações consideradas e $\lambda = 2, 5$. Essa aproximação será considerada no trabalho.

Figura 3.2: Funções de penalidades e suas aproximações quadráticas.

3.4 Estimação do parâmetro λ por validação cruzada

Nesta seção, discute-se a estimação do parâmetro de suavização λ , para a maximização de β em (3.12) e (3.27). Considera-se também a estimação do parâmetro a quando a função de penalidade utilizada é a SCAD. Como visto nas definições das funções de penalidade, o parâmetro de suavização é responsável pela determinação de quais valores de β serão penalizados. Se $\lambda \rightarrow 0$, a estimativa de máxima verossimilhança penalizada coincide com a estimativa de máxima verossimilhança original, caso $\lambda \rightarrow +\infty$, os β 's serão virtualmente zero. Na prática, não é fácil escolher o valor de λ apropriadamente e algumas metodologias têm sido propostas na literatura (Fan e Li, 2001). Neste trabalho, considera-se uma metodologia para a escolha do parâmetro de suavização das penalidades baseada no valor preditivo do modelo, conforme Verweij e Van Houwelingen (1994), isto é, o objetivo é obter $\hat{\lambda}$ que forneça previsões razoáveis para novas observações.

Considere um modelo de regressão linear com variável resposta Y_i e vetor de covariáveis X_i . Em modelos lineares, o valor preditivo pode ser medido pela soma de quadrados de predição (SQP), que é função do parâmetro de suavização λ , dada por

$$\text{SQP}(\lambda) = \sum_{i=1}^n \left(Y_i - X_i \hat{\beta}_{\lambda}^{(-i)} \right)^2,$$

em que $\hat{\beta}_{\lambda}^{(-i)}$ é a estimativa para vetor de coeficientes da regressão quando o indivíduo i é desconsiderado da amostra. Essa é uma medida de quão bem os valores ajustados, por meio de um modelo candidato, podem prever as variáveis respostas observadas. Para modelos baseados em verossimilhança parcial, Verweij e Van Houwelingen (1993) sugerem uma generalização da SQP dada pela log-verossimilhança parcial de validação cruzada

$$\text{LVC}(\lambda) = \sum_{i=1}^n l_i(\hat{\beta}_{\lambda}^{(-i)}) = \sum_{i=1}^n \left[l(\hat{\beta}_{\lambda}^{(-i)}) - l_{(-i)}(\hat{\beta}_{\lambda}^{(-i)}) \right],$$

em que $l_i(\beta)$ é a contribuição do indivíduo i para a log-verossimilhança parcial, $l(\cdot)$ é a log-verossimilhança parcial usual e $l_{(-i)}(\cdot)$ é log-verossimilhança parcial desconsiderando o indivíduo

i. O parâmetro de suavização λ que será utilizado na log-verossimilhança parcial penalizada para encontrar a estimativa do vetor β será aquele que maximiza a LVC.

Neste capítulo foram apresentados o modelo de riscos proporcionais de Cox e o modelo de Cox com fragilidade gama, bem como a obtenção das respectivas log-verossimilhanças e a inclusão das funções de penalidade nas mesmas. Ainda foram consideradas uma aproximação para as funções de penalidade e a estimação de λ pelo método de validação cruzada.

Capítulo 4

Estudo de Simulação e Aplicação

Neste capítulo são apresentados e discutidos os resultados de estudos de simulação conduzidos para ilustrar e avaliar o desempenho dos métodos de seleção de covariáveis considerados. Os estudos foram feitos com base nos métodos de seleção de covariáveis apresentados no Capítulo 2, no critério AIC e no método que estima os coeficientes quando somente as variáveis diferentes de zero estão no modelo (ótimo). Este último será utilizado como referência nas comparações a serem feitas. As comparações basearam-se nos coeficientes estimados e seus erros padrões. Para finalizar, considera-se uma aplicação dos métodos a um conjunto de dados reais.

4.1 Estudo de simulação

Para a comparação dos diversos métodos, consideraram-se dois exemplos. O primeiro deles refere-se ao modelo de riscos proporcionais de Cox e o segundo ao modelo de Cox com fragilidade gama. Todos os métodos de seleção de covariáveis foram comparados segundo as medidas descritas a seguir:

- Erro Relativo do Modelo (ERM)

O ERM é definido como a razão dos erros para modelos ajustados segundo duas abordagens distintas. É uma medida que quantifica o ganho por usar-se uma abordagem em relação à outra. O erro do modelo (EM), como definido em Fan e Li (2002), para o modelo de riscos

proporcionais de Cox e modelo de Cox com fragilidade gama é

$$E[\exp(-\mathbf{x}^T \hat{\boldsymbol{\beta}}) - \exp(-\mathbf{x}^T \boldsymbol{\beta}_0)]^2, \quad (4.1)$$

em que $\hat{\boldsymbol{\beta}}$ é a estimativa do parâmetro fornecida pelo modelo utilizado e $\boldsymbol{\beta}_0$ é o vetor de parâmetros conhecido. Nos exemplos, comparam-se as medianas dos erros relativos do modelo de todos os métodos de seleção, seguindo Tibshirani (1996), pois, conforme apontado por Fan e Li (2002), algumas funções de penalidade, por exemplo a HT, apresentam certa instabilidade que influenciaria o comportamento da média do ERM.

- Desvio Mediano Absoluto (DMA)

O DMA refere-se à mediana dos desvios absolutos da mediana (Hoaglin et al., 2000) e é definido por

$$\text{DMA}(X) = 1,4826 \text{ Med}_i(|X_i - \text{Med}_l(X_l)|). \quad (4.2)$$

Trata-se de uma medida de variabilidade fácil de ser calculada, que é computacionalmente simples e robusta. A constante 1,4826 é utilizada historicamente e faz com que, sob a suposição $X_i \sim N(\mu, \sigma^2)$, $E[\text{DMA}(X_1, X_2, \dots, X_n)] \xrightarrow{n \rightarrow +\infty} \sigma^2$. Nas simulações, compararam-se os DMAs das estimativas de β_j 's diferentes de zero e erros-padrões associados.

- Número Médio de Coeficientes nulo

Considera-se o número médio de coeficientes nulos corretos e incorretos pelos diversos métodos de seleção de covariáveis.

Neste trabalho foram utilizadas rotinas existentes no programa R, *coxph* e *frailty.gamma*, juntamente com outras apresentadas no Apêndice A, para a obtenção das estimativas dos parâmetros de interesse $\boldsymbol{\beta}$. A escolha de λ ótimo foi realizada por meio da metodologia apresentada na Seção 3.4. Todas as simulações foram realizadas utilizando códigos escritos em R (R Development Core Team, 2006).

4.1.1 Simulação para o modelo riscos proporcionais de Cox

Neste primeiro caso considera-se $R = 1.000$ réplicas. Cada uma possui $n = 100$ observações geradas a partir de um modelo exponencial com taxa de falha dada por

$$h(t|\mathbf{x}) = \exp(\mathbf{x}^T \boldsymbol{\beta}),$$

em que $\boldsymbol{\beta}_0^T = (0,2; 0; 0,8; 0; 0; 0; 0; 0,6)$. O vetor de covariáveis $\mathbf{x}_i$ foi gerado a partir de uma distribuição normal padronizada, supondo-se que a correlação entre $\mathbf{x}_i$ e $\mathbf{x}_j$ era $\rho^{|i-j|}$. Aqui tomou-se $\rho = 0,5$. Com o propósito de avaliar o efeito do percentual de censura sobre a seleção das covariáveis, realizou-se uma comparação entre os cenários 0%, 30% e 60% de censuras. Os tempos de censura foram gerados usando uma distribuição exponencial com média $C\exp(\mathbf{x}^T \boldsymbol{\beta})$, em que C foi escolhido apropriadamente para produzir a proporção média de censuras desejada. Com exceção do cenário 0% de censura, somente as amostras que estavam dentro de uma margem de 5% para mais ou para menos do percentual de censura desejado foram consideradas, ou seja, no caso da geração da amostra com 30% de censura, se uma réplica possuísse uma proporção de censura inferior a 25% ou superior a 35%, ela era descartada e uma nova era gerada, prosseguindo-se assim até atingir o número de réplicas desejadas. Os erros dos modelos (EM) para os procedimentos estudados neste trabalho são calculados por

$$E[\exp(-\mathbf{x}_{(j)}^T \hat{\boldsymbol{\beta}}_{\delta(j)}) - \exp(-\mathbf{x}_{(j)}^T \boldsymbol{\beta}_{0(j)})]^2 \approx \frac{1}{n} \sum_{i=1}^n \left[\exp(-\mathbf{x}_{i(j)}^T \hat{\boldsymbol{\beta}}_{\delta(j)}) - \exp(-\mathbf{x}_{i(j)}^T \boldsymbol{\beta}_{0(j)}) \right]^2 = e_{\delta(j)}, \quad (4.3)$$

$j = 1, \dots, R$, em que $\delta = \text{Ótimo, AIC, HT, } L_1, \text{ SCAD e EMVP}$, sendo esta última associada com o estimador de máxima verossimilhança parcial. O ERM será definido sempre em comparação ao EM calculado para o EMVP e será estimado por

$$\widehat{\text{ERM}}_{\delta(j)} = \frac{e_{\delta(j)}}{e_{\text{EMVP}(j)}}, \quad j = 1, \dots, R.$$

Conseqüentemente, a mediana do erro relativo do modelo (MERM) é estimada por

$$\widehat{\text{MERM}} = \text{Med}_{1 \leq j \leq R}(\widehat{\text{ERM}}_{\delta(j)}).$$

Percentual de censura	Método	$\widehat{\text{MERM}}$	Número médio de coeficientes nulos	
			Corretos	Incorretos
0%	Ótimo	0,3934	5,00	0,00
	AIC	0,7994	4,03	0,36
	L_1	0,5096	3,51	0,20
	HT	0,5100	3,99	0,22
	SCAD	0,4934	3,71	0,09
30%	Ótimo	0,3859	5,00	0,00
	AIC	0,7749	4,02	0,49
	L_1	0,5550	3,52	0,33
	HT	0,5291	3,91	0,38
	SCAD	0,5050	3,59	0,17
60%	Ótimo	0,3279	5,00	0,00
	AIC	0,7030	4,10	0,58
	L_1	0,4257	3,58	0,51
	HT	0,4031	4,02	0,52
	SCAD	0,4106	3,91	0,21

Tabela 4.1: MERM estimado e número médio de coeficientes nulos para o modelo de riscos proporcionais de Cox.

A Tabela 4.1 apresenta as MERM's dos métodos de seleção de covariáveis avaliados. O número de coeficientes estimados como zero também é apresentado, sendo dividido em duas colunas, uma intitulada “corretos”, que apresenta o número médio de coeficientes que deveriam ser zero. e outra chamada de “incorretos”, que apresenta o número médio de coeficientes que não deveriam ser nulos, mas cujos métodos de seleção os indicaram como tal. Quanto mais próximos de 5 e 0 forem os valores nas colunas “corretos” e “incorretos”, respectivamente, melhores os resultados.

Inicialmente, compararam-se os diversos métodos dentro de cada cenário de censuras. Para o cenário sem censura, o AIC obteve melhores resultados que os métodos baseados na verossimilhança penalizada, quando comparado o número médio de coeficientes indicados como zero corretamente;

entretanto, foi o que obteve o pior desempenho em relação aos incorretos. Convém notar que a superioridade ou inferioridade com relação a esse quesito é muito pequena, podendo-se afirmar que os métodos são comparáveis com relação ao número médio de coeficiente indicado como zero. O mesmo padrão é observado quando são avaliados os cenários com um médio (30%) e alto (60%) percentual de censuras. Vale ressaltar que, dentre os métodos de seleção de covariáveis baseados na verossimilhança penalizada, o HT obteve um melhor desempenho para indicar os coeficientes nulos do modelo.

Da Tabela 4.1 pode-se concluir que, dentro de cada cenário de censura, o método SCAD foi aquele que forneceu os melhores resultados, sendo comparável com o método ótimo. Note, também, que todos os métodos baseados na verossimilhança penalizada apresentaram MERM menores que o AIC.

Igualmente se nota que os métodos de seleção avaliados neste trabalho apresentaram um padrão de diminuição da MERM à medida que a proporção de censuras aumentou para o modelo de riscos proporcionais de Cox. Isto indica que, no cenário com a maior proporção de censuras, o efeito da seleção de covariáveis de todos os métodos mostrou-se mais efetivo quando comparado com as estimativas de máxima verossimilhança com todas as variáveis presentes no modelo.

O DMA foi utilizado para uma avaliação da variação das estimativas e erros padrões dos métodos abordados. Esta medida de variabilidade, como descrita em (4.2) para os coeficientes não nulos e seus erros padrões, é dada por

$$\begin{aligned} \text{SD} &= \widehat{\text{DMA}}_{1 \leq i, l \leq n}(\beta_j) = 1,4826 \text{ Med}_i(|\hat{\beta}_{ji} - \text{Med}_l(\hat{\beta}_{jl})|), \\ &\text{e} \\ \text{SD}_{\text{dma}} &= \widehat{\text{DMA}}_{1 \leq i, l \leq n}(\text{EP}(\beta_j)) = 1,4826 \text{ Med}_i(|\text{EP}(\hat{\beta}_{ji}) - \text{Med}_l(\text{EP}(\hat{\beta}_{jl}))|), \quad j = 1, 3, 8. \end{aligned} \quad (4.4)$$

Na Tabela 4.2 encontram-se os DMA's dos coeficientes diferentes de zero e de seus erros padrões para o modelo de riscos proporcionais de Cox.

Percentual de censura	$\hat{\beta}$	SD (SD_{dma})				
		Ótimo	AIC	L_1	HT	SCAD
0%	$\hat{\beta}_1$	0,101 (0,010)	0,184 (0,026)	0,111 (0,012)	0,115 (0,013)	0,120 (0,012)
	$\hat{\beta}_3$	0,130 (0,011)	0,155 (0,015)	0,144 (0,014)	0,145 (0,016)	0,162 (0,015)
	$\hat{\beta}_8$	0,121 (0,011)	0,141 (0,015)	0,128 (0,014)	0,131 (0,016)	0,162 (0,015)
30%	$\hat{\beta}_1$	0,136 (0,013)	0,284 (0,075)	0,147 (0,015)	0,121 (0,017)	0,158 (0,015)
	$\hat{\beta}_3$	0,159 (0,014)	0,190 (0,019)	0,166 (0,018)	0,166 (0,021)	0,204 (0,019)
	$\hat{\beta}_8$	0,130 (0,013)	0,158 (0,017)	0,145 (0,017)	0,146 (0,019)	0,163 (0,018)
60%	$\hat{\beta}_1$	0,174 (0,024)	0,101 (0,023)	0,095 (0,023)	0,084 (0,027)	0,184 (0,026)
	$\hat{\beta}_3$	0,209 (0,026)	0,225 (0,028)	0,215 (0,024)	0,229 (0,028)	0,225 (0,027)
	$\hat{\beta}_8$	0,190 (0,025)	0,190 (0,026)	0,195 (0,023)	0,200 (0,026)	0,206 (0,025)

Tabela 4.2: SD e SD_{dma} dos coeficientes não nulos para o modelo de riscos proporcionais de Cox.

As duas medidas apresentadas na Tabela 4.2 podem ser analisadas como medidas de acurácias das estimativas dos parâmetros e de seus erros padrões. No cenário com 0% de censura, os métodos de seleção HT e L_1 possuem desempenho comparável ao ótimo em relação às duas medidas, apresentando melhor resultado que os demais métodos. Nos cenários 30% e 60% de observações censuradas, os métodos são mais irregulares, o que gera um equilíbrio maior entre os desempenhos observados. As estimativas das variâncias, avaliadas pelos DMAs, mostraram-se bastante razoáveis para todos os métodos, quando comparadas com o cenário ótimo.

Conforme citado anteriormente, o AIC é o método que melhor seleciona o número de zeros corretamente; entretanto, vale destacar que ele sofre com o excesso de variabilidade em todos os cenários para as estimativas dos coeficientes nulos, como pode ser observado pelas figuras (4.1), (4.2) e (4.3).

Considerando o cenário com 0% de censura, observou-se um tempo médio de 59 segundos, 2 minutos e 9 segundos e 2 minutos e 52 segundos para que os métodos L_1 , HT e SCAD concluíssem a seleção de variáveis, respectivamente. Para o cenário com 30% de censuras, observou-se o mesmo padrão do cenário anterior, no qual o método L_1 , com 49 segundos, foi o mais rápido para a conclusão do processo de seleção de variáveis, enquanto o método HT precisou de 1 minuto e 7 segundos

e o método SCAD de 2 minutos, em média, para a escolha das variáveis. No cenário com um alto percentual de censura (60%), foram observados tempos médios de 36 segundos, 1 minuto e 20 segundos e 1 minuto e 39 segundos para a composição do modelo final.

Figura 4.1: Boxplot do Modelo de riscos proporcionais de Cox – 0% de censura.

Figura 4.2: Boxplot do Modelo de riscos proporcionais de Cox – 30% de censura.

Figura 4.3: Boxplot do Modelo de riscos proporcionais de Cox – 60% de censura.

4.1.2 Simulação para o modelo de Cox com fragilidade gama

O segundo caso procura avaliar o uso da metodologia para a situação em que os dados são dependentes. Para tanto, considerou-se aqui também $R = 1.000$ réplicas. Cada réplica foi composta por $n = 100$ grupos (famílias), cada um com 2 componentes. Cada família teve os tempos de sobrevivência (no caso um vetor bivariado) gerados a partir do modelo exponencial bivariado de Sarkar (Sarkar, 1987), que demonstrou o seguinte teorema

Teorema 4.1 *Sejam Z_1, Z_2, Z_3 e V variáveis aleatórias independentes, com densidades $e(\lambda_1), e(\lambda_2), e(\lambda_{12})$ e $U(0, 1)$, respectivamente. Defina*

$$\begin{aligned} T_1 &= X + \{1 - I(Z_1 > Z_2)\} \{\lambda_1^{-1} A^{-1}(V^{1/(1+\gamma)} A(\lambda_1 Y)) - X\}, \\ T_2 &= Y + I(Z_1 > Z_2) \{\lambda_2^{-1} A^{-1}(V^{1/(1+\gamma)} A(\lambda_2 X)) - Y\}, \end{aligned} \quad (4.5)$$

em que $X = \min(Z_1, Z_3)$, $Y = \min(Z_2, Z_3)$, $A(z) = [1 - \exp(-z)]$ para $z > 0$, $A^{-1}(z) = -\log(1 - z)$ para $0 < z < 1$ e $\gamma = \lambda_{12}/(\lambda_1 + \lambda_2)$. Então, $(T_1, T_2) \sim ACBVE_2(\lambda_1, \lambda_2, \lambda_{12})$.

O Teorema 4.1 mostra como construir uma distribuição exponencial bivariada ($ACBVE_2$), a partir de distribuições exponenciais e uniforme independentes. A prova do teorema e maiores detalhes sobre a distribuição podem ser encontrados em Sarkar (1987). Sendo assim, esse teorema torna-se uma ferramenta útil para simulação.

Com base no resultado anterior, foram gerados, para cada família, 3 tempos exponenciais (z_1, z_2, z_3) e 1 uniforme (v). Os tempos z_1 e z_2 foram gerados com taxas de falhas

$$h_{ij}(t) = \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}_0), \quad i = 1, \dots, n \quad \text{e} \quad j = 1, 2,$$

em que $\boldsymbol{\beta}_0^T = (0, 2; 0; 0; 0, 8)$ e $\mathbf{x}$, como antes, foi gerado a partir de distribuições normais padronizadas. Calculou-se a taxa de falha do tempo z_3 (λ_{12}) de forma que a correlação fosse incluída com base nos resultados a seguir. Para controlar a correlação, utilizou-se o fato de que, conforme Sarkar (1987),

$$\rho(T_1, T_2) = \frac{\lambda_{12}}{\lambda} + \frac{\lambda_{12}}{\lambda} \sum_{j=1}^{\infty} \frac{j!}{(\lambda + \lambda_1 + \lambda_2)^j} \left\{ \lambda_1^j \prod_{k=1}^j (\lambda + k\lambda_1)^{-1} + \lambda_2^j \prod_{k=1}^j (\lambda + k\lambda_2)^{-1} \right\}, \quad (4.6)$$

em que T_1 e T_2 são os tempos de sobrevivência dos membros da família e $\lambda = \lambda_1 + \lambda_2 + \lambda_{12}$. Como a expressão (4.6) envolve uma soma infinita, seu valor foi truncado, desconsiderando-se a soma. Dessa

forma, a expressão (4.6) foi reescrita como

$$\rho(T_1, T_2) = \frac{\lambda_{12}}{\lambda_1 + \lambda_2 + \lambda_{12}}. \quad (4.7)$$

Uma vez definida a correlação desejada, encontrava-se o valor do parâmetro λ_{12} em (4.7) e gerava-se o tempo z_3 . O tempo v era gerado de uma distribuição $U(0, 1)$. Após gerados os tempos z_1 , z_2 , z_3 e v , era utilizado, então, o resultado do teorema para gerar os tempos T_1 e T_2 .

Nesta simulação censurou-se a família em vez de seus elementos. Isso foi feito gerando-se tempos de censura (para cada família) com distribuição exponencial de taxa de falha $C\exp(\mathbf{x}^T\beta_0)$, em que C foi escolhido apropriadamente para produzir em média a proporção de censura desejada. O tempo de censura foi comparado com o $\min(T_1, T_2)$. Caso fosse menor, os tempos dos elementos da família eram considerados censurados. Serão consideradas 0% e 30% como proporções de censura e correlações de 0, 2 e 0, 5. Assim como no exemplo anterior, considerou-se uma margem de 5% para o percentual de censura desejado como critério para a inclusão da réplica na simulação.

Os resultados para o modelo de Cox com fragilidade gama são apresentados na Tabela 4.3. Novamente são avaliados o MERM e o número médio de coeficientes zero “corretos” e “incorretos”. Vale ressaltar que, nesse caso, o modelo possui somente quatro variáveis, das quais duas têm efeito significativo. A diminuição do número de variáveis no modelo deve-se ao fato do método de estimação dos coeficientes utilizado apresentar alguns problemas de instabilidade numérica, o que causa a não convergência de o algoritmo. Foi necessário substituir réplicas para as quais não ocorria a convergência; isso ocorreu numa proporção maior com os métodos baseados na verossimilhança penalizada e em menor proporção com o AIC e o método ótimo. Observou-se também que, quanto maior a correlação, maior o problema de convergência. As dificuldades de estimação em modelos de Cox com fragilidade já foram citadas por diversos autores na literatura, como, por exemplo, Therneau e Grambsch (2000) e Chicarino (1999). Apesar de todas as dificuldades, o objetivo é avaliar, no cenário de dados correlacionados, o impacto das proporções de censuras e da intensidade da relação entre as observações, como feito por Fan e Li (2002).

Correlação	Percentual de censura	Método	$\widehat{\text{MERM}}$	Número médio de coeficientes nulos	
				Corretos	Incorretos
0,2	0%	Ótimo	0,6764	2,00	0,00
		AIC	0,7408	1,29	0,21
		L_1	1,0298	0,99	0,18
		HT	0,7321	1,18	0,35
		SCAD	0,6977	1,02	0,33
	30%	Ótimo	0,6703	2,00	0,00
		AIC	0,7653	1,29	0,17
		L_1	1,0489	0,95	0,17
		HT	0,7522	1,15	0,30
		SCAD	0,7100	1,01	0,32
0,5	0%	Ótimo	0,8468	2,00	0,00
		AIC	1,0640	1,27	0,38
		L_1	1,2245	1,01	0,32
		HT	1,1087	1,22	0,45
		SCAD	0,8798	1,13	0,35
	30%	Ótimo	0,9509	2,00	0,00
		AIC	1,0323	1,37	0,43
		L_1	1,2833	1,05	0,45
		HT	1,1114	1,30	0,52
		SCAD	0,9534	1,21	0,39

Tabela 4.3: MERM estimado e número médio de coeficientes nulos para o modelo de Cox com fragilidade gama.

Inicialmente cada cenário foi avaliado de forma isolada. Na situação em que a correlação é fraca (correlação igual a 0,2), os métodos se equivalem com relação ao número de coeficientes indicados como zero; porém, a performance do SCAD na MERM é comparável com o ótimo, enquanto os demais métodos apresentam desempenho inferior. Note que esse é o comportamento padrão de todos os métodos, em todos os cenários avaliados. Vale ressaltar que o L_1 obteve o pior desempenho, tanto na avaliação da MERM como no número médio de coeficientes indicados como zero.

Avaliando os métodos de seleção para uma mesma correlação, em que se variou o percentual de censura, não se observaram grandes mudanças nos valores da MERM. Por exemplo, no cenário em que a correlação é igual a 0,2, o método de seleção SCAD forneceu MERM igual a 0,6977 e 0,7100

para 0% e 30% de censura, respectivamente. Pode-se observar na Tabela 4.3 que todos os métodos possuem esse mesmo padrão, isto é, eles apresentaram valores muito semelhantes para a MERM nos dois cenários de censuras considerados, quando avaliados dentro de uma mesma correlação. Quando os métodos foram analisados dentro de um mesmo percentual de censura, variando-se a correlação, observou-se que à medida em que os dados eram mais correlacionados, eles apresentavam resultados piores em relação ao modelo de Cox com fragilidade gama com todas as variáveis no modelo, pois ocorreu um aumento no valor da MERM.

Na Tabela 4.2 encontram-se os DMAs, como descrito em (4.4), para as estimativas dos coeficientes diferentes de zero e de seus erros padrões para o modelo de Cox com fragilidade gama.

Ao analisar o SD e o SD_{dma} para o modelo de Cox com fragilidade gama, observam-se que os valores são comparáveis. Da mesma forma que no caso da seção 4.2.1, o AIC, apesar de ser aquele que melhor seleciona o número de zeros corretamente, sofre com o excesso de variabilidade em todos os cenários para as estimativas dos coeficientes não nulos. Os boxplots dos coeficientes ilustram essa variabilidade, como pode ser observado pelas figuras 4.4 a 4.7.

Avaliando a performance dos métodos com relação ao tempo computacional, observou-se que, assim como no modelo de riscos proporcionais de Cox, o método L_1 obteve melhores resultados que os outros métodos baseados em verossimilhança penalizada na maioria dos cenários avaliados. Considerando o cenário com $\rho = 0,2$ e 0% de censura, observou-se um tempo médio de 8 minutos e 30 segundos, 7 minutos e 46 segundos e 14 minutos e 36 segundos para que os métodos L_1 , HT e SCAD concluíssem a seleção de variáveis, respectivamente. Para o cenário com $\rho = 0,2$ e 30% de censuras, observou-se que diferentemente do cenário anterior, o método L_1 , com 4 minutos e 54 segundos, foi o mais rápido para a conclusão do processo de seleção de variáveis, enquanto o método HT precisou de 6 minutos e 32 segundos e o método SCAD de 8 minutos e 24 segundos, em média, para a escolha das variáveis.

Nos cenários com maior correlação entre os tempos de sobrevivência de uma família, isto é, quando se considerou $\rho = 0,5$, observou-se um crescimento nos tempos médios para a seleção de variáveis quando comparados com o cenários em que $\rho = 0,2$. Observou-se um tempo médio de 14

min e 30 segundos, 16 minutos e 33 segundos e 21 minutos e 42 segundos para os métodos L_1 , HT e SCAD, respectivamente, quando se considerou 0% de censura. No cenário com 30% de censuras o método L_1 precisou, em média, de 12 minutos e 5 segundos para completar o processo de seleção das variáveis, enquanto o método HT gastou 13 minutos e 14 segundos e o SCAD 18 minutos e 14 segundos, em média.

Correlação	Percentual de censura	$\hat{\beta}$	SD (SD_{dma})				
			Ótimo	AIC	L_1	HT	SCAD
0,2	0%	$\hat{\beta}_1$	0,101 (0,053)	0,123 (0,015)	0,125 (0,058)	0,121 (0,055)	0,121 (0,080)
		$\hat{\beta}_4$	0,100 (0,058)	0,116 (0,055)	0,119 (0,059)	0,101 (0,056)	0,117 (0,059)
	30%	$\hat{\beta}_1$	0,116 (0,062)	0,146 (0,065)	0,139 (0,065)	0,117 (0,061)	0,119 (0,089)
		$\hat{\beta}_4$	0,132 (0,069)	0,150 (0,052)	0,152 (0,061)	0,130 (0,081)	0,150 (0,072)
0,5	0%	$\hat{\beta}_1$	0,112 (0,056)	0,136 (0,073)	0,149 (0,079)	0,127 (0,078)	0,143 (0,067)
		$\hat{\beta}_4$	0,113 (0,054)	0,148 (0,075)	0,124 (0,078)	0,134 (0,081)	0,131 (0,069)
	30%	$\hat{\beta}_1$	0,137 (0,087)	0,206 (0,055)	0,189 (0,065)	0,259 (0,082)	0,172 (0,092)
		$\hat{\beta}_4$	0,121 (0,088)	0,159 (0,032)	0,170 (0,078)	0,155 (0,049)	0,132 (0,076)

Tabela 4.4: SD e SD_{dma} dos coeficientes não nulos para o modelo de Cox com fragilidade gama.

Figura 4.4: Boxplot do Modelo de Cox com fragilidade gama – 0% de censura e correlação 0,2.

Figura 4.5: Boxplot do Modelo de Cox com fragilidade gama – 0% de censura e correlação 0,5.

Figura 4.6: Boxplot do Modelo de Cox com fragilidade gama – 30% de censura e correlação 0,2.

Figura 4.7: Boxplot do Modelo de Cox com fragilidade gama – 30% de censura e correlação 0,5.

4.2 Aplicação a dados reais

Os métodos estudados neste trabalho também foram aplicados a um conjunto de dados reais envolvendo pacientes com insuficiência cardíaca grave. O banco de dados é proveniente de um estudo conduzido no Instituto do coração do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (INCOR - HC - FMUSP). O banco de dados possui 1220 pacientes, que foram acompanhados desde janeiro de 1991 até dezembro de 2000. Para simplificar a aplicação dos métodos, foram selecionados somente os pacientes que não possuíam dados faltantes, resultando num total de 381 pacientes.

O objetivo principal do estudo é avaliar a influência da massa ventricular esquerda na sobrevida dos pacientes, considerando-se também outros fatores clínicos e sócio-demográficos. Ao todo observaram-se 132 falhas, ou seja, 132 pacientes vieram a óbito, o que resultou em 251 censuras (71%).

Os pesquisadores selecionaram as seguintes variáveis para a avaliação do efeito na sobrevivência.

- tempo de sobrevida do paciente (meses);
- indicador de óbito;
- idade dos pacientes (anos);
- gênero;
- classificação do indivíduo quanto à doença cardíaca de base - etiologia (chagástica ou não chagástica);
- massa ventricular esquerda corrigida (g/m);
- diâmetro interno do átrio esquerdo (cm);
- altura (cm);
- peso (kg).

Para analisar a sobrevida dos pacientes, utilizou-se o modelo de riscos proporcionais de Cox. Para reduzir um possível viés na modelagem, foram consideradas as interações entre massa ventricular esquerda com as variáveis gênero, idade e diâmetro interno do átrio esquerdo, fornecendo um total de 10 variáveis explicativas. Consideraram-se os métodos de máxima verossimilhança parcial, AIC, L_1 , HT e SCAD. A Tabela 4.5 apresenta as estimativas para os coeficientes e seus erros padrões.

Variáveis	EMV	AIC	L_1	HT	SCAD
Idade	0,2144 (0,0853)	0,2200 (0,0836)	0,1909 (0,0868)	0,2190 (0,0854)	0,2201 (0,0851)
Diâmetro átrio	-0,2263 (1,2102)	0 (-)	0 (-)	0 (-)	0 (-)
Massa	0,0277 (0,0164)	0,0272 (0,0086)	0,0247 (0,0122)	0,0270 (0,0088)	0,0270 (0,0089)
Gênero	1,7865 (2,2950)	0,3833 (0,2220)	0,3145 (0,1536)	0, (-)	0,3531 (0,1801)
Etiologia	1,1538 (0,2369)	1,2102 (0,2225)	1,3566 (0,2263)	1,1544 (0,2373)	1,1521 (0,2347)
Peso	-0,0046 (0,0077)	0 (-)	0 (-)	0 (-)	0 (-)
Altura	-0,0031 (2,8970)	0 (-)	0 (-)	0 (-)	0 (-)
Massa:Gênero	-0,0030 (0,0045)	0 (-)	0 (-)	0 (-)	0 (-)
Massa:Diâmetro	0,0003 (0,0032)	0 (-)	0 (-)	0 (-)	0 (-)
Massa:Idade	-0,0004 (0,0001)	-0,0005 (0,0002)	-0,0039 (0,0051)	-0,0042 (0,0074)	-0,0008 (0,0003)

Tabela 4.5: Coeficientes e Erros padrões estimados.

O método de validação cruzada estimou os seguintes valores para o parâmetro de suavização λ das funções de penalidade: 1,5789; 3,3832 e 2,3451 para o L_1 , HT e SCAD, respectivamente.

Nenhum dos métodos de seleção de variáveis selecionou a variável altura. Das interações incluídas no modelo, somente a interação entre massa e idade foi selecionada. Ao analisar os modelos selecionados pelos diversos métodos, note que o AIC, SCAD e L_1 selecionaram as mesmas variáveis; porém, as estimativas dos coeficientes do AIC e SCAD são mais parecidas entre si. O método HT selecionou uma variável a menos que os outros modelos, o que de certo modo era esperado, tendo em vista que este método é resultado de uma função de penalidade que é mais restrita da log-verossimilhança, gerando, assim, um número maior de coeficientes indicados como zero, como pode ser observado nas simulações.

Pela observação dos erros padrões, pode-se verificar, finalmente, que a precisão das estimativas para todos os métodos são comparáveis.

Com base nos resultados das simulações e da aplicação pode-se concluir que, apesar do tempo computacional superior aos outros métodos, o método SCAD é o mais indicado, dentre os métodos que compõem este trabalho, para os modelos em questão, isto é, essa técnica mostrou-se uma ferramenta útil no processo de seleção, pois obteve, em quase todas as simulações, os melhores resultados em relação à MERM, vale ressaltar que, esses resultados eram sempre comparáveis ao método ótimo, e em relação à precisão das estimativas, não obteve resultados discrepantes dos outros métodos.

Capítulo 5

Considerações Finais e Estudos Futuros

O problema de se selecionar um subconjunto de covariáveis para a construção de um modelo parcimonioso foi abordado neste trabalho por meio de vários métodos de seleção de covariáveis baseados em verossimilhanças penalizadas. Foram considerados os modelos de riscos proporcionais de Cox e modelo de Cox com fragilidade gama. Funções de penalidade interessantes para o problema foram apresentadas, e suas características, discutidas. O desempenho das diferentes penalidades foi avaliado em cenários com diferentes proporções de censuras e, no caso do modelo de Cox com fragilidade gama, ainda se consideraram diferentes correlações entre as observações de um mesmo grupo.

Os métodos discutidos na dissertação são facilmente implementados no programa R com a criação de funções definidas em Therneau et al. (2003). Contudo, o problema da estimação do parâmetro de suavização λ das funções de penalidade deve ser considerado. A grande diferença entre os tempos de execução dos métodos baseados em verossimilhança penalizada e o AIC refere-se à estimação de λ . Seria muito produtivo, em um trabalho futuro, comparar o impacto de diferentes abordagens para a seleção desse parâmetro, considerando, além do método de validação cruzada simples, sua generalização considerada por (Fan e Li, 2002) e outros métodos mais recentes, como apresentado em Huang e Harrington (2002), que utilizam um método *bootstrap* para a escolha do parâmetro.

Com base nas simulações analisadas para o modelo de riscos proporcionais de Cox, foi possível

perceber que as funções de penalidade L_1 , HT e SCAD têm melhor desempenho em relação à mediana do erro relativo do modelo (MERM) e resultados comparáveis no que diz respeito ao número médio de coeficientes indicados como zero. Convém destacar que a SCAD apresentou os melhores resultados dentre os métodos de verossimilhança penalizada, resultados comparáveis com o ótimo.

Uma questão importante diz respeito à obtenção das estimativas no modelo de fragilidade gama. Neste trabalho, utilizaram-se funções já existentes no programa R. Problemas de não convergência foram encontrados, principalmente nos cenários com maior presença de censura e maior correlação entre as observações de um mesmo grupo. Para conseguir resultados para esse modelo, foi necessário diminuir bastante o número de covariáveis. Todavia, foi observado que o método SCAD apresentou um melhor desempenho quando comparado com as demais técnicas de seleção de variáveis e resultados comparáveis com o método que estima os coeficientes, quando somente as variáveis diferentes de zero estão no modelo (ótimo).

Apesar da dificuldade prática de escolher o parâmetro de suavização λ ótimo, a simulação sugere que os métodos de seleção de covariáveis baseados em verossimilhança penalizada fornecem resultados compatíveis com o AIC no sentido de selecionar covariáveis com efeito significativo corretamente, e resultados melhores que o AIC quando considera-se a mediana relativa do modelo.

Por fim, o método é uma alternativa atraente por possibilitar a escolha das covariáveis e a estimação dos efeitos significativos em um único passo, facilitando a obtenção de um modelo parcimonioso.

Apêndice A

Aspectos Computacionais

Para o ajuste dos modelos estudados neste trabalho foram utilizadas rotinas existentes no programa R, juntamente com outras propostas por Therneau e Grambsch (1998). Abaixo encontram-se os códigos das rotinas que viabilizam o acréscimo de funções de penalidade nas rotinas usualmente utilizadas para a maximização do vetor de parâmetros β , em problemas de sobrevivência. A seguir são descritas essas rotinas para a função de penalidade L_1 .

```
pfun.l1 <- function(coef, coef0, theta, nevent, vars){  
  list(penalty = sum(abs(coef0)*theta) + sum(coef^2-coef0^2)*(theta/(2*abs(coef0))),  
  first = (theta/abs(coef0))*coef,  
  second = theta/abs(coef0),  
  flag = F)  
},  
  
cfun.l1 <- function(parms, ...) list (theta=parms$theta, done = T),
```

em que θ é o parâmetro de suavização, coef é o vetor de parâmetros a serem estimados e coef0 é um vetor de valores próximos a coef . O processo iterativo depende dessas duas funções. Na função *pfun*, estão envolvidas a função de penalidade e suas derivadas primeira e segunda. Note que aqui, está sendo utilizada a aproximação quadrática da função de penalidade L_1 , como sugerida na Seção

3.3. A função controle *cfun* retorna o valor de β e o seu sinal. As rotinas estão sendo definidas de forma que será considerada uma mesma função de penalidade para todos os componentes de β . A seguir é descrita uma função, chamada de *l1*, que faz o uso de *cfun* e *pfun*.

```
l1 <- function(..., theta = 0){
  x <- cbind(...)
  xname <- as.character(parse(text=substitute(cbind(...))))[-1]
  vars <- apply(x, 2, function(z) var(z[!is.na(z)]))
  class(x) <- 'coxph.penalty'
  temp <- list(pfun = pfun.l1,
    cfun = cfun.l1,
    diag = T,
    cparm = list(theta=theta),
    varname=paste('l1(', xname, ')', sep=''))
  attributes(x) <- c(attributes(x), temp)
  x
}
```

Para maiores detalhes das rotinas, ver Therneau e Grambsch (1998). Para exemplificar a utilização de *cpun*, *pfun* e *l1*, considere o caso definido na Seção 4.1.1, em que deseja estimar o vetor de parâmetros β para o modelos de riscos proporcionais de Cox. Lembre-se que, para a obtenção das estimativas é necessário $\hat{\lambda}$. Considere, então, que a estimativa de λ fora obtida pelo LVC, denotada aqui por *lam_est*. É possível obter as estimativas de máxima verossimilhança parcial penalizada com o seguinte comando

```
fit <- coxph(Surv(tempo, ind) ~ l1(x1,x2,x3,x4,x5,x6,x7,x8, theta=lam_est),
  iter.max=500000)
```

Para o modelo de Cox com fragilidade gama foi utilizado a rotina do R chamada *frailty.gamma* juntamente com as funções definidas nesta seção.

Referências Bibliográficas

- Andersen, P. K. e Gill, R. D. (1982). Cox's regression model for counting process: a large sample study, *The Annals of Statistics* **10**: 1100–1120.
- Antoniadis, A. (1997). Wavelets in statistics: a review (with discussion), *Journal of the Italian Statistical Society* **6**: 97–144.
- Breslow, N. (1972). Contribuição à discussão do artigo de D. R. Cox, *Journal of the Royal Statistical Society, Ser. B* **34**: 216–217.
- Chicarino, M. P. Z. (1999). Modelo semiparamétrico de fragilidade gama, Master's thesis, Instituto de Matemática e Estatística, Universidade de São Paulo.
- Clayton, D. G. e Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model, *Journal of the Royal Statistical Society, Ser. A* **148**: 82–117.
- Cox, D. R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Ser. B* **34**: 187–220.
- Cox, D. R. (1975). Partial likelihood, *Biometrika* **62**: 269–276.
- Fan, J. (1997). Comment on “wavelets in statistics: a review” by Antoniadis, *Journal of the Italian Statistical Society* **6**: 131–138.
- Fan, J. e Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association* **96**(456): 1348–1360.

- Fan, J. e Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model, *The Annals of Statistics* **30**(1): 74–99.
- Good, I. J. M. e Gaskins, R. A. (1971). Non-parametric roughness penalties for probability densities, *Biometrika* **58**(2): 255–277.
- Grey, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis, *Journal of the American Statistical Association* **87**: 255–277.
- Heiser, W. J. (1995). Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis, *Recent Advances in Descriptive Multivariate Analysis* (W.J. Krzanowski ed.) pp. 157–189. Clarendon Press, Oxford.
- Hoaglin, D. C., Mosteller, F. e Tukey, J. W. (2000). *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons.
- Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer-Verlag.
- Huang, J. e Harrington, D. (2002). Penalized partial likelihood regression for right-censored data with bootstrap selection of the penalty parameter, *Biometrics* **58**: 781–791.
- Hunter, D. R. e Li, R. (2005). Variable selection using MM algorithms, *The Annals of Statistics* **33**(4): 1617–1642.
- Kalbfleisch, J. D. e Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2 edn, John Wiley and Sons.
- Klein, J. P. (1992). Semiparametric estimation of random effects using Cox model based on the EM algorithm, *Biometrics* **48**: 795–806.
- Klein, J. P. e Moeschberger, M. L. (1997). *Survival Analysis - Techniques for Censored and Truncated Data*, Springer-Verlag.
- Lange, K., Hunter, D. R. e Yang, I. (2000). Optimization transfer using surrogate objective functions (with discussion), *Journal of Computational and Graphical Statistics* **9**: 1–59.

- Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, Wiley.
- Murphy, S. A. e Van Der Vaart, A. W. (2000). On profile likelihood, *Journal of the American Statistical Association* **95**: 449–465.
- Nielsen, G. G., Gill, R. D., Andersen, P. K. e Sorensen, T. I. A. (1992). A counting process approach to maximum likelihood estimation in frailty models, *Scandinavian Journal of Statistics* **19**: 755–762.
- Oishi, J. (1983). Regressão sobre Cristas, Master’s thesis, Instituto de Matemática e Estatística, Universidade de São Paulo.
- R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*<http://www.R-project.org>
- Sarkar, S. K. (1987). A continuous bivariate exponential distribution, *Journal of the American Statistical Association* **82**: 667–675.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall.
- Simonoff, J. S. (2008). *Encyclopedia of Biostatistics*, 2 edn, John Wiley & Sons.
- Therneau, T. M. e Grambsch, P. M. (1998). Penalized Cox models and frailty, *Relatório Técnico não publicado* .
*<http://mayoresearch.mayo.edu/mayo/research/bioestat/upload/frailty.pdf>
- Therneau, T. M. e Grambsch, P. M. (2000). *Modeling Survival Data - Extending the Cox Model*, Springer-Verlag.
- Therneau, T. M., Grambsch, P. M. e Pankratz, V. S. (2003). Penalized survival models and frailty, *Journal of Computational and Graphical Statistics* **12**(1): 156–175.
- Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society, Ser. B* **58**: 267–288.

- Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model, *Statistics in Medicine* **16**: 385–395.
- Verweij, P. J. M. e Van Houwelingen, H. C. (1993). Cross-validation in survival analysis, *Statistics in Medicine* **12**: 2305–2314.
- Verweij, P. J. M. e Van Houwelingen, H. C. (1994). Penalized likelihood in Cox regression, *Statistics in Medicine* **13**: 2427–2436.