Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Dissertação de Mestrado

Tópicos em regularização com uma aplicação em Seleção Genômica.

Pedro Henrique Toledo de Oliveira Sousa

Brasília, 21 de março de 2019

Universidade de Brasília

Dissertação de Mestrado

Tópicos em regularização com uma aplicação em Seleção Genômica.

Autor:

Pedro Henrique T.O Sousa

Orientadora:

Profa. Dra. Joanlise M.L Andrade

Coorientador:

Prof. Dr. Bernardo B. Andrade

Dissertação apresentada ao Departamento de Estatística da Universidade de Brasília, como requisito parcial para obtenção do título de Mestre em Estatística.

Brasília, 21 de março de 2019

Agradecimentos

A minha experiência na Universidade de Brasília (UnB) tem sido maravilhosa e enriquecedora desde o meu ingresso na graduação. Em todas as minhas conquistas neste período tive pessoas que, em maior ou em menor escala, me acompanharam e me incentivaram.

Primeiramente, dedico um agradecimento muito especial à minha mãe e ao meu avô materno (*in memoriam*). Muito do que sou hoje é produto dos seus esforços e ensinamentos.

Gostaria de agradecer à minha orientadora, Professora Doutora Joanlise Marco de Leon Andrade, e ao meu coorientador, Professor Doutor Bernardo Borba de Andrade, por todo o apoio, paciência e orientação.

Desejo igualmente agradecer a todos os professores do departamento de estatística da UnB. Cada um de vocês contribuiu significativamente para a minha vida profissional, com ensinamentos dentro e fora da sala de aula. Aproveito para agradecer aos funcionários da secretaria, segurança e limpeza que sempre foram tão atenciosos e prestativos comigo.

Agradeço ao Dr. Dario Grattapaglia por nos ceder os dados e nos auxiliar em diversos aspectos deste trabalho.

Por fim, agradeço à Ana Gabriela Pereira de Vasconcelos e aos pesquisadores Joseane Padilha da Silva, Orzenil Bonfim da Silva Junior, Rafael Tassinari Resende e Bruno Marco de Lima, pelas reuniões e ensinamentos sobre as aplicações dos métodos estatísticos em genética.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Código de Financiamento 001.

"Our quest for discovery fuels our creativity in all fields, not just science. If we reached the end of the line, the human spirit would shrivel and die."

(Stephen Hawking)

Resumo

Os métodos de regularização foram desenvolvidos para contornar problemas de *overfitting* e são amplamente utilizados em modelagens preditivas. Neste trabalho realiza-se uma breve introdução sobre a álgebra de matrizes relacionada a tais métodos, com ênfase nas inversas generalizadas, no posto e nas possíveis dimensões dessas matrizes, bem como apresentar uma solução geral, para sistemas lineares consistentes e inconsistentes. Em seguida, as decomposições de matrizes SVD (Singular Value Decomposition) e GSVD (Generalized Singular Value Decomposition) são utilizadas para a implementação dos modelos de regularização Tikhonov e TSVD e, posteriormente, analisa-se outros dois métodos de regularização (LASSO e LASSO Bayesiano), que estimam os coeficientes e simultaneamente realizam a seleção de variáveis. Como aplicação, realiza-se uma avaliação da qualidade preditiva dos modelos de regularização no contexto de Seleção Genômica em dados genéticos superdimensionados e de alta complexidade. Os referidos dados caracterizam-se por conter informações do DNA (genótipos) de plantas de eucalipto e a finalidade da análise é desenvolver uma abordagem alternativa aos programas de melhoramento genético tradicionais. Em resumo, os resultados mostram que os modelos para fenótipos com maior herdabilidade apresentam medidas de previsão superiores. Por fim, os métodos que conduzem a seleção de variáveis se mostraram superioridade nas tarefas preditivas em todos os casos avaliados.

Palavras-chave: Regularização, *overfitting*, seleção de variáveis, validação cruzada, Seleção Genômica, eucalipto, melhoramento genético.

Abstract

Regularization methods have been developed to overcome overfitting and are widely used in predictive modeling. This study introduces the matrix algebra related to such methods, with emphasis on the generalized inverse, the rank and the possible dimensions of those matrices, while presenting a general solution for consistent and inconsistent linear systems. Next, it employs the SVD (Singular Value Decomposition) and GSVD (Generalized Singular Value Decomposition) matrix decompositions to implement the Tikhonov and TSVD regularization models, and then analyzes two other regularization methods – namely, LASSO and Bayesian LASSO – that estimate the coefficients and simultaneously perform the variable selection. In addition, the study conducts an evaluation of the predictive accuracy of the models applied to complex high-dimensional data in the context of Genomic Selection. The data contains DNA information (genotypes) from eucalyptus plants, and the purpose of the analysis is to develop an alternative approach to the traditional programs for genetic improvement of species. In summary, the results show that models which were applied using phenotypes with higher heritability have better predictive ability. The methods that conduct variable selection were superior in the predictive tasks for all evaluated cases.

Keywords: Regularization, *overfitting*, variable selection, cross validation, Genomic Selection, eucalyptus, genetic improvement.

Lista de siglas e abreviações

Best Approximate Solution	BAS
Best Linear Unbiased Predictor	BLUP
Deoxyribonucleic Acid	DNA
Erro Quadrático Médio	EQM
Erro Quadrático Médio da validação cruzada	EQMcv
Generalized Singular Value Decomposition	GSVD
Genome-Wide Selection	GWS
Genomic Breeding Values	GBV's
Genomic Estimated Breeding Values	GEBV's
Genomic Selection	GS
Independentes e Identicamente Distribuídas	iid's
Least Absolute Shrinkage and Selection Operator	LASSO
Linkage Disequilibrium	LD
Minor Allele Frequency	MAF
Marker Assisted Selection	MAS
Markov Chain Monte Carlo	MCMC
Mínimos Quadrados Ordinários	MQO
Near Infrared Reflectance Spectroscopy	NIRS
Quantitative Trait Loci	QTL
Single Nucleotide Polymorphism	SNP
Singular Value Decomposition	SVD
Truncated Singular Value Decomposition	TSVD

Sumário

Capítulo 1

Introdução

Embora as bases de dados modernas estejam crescendo com respeito ao número de amostras, o aumento do número de variáveis é ainda mais significativo. Dados *superdimensionados* envolvem um número de parâmetros desconhecidos menor que o número de amostras ($p >> n$). Dados com essa característica são frequentemente encontrados em bases nas quais as amostras são baseadas em séries temporais, informações genéticas, imagens, entre outros exemplos.

Em um estudo cujo objetivo é explicar o comportamento ou simplesmente prever os valores de uma determinada variável $\boldsymbol{Y}$ por meio de um modelo de regressão, a existência de um número de variáveis significativamente maior que o número de amostras, ou seja, a existência de uma matriz de dados com mais colunas que linhas traz inconveniências como o fenômeno da multicolinearidade, a existência de mais parâmetros que observações e problemas de inversão de matrizes.

Muitos dos métodos estatísticos convencionais não podem ser utilizados em dados superdimensionados e, para que a estimação de parâmetros seja possível, é preciso estabelecer suposições adicionais ou certas restrições aos modelos matemáticos.

O matemático russo Tikhonov foi um dos pioneiros nos estudos sobre métodos de regularização e marcou o início de uma formulação matemática, baseada no controle da estabilidade de soluções de sistemas lineares e que veio a ser útil para aplicações em dados superdimensionados. A partir dos seus estudos, diversas outras

técnicas foram desenvolvidas.

Os chamados métodos de regularização representam um importante conceito estatístico e frequentemente são utilizados para previsão de variáveis devido à possibilidade de se evitar o *overfitting* e de se obter erros quadráticos médios menores, relativamente aos obtidos através do método de mínimos quadrados. Embora já possuam utilidade em dados com mais observações que parâmetros, os métodos de regularização se destacam pela sua versatilidade ao poderem ser aplicados em dados superdimensionados já que são capazes de contornar as inconveniências da multicolinearidade e o problema do posto incompleto da matriz de dados.

Diante do contexto de aprendizado estatístico no qual a regularização se enquadra, o objetivo deste trabalho é abordar os principais métodos de regularização já consagrados na literatura estatística, assim como apresentar alguns métodos alternativos de regularização que ainda não são tão utilizados na comunidade estatística. Assim sendo, serão apresentados conceitos metodológicos, algoritmos computacionais, a teoria estatística sobre regularização e vantagens e desvantagens de determinados métodos. Adicionalmente, aplica-sa algumas das técnicas em dados genéticos superdimensionados, avaliando a qualidade preditiva com métricas específicas da área genômica.

Através da Seleção Genômica (GS), caracterizada pela utilização conjunta e direta das informações presentes na fita de DNA, será utilizada uma base de dados genéticos de eucaliptos no treinamento e validação de modelos estatísticos, posteriormente utilizados como ferramentas que auxiliam pesquisadores a compreender a conexão existente entre elementos genômicos e fenótipos de interesse comercial.

No capítulo dois da dissertação será introduzido o conceito de regularização através de uma perspectiva matemática baseada na decomposição em valores singulares. O capítulo três irá apresentar o modelo de regressão linear clássico de mínimos quadrados, ressaltando suas limitações em dados superdimensionados e, posteriormente, irá destrinchar a teoria de regularização tal como é abordada na literatura estatística, introduzindo três métodos bastante conhecidos: LASSO, Ridge e LASSO Bayesiano. No capítulo quatro, será realizada uma breve descrição da metodologia

e dos termos utilizados em Seleção Genômica, assim como será realizada a aplicação de alguns desses métodos nos dados genéticos.

Capítulo 2

Problemas Inversos e Regularização

Neste Capítulo são apresentados alguns métodos de regularização que originalmente foram desenvolvidos para resolver problemas específicos de áreas da ciência que não necessariamente estão ligadas a fenômenos aleatórios. No entanto, estes métodos podem ser de grande utilidade para o desenvolvimento de estudos envolvendo previsão.

Primeiramente, os conceitos de consistência de um sistema linear e existência e unicidade de uma solução serão apresentados, sendo precedidos dos modelos de regularização que se baseiam nos conceitos de inversa generalizada e decomposição em valores singulares (SVD). Além de discorrer sobre sistemas consistentes e soluções aproximadas por mínimos quadrados, serão apresentados quatro cenários possíveis com sistemas que se diferenciam pela natureza da matriz $\boldsymbol{X}$ no tocante a sua dimensão e ao seu posto.

2.1 Problemas Inversos

Problemas inversos contrastam os chamados problemas diretos. Em problemas diretos, os valores assumidos por uma variável resposta $\boldsymbol{Y}$ são calculados a partir de modelos determinísticos como equações físicas ou alguma equação diferencial envolvendo o uso de constantes (parâmetros) conhecidas, que fornecem informações acerca de um determinado fenômeno. Todavia, em problemas inversos, os valores

assumidos pela variável resposta $\boldsymbol{Y}$ são observados, enquanto os parâmetros do modelo são desconhecidos. Dessa forma, problemas estatísticos de inferência podem ser considerados uma classe de problemas inversos.

Muitas complicações derivam dos problemas inversos, a citar como exemplo a dificuldade, ou até mesmo a impossibilidade, de se inverter uma matriz a fim de se manipular os elementos de um sistema linear para se isolar os parâmetros (ou constantes) desejados.

A definição de problema mal-posto engloba a grande maioria dos problemas inversos e contribui para a melhor compreensão acerca das peculiaridades dos problemas inversos. O matemático J. Hadamard trabalhou na definição do que vem a ser um problema bem-posto, estabelecendo três condições necessárias:

1. existência da solução;

2. unicidade da solução;

3. estabilidade da solução (a solução depende continuamente de $\boldsymbol{y}$).

Diz-se que uma solução é instável se uma pequena variação na amostra da variável resposta acarretar na existência de soluções muito diferentes.

Qualquer problema que deixe de satisfazer alguma dessas três condições é classificado como mal-posto.

Dentre as três condições de Hadamard, a existência e unicidade da solução serão aqui discutidas detalhadamente.

Seja um sistema linear do tipo $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{y}$, em que $\boldsymbol{X}$ é uma matriz com n linhas e p colunas, $\boldsymbol{b}$ é um vetor com p componentes e $\boldsymbol{y}$ um vetor com n componentes. Há muitas situações em que este sistema não possui solução. Neste caso, o sistema é comumente denominado como sistema inconsistente ou impossível, enquanto que quando o sistema possui alguma solução pode ser classificado como sistema possível e determinado (a solução é única) ou como sistema possível e indeterminado (existem infinitas soluções).

Embora em sistemas impossíveis a solução exata não exista, é possível se obter

soluções aproximadas para estes problemas, como será mostrado mais adiante. Se para um determinado sistema existe uma solução exata, então diz-se que este sistema é consistente. Existem diversas formas de se verificar a consistência de um sistema, dentre elas, cita-se quatro:

1. Uma condição necessária e suficiente para que o sistema $\boldsymbol{Xb} = \boldsymbol{y}$ seja consistente é $\boldsymbol{y}$ pertencer ao espaço coluna de $\boldsymbol{X}$. Assim, uma maneira de se verificar a consistência é comparar o posto de $\boldsymbol{X}$ com o posto da matriz $\boldsymbol{X}$ aumentada de $\boldsymbol{y}$. Se ambos forem iguais, então o sistema é consistente.

$$\boldsymbol{Xb} = \boldsymbol{y} \text{ é consistente} \Leftrightarrow \text{posto}(\boldsymbol{X}) = r = \text{posto}(\boldsymbol{X}\vdots\boldsymbol{y}). \tag{2.1}$$

2. Se a matriz $\boldsymbol{X}$ for uma matriz não singular, ou seja, se $\boldsymbol{X}$ possuir uma inversa $\boldsymbol{X^{-1}}$, então o sistema $\boldsymbol{Xb} = \boldsymbol{y}$ é consistente.

3. Uma condição suficiente para que o sistema $\boldsymbol{Xb} = \boldsymbol{y}$ seja consistente é $\text{posto}(\boldsymbol{X}) = n$.

4. Um sistema $\boldsymbol{Xb} = \boldsymbol{y}$ é consistente se, e somente se, $\text{N}(\boldsymbol{X'})$, o espaço nulo da matriz transposta de $\boldsymbol{X}$, for trivial.

Estendendo a análise das propriedades de um sistema para além da sua consistência, o estudo da unicidade da solução também é algo relevante na abordagem dos problemas inversos.

Não é uma tarefa tão complicada estabelecer uma relação entre o posto de $\boldsymbol{X}$ e a existência ou não de múltiplas soluções. No caso específico em que a matriz $\boldsymbol{X}$ possui posto menor do que p, o problema inverso $\boldsymbol{Xb} = \boldsymbol{y}$ é dito mal-posto por não satisfazer a condição de unicidade. Detalhadamente, se a matriz de dados possui tal característica, tem-se uma dependência linear entre as variáveis, implicando que a equação vetorial

$$b_0\mathbf{1} + b_1\boldsymbol{X_{.1}} + b_2\boldsymbol{X_{.2}} + \cdots + b_p\boldsymbol{X_{.p}} = \mathbf{0} \tag{2.2}$$

possua soluções diferentes da solução trivial $\boldsymbol{b} = \mathbf{0}$. Isso equivale a dizer que o sistema homogêneo $\boldsymbol{X}\boldsymbol{b} = \mathbf{0}$ possui soluções não triviais gerando um espaço nulo de $\boldsymbol{X}$ com mais elementos do que apenas $\boldsymbol{b} = \mathbf{0}$. Consequentemente, se o sistema homogêneo possui mais de uma solução, ele possui infinitas soluções.

De fato, seja $\mathrm{N}(\boldsymbol{X})$ o espaço nulo da matriz de dados e seja $\boldsymbol{b_a}$ e $\boldsymbol{b_b}$ dois elementos de $\mathrm{N}(\boldsymbol{X})$, então

$$\boldsymbol{b_c} = \eta \boldsymbol{b_a} + \gamma \boldsymbol{b_b} \tag{2.3}$$

também será uma solução para $\boldsymbol{X}\boldsymbol{b} = \mathbf{0}$ para todo valor real de η e γ, pois

$$\begin{aligned} \boldsymbol{X}\boldsymbol{b_c} &= \boldsymbol{X}[\eta \boldsymbol{b_a} + \gamma \boldsymbol{b_b}] \\ &= \eta \boldsymbol{X}\boldsymbol{b_a} + \gamma \boldsymbol{X}\boldsymbol{b_b} \\ &= \eta \mathbf{0} + \gamma \mathbf{0} \\ &= \mathbf{0}. \end{aligned} \tag{2.4}$$

Todavia, existe uma quantidade limitada de soluções linearmente independentes conforme postula o Teorema 2.1.1.

Teorema 2.1.1. *O sistema linear homogêneo* $\boldsymbol{X}\boldsymbol{b} = \mathbf{0}$, *com* $\mathrm{posto}(\boldsymbol{X}) = r$, *tem exatamente* $p - r$ *vetores solução linearmente independentes.*

Por fim, uma vez que todo vetor solução de um sistema $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{y}$ pode ser obtido através da soma de um vetor solução específico deste sistema e uma combinação linear das $p-r$ soluções linearmente independentes do sistema homogêneo associado, tem-se que se $\mathrm{posto}(\boldsymbol{X}) < p$, então a solução de $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{y}$ não é única, existindo exatamente $p - r + 1$ soluções linearmente independentes.

2.2 Sistemas Consistentes

Em um sistema consistente $\boldsymbol{Xb} = \boldsymbol{y}$, o estudo da inversão de $\boldsymbol{X}$ é importante para a obtenção e compreensão de certas propriedades da solução $\boldsymbol{b}$. Quando a matriz $\boldsymbol{X}$ é quadrada e de posto completo e, portanto, não singular, descrever $\boldsymbol{b}$ através de $\boldsymbol{X}^{-1}$ e de $\boldsymbol{y}$ é uma tarefa simples. No entanto, quando $\boldsymbol{X}$ é singular ou não é quadrada, a solução do sistema consistente pode ser estabelecida por meio da utilização da chamada inversa generalizada.

Definição 2.2.1. *A matriz $\boldsymbol{X}^{-}$ é uma inversa generalizada (g-inversa) de $\boldsymbol{X}$ se, e somente se, $\boldsymbol{X}\boldsymbol{X}^{-}\boldsymbol{X} = \boldsymbol{X}$.*

Através da Definição 2.2.1 é possível mostrar que a g-inversa de uma matriz sempre existe (Souza, 1998), embora nem sempre esta seja única. A Definição 2.2.2 apresenta as condições necessárias para a obtenção da chamada inversa de Moore-Penrose, que é um caso específica de g-inversa que, além de sempre existir, é única.

Definição 2.2.2. *Seja $\boldsymbol{X}$ uma matriz n por p. Existe uma única matriz $\boldsymbol{X}^{+}$ tal que:*

1. *$\boldsymbol{X}^{+}$ é uma inversa generalizada de $\boldsymbol{X}$;*
2. *$\boldsymbol{X}$ é uma inversa generalizada de $\boldsymbol{X}^{+}$;*
3. *$\boldsymbol{X}\boldsymbol{X}^{+}$ e $\boldsymbol{X}^{+}\boldsymbol{X}$ são projeções ortogonais.*

Satisfeitas as três condições, $\boldsymbol{X}^{+}$ é chamada de inversa de Moore-Penrose.

A aplicação da inversa de Moore-Penrose, que frequentemente é chamada de pseudoinversa, é bastante conveniente já que é única e está atrelada à decomposição em valores singulares (SVD).

Como na versão compacta da decomposição SVD, apresentada de forma mais detalhada no Apêndice A, as matrizes $\boldsymbol{U_r}$ e $\boldsymbol{V_r}$ são ambas ortogonais, tem-se que a obtenção da matriz pseudoinversa de $\boldsymbol{X}$ pode ser viabilizada tal que $\boldsymbol{X}^{+} = \boldsymbol{V_r}\boldsymbol{S_r^{-1}}\boldsymbol{U_r'}$.

Definido $\boldsymbol{X}^{+}$, é possível obter uma quantidade $\boldsymbol{b}_{+}$ tal que:

$$\boldsymbol{b}_{+} = \boldsymbol{X}^{+}\boldsymbol{y},$$

$$\boldsymbol{b}_{+} = \boldsymbol{V}_{\boldsymbol{r}}\boldsymbol{S}_{\boldsymbol{r}}^{-\mathbf{1}}\boldsymbol{U}_{\boldsymbol{r}}'\boldsymbol{y} = \sum_{j=1}^{r} \frac{\boldsymbol{U}_{\cdot\boldsymbol{j}}'\boldsymbol{y}}{s_j} V_{\cdot j}. \tag{2.5}$$

Embora o termo $\boldsymbol{b}_{+}$ esteja bem definido pela decomposição SVD, resta mostrar que este é, de fato, uma solução possível para todo sistema consistente.

O Teorema 2.2.1 fornece um resultado muito importante que é utilizado para provar que $\boldsymbol{b}_{+}$ é uma solução para os sistemas consistentes.

Teorema 2.2.1. *O sistema* $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{y}$ *é consistente se, e somente se, a igualdade* $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{X}^{-}\boldsymbol{y}$ *é verdadeira.*

Seja $\boldsymbol{b}_{+}$ um candidato a vetor solução de um sistema consistente. Se $\boldsymbol{b}_{+}$ for pré-multiplicado por $\boldsymbol{X}$, tem-se que

$$\boldsymbol{X}\boldsymbol{b}_{+} = \boldsymbol{X}\boldsymbol{X}^{+}\boldsymbol{y}. \tag{2.6}$$

Considerando a consistência do sistema, lembrando que a inversa de Moore-Penrose é um caso específico de g-inversa e recorrendo ao Teorema 2.2.1, tem-se que

$$\boldsymbol{X}\boldsymbol{b}_{+} = \boldsymbol{y} \tag{2.7}$$

é uma igualdade válida e, portanto, fica demostrado que $\boldsymbol{b}_{+}$ é, de fato, um vetor solução para o sistema consistente $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{y}$.

Dessa forma, a obtenção de $\boldsymbol{b}_{+}$, chamada de solução pseudoinversa, é uma alternativa conveniente e bastante útil para os casos em que $\boldsymbol{X}^{-\mathbf{1}}$ não existe. Ademais, a solução pseudoinversa possui ótimas propriedades, herdadas da decomposição SVD e que serão abordadas nas próximas seções.

2.3 Sistemas Inconsistentes

Em muitos problemas práticos, não existe uma solução exata para o sistema de equações de tal forma que este se caracterize pela sua inconsistência. Na análise estatística de experimentos, assim como na modelagem de dados observacionais, a inconsistência é interpretada como uma consequência do efeito de variações exógenas ao modelo, ou seja, quantidades que não estão contempladas e nem associadas com as informações contidas nas colunas da matriz $\boldsymbol{X}$. Devido a essas variações exógenas, os valores observados para o vetor $\boldsymbol{y}$, em um dado fenômeno, irão ser ligeiramente diferentes do que deveriam ser caso elas não existissem. Assim sendo, o sistema $\boldsymbol{Xb} = \boldsymbol{y}$ é perturbado o suficiente a ponto de se criar a inconsistência.

Em tais situações, procura-se um valor de $\boldsymbol{b}$ que chegue o mais próximo possível do que seria uma solução exata para $\boldsymbol{Xb} = \boldsymbol{y}$. O método de mínimos quadrados ordinários (MQO) é uma das técnicas, dentre diversas outras, que possibilita a obtenção de alguma informação desses sistemas inconsistentes através de soluções aproximadas baseadas na minimização da quantidade $\| \boldsymbol{Xb} - \boldsymbol{y} \|_2^2$.

Definição 2.3.1. *Um vetor $\boldsymbol{b_{MQ}}$ é definido como um vetor solução aproximada de mínimos quadrados para o sistema inconsistente $\boldsymbol{Xb} = \boldsymbol{y}$ se, e somente se,*

$$\| \boldsymbol{Xb_{MQ}} - \boldsymbol{y} \|_2^2 \leq \| \boldsymbol{Xb_o} - \boldsymbol{y} \|_2^2 \, , \tag{2.8}$$

para qualquer outra solução aproximada $\boldsymbol{b_o}$ possível.

Como, no sistema inconsistente, $\boldsymbol{y}$ não pertence ao espaço coluna de $\boldsymbol{X}(\boldsymbol{y} \notin C(\boldsymbol{X}))$, o que o método de mínimos quadrados faz é, em outras palavras, resolver um sistema de equações no qual, ao invés de $\boldsymbol{y}$, tem-se sua projeção no espaço coluna de $\boldsymbol{X}$:

$$\boldsymbol{Xb} = \underset{C(\boldsymbol{X})}{\text{proj}}\, \boldsymbol{y}. \tag{2.9}$$

Uma abordagem alternativa, equivalente e computacionalmente conveniente do problema é obter a solução de mínimos quadrados resolvendo o chamado sistema

de equações normais, que é consistente e associado a $\boldsymbol{Xb} = \boldsymbol{y}$.

Definição 2.3.2. *Seja $\boldsymbol{Xb} = \boldsymbol{y}$ um sistema linear. O sistema $\boldsymbol{X'Xb} = \boldsymbol{X'y}$, dito sistema de equações normais, é sempre consistente e um vetor $\boldsymbol{b}$ qualquer é solução de $\boldsymbol{X'Xb} = \boldsymbol{X'y}$ se, e somente se, $\boldsymbol{b}$ é uma solução de mínimos quadrados de $\boldsymbol{Xb} = \boldsymbol{y}$.*

A Definição 2.3.3 apresenta um tipo específico de inversa generalizada que está diretamente ligada ao vetor solução $\boldsymbol{b_{MQ}}$.

Definição 2.3.3. *Seja $\boldsymbol{X}$ uma matriz de posto r qualquer e $\boldsymbol{X^*}$ uma matriz que satisfaz as condições:*

1. $\boldsymbol{XX^*X} = \boldsymbol{X}$;
2. $\boldsymbol{XX^*} = (\boldsymbol{XX^*})'$.

Então, $\boldsymbol{X^*}$ é dita inversa generalizada de mínimos quadrados de $\boldsymbol{X}$.

A matriz $\boldsymbol{X^*}$ pode ser expressa em termos de uma g-inversa qualquer de tal forma que toda matriz $\boldsymbol{X^*} = (\boldsymbol{X'X})^{-}\boldsymbol{X'}$ é uma inversa de mínimos quadrados de $\boldsymbol{X}$. A solução aproximada pode, então, ser descrita em termos de $\boldsymbol{X^*}$ tal que $\boldsymbol{b_{MQ}} = \boldsymbol{X^*y}$.

Embora $\boldsymbol{b_{MQ}}$ seja uma solução ótima, já que minimiza $\| \boldsymbol{Xb} - \boldsymbol{y} \|_2^2$, esta não necessariamente é única, podendo existir outras que satisfaçam a mesma condição. Diante disso, a Definição 2.3.4 apresenta o conceito de melhor solução aproximada (BAS:*Best Approximate Solution*).

Definição 2.3.4. *Um vetor $\boldsymbol{b}_+$ é definido como a melhor solução aproximada (BAS) de um sistema de equações lineares inconsistente se, e somente se, para qualquer outra solução aproximada $\boldsymbol{b_o}$:*

1. $\| \boldsymbol{Xb}_+ - \boldsymbol{y} \|_2^2 \leq \| \boldsymbol{Xb_o} - \boldsymbol{y} \|_2^2$;
2. $\| \boldsymbol{b}_+ \| \leq \| \boldsymbol{b_o} \|$.

Assim, a melhor solução aproximada é única, além de ser um caso específico de solução de mínimos quadrados.

Teorema 2.3.1. *A melhor solução aproximada de um sistema inconsistente é* $\boldsymbol{b}_+ = \boldsymbol{X}^+\boldsymbol{y}$.

Pelo Teorema 2.3.1, constata-se que a solução pseudoinversa, que já foi apresentada na seção anterior como uma solução possível para os sistemas consistentes, também é útil em sistemas inconsistentes, para os quais ela configura-se uma solução aproximada de norma mínima.

Uma vez que a solução pseudoinversa também resolve um sistema de equações normais, pode-se demonstrar facilmente que a norma da solução pseudoinversa é a menor dentre todas as outras normas de soluções de mínimos quadrados.

Dessa forma,

$$(\boldsymbol{X'X})\boldsymbol{b_{mq}} = \boldsymbol{X'y}, \tag{2.10}$$

assim como

$$(\boldsymbol{X'X})\boldsymbol{b}_+ = \boldsymbol{X'y} \tag{2.11}$$

e, portanto,

$$(\boldsymbol{X'X})(\boldsymbol{b_{mq}} - \boldsymbol{b}_+) = (\boldsymbol{X'X})\boldsymbol{b_{Null}} = \boldsymbol{0}, \tag{2.12}$$

em que $\boldsymbol{b_{Null}} \in \mathrm{N}(\boldsymbol{X})$ haja vista $\mathrm{N}(\boldsymbol{X'X}) \equiv \mathrm{N}(\boldsymbol{X})$. Baseando-se na Equação (2.12), $\boldsymbol{b_{mq}}$ pode ser descrito pela soma de $\boldsymbol{b}_+$ e um valor arbitrário pertencente ao espaço nulo de $\boldsymbol{X}$ tal que

$$\boldsymbol{b_{mq}} = \boldsymbol{b}_+ + \boldsymbol{b_{Null}}. \tag{2.13}$$

Em uma decomposição SVD, em que $\boldsymbol{X}$ tem posto r, $\boldsymbol{V_0} = [\boldsymbol{V_{.(r+1)}}, \boldsymbol{V_{.(r+2)}}, \ldots, \boldsymbol{V_{.p}}]$ é uma matriz cujas colunas formam uma base ortonormal para o espaço nulo de $\boldsymbol{X}$. Substituindo $\boldsymbol{b_{Null}}$ por uma combinação linear das colunas de $\boldsymbol{V_0}$, tem-se

$$\boldsymbol{b_{mq}} = \boldsymbol{b}_+ + \sum_{i=r+1}^{p} \alpha_i \boldsymbol{V_{.i}}, \tag{2.14}$$

em que $\alpha_i \in \mathbb{R}$.

Como as colunas de $\boldsymbol{V}$ são ortonormais, $\boldsymbol{b_+}$ e $\boldsymbol{b_{Null}}$ são independentes e a norma da soma é igual à soma das normas, valendo a mesma regra para o quadrado da norma. A demonstração deste resultado encontra-se no Apêndice B. Diante disso,

$$\| \boldsymbol{b_{mq}} \|_2^2 = \| \boldsymbol{b_+} \|_2^2 + \| \sum_{i=r+1}^{p} \alpha_i \boldsymbol{V_{.i}} \|_2^2 \tag{2.15}$$

$$= \| \boldsymbol{b_+} \|_2^2 + \sum_{i=r+1}^{p} \alpha_i^2 \| \boldsymbol{V_{.i}} \|_2^2 = \| \boldsymbol{b_+} \|_2^2 + \sum_{i=r+1}^{p} \alpha_i^2. \tag{2.16}$$

Portanto,

$$\| \boldsymbol{b_{mq}} \|_2^2 = \| \boldsymbol{b_+} \|_2^2 + \sum_{i=r+1}^{p} \alpha_i^2 \geq \| \boldsymbol{b_+} \|_2^2 . \tag{2.17}$$

Quando $r = p$, o único $\boldsymbol{b_{Null}}$ possível é $\mathbf{0}$ e existe uma única solução de mínimos quadrados: $\boldsymbol{b_{mq}} = \boldsymbol{b_+}$. Já quando $r < p$, a solução $\boldsymbol{b_+}$ será a que possui a menor norma dentre todas as outras possíveis soluções de mínimos quadrados existentes.

2.4 Situações Gerais Possíveis

Em linhas gerais, a decomposição SVD fornece algumas informações a respeito de um sistema linear em estudo, assim como estabelece certas propriedades para a respectiva solução. Avaliando as dimensões da matriz $\boldsymbol{X}$ e o seu posto, pode-se definir quatro cenários possíveis para um sistema linear:

1. Quando $\boldsymbol{X}$ é uma matriz quadra de posto completo, $n = p = r$, tem-se que $\boldsymbol{U_r'} = \boldsymbol{U_r^{-1}}$ e $\boldsymbol{V_r'} = \boldsymbol{V_r^{-1}}$ tais que

$$\boldsymbol{X^+} = \boldsymbol{V_r S_r^{-1} U_r'}$$

$$= (\boldsymbol{U_r S_r V_r'})^{-1}$$

$$= \boldsymbol{X}^{-1}. \tag{2.18}$$

Como N($\boldsymbol{X'}$) é trivial, o sistema é consistente, permitindo que a solução $\boldsymbol{b_+} =$

$\boldsymbol{X}^{+}\boldsymbol{y}$ ajuste-se com exatidão aos dados. Como verificado anteriormente, a existência de múltiplas soluções para um sistema linear depende estritamente do fato de N($\boldsymbol{X}$) ser trivial ou não, ou seja, de $\boldsymbol{V_r}$ ser uma matriz quadrada p por p. Uma vez que N($\boldsymbol{X}$) é trivial, tem-se que o vetor solução $\boldsymbol{b}_+$ é o único possível para o sistema deste cenário.

2. Se $n = r$ e $r < p$, então N($\boldsymbol{X'}$) é trivial, enquanto N($\boldsymbol{X}$) não. Outra consequência desta possível configuração é que $\boldsymbol{U_r}$ é uma matriz quadrada, e portanto, $\boldsymbol{U_r'} = \boldsymbol{U_r^{-1}}$, mas $\boldsymbol{V_r}$ não é quadrada e só é possível obter a igualdade $\boldsymbol{V_r'V_r} = \boldsymbol{I_r}$. Assim,

$$\begin{aligned}\boldsymbol{b}_+ &= \boldsymbol{V_r S_r^{-1} U_r' y}\\ &= \boldsymbol{V_r S_r U_r' U_r S_r^{-1} S_r^{-1} U_r' y}\\ &= \boldsymbol{X'(U_r S_r^{-1} S_r^{-1} U_r')y = X'(XX')^{-1}y}. \end{aligned} \tag{2.19}$$

Dessa forma, o sistema é consistente e $\boldsymbol{b}_+$ é uma solução exata, embora esta não seja única. Neste cenário com múltiplas soluções para $\boldsymbol{Xb} = \boldsymbol{y}$, $\boldsymbol{b}_+$ é a solução exata que possui a menor norma.

3. Se $n > r$ e $r = p$, então N($\boldsymbol{X}$) é trivial e N($\boldsymbol{X'}$) não. Em outras palavras, o sistema é inconsistente e não existe uma solução exata que ajuste os dados com perfeição. Como N($\boldsymbol{X}$) é trivial, existe apenas uma única solução aproximada para o problema. Sabendo que $\boldsymbol{V_r'} = \boldsymbol{V_r^{-1}}$, $\boldsymbol{U_r'U_r} = \boldsymbol{I_r}$ e $(\boldsymbol{X'X})^{-1} = \boldsymbol{V_r S_r^{-1} S_r^{-1} V_r'}$, tem-se que

$$\begin{aligned}\boldsymbol{b}_+ &= \boldsymbol{X}^{+}\boldsymbol{y} = \boldsymbol{V_r S_r^{-1} U_r' y}\\ &= \boldsymbol{V_r S_r^{-1} S_r^{-1} V_r' V_r S_r U_r' y}\\ &= \boldsymbol{(X'X)^{-1} X'y}. \end{aligned} \tag{2.20}$$

Vale lembrar que, por $\boldsymbol{b}_+$ ser uma solução aproximada neste cenário, $\boldsymbol{Xb}_+$ projeta $\boldsymbol{y}$ no espaço coluna de $\boldsymbol{X}$, fornecendo o ponto do espaço coluna que é

mais próximo de $\boldsymbol{y}$.

4. Se $r < n$ e $r < p$, então tanto N($\boldsymbol{X}$) como N($\boldsymbol{X'}$) não são triviais. Assim, a solução pseudoinversa é uma solução aproximada, embora não seja a única solução aproximada possível. Todavia, a solução pseudoinvesa é a melhor solução aproximada (BAS) já que é baseada também na minimização de $\| \boldsymbol{b} \|$. Assim, como no terceiro cenário em que $n > r$ e $r = p$, tem-se que

$$\boldsymbol{X}\boldsymbol{b}_+ = \boldsymbol{U}_r\boldsymbol{S}_r\boldsymbol{V}_r'\boldsymbol{V}_r\boldsymbol{S}_r^{-1}\boldsymbol{U}_r'\boldsymbol{y} = \boldsymbol{U}_r\boldsymbol{U}_r'\boldsymbol{y} = \underset{\mathrm{C}(\boldsymbol{X})}{\mathrm{proj}}\,\boldsymbol{y}, \tag{2.21}$$

em que C($\boldsymbol{X}$) é o espaço coluna de $\boldsymbol{X}$.

Tomando como base os quatro cenários acima, fica explícito que a solução pseudoinversa é uma solução bastante versátil por resolver sistemas consistentes e inconsistentes nas suas mais variadas formas.

2.5 O Método TSVD

Além da existência e unicidade de uma solução, a sua estabilidade também é um fator importante e está diretamente ligada ao conceito de *overfitting*. De fato, falar em instabilidade de uma solução é equivalente a dizer que a solução do sistema $\boldsymbol{X}\boldsymbol{b} = \boldsymbol{y}$ está muito sensível às variações em $\boldsymbol{y}$, ou seja, a solução é útil apenas para aqueles dados observados em $\boldsymbol{y}$ e será de pouca utilidade para previsões.

A estabilidade da solução pseudoinversa é um fator que depende das características da sequência de valores singulares. Pequenos valores singulares geram soluções $\boldsymbol{b}_+$ instáveis. Se a decomposição SVD for truncada de tal forma que se tenha t valores singulares e $t < r$, então é possível obter uma solução TSVD (*Truncated Singular Value Decomposition*) que será mais estável se comparada à solução pseudoinversa baseada no SVD compacto, em que se utiliza todos os valores singulares. Todavia, o ajuste dos dados baseado na solução TSVD pode ser pior devido à redução do subespaço no qual a solução recai (Aster et al., 2011). A solução TSVD pode, então,

ser descrita da seguinte forma:

$$\boldsymbol{b_{TSVD}} = \boldsymbol{V_t S_t^{-1} U_t' y} = \sum_{j=1}^{t} \frac{\boldsymbol{U_{.j}' y}}{s_j} V_{.j}. \tag{2.22}$$

Embora, a solução TSVD não ajuste tão bem quanto a solução pseudoinversa baseada no SVD compacto, essa é melhor em tarefas de previsão já que evita o *overfitting*. O *trade-off*, então, entre a qualidade do ajuste e o *overfitting*, é estabelecido e a escolha do melhor t a ser utilizado basear-se-á em métodos que avaliam o comportamento destes dois fatores conjuntamente.

2.6 A Regularização de Tikhonov

O *trade-off* entre o *overfitting* e a qualidade do ajuste que um determinado modelo fornece é a principal característica em comum dentre os métodos de regularização. O que diferencia estes diversos métodos é a forma com a qual os cálculos são realizados para se obter os valores preditos e a regra de decisão utilizada para definir o ponto ótimo deste *trade-off*.

A regularização de Tikhonov foi desenvolvida para lidar com sistemas lineares do tipo $\boldsymbol{Xb} = \boldsymbol{y}$ que são mal-postos. Como já discutido nas seções anteriores, isso pode acontecer quando o sistema é inconsistente, admite infinitas soluções ou apresenta-se instável. Objetivando-se obter resultados com propriedades apropriadas, a regularização de Tikhonov de ordem zero fornece uma solução obtida a partir do seguinte problema de otimização:

$$\min \left\{ \parallel \boldsymbol{Xb} - \boldsymbol{y} \parallel_2^2 + \alpha^2 \parallel \boldsymbol{b} \parallel_2^2 \right\}, \tag{2.23}$$

para algum $\alpha \geq 0$. Na Equação (2.23) a constante α deve ser definida por quem está implementando o modelo e é dita um parâmetro de regularização justamente por controlar o *trade-off* existente.

Assim sendo, a regularização de Tikhonov de ordem zero é equivalente à re-

gressão Ridge, técnica amplamente difundida na literatura estatística e que será discutida no Capítulo 3. Todavia, existe uma definição mais abrangente para a regularização de Tikhonov tal que a regularização imposta no problema seja definida por um α e por uma norma euclidiana $\| \boldsymbol{Lb} \|_2$. Assim, o método de regularização proposto por Tikhonov baseia-se em resolver o seguinte problema de mínimos quadrados regularizado:

$$\min \left\{ \| \boldsymbol{Xb} - \boldsymbol{y} \|_2^2 + \alpha^2 \| \boldsymbol{Lb} \|_2^2 \right\}, \tag{2.24}$$

em que $\boldsymbol{L}$ é uma matriz $(p - Ord)$ por p (Ord é a ordem da regularização), baseada no conceito de diferenciação do vetor solução b. Quando $\boldsymbol{L} = \boldsymbol{I_{n \times n}}$, tem-se a regularização de Tikhonov de ordem zero, apresentada em (2.23).

Na regularização de Tikhonov de primeira ordem, o elemento de regularização $\| \boldsymbol{Lb} \|_2^2$ é descrito pela soma de quadrados das primeiras diferenças $(b_2 - b_1)$, $(b_3 - b_2)$, $\ldots$, $(b_n - b_{n-1})$ de tal forma que

$$\| \boldsymbol{Lb} \|_2^2 = (\boldsymbol{Lb})' \boldsymbol{Lb} = \sum_{i=1}^{n-1} (b_{i+1} - b_i)^2, \tag{2.25}$$

em que

$$\boldsymbol{L} = \begin{bmatrix} -1 & 1 & 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \ldots & 0 & 0 & \ldots & -1 & 1 & 0 \\ 0 & \ldots & 0 & 0 & \ldots & 0 & -1 & 1 \end{bmatrix}_{(p-1) \times p} . \tag{2.26}$$

Já quando

$$\boldsymbol{L} = \begin{bmatrix} -1 & 2 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & 0 & \dots & -1 & 2 & 1 & 0 \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & -1 & 2 & 1 \end{bmatrix}_{(p-2)\times p}, \tag{2.27}$$

tem-se uma regularização de Tikhonov de segunda ordem e a quantidade $\| \boldsymbol{Lb} \|_2^2$ é baseada na diferença das primeiras diferenças: $(b_3 - b_2) - (b_2 - b_1)$, $(b_4 - b_3) - (b_3 - b_2)$, $\dots$, $(b_n - b_{n-1}) - (b_{n-1} - b_{n-2})$. Do mesmo modo, as chamadas regularização de Tikhonov de ordem superior seguem a mesma linha de raciocínio desenvolvida para as regularizações de primeira e segunda ordem.

Quanto maior a ordem da regularização de Tikhonov, mais próximas umas das outras serão as componentes do vetor solução, permitindo uma contribuição para a previsão de $\boldsymbol{y}$ mais igualitária entre as variáveis do modelo. Quando $\boldsymbol{X}$ possui muitas variáveis muito correlacionadas, a regularização de ordem maior do que zero torna-se uma ferramente bastante útil e recomendada.

Portanto, a regularização de Tikhonov explora o procedimento de mínimos quadrados adicionando um fator de regularização com o objetivo de reduzir o *overfitting*, que muitas vezes acabam por dominar as soluções que são baseadas no procedimento padrão de mínimos quadrados.

O problema de minimização (2.24) é equivalente a um problema linear de mínimos quadrados padrão (Aster et al., 2011):

$$\min \left\| \begin{pmatrix} \boldsymbol{X} \\ \alpha \boldsymbol{L} \end{pmatrix} \boldsymbol{b} - \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} \right\|_2^2. \tag{2.28}$$

O espaço nulo da matriz $\begin{pmatrix} \boldsymbol{X} \\ \alpha \boldsymbol{L} \end{pmatrix}$ é equivalente à interseção entre os espaços nulos de $\boldsymbol{X}$ e de $\boldsymbol{L}$. Assim, se $\mathrm{N}(\boldsymbol{X}) \cap \mathrm{N}(\boldsymbol{L})$ for trivial, existirá uma única solução

para (2.28).

Diante do problema de minimização estabelecido em (2.28),

$$\boldsymbol{b_{Tk}} = \left[\begin{pmatrix} \boldsymbol{X'} & \alpha \boldsymbol{L'} \end{pmatrix} \begin{pmatrix} \boldsymbol{X} \\ \alpha \boldsymbol{L} \end{pmatrix} \right]^{-1} \begin{pmatrix} \boldsymbol{X'} & \alpha \boldsymbol{L'} \end{pmatrix} \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} \tag{2.29}$$

$$= (\boldsymbol{X'X} + \alpha^2 \boldsymbol{L'L})^{-1} \boldsymbol{X'y}. \tag{2.30}$$

2.6.1 A Implementação SVD

Através de uma abordagem baseada na decomposição SVD, a regularização de Tikhonov de ordem zero, diferentemente do método TSVD apresentado na seção 2.5, não deixa de levar em consideração os valores singulares que são pequenos demais. Ao invés disso, ela utiliza uma informação a priori que controla o nível de regularização, reduzindo, e não excluindo, o impacto dos valores singulares pequenos na obtenção da solução do sistema linear.

Seja $\boldsymbol{b_{Tk_0}}$ o vetor solução para a regularização de ordem zero,

$$\boldsymbol{b_{Tk_0}} = (\boldsymbol{X'X} + \alpha^2 \boldsymbol{I})^{-1} \boldsymbol{X'y}.$$

Como se tem a garantia da existência da inversa de $(\boldsymbol{X'X} + \alpha^2 \boldsymbol{I})$, mas não se tem a garantia da existência de $(\boldsymbol{X'X})^{-1}$, passa-se a matriz $(\boldsymbol{X'X} + \alpha^2 \boldsymbol{I})$ para o lado esquerdo da equação a fim de se utilizar a decomposição SVD e se obter uma versão simplificada da solução baseada nos valores singulares:

$$\left[\boldsymbol{VS'U'USV'} + \alpha^2 \boldsymbol{I}\right] \boldsymbol{b_{Tk_0}} = \boldsymbol{VS'U'y},$$

$$\left[\boldsymbol{VS'SV'} + \alpha^2 \boldsymbol{I}\right] \boldsymbol{b_{Tk_0}} = \boldsymbol{VS'U'y},$$

$$\left[\boldsymbol{VS'SV'} + \alpha^2 \boldsymbol{VV'}\right] \boldsymbol{b_{Tk_0}} = \boldsymbol{VS'U'y},$$

$$\boldsymbol{V}\left[\boldsymbol{S'S} + \alpha^2\right] \boldsymbol{V'b_{Tk_0}} = \boldsymbol{VS'U'y},$$

$$\left[\boldsymbol{S'S} + \alpha^2\right] \boldsymbol{V'b_{Tk_0}} = \boldsymbol{S'U'y},$$

$$\boldsymbol{V'b_{Tk_0}} = \left[\boldsymbol{S'S} + \alpha^2\right]^{-1} \boldsymbol{S'U'y},$$

$$\boldsymbol{b_{Tk_0}} = \boldsymbol{V}\left[\boldsymbol{S'S} + \alpha^2\right]^{-1} \boldsymbol{S'U'y}$$

$$= \boldsymbol{V}\left[\boldsymbol{S'S} + \alpha^2\right]^{-1} \boldsymbol{S'U'y}. \tag{2.31}$$

Dessa maneira,

$$\boldsymbol{b_{Tk_0}} = \boldsymbol{V_{p\times p}} \begin{bmatrix} \frac{s_1}{s_1^2+\alpha^2} & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & \frac{s_2}{s_2^2+\alpha^2} & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{s_{\min(n,p)}}{s^2_{\min(n,p)}+\alpha^2} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & 0 \end{bmatrix}_{p\times n} \boldsymbol{U'_{n\times n}y_{n\times 1}} \tag{2.32}$$

$$= \begin{bmatrix} V_{11}\frac{s_1}{s_1^2+\alpha^2} & V_{12}\frac{s_2}{s_2^2+\alpha^2} & \dots & V_{1\,\min(n,p)}\frac{s_{\min(n,p)}}{s^2_{\min(n,p)}+\alpha^2} & 0 & \dots & 0 \\ V_{21}\frac{s_1}{s_1^2+\alpha^2} & V_{22}\frac{s_2}{s_2^2+\alpha^2} & \dots & V_{2\,\min(n,p)}\frac{s_{\min(n,p)}}{s^2_{\min(n,p)}+\alpha^2} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ V_{p1}\frac{s_1}{s_1^2+\alpha^2} & V_{p2}\frac{s_2}{s_2^2+\alpha^2} & \dots & V_{p\,\min(n,p)}\frac{s_{\min(n,p)}}{s^2_{\min(n,p)}+\alpha^2} & 0 & \dots & 0 \end{bmatrix}_{p\times n} \begin{bmatrix} \langle \boldsymbol{U_{.1}}, \boldsymbol{y}\rangle \\ \langle \boldsymbol{U_{.2}}, \boldsymbol{y}\rangle \\ \vdots \\ \langle \boldsymbol{U_{.n}}, \boldsymbol{y}\rangle \end{bmatrix}_{n\times 1} \tag{2.33}$$

$$= \begin{bmatrix} \sum_{j=1}^{n} \frac{s_j}{s_j^2+\alpha^2}\langle \boldsymbol{U_{.j}}, \boldsymbol{y}\rangle V_{1j} \\ \sum_{j=1}^{n} \frac{s_j}{s_j^2+\alpha^2}\langle \boldsymbol{U_{.j}}, \boldsymbol{y}\rangle V_{2j} \\ \vdots \\ \sum_{j=1}^{n} \frac{s_j}{s_j^2+\alpha^2}\langle \boldsymbol{U_{.j}}, \boldsymbol{y}\rangle V_{pj} \end{bmatrix}_{n\times 1} = \sum_{j=1}^{n} \frac{s_j\langle \boldsymbol{U_{.j}}, \boldsymbol{y}\rangle}{s_j^2+\alpha^2} V_{.j} = \sum_{j=1}^{n} \frac{s_j \boldsymbol{U'_{.j}y}}{s_j^2+\alpha^2} V_{.j}. \tag{2.34}$$

Como só existem no máximo $\min(n,p)$ valores singulares diferentes de zero, a solução $\boldsymbol{b_{Tk_0}}$ pode ser expressa por

$$\boldsymbol{b_{Tk_0}} = \sum_{j=1}^{\min(n,p)} \frac{s_j \boldsymbol{U'_{.j}y}}{s_j^2+\alpha^2} V_{.j}, \tag{2.35}$$

ou equivalentemente,

$$\boldsymbol{b_{Tk_0}} = \sum_{j=1}^{\min(n,p)} \frac{s_j^2}{s_j^2+\alpha^2} \frac{\boldsymbol{U'_{.j}y}}{s_j} V_{.j} = \sum_{j=1}^{\min(n,p)} f_j \frac{\boldsymbol{U'_{.j}y}}{s_j} V_{.j}, \tag{2.36}$$

em que f_j é uma quantidade chamada de filtro.

A Equação (2.36) proporciona clareza na interpretação do papel da componente de regularização α. A ideia da regularização é filtrar os valores singulares de tal forma que apenas os melhores influenciem a solução. Quanto mais os valores singulares são maiores do que α, mais próximo de um o filtro será e irá permitir que tal valor influencie na solução, enquanto que quanto mais os valores singulares são menores do que α, mais próximo de zero o filtro será, amenizando o efeito destes valores singulares pequenos. Ademais, fica evidente que ao adicionar α^2 no denominador da Equação (2.35), contorna-se problemas gerados com matrizes de posto incompleto, pois mesmo que exista algum valor singular dentro deste somatório que seja igual a zero, devido à deficiência no rank da matriz, a solução estará bem definida.

2.6.2 A Implementação GSVD

Os problemas envolvendo a regularização de Tikhonov de ordem superior também podem ter soluções caracterizadas pelo somatório de um produto entre um filtro f_i e uma quantidade específica com informações advindas de uma decomposição de matrizes. Como além da matriz $\boldsymbol{X}$ tem-se a matriz $\boldsymbol{L}$, a decomposição SVD não é viável, fazendo-se necessário recorrer à decomposição em valores singulares generalizada (GSVD).

Definição 2.6.1. *Seja $\boldsymbol{X}_{n\times p}$ e $\boldsymbol{L}_{l\times p}$, em que $l = (p - Ord)$ e $\boldsymbol{L}$ é uma matriz de posto completo. Então, existe $\boldsymbol{U_{n\times n}}$ e $\boldsymbol{V_{l\times l}}$, ambas matrizes ortogonais, e $\boldsymbol{D_{p\times p}}$, que é uma matriz não singular, tais que:*

1. $\boldsymbol{X} = \boldsymbol{U\Lambda D'}$, *em que $\boldsymbol{\Lambda}$ é uma matriz esparsa n por p com elementos diagonais $0 \leq \Lambda_{1,k+1} \leq \Lambda_{2,k+2} \leq \cdots \leq \Lambda_{\min(n,p),k+\min(n,p)}$, sendo $k = 0$ quando $n > p$ e $k = p - n$ quando $n \leq p$.*

2. $\boldsymbol{L} = \boldsymbol{VMD'}$, *em que $\boldsymbol{M}$ é uma matriz diagonal l por p cujos elementos se dispõem em ordem decrescente tal que $M_{1,1} \geq M_{2,2} \geq \cdots \geq M_{l,l} \geq 0$.*

3. $\boldsymbol{M'M} + \boldsymbol{\Lambda'\Lambda} = \boldsymbol{I}$.

Definida a decomposição GSVD de $\boldsymbol{X}$ e $\boldsymbol{L}$ e seja $\boldsymbol{\lambda} = \sqrt{\text{diag}(\boldsymbol{\Lambda'\Lambda})}$ e $\boldsymbol{\mu} = \sqrt{\text{diag}(\boldsymbol{M'M})}$, os valores singulares generalizados compõem um vetor $\boldsymbol{\gamma}$ com p componentes tal que

$$\boldsymbol{\gamma} = \frac{\boldsymbol{\lambda}}{\boldsymbol{\mu}}. \tag{2.37}$$

Aplicando o GSVD na equação normal:

$$(\boldsymbol{X'X} + \alpha^2\boldsymbol{L'L})\boldsymbol{b_{Tkl}} = \boldsymbol{X'y},$$

$$(\boldsymbol{D\Lambda'U'U\Lambda D'} + \alpha^2\boldsymbol{DM'V'VMD'})\boldsymbol{b_{Tkl}} = \boldsymbol{D\Lambda'U'y},$$

$$(\boldsymbol{D\Lambda'\Lambda D'} + \alpha^2\boldsymbol{DM'MD'})\boldsymbol{b_{Tkl}} = \boldsymbol{D\Lambda'U'y},$$

$$\boldsymbol{D}(\boldsymbol{\Lambda'\Lambda} + \alpha^2\boldsymbol{M'M})\boldsymbol{D'b_{Tkl}} = \boldsymbol{D\Lambda'U'y},$$

$$(\boldsymbol{\Lambda'\Lambda} + \alpha^2\boldsymbol{M'M})\boldsymbol{D'b_{Tkl}} = \boldsymbol{\Lambda'U'y},$$

$$\boldsymbol{D'b_{Tkl}} = (\boldsymbol{\Lambda'\Lambda} + \alpha^2\boldsymbol{M'M})^{-1}\boldsymbol{\Lambda'U'y},$$

$$\boldsymbol{b_{Tkl}} = (\boldsymbol{D'})^{-1}(\boldsymbol{\Lambda'\Lambda} + \alpha^2\boldsymbol{M'M})^{-1}\boldsymbol{\Lambda'U'y}, \tag{2.38}$$

ou equivalentemente,

$$\boldsymbol{b_{Tkl}} = (\boldsymbol{D}^{-1})'(\boldsymbol{\Lambda'\Lambda} + \alpha^2\boldsymbol{M'M})^{-1}\boldsymbol{\Lambda'U'y}. \tag{2.39}$$

Fazendo-se $\boldsymbol{T} = (\boldsymbol{D}^{-1})'$, tem-se, assim como para a regularização de ordem zero, uma solução expressa em termos de um somatório e dos filtros:

$$\boldsymbol{b_{Tkl}} = \sum_{j=1}^{p} \frac{\lambda_j \boldsymbol{U'_{.,j+k}y}}{\lambda_j^2 + \alpha^2\mu_j^2}\boldsymbol{T_{.,j}} = \sum_{j=1}^{p} \frac{\gamma_j^2}{\gamma_j^2 + \alpha^2}\frac{\boldsymbol{U'_{.,j+k}y}}{\lambda_j}\boldsymbol{T_{.,j}}, \tag{2.40}$$

em que $f_j = \frac{\gamma_j^2}{\gamma_j^2+\alpha^2}$ são filtros análogos aos obtidos para a regularização de ordem zero.

Na Equação (2.40) pode acontecer casos em que $\lambda_j = \gamma_j = 0$, gerando uma indeterminação do tipo $\frac{0^2}{0}$. Nesses casos, os termos devem ser considerados como

zero. Ademais, podem existir casos em que γ_j é infinito, de tal forma que o filtro seja igual a 1 (Aster et al., 2011).

2.7 Exemplo Prático do TSVD e de Tikhonov

Para ilustrar a utilidade, em situações aplicadas, dos métodos até então apresentados, far-se-á uso de duas bases de dados. A primeira é uma base de dados de próstata que é utilizada nos exemplos do livro Hastie et al. (2013) e que pode ser obtida através da internet: `https://web.stanford.edu/~hastie/ElemStatLearn/data.html`. Já a segunda base é utilizada nos exemplos do livro Kuhn and Johnson (2013) e é composta por dados de origem química, que podem ser utilizados em modelos de regularização para se realizar previsões da solubilidade molecular.

2.7.1 Dados de Próstata

A base de dados de próstata (`PROS`) provém de um estudo realizado em 97 homens sobre os quais foram realizadas medições (clínicas e demográficas) de 8 características diferentes. Tais informações pode ser utilizadas para modelar, através de um sistema linear, quantidades relacionadas a um antígeno chamado PSA (Prostate Specific Antigen), que é um fator de risco associado ao câncer de próstata. Quanto maior o valor do PSA, maior o risco de um indivíduo ter o câncer de próstata.

Em conformidade com Hastie et al. (2013), o sistema linear associado ao problema é definido pela matriz de dados $\boldsymbol{X}$, com 97 linhas e 8 colunas, e pela variável resposta y, que é o logaritmo do PSA. Assim, o objetivo da aplicação dos métodos TSVD e Tikhonov é obter soluções aproximadas que se ajustem bem ao sistema, ao passo que sejam também capazes de realizar boas previsões da variável resposta para futuros indivíduos, ou seja, soluções aproximadas viáveis que evitam o *overfitting*.

A Tabela 2.1 apresenta todas as oito características medidas no estudo para exercerem o papel de variáveis preditoras, assim como o logaritmo do PSA, que é a

variável resposta y.

Significado	Abreviações
Logaritmo do volume do câncer	lcavol
Logaritmo do peso da próstata	lweight
Idade	age
Logaritmo da quantidade de hiperplasia prostática benigna	lbph
Invasão da vesícula seminal	svi
Logaritmo da penetração capsular	lcp
Gleason score	gleason
Percentual do *Gleason score*	pgg45
Logaritmo do antígeno PSA	lPSA

Tabela 2.1: Variáveis que compõem o sistema linear (base `PROS`).

A matriz de correlações apresentada na Figura 2.1 evidencia que o logaritmo do PSA possui correlações significativamente altas com grande parte das variáveis preditoras. Correlações significativas também são encontradas entre as variáveis que compõem as colunas de $\boldsymbol{X}$.

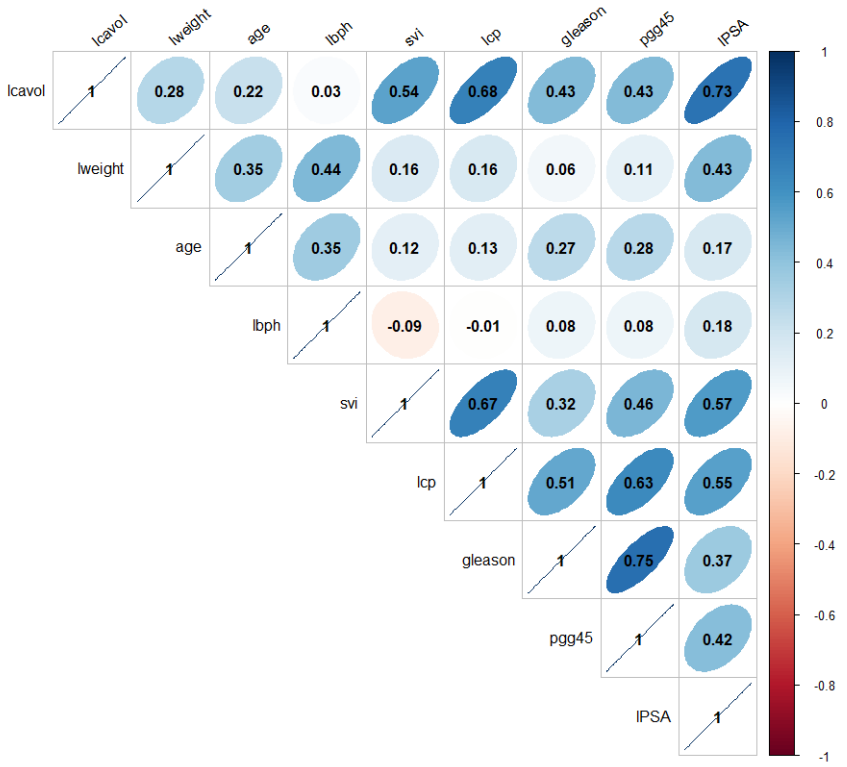


Figura 2.1: Matriz de correlações das variáveis do sistema linear (base `PROS`).

Para a avaliação do desempenho dos modelos de regularização TSVD e Tikhonov, foi implementado o procedimento de validação cruzada 5-fold, através do qual calculou-se o Erro Quadrático Médio (EQM) sobre o conjunto de teste em cada uma das cinco etapas da validação. Através da média desses cinco EQM's, tem-se uma medida de avaliação de desempenho, que aqui será chamada de Erro Quadrático Médio da validação cruzada (EQMcv).

No que diz respeito ao intercepto de cada modelo, realizou-se a padronização das variáveis preditoras. Assim, a estimativa do intercepto passa a ser a média da variável resposta. Como para o procedimento de validação cruzada é necessário dividir o banco de dados, optou-se por realizar a padronização do conjunto de treinamento e a subsequente padronização do conjunto de teste em cada uma das cinco etapas da validação.

A Tabela 2.2 apresenta os EQMcv's para cada parâmetro t do TSVD.

Parâmetro TSVD	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$
EQMcv	0.752	0.742	0.639	0.645	0.631	0.621	0.603	0.597

Tabela 2.2: EQMcv's segundo a quantidade de valores singulares utilizada no modelo (base `PROS`).

Como quanto menor o EQMcv, melhor é o modelo, tem-se que o melhor parâmetro do TSVD para este conjunto de dados é $t = 8$. Em outras palavras, como para $t = 8$ utiliza-se todos os valores singulares, o resultado da validação cruzada aponta que a solução pseudoinversa é a mais apropriada quando comparada com as outras possíveis soluções TSVD.

Um EQMcv próximo do retornado pela solução pseudoinversa é observado quando $t = 7$. Todavia, para regularizações mais intensas com valores de $t < 7$, a performance preditiva dos respectivos modelos não foi igualmente satisfatória.

Já para a regularização de Tikhonov, foi realizada uma implementação computacional baseada nas decomposições SVD e GSVD, sendo testados os métodos de ordem zero, um e dois. Como o parâmetro de regularização α pode assumir qualquer valor tal que $\alpha \geq 0$, tem-se a necessidade de se definir um *grid* de busca. Assim, o

grid escolhido abrange uma sequência que vai de 0 a 7, variando 0.025 de elemento a elemento.

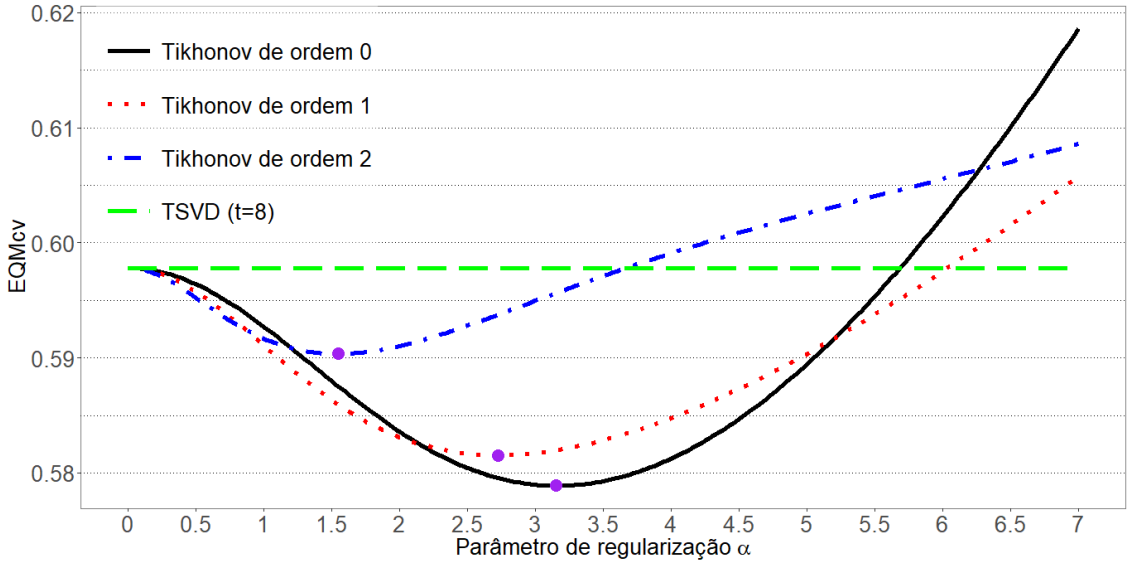


Figura 2.2: Sensibilidade da métrica de avaliação de desempenho segundo os valores de α para cada regularização de Tikhonov (base `PROS`).

A Figura 2.2 apresenta os EQMcv's para os respectivos parâmetros de regularização em cada método Tikhonov, assim como o EQMcv obtido a partir da utilização do método TSVD com $t = 8$.

Como para $\alpha = 0$ a regularização de Tikhonov, independente de sua ordem, é equivalente ao procedimento de mínimos quadrados, tem-se o mesmo EQMcv para as três curvas no ponto inicial. Posteriormente, as curvas decaem até um ponto ótimo de regularização (ponto destacado sobre a curva do gráfico) e, logo em seguida, voltam a subir. Se fossem realizados cálculos para valores de α maiores do que 7, seria registrada uma ascensão das três curvas. Esse padrão de curva é uma característica dos métodos de regularização e será melhor abordado mais adiante.

Um aspecto interessante a ser observado é que, após o ponto ótimo, conforme a ordem da regularização de Tikhonov aumenta, a curva tende a subir de forma menos intensa.

A Tabela 2.3 apresenta os parâmetros ótimos em cada um dos três métodos juntamente com os respectivos EQMcv's.

	Tikhonov de ordem 0	Tikhonov de ordem 1	Tikhonov de ordem 2
Parâmetro α	3.150	2.725	1.550
EQMcv	0.578	0.581	0.590

Tabela 2.3: EQMcv's segundo o parâmetro ótimo de cada método (base `PROS`).

Diante da análise realizada, percebe-se que a regularização de Tikhonov de ordem zero foi a que mostrou-se com maior poder preditivo. Ademais, independente da ordem, as regularizações de Tikhonov performaram melhor do que o método TSVD.

2.7.2 Dados de Quimiometria

A base de dados de origem química (`QUIM`), contém 1267 observações e 228 preditores. Dentre os preditores, 208 são atributos binários que indicam a presença ou não de subestruturas químicas específicas, 16 são atributos de contagem e os 4 restantes são atributos contínuos. Tais informações serão utilizadas para prever a solubilidade de uma determinada substância química.

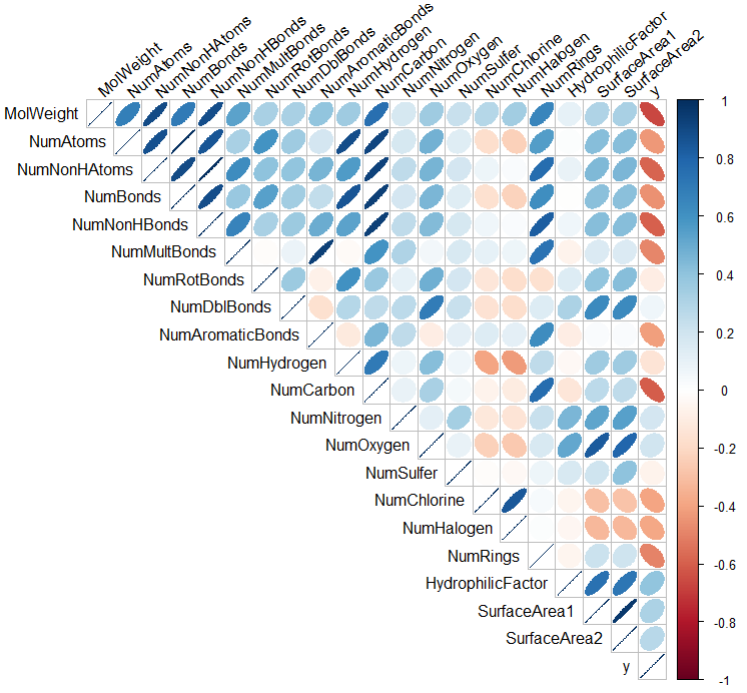


Figura 2.3: Matriz de correlações das variáveis de contagem e contínuas do sistema linear, incluindo a variável resposta y (base `QUIM`).

A matriz presente na Figura 2.3 apresenta a estrutura de correlação entre os preditores de solubilidade que não são binários, assim como suas correlações com a variável resposta.

Existem algumas correlações muito fortes e positivas entre alguns preditores que podem inviabilizar a utilização de métodos convencionais como a regressão linear.

Assim como na aplicação em dados de próstata, antes de se submeter os dados aos modelos de regularização, foi realizada a padronização das variáveis preditoras a fim de se estimar o intercepto com base na média da variável resposta, facilitando, assim, a implementação computacional dos modelos. As padronizações foram realizadas nos conjuntos de treinamento e teste em cada uma das cinco etapas da validação cruzada, conforme justificado anteriormente.

A Figura 2.4 apresenta os EQMcv's segundo cada parâmetro de regularização do método TSVD.

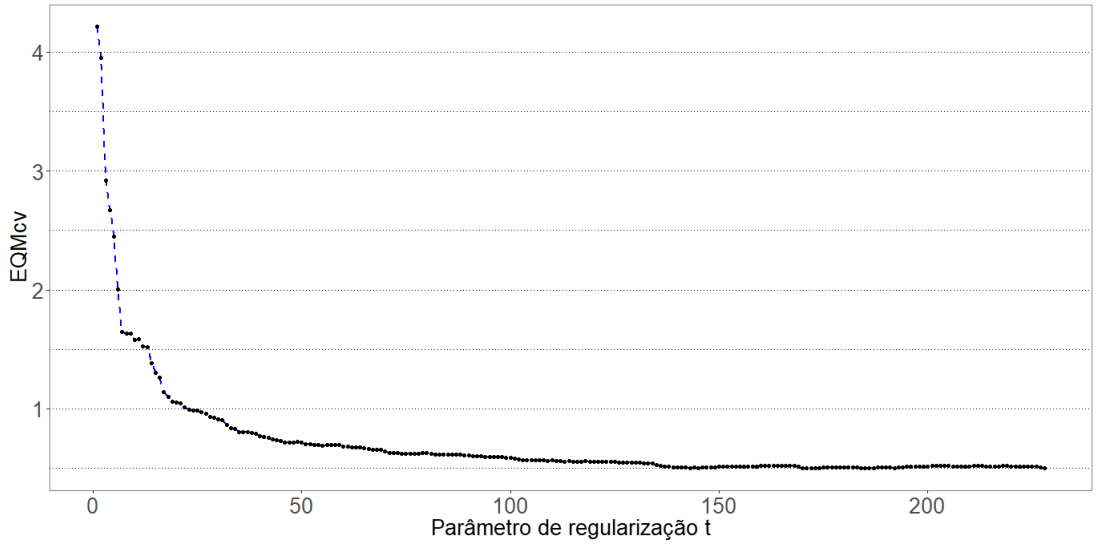


Figura 2.4: Sensibilidade da métrica de avaliação de desempenho segundo os valores de t (base `QUIM`).

O EQMcv da solução pseudoinversa ($t = 228$) foi de 0.499. Embora o menor EQMcv existente no gráfico seja de 0.497, obtido com $t = 173$, a Figura 2.4 apresenta uma tendência clara, revelando a existência de diferenças irrisórias na métrica

para grandes variações de t quando este é maior do que 135, ao passo que tem-se a existência de diferenças significativas para pequenas variações do parâmetro t quando este é pequeno.

Para os métodos de Tikhonov de ordem zero, um e dois, definiu-se um *grid* de busca abrangendo uma sequência de 0 a 10, variando 0.025 de elemento a elemento. A Figura 2.5 apresenta os EQMcv's resultantes da validação cruzada para cada parâmetro de regularização em cada método Tikhonov, além de apresentar o melhor EQMcv obtido pelo método TSVD.

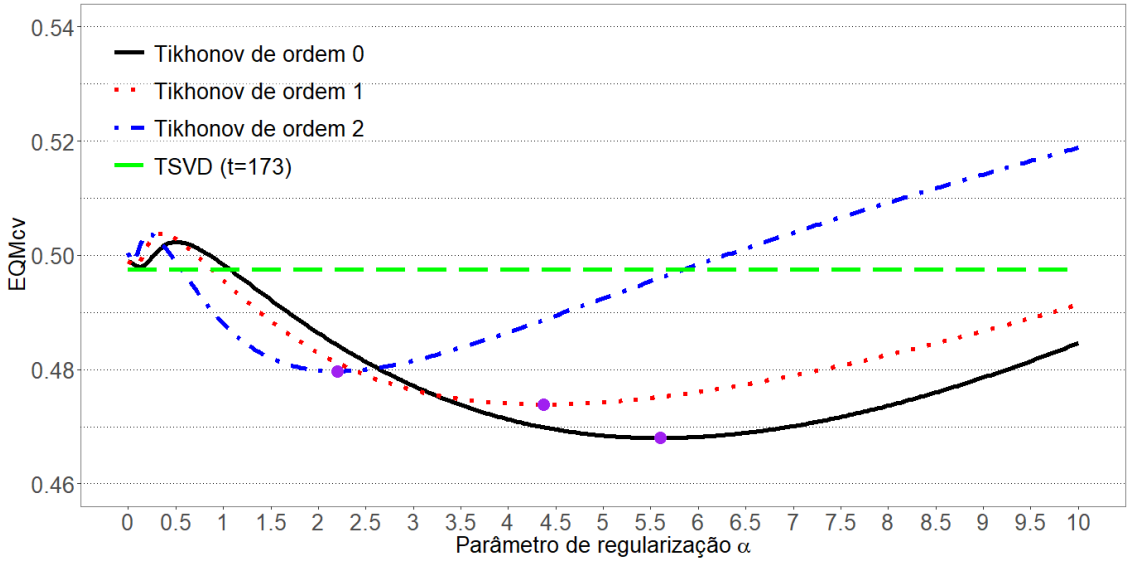


Figura 2.5: Sensibilidade da métrica de avaliação de desempenho segundo os valores de α para cada regularização de Tikhonov (base `QUIM`).

Novamente, todos os métodos de Tikhonov alcançaram resultados melhores do que o melhor resultado obtido pelo TSVD, sendo a regularização de Tikhonov de ordem zero a que retornou o menor EQMcv.

A Tabela 2.4 apresenta os parâmetros ótimos em cada um dos três métodos de Tikhonov juntamente com os respectivos EQMcv's.

	Tikhonov de ordem 0	Tikhonov de ordem 1	Tikhonov de ordem 2
Parâmetro α	5.600	4.375	2.200
EQMcv	0.468	0.473	0.479

Tabela 2.4: EQMcv's segundo o parâmetro ótimo de cada método (base `QUIM`).

Um padrão observado nas Figuras 2.2 e 2.5 é a compensação do efeito de regularização. Como a regularização é naturalmente mais intensa nos métodos de ordem superior, a região de menor EQMcv de cada curva tende a se localizar em valores menores de α conforme a ordem aumenta.

O pico presente no início das três curvas não era esperado. Para verificar a origem do problema, foi realizada uma comparação da curva gerada pela implementação computacional do método Tikhonov de ordem zero, com as curvas geradas pelas regressões Ridge disponíveis nos pacotes *MASS* e *glmnet*, ambos provenientes do *software* R. Vale ressaltar que para inserir o parâmetro de regularização na função do *glmnet* e obter equivalência nos resultados é necessário realizar uma pequena transformação:

$$\lambda_{glmnet} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}\frac{\alpha^2}{n} = s_y\frac{\alpha^2}{n}, \tag{2.41}$$

em que λ_{glmnet} é o parâmetro a ser inserido na função do *glmnet*, s_y é o desvio padrão da variável resposta, n é o número de observações e α é o parâmetro de regularização já apresentado na seção 2.6. Na subseção 3.4 será explicado como a função *glmnet* do respectivo pacote opera e, então, será possível entender o porquê dessa transformação.

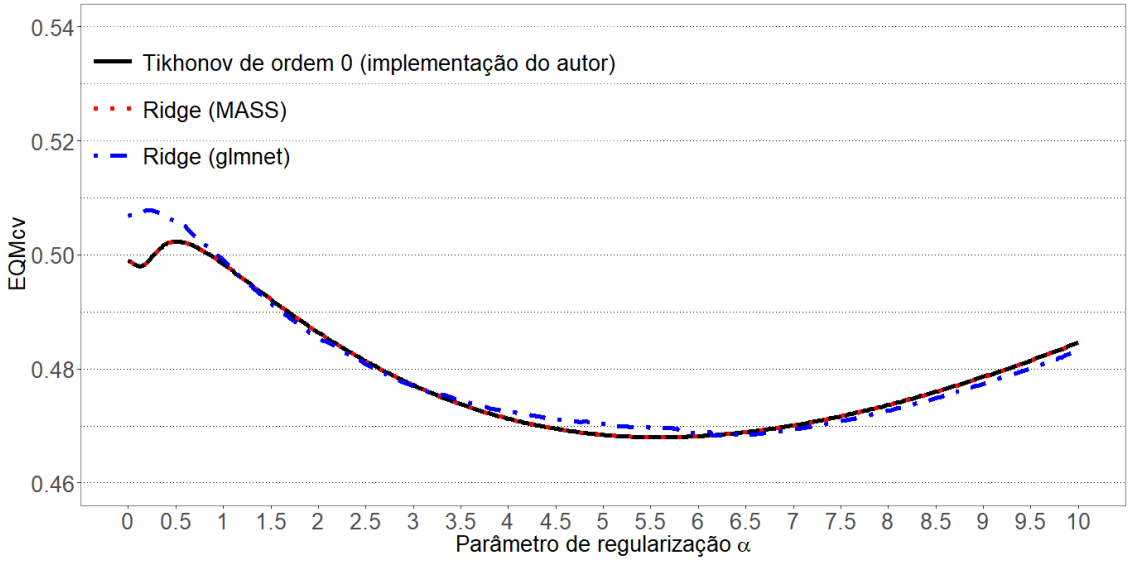


Figura 2.6: Curvas de validação cruzada segundo a implementação computacional (base `QUIM`).

No caso dos dados de próstata, as três curvas coincidem. Todavia, para os dados de quimiometria, a comparação com o pacote *glmnet* evidencia algumas diferenças, principalmente quando α está entre 0 e 1. Para este intervalo, a curva do *glmnet* já começa um pouco mais acima do que as outras e, embora também tenha uma pico inesperado nos valores iniciais, este é menos evidente.

Até onde pôde-se constatar, essas diferenças são consequentes de inconsistências numéricas.

Ademais, para este conjunto de dados, o Ridge do pacote *MASS* não gerou um resultado coerente quando $\alpha = 0$, retornando um EQMcv demasiadamente alto, que não foi apresentado na Figura 2.6.

Capítulo 3

A Regularização Via Otimização

Este Capítulo volta a atenção à descrição geral dos principais métodos de regularização disponíveis na literatura estatística, embora tenha como principal objetivo descrever a implementação do LASSO Bayesiano (Park and Casella, 2008) através do Amostrador de Gibbs. Resultados relacionados às distribuições condicionais a serem utilizadas na implementação que foram descritos em Park and Casella (2008) serão aqui explorados detalhadamente para uma maior compreensão sobre como o LASSO Bayesiano pode ser computacionalmente implementado.

3.1 Modelo de Regressão Linear

O objetivo de uma análise de regressão convencional é explicar o comportamento ou prever os valores de um determinado vetor de variáveis aleatórias $\boldsymbol{Y}$ através de uma relação funcional $\boldsymbol{y} = \mathrm{f}(\boldsymbol{X})$. Uma das relações funcionais mais utilizadas é o modelo de regressão linear

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} \ + \ \boldsymbol{\epsilon} \, , \tag{3.1}$$

no qual $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ é o vetor de parâmetros da regressão, $\boldsymbol{\epsilon}$ é um vetor de erros do modelo, $\boldsymbol{y}$ é o valor em um conjunto numérico que representa a variável resposta e $\boldsymbol{X}_{n\times(p+1)}$ é uma matriz de p preditores lineares com n observações, sendo

p menor do que n, e a primeira coluna atribuída ao intercepto.

Tal modelo, apesar de sua grande usabilidade, possui algumas limitações que o faz menos desejável ou até mesmo inviável. Em estudos de dados genéticos o número de amostras envolvidas frequentemente é menor do que o número de características, i.e $p > n$. Como resultado, a matriz $\boldsymbol{X}$ — dita uma matriz de alta dimensão — possui covariáveis que sabidamente possuem algum nível de colinearidade entre si.

O fenômeno da colinearidade é indesejável e implica que o espaço coluna de $\boldsymbol{X}$, espaço sobre o qual obtém-se as projeções ortogonais $\hat{\boldsymbol{y}}$ de $\boldsymbol{y}$, possua uma dimensão menor que p, dificultando a detecção da real contribuição das covariáveis no processo de explanação da variabilidade proveniente da variável resposta $\boldsymbol{Y}$.

Como consequência, tais problemas frequentemente refletem ajustes com grandes incertezas nas estimativas dos componentes de $\boldsymbol{\beta}$ que se relacionam com as variáveis colineares.

Além da inconveniência da multicolinearidade, a saturação do modelo é uma característica dos dados de alta dimensão e dificulta a implementação de uma relação funcional $\boldsymbol{y} = \mathrm{f}(\boldsymbol{X})$. Métodos alternativos como as regressões regularizadas foram desenvolvidos para contornar algumas complicações como as detalhadas anteriormente.

3.2 Métodos de Regularização

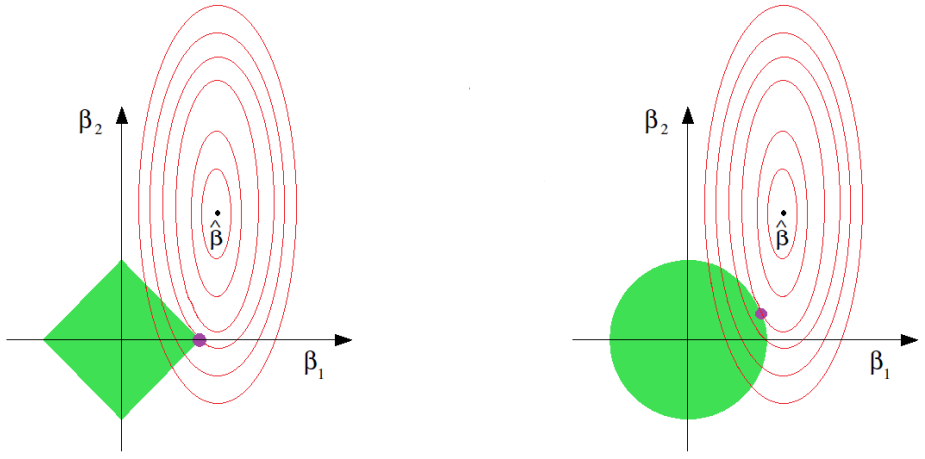
Como mencionado previamente, em modelos de regressão linear cujas variáveis são colineares, os coeficientes tornam-se pouco confiáveis e, uma vez que a matriz de dados $\boldsymbol{X}$ não possui posto completo, a matriz $(\boldsymbol{X}'\boldsymbol{X})$ passa a ser singular, inviabilizando a sua inversão. Por mais que, em muitos casos, os programas implementados em *software* estatísticos consigam calcular a sua inversa, objetivando usá-la para estimar os coeficientes e suas variâncias, o processo de inversão dos dados é instável e induz estimativas com altas variâncias. A imposição de uma restrição no tamanho dos coeficientes alivia o efeito de problemas provenientes do conflito entre coeficientes de variáveis correlatas (Hastie et al., 2013).

O estimador em um método de regularização pode ser genericamente descrito como sendo:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{q}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \tag{3.2}$$

$$\text{sujeito a} \sum_{j=1}^{p} |\beta_j|^q \leq t. \tag{3.3}$$

A restrição imposta à Equação (3.2) deixa clara a habilidade de tais métodos de controlar o quão grande pode ser o valor assumido por β_j. Dois fatores são determinantes nesse controle, ou regularização: q e t. Enquanto q possui uma interpretação um pouco mais rebuscada, a interpretação da influência do elemento t na regularização é direta: quanto menor o valor de t, mais intensa será a regularização.



(a) `LASSO` ($q = 1$) (b) `RIDGE` ($q = 2$)

Figura 3.1: Ilustração da estimação dos parâmetros na regularização e a sensibilidade dos resultados com relação ao valor de q.

As Figuras 3.1(a) e 3.1(b) exemplificam o processo de estimação para um modelo com duas variáveis explicativas. O vetor $\hat{\boldsymbol{\beta}}$ é o estimador de mínimos quadrados da regressão sem regularização. As linhas em vermelho representam as curvas de nível da função de erro por mínimos quadrados, enquanto que as figuras geométricas centradas na origem do gráfico representam as regiões definidas pelas restrições

$|\beta_1| + |\beta_2| \leq t$ e $\beta_1^2 + \beta_2^2 \leq t$ respectivamente. O ponto que representa a interseção entre a região de restrição e a menor curva de nível existente, dentre as curvas que intersectam a região, é a estimativa proveniente do método de regularização.

Para a implementação desses modelos de regularização, é interessante, sempre que possível, realizar uma padronização prévia dos dados a fim de se evitar que a escala das covariáveis seja determinante no resultado final.

Especificamente, quando $q = 1$ e $q = 2$ o método de regularização recebe o nome de regressão LASSO (*Least Absolute Shrinkage and Selection Operator*) e regressão Ridge respectivamente (Hastie et al., 2013). Ambos ganharam destaque por possuírem propriedades relevantes.

A componente q da Equação (3.3) pode assumir qualquer valor estritamente maior que 0 na regularização. No entanto, é aconselhável que se escolha um valor de q tal que $q \geq 1$, caso contrário, a região de restrição torna-se não-convexa, dificultando o problema de otimização (Hastie et al., 2013). Quando $q = 1$ a região de restrição ainda é convexa, mas $|\beta_j|^q$ não é diferenciável em 0, implicando em uma das principais propriedades da regressão LASSO que é forçar alguns $\beta_j's$ a serem estimados exatamente em 0. Já quando $q = 2$, o método perde esse propriedade, mas continua controlando a participação das covariáveis no modelo, mesmo que de forma mais suavizada. Hastie et al. (2013) apresenta uma relação entre a regressão Ridge com componentes principais através da decomposição em valores singulares da matriz de dados $\boldsymbol{X}$, mostrando que o Ridge projeta os valores de Y nesses componentes controlando com mais fervor os coeficientes dos componentes de menor variância e do que os coeficientes dos componentes de maior variância.

A Equação (3.2) sujeita à restrição descrita em (3.3) pode ser reformulada e reescrita em sua forma Lagrangeana

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{q}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}, \tag{3.4}$$

na qual $\lambda \geq 0$ é o parâmetro de regularização, controlando o quão grande pode ser cada β_j. Dessa forma, λ possui uma interpretação oposta à interpretação dada a t,

ou seja, quanto maior o valor de λ, os $\beta_j's$ serão estimados mais próximos de zero. Já quando λ se aproxima de 0, as estimativas da regressão regularizada se aproximam das estimativas da regressão por mínimos quadrados convencional.

Com a regularização explicitada na sua forma Lagrangeana, a regressão Ridge possui solução explicita para o cálculo das estimativas:

$$\hat{\boldsymbol{\beta}}_{\mathbf{2}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y} \ , \tag{3.5}$$

em que $\boldsymbol{I}$ é a matriz identidade $p \times p$. Perceba que na regressão Ridge as estimativas se assemelham às estimativas de mínimos quadrados, possuindo uma relação linear com $\boldsymbol{y}$ e tendo como diferença a soma do parâmetro de regularização na diagonal principal da matriz $(\boldsymbol{X}'\boldsymbol{X})$, que faz da estimação um problema não-singular mesmo se $\boldsymbol{X}$ não for de posto completo.

O estimador Ridge também pode ser reescrito como uma função do estimador de mínimos quadrados ($\boldsymbol{\beta}$). Seja

$$\boldsymbol{W} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}, \tag{3.6}$$

o estimador Ridge pode ser expresso por:

$$\boldsymbol{W}\hat{\boldsymbol{\beta}} = \boldsymbol{W}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

$$= (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

$$= (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_2.$$

Portanto,

$$\boldsymbol{W}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_2. \tag{3.7}$$

Em outras palavras, o operador linear $\boldsymbol{W}$ transforma o estimador de mínimos quadrados de $\boldsymbol{\beta}$ no estimador Ridge. Como resultado,

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathbf{2}}) = \mathrm{Var}(\boldsymbol{W}\hat{\boldsymbol{\beta}}) = \boldsymbol{W}\mathrm{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{W}' = \sigma^2\boldsymbol{W}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{W}', \tag{3.8}$$

em que σ^2 é a variância do erro.

Por ter uma variância definida explicitamente, torna-se possível compará-la com a variância do estimador de mínimos quadrados. Ao se realizar a comparação, tem-se:

$$\text{Var}(\hat{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}}_{\boldsymbol{2}}) = \sigma^2 \left[\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}\right]^{-1} \left[2\lambda \boldsymbol{I} + \lambda^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}\right] \left[(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\right]'. \tag{3.9}$$

A demostração da igualdade (3.9) encontra-se no Apêndice C. Para que a matriz $\boldsymbol{X}'\boldsymbol{X}$ seja inversível, é preciso que ela seja positiva definida, ou seja, $\boldsymbol{X}$ precisa ser de posto completo. Uma vez que $\boldsymbol{I}$ também é positiva definida, tem-se que $[(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I})^{-1}]'$ é positiva definida, assim como $(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})$ também é positiva definida, desde que $\lambda \geq 0$. Especificamente em $[2\lambda \boldsymbol{I} + \lambda^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}]$, a matriz é positiva definida para todo $\lambda > 0$, mas é semi-definida positiva quando $\lambda = 0$. Assim, $\text{Var}(\hat{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}}_{\boldsymbol{2}}) \succeq 0$ para todo $\lambda \geq 0$.

A Figura 3.2, elaborada de forma a reproduzir os padrões das curvas em Hoerl and Kennard (1970, Fig. 1), é uma representação do comportamento que os erros quadráticos médios (EQM's) podem assumir em um determinado conjunto de dados e mostra a relação entre a variância, o viés ao quadrado e o parâmetro λ. Percebe-se que existe um *trade-off* entre viés e variância: a variância total decresce conforme λ aumenta, enquanto o viés ao quadrado aumenta. A linha curvada sólida representa o erro quadrático médio da regressão Ridge e evidencia a superioridade do Ridge, para determinados valores de λ, em relação ao método de mínimos quadrados.

Theobald (1974) mostra que existe um valor λ_0 tal que, para todo $0 < \lambda < \lambda_0$, o estimador Ridge retorna um EQM menor que o EQM obtido via MQO. Posteriormente, Farebrother (1976) generaliza o resultado para quando $\boldsymbol{X}$ não possui posto completo.

Hoerl and Kennard (1970) também mostram que, através da análise das derivadas das curvas do viés ao quadrado e da variância, chega-se à conclusão de que é possível variar o valor de $\lambda \geq 0$ de tal forma que as estimativas passem a ter um

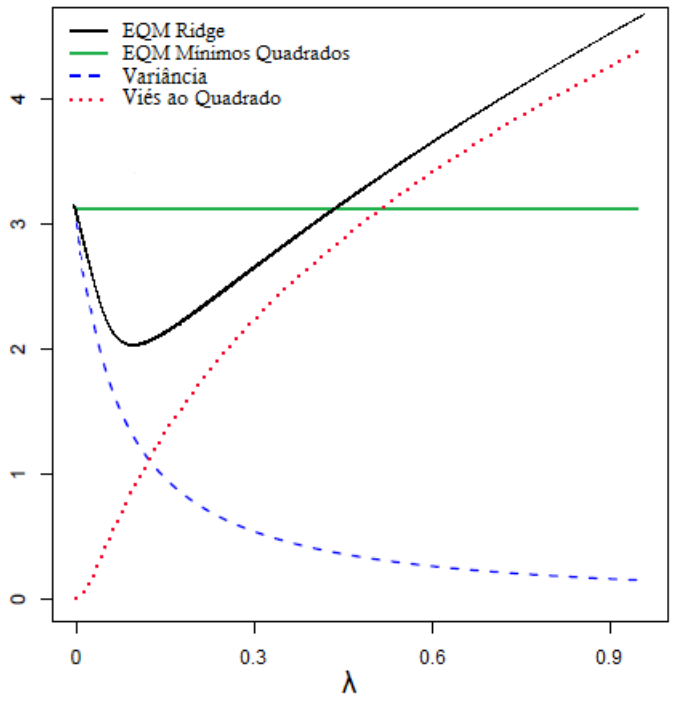


Figura 3.2: Ilustração do Erro Quadrático Médio e do *trade-off* entre o quadrado do viés e a variância.

pequeno viés e uma consequente e substancial redução na variância, reduzindo o erro quadrático médio e melhorando a qualidade das predições.

Na prática, o valor ótimo de λ nos métodos de regularização frequentistas é usualmente determinado empiricamente por meio de métricas de desempenho de validação cruzada, sendo escolhido o valor de λ que na média apresentar o melhor desempenho nos conjuntos de teste. Uma vez que os dados são subdivididos em teste e treinamento, a avaliação da qualidade da predição depende do corte que foi realizado nos dados. Ao se dividir os dados em várias partes de tal forma que se tenha vários conjuntos de treinamento e teste, a validação cruzada diminui esse efeito de corte.

3.3 LASSO Bayesiano

Métodos bayesianos também podem ser utilizados para seleção de variáveis e regularização das estimativas. Muitas estimativas penalizadas provenientes de métodos frequentistas são equivalentes a algumas estatísticas resumo de distribuições a posteriori de modelos bayesianos específicos. Como exemplo, tem-se a correspondência entre a estimativa LASSO com a moda a posteriori de um modelo bayesiano com prioris iid's (Independentes e Identicamente Distribuídas) Laplace nos coeficientes e a correspondência entre a estimativa Ridge com a moda a posteriori de um modelo com prioris iid's Normal (Hastie et al., 2013).

Ao se implementar métodos bayesianos, é mais comum utilizar-se a média a posteriori no lugar da moda. No entanto, ao se utilizar a média como estimativa bayesiana, a correspondência com o LASSO deixa de existir, enquanto que a regressão Ridge também possui correspondência com a média a posteriori (Hastie et al., 2013). Além disso, quando utilizadas prioris iid's Laplace nos coeficientes, diferentemente da moda a posteriori, a média a posteriori não possui a propriedade de estimar alguns coeficientes exatamente no 0.

Avaliar o quão concentrada a priori é na vizinhança de 0 e se ela tem caudas pesadas ou não são fatores relevantes que auxiliam na compreensão da influência da priori no modelo. Dessa forma, a densidade a priori dos $\beta_j's$ será responsável por definir se o modelo irá induzir seleção de variáveis e controlar o tamanho dos coeficientes ou apenas induzir este último.

O LASSO Bayesiano de Park and Casella (2008) é descrito a seguir. Como descrito anteriormente, ao se definir uma priori Laplace

$$\pi(\boldsymbol{\beta}) = \prod_{j=1}^{p} \frac{\lambda}{2} \exp\left\{-\lambda|\beta_j|\right\}, \text{ com } \beta_j \in \mathbb{R}\, \forall j, \tag{3.10}$$

e uma priori independente $\pi(\sigma^2)$ com $\sigma^2 > 0$, a distribuição a posteriori $\pi(\boldsymbol{\beta}, \sigma^2|\tilde{\boldsymbol{y}})$ é proporcional a

$$\pi(\boldsymbol{\beta}, \sigma^2|\tilde{\boldsymbol{y}}) \propto \pi(\sigma^2)(\sigma^2)^{-\frac{(n-1)}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) - \lambda\sum_{j=1}^{p}|\beta_j|\right\}, \tag{3.11}$$

em que $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \bar{\boldsymbol{y}}$. Como consequência, as estimativas irão depender de λ e da escolha da priori de σ^2.

Apesar de interessante, o modelo implementado com a priori (3.10) e uma priori independente $\pi(\sigma^2)$ permite a geração de distribuições a posteriori proporcionais à quantidade (3.11) que possuem mais de uma moda. Ao se fixar σ^2 e condicionar $\pi(\boldsymbol{\beta})$ no valor fixo σ^2, Park and Casella (2008) mostram que a distribuição a posteriori resultante é sabidamente unimodal. Assim, Park and Casella (2008) utilizam uma variação da priori definida em (3.10):

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p}\frac{\lambda}{2\sqrt{\sigma^2}}\exp\left\{-\lambda\frac{|\beta_j|}{\sqrt{\sigma^2}}\right\}, \text{ com } \beta_j \in \mathbb{R}\,\forall j. \tag{3.12}$$

Considerando que σ^2 possua uma distribuição a priori gama inversa com parâmetros α e γ tal que,

$$\pi(\sigma^2) = \frac{\gamma^a}{\Gamma(a)}(\sigma^2)^{-a-1}\exp\left\{-\frac{\gamma}{\sigma^2}\right\} \quad \text{com} \quad \sigma^2 > 0 \quad \text{e} \quad (a > 0, \gamma > 0), \tag{3.13}$$

e utilizando a priori (3.12), a distribuição a posteriori de $\boldsymbol{\beta}$ e σ^2 é proporcional a

$$\pi(\boldsymbol{\beta}, \sigma^2|\tilde{\boldsymbol{y}}) \propto (\sigma^2)^{-\frac{(n+p-1)}{2}-a-1}\exp\left\{-\frac{1}{\sigma^2}\left(\frac{(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})}{2} + \gamma\right) - \lambda\sum_{j=1}^{p}\frac{|\beta_j|}{\sqrt{\sigma^2}}\right\}. \tag{3.14}$$

Assim, ao se utilizar uma priori para σ^2 degenerada com $a = 0$ e $\gamma = 0$, tem-se uma equivalência com a utilização de uma priori não informativa *scale-invariant* $\frac{1}{\sigma^2}$. Dessa forma, tem-se a garantia de uma posteriori unimodal (Park and Casella, 2008).

Essa formulação do LASSO Bayesiano traz um benefício a mais, pois uma vez que os coeficientes $\beta_j's$ em (3.14), na parte responsável pela regularização, são padronizados, o parâmetro de regularização λ não é mais sensível a mudanças de escala em $\boldsymbol{y}$, ou seja, uma mudança na unidade de medida de $\boldsymbol{y}$ não exige uma mudança no valor de λ para produzir o mesmo modelo de LASSO Bayesiano.

3.3.1 O Modelo Hierárquico

Resumir a informação contida na distribuição a posteriori pode não ser uma tarefa fácil. Em teoria, uma variável aleatória X com valores no conjunto χ possui densidade $\mathrm{f}(x) = K\,\mathrm{g}(x)$, cujo K é um valor único que representa uma constante de normalização e $\mathrm{g}(x)$ é dito o núcleo de $\mathrm{f}(x)$. A expressão (3.14) nada mais é do que o núcleo da distribuição a posteriori conjunta de $\boldsymbol{\beta}$ e σ^2 e para se amostrar os valores dessa distribuição por métodos convencionais é preciso, primeiro, descobrir qual é a constante de normalização K. Na prática, calcular o valor de K pode ser uma tarefa árdua e gerar valores aleatórios de distribuições multidimensionais sem supor independência pode ser mais complicado ainda.

Nesse contexto, vários métodos ditos algoritmos MCMC (*Monte Carlo via Cadeia de Markov*) foram desenvolvidos para contornar e fazer da geração de amostras dessas distribuições complexas uma operação viável. O Amostrador de Gibbs é um dos exemplos mais simples de método MCMC e para utilizá-lo basta se obter os núcleos das distribuições condicionais.

Especialmente para o problema de regularização sobre o qual a posteriori proporcional à quantidade (3.14) foi construída, existe um método razoavelmente simples com uma estratégia computacional para se obter a média ou qualquer outra estatística resumo da distribuição a posteriori. A estratégia consiste em formular um modelo hierárquico cuja finalidade é fornecer as distribuições condicionais necessárias para a implementação do amostrador de Gibbs, que irá gerar as amostras da posteriori para a estimação das estatísticas resumo.

Por conveniência computacional, a distribuição a priori Laplace $\pi(\boldsymbol{\beta}|\boldsymbol{\sigma^2})$ passa

a ser representada como uma mistura infinita de densidades Normal (Andrews and Mallows, 1974):

$$\frac{a}{2}\exp\{-a|z|\} = \int_0^\infty \frac{1}{\sqrt{2\pi s}}\exp\left\{-\frac{z^2}{2s}\right\}\frac{a^2}{2}\exp\left\{-a^2\frac{s}{2}\right\}\mathrm{ds}\,,\ \text{com}\ \ a = \frac{\lambda}{\sqrt{\sigma^2}} > 0. \tag{3.15}$$

A formulação hierárquica proposta por Park and Casella (2008) é, então, dada por:

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{X},\boldsymbol{\beta},\sigma^2 &\sim \mathrm{N}_n\left(\boldsymbol{\mu}\mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n\right), \qquad (3.16)\\
\boldsymbol{\beta}|\tau_1^2,\ldots,\tau_p^2,\sigma^2 &\sim \mathrm{N}_p\left(\mathbf{0}_p, \sigma^2\boldsymbol{D}_\tau\right), \quad \boldsymbol{D}_\tau = \mathrm{diag}(\tau_1^2,\ldots,\tau_p^2),\\
\tau_1^2,\ldots,\tau_p^2 &\sim \prod_{j=1}^{p}\frac{\lambda^2}{2}\exp\left\{-\lambda^2\frac{\tau_j^2}{2}\right\}\mathrm{d}\tau_j^2, \quad \tau_1^2,\ldots,\tau_p^2 > 0,\\
\sigma^2 &\sim \pi(\sigma^2)\mathrm{d}\sigma^2.
\end{aligned}$$

Assim, $\tau_1^2,\ldots,\tau_p^2$, que devem ser independentes de σ^2, são parâmetros auxiliares utilizados para que se possa reescrever a priori (3.12) como uma mistura de distribuições normal. Para facilitar a compreensão da utilização da expressão (3.15) basta pensar que $\pi(\beta_j|\sigma^2)$ é basicamente a distribuição marginal encontrada após integrar a distribuição conjunta de β_j, τ_j^2 e σ^2 sobre os valores de τ_j^2, ou seja,

$$\pi(\beta_j|\sigma^2) = \int_T \pi(\beta_j,\tau_j^2|\sigma^2)\,\mathrm{d}\tau_j^2 = \int_T \frac{\pi(\beta_j|\tau_j^2,\sigma^2)\pi(\tau_j^2,\sigma^2)}{\pi(\sigma^2)}\,\mathrm{d}\tau_j^2 = \int_T \pi(\beta_j|\tau_j^2,\sigma^2)\pi(\tau_j^2)\,\mathrm{d}\tau_j^2. \tag{3.17}$$

A formulação hierárquica, assim construída, torna possível os cálculos das distribuições condicionais dos parâmetros do modelo.

3.3.2 A Implementação Computacional

Através da formulação hierárquica (3.16), considerando uma priori independente não informativa para μ e supondo que σ^2 tenha uma distribuição gama inversa, a densidade conjunta pode ser expressa por

$$\mathrm{f}(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\beta},\sigma^2)\pi(\sigma^2)\pi(\boldsymbol{\mu})\prod_{j=1}^{p}\pi(\beta_j|\tau_j^2,\sigma^2)\pi(\tau_j^2) =$$

$$\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left\{\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{\mu}\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{y}-\boldsymbol{\mu}\mathbf{1}_n-\boldsymbol{X}\boldsymbol{\beta}\right)\right\}\frac{\gamma^a}{\Gamma(a)}(\sigma^2)^{-a-1}\exp\left\{-\frac{\gamma}{\sigma^2}\right\}\times$$

$$\prod_{j=1}^{p}\frac{1}{(2\sigma^2\tau_j^2)^{\frac{1}{2}}}\exp\left\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2\right\}\frac{\lambda^2}{2}\exp\left\{-\lambda^2\frac{\tau_j^2}{2}\right\}. \tag{3.18}$$

Para facilitar as contas, o parâmetro μ é retirado da expressão, sem perda de generalidade, fazendo-se $\tilde{\boldsymbol{y}}=\boldsymbol{y}-\bar{\boldsymbol{y}}$. Então, a densidade conjunta (3.18) passa a ser uma conjunta que é marginal apenas com relação a μ e que é proporcional a

$$\propto\ \frac{1}{(\sigma^2)^{\frac{(n-1)}{2}}}\exp\left\{-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})\right\}(\sigma^2)^{-a-1}\exp\left\{-\frac{\gamma}{\sigma^2}\right\}\times$$

$$\prod_{j=1}^{p}\frac{1}{(\sigma^2\tau_j^2)^{\frac{1}{2}}}\exp\left\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2\right\}\exp\left\{-\lambda^2\frac{\tau_j^2}{2}\right\}. \tag{3.19}$$

Eliminando os termos que não envolvem $\boldsymbol{\beta}$ de (3.19) e considerando $\boldsymbol{D}_\tau$ como sendo uma matriz diagonal cujos elementos da diagonal principal são τ_j^2, tem-se

$$-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})-\frac{1}{2\sigma^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\tau_j^2}=-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})-\frac{1}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{D}_\tau^{-1}\boldsymbol{\beta}$$

$$=-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}-2\tilde{\boldsymbol{y}}'\boldsymbol{X}\boldsymbol{\beta}+\tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}}+\boldsymbol{\beta}'\boldsymbol{D}_\tau^{-1}\boldsymbol{\beta}\right)$$

$$=-\frac{1}{2\sigma^2}\left[\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{D}_\tau^{-1})\boldsymbol{\beta}-2\tilde{\boldsymbol{y}}'\boldsymbol{X}\boldsymbol{\beta}+\tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}}\right]$$

$$=-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}'\boldsymbol{A}\boldsymbol{\beta}-2\tilde{\boldsymbol{y}}'\boldsymbol{X}\boldsymbol{\beta}+\tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}}\right) \tag{3.20}$$

em que $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_\tau^{-1})$.

Levando em consideração o fato de A ser uma matriz ortogonal, tem-se que a quantidade

$$(\boldsymbol{\beta} - \boldsymbol{A}^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}})'\boldsymbol{A}(\boldsymbol{\beta} - \boldsymbol{A}^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}}) = (\boldsymbol{\beta} - \boldsymbol{A}^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}})'(\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{X}'\tilde{\boldsymbol{y}})$$

$$= \left[\boldsymbol{\beta}' - \tilde{\boldsymbol{y}}'\boldsymbol{X}(\boldsymbol{A}^{-1})'\right](\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{X}'\tilde{\boldsymbol{y}}) = \boldsymbol{\beta}'\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{y}}'\boldsymbol{X}(\boldsymbol{A}^{-1})'\boldsymbol{A}\boldsymbol{\beta} + \tilde{\boldsymbol{y}}'\boldsymbol{X}(\boldsymbol{A}^{-1})'\boldsymbol{X}'\tilde{\boldsymbol{y}}$$

$$= \boldsymbol{\beta}'\boldsymbol{A}\boldsymbol{\beta} - (\tilde{\boldsymbol{y}}'\boldsymbol{X}\boldsymbol{\beta})' - \tilde{\boldsymbol{y}}'\boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{y}}'\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}} \tag{3.21}$$

e, como $\tilde{\boldsymbol{y}}'\boldsymbol{X}\boldsymbol{\beta}$ possui dimensão 1×1, $(\tilde{\boldsymbol{y}}'\boldsymbol{X}\boldsymbol{\beta})' = \tilde{\boldsymbol{y}}'\boldsymbol{X}\boldsymbol{\beta}$, consequentemente

$$\boldsymbol{\beta}'\boldsymbol{A}\boldsymbol{\beta} - 2\tilde{\boldsymbol{y}}'\boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{y}}'\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}}. \tag{3.22}$$

Com a utilização do quadrado (3.22), a expressão (3.20) pode ser reescrita como sendo

$$-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}'\boldsymbol{A}\boldsymbol{\beta} - 2\tilde{\boldsymbol{y}}'\boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}}\right)$$

$$= -\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}'\boldsymbol{A}\boldsymbol{\beta} - 2\tilde{\boldsymbol{y}}'\boldsymbol{X}\boldsymbol{\beta} + \tilde{\boldsymbol{y}}'\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{y}}'\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}} + \tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}}\right)$$

$$= -\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta} - \boldsymbol{A}^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}})\boldsymbol{A}(\boldsymbol{\beta} - \boldsymbol{A}^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}}) + \tilde{\boldsymbol{y}}'(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}')\tilde{\boldsymbol{y}}\right]. \tag{3.23}$$

Como $\tilde{\boldsymbol{y}}'(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}')\tilde{\boldsymbol{y}}$ não depende de $\boldsymbol{\beta}$, tem-se que a parte da distribuição conjunta (3.19) que depende de $\boldsymbol{\beta}$ é descrita pela quantidade

$$-\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta} - \boldsymbol{A}^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}})\boldsymbol{A}(\boldsymbol{\beta} - \boldsymbol{A}^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}})\right], \tag{3.24}$$

que é exatamente o núcleo da distribuição Normal Multivariada. Assim, sabe-se que a distribuição de $\boldsymbol{\beta}$ condicionado aos outros termos é uma Normal Multivariada com média $\boldsymbol{A}^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}}$ e variância $\sigma^2\boldsymbol{A}^{-1}$.

A construção da distribuição condicional de σ^2 já é mais direta. Levando em

consideração que

$$\sigma^{2\left(-\frac{n-1}{2}-\frac{p}{2}-\alpha-1\right)}\exp\left\{-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})+\frac{1}{2\sigma^2}\boldsymbol{\beta}'\boldsymbol{D}_\tau^{-1}\boldsymbol{\beta}+\frac{\gamma}{\sigma^2}\right\} \tag{3.25}$$

contém todos os termos de (3.19) que englobam σ^2 e realizando-se uma comparação com uma densidade gama inversa com parametrização tal que

$$\mathrm{f}(x)=\frac{s^r}{\Gamma(r)}x^{-r-1}\exp\left\{-\frac{s}{x}\right\}, \tag{3.26}$$

percebe-se que a distribuição de σ^2 condicionada a todo o resto é uma gama inversa com $r=\frac{n-1}{2}+\frac{p}{2}+\alpha$ e $s=\frac{(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})}{2}+\frac{\boldsymbol{\beta}'\boldsymbol{D}_\tau^{-1}\boldsymbol{\beta}}{2}+\gamma$.

Por último, resta encontrar as distribuições condicionais dos parâmetros que foram criados para tornar possível a implementação do LASSO Bayesiano através do Amostrador de Gibbs. Sabendo que a parte da distribuição conjunta (3.19) que contém τ_j^2 é

$$(\tau_j^2)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\frac{\beta_j^2}{\tau_j^2\sigma^2}+\lambda^2\tau_j^2\right)\right\}, \tag{3.27}$$

Park and Casella (2008) apresentam que $\eta_j^2=\frac{1}{\tau_j^2}$, quando condicionado aos demais termos da formulação hierárquica, possui distribuição Normal-Inversa com parametrização tal que

$$\mathrm{f}(x)=\sqrt{\frac{\lambda'}{2\pi}}x^{-\frac{3}{2}}\exp\left\{-\frac{\lambda'(x-\mu')^2}{2(\mu')^2x}\right\}, \text{ com } x>0. \tag{3.28}$$

Para verificar isto é preciso, primeiro, reescrever a quantidade (3.27) em termos de η_j^2 ao invés de τ_j^2. Visto que o jacobiano proveniente desta transformação de variável é $-\frac{1}{(\eta_j^2)^2}$, a densidade de η_j^2 é proporcional a

$$\propto (\eta_j^2)^{-\frac{3}{2}}\exp\left\{-\frac{1}{2}\left(\frac{\beta_j^2}{\sigma^2}\eta_j^2+\frac{\lambda^2}{\eta_j^2}\right)\right\}. \tag{3.29}$$

Para que a quantidade (3.29) fique no formato de uma distribuição Normal-Inversa com a parametrização de (3.28) é preciso completar um quadrado. Como

$\left(-2\beta_j^2\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}\right)/\sigma^2$ não depende de η_j^2, pode-se reescrever (3.29) como sendo proporcional a

$$
\begin{aligned}
&(\eta_j^2)^{-\frac{3}{2}}\exp\left\{-\frac{1}{2}\left(\frac{\beta j^2}{\sigma^2}\eta_j^2 - 2\frac{\beta j^2\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}}{\sigma^2} + \frac{\lambda^2}{\eta_j^2}\right)\right\}\\
&= (\eta_j^2)^{-\frac{3}{2}}\exp\left\{-\beta_j^2\left(\frac{\eta_j^2}{2\sigma^2} - \frac{\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}}{\sigma^2} + \frac{\lambda^2}{2\eta_j^2\beta_j^2}\right)\right\}\\
&= (\eta_j^2)^{-\frac{3}{2}}\exp\left\{-\frac{\beta_j^2\left[(\eta_j^2)^2 - 2\eta_j^2\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}} + \frac{\lambda^2\sigma^2}{\beta_j^2}\right]}{2\sigma^2\eta_j^2}\right\}\\
&= (\eta_j^2)^{-\frac{3}{2}}\exp\left\{-\frac{\beta_j^2\left(\eta_j^2 - \sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}\right)^2}{2\sigma^2\eta_j^2}\right\}. \qquad (3.30)
\end{aligned}
$$

Assim, de fato, a distribuição de $\eta_j^2 = \frac{1}{\tau_j^2}$ é Normal-Inversa com $\mu' = \sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}$ e $\lambda' = \lambda^2$.

Calculadas as distribuições condicionais, o Amostrador de Gibbs é capaz de gerar amostras aleatórias da distribuição a posteriori (3.14), que serão utilizadas para se calcular os parâmetros do LASSO Bayesiano através de estatísticas resumo como a moda ou a média.

3.3.3 Lidando Com o Parâmetro de Regularização

O parâmetro de regularização λ pode ser calibrado de diversas maneiras. Embora o método de validação cruzada seja frequentemente utilizado para tal calibração, Park and Casella (2008) propõe outras duas alternativas bayesianas: o uso do método chamado Bayes Empírico via Máxima Verossimilhança Marginal e o uso de uma hiperpriori para λ. Para a aplicação do LASSO Bayesiano nos dados genéticos será utilizada a segunda alternativa bayesiana. Dessa forma, a escolha de um λ específico não será mais necessária, ao passo que a escolha da hiperpriori deve ser realizada com bastante cuidado de modo a evitar inconveniências na distribuição a

posteriori, a citar como exemplo a ocorrência de múltiplas modas (Park and Casella, 2008).

Tal como sugerido pelos autores do método, a λ^2 foi atribuída uma hiperpriori gama da seguinte forma:

$$\pi(\lambda^2) = \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}e^{-\delta\lambda^2}, \quad \lambda^2 > 0 \quad \text{e} \quad r, \delta > 0. \tag{3.31}$$

A definição da hiperpriori em λ^2 e não em λ é realizada por pura conveniência matemática.

Assim, a densidade conjunta resultante da formulação hierárquica passa a ser expressa por

$$\mathrm{f}(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\beta}, \sigma^2)\pi(\sigma^2)\pi(\boldsymbol{\mu}) \prod_{j=1}^{p} \pi(\beta_j|\tau_j^2, \sigma^2)\pi(\tau_j^2)\pi(\lambda^2) =$$

$$\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{\frac{1}{2\sigma^2}\left(\boldsymbol{y} - \boldsymbol{\mu}\mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta}\right)'\left(\boldsymbol{y} - \boldsymbol{\mu}\mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta}\right)\right\} \frac{\gamma^a}{\Gamma(a)}(\sigma^2)^{-a-1} \exp\left\{-\frac{\gamma}{\sigma^2}\right\} \times$$

$$\prod_{j=1}^{p} \frac{1}{(2\sigma^2\tau_j^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2\right\} \frac{\lambda^2}{2} \exp\left\{-\lambda^2\frac{\tau_j^2}{2}\right\} \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1} \exp\{-\delta\lambda^2\}. \tag{3.32}$$

Eliminando-se os termos que não envolvem λ^2 na Equação (3.32), tem-se

$$\prod_{j=1}^{p}\left[\frac{\lambda^2}{2} \exp\left\{-\lambda^2\frac{\tau_j^2}{2}\right\}\right] \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1} \exp\{-\delta\lambda^2\}$$

$$(\lambda^2)^p \exp\left\{-\lambda^2\frac{1}{2}\sum_{j=1}^{p}\tau_j^2\right\}(\lambda^2)^{r-1} \exp\{-\delta\lambda^2\}$$

$$(\lambda^2)^{p+r-1} \exp\left\{-\lambda^2\left(\delta + \frac{1}{2}\sum_{j=1}^{p}\tau_j^2\right)\right\}. \tag{3.33}$$

Dessa forma, a distribuição condicional de λ^2 é gama com parâmetro de forma igual a $p + r$ e parâmetro de taxa igual a $\sum_{j=1}^{p}\frac{\tau_j^2}{2} + \delta$.

O hiperparâmetro δ deve ser suficientemente maior do que 0 para evitar problemas conceituais e computacionais. Da mesma forma, r deve ser diferente de 0

para evitar complicações na distribuição a posteriori (Park and Casella, 2008).

3.4 Exemplo Prático do LASSO e LASSO Bayesiano

Para uma rápida comparação do LASSO e LASSO Bayesiano com os métodos apresentados no Capítulo 2, ambos foram aplicados nas mesmas bases de dados utilizadas na seção 2.7. Para a aplicação do LASSO utilizou-se o pacote *glmnet* (Friedman et al., 2010) e para a aplicação do LASSO Bayesiano utilizou-se o pacote *monomvn* (Gramacy, 2018), ambos provêm do *software* R.

Em ambos os casos realizou-se a padronização das variáveis preditoras antes de se submeter os dados ao modelo.

No caso da aplicação do LASSO, a função *glmnet* do respectivo pacote trabalha no procedimento de otimização da seguinte função objetivo:

$$\min_{(\beta_0,\boldsymbol{\beta})} \frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{x_i'}\boldsymbol{\beta})^2 + \lambda_{glmnet}\left[\frac{(1-a)}{2} \parallel \boldsymbol{\beta} \parallel_2^2 + a \parallel \boldsymbol{\beta} \parallel_2\right], \tag{3.34}$$

em que $\lambda_{glmnet} \geq 0$ é o parâmetro de regularização utilizado como *input* na função e $0 \leq a \leq 1$. Quando $a = 0$, tem-se a regressão Ridge e quando $a = 1$, tem-se o LASSO.

A divisão por duas vezes o número de observações ($2n$) e uma padronização implícita da variável resposta realizada pela função *glmnet*, fazem com que o parâmetro λ_{glmnet} seja diferente do λ da Equação (3.4). Quando $a = 0$ ou $a = 1$, é possível se estabelecer uma relação de correspondência entre tais parâmetros. Neste contexto, a função *glmnet* primeiro padroniza $\boldsymbol{y}$, dividindo-o por $s_y = \sqrt{\sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{n}}$, para posteriormente minimizar

$$\frac{1}{2n}\sum_{i=1}^{n}(y_{s_i} - \boldsymbol{x_i'}\boldsymbol{\gamma})^2 + \eta\sum_{j=1}^{p}\left[\frac{(1-a)}{2}\gamma_j^2 + a|\gamma_j|\right], \tag{3.35}$$

em que $\boldsymbol{y_s} = \frac{\boldsymbol{y}}{s_y}$, $\gamma = \frac{\boldsymbol{\beta}}{s_y}$ e $\eta = \frac{\lambda_{glmnet}}{s_y}$. De modo equivalente, pode-se minimizar

$$\frac{1}{2ns_y^2}\sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{x_i'}\boldsymbol{\beta})^2 + \frac{\lambda_{glmnet}}{s_y}\frac{a}{s_y}\sum_{j=1}^{p}|\beta_j| + \frac{\lambda_{glmnet}}{s_y}\frac{(1-a)}{2s_y^2}\sum_{j=1}^{p}\beta_j^2, \tag{3.36}$$

ou equivalentemente, minimizar

$$\sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{x_i'}\boldsymbol{\beta})^2 + 2n\lambda_{glmnet}a\sum_{j=1}^{p}|\beta_j| + n\frac{\lambda_{glmnet}}{s_y}(1-a)\sum_{j=1}^{p}\beta_j^2. \tag{3.37}$$

Desse modo, para aplicar o LASSO ou o Ridge de tal forma que estes estejam em conformidade com a Equação (3.4), é preciso, inicialmente, realizar uma transformação no parâmetro de regularização que será utilizado no *input* da função:

1. No caso do LASSO, tem-se $\lambda = 2n\lambda_{glmnet}$ e, portanto, $\lambda_{glmnet} = \frac{\lambda}{2n}$.

2. No caso do Ridge, tem-se $\lambda = n\frac{\lambda_{glmnet}}{s_y}$ e, portanto, $\lambda_{glmnet} = s_y\frac{\lambda}{n}$.

Dessa forma, definiu-se, para cada base de dados, um *grid* de valores de λ a partir do respectivo *grid* utilizado na seção 2.7, já que $\lambda = \alpha^2$. Posteriormente, calculou-se o λ_{glmnet} a ser utilizado na função.

Para a aplicação do LASSO Bayesiano, a função do pacote utilizada é a *blasso*, que possui um algoritmo implementado idêntico ao descrito em Park and Casella (2008).

Para a base de próstata, foram coletadas 100.000 amostras MCMC, utilizando um *thin* igual a 100 (número de amostras geradas a serem descartadas no intervalo entre as coletas). Para o modelo hierárquico, os hiperparâmetros das distribuições a priori do λ^2 e do σ^2 são valores não informativos provenientes do *default* da função. Para a estimação das médias das distribuições a posteriori dos coeficientes, desconsiderou-se as 10.000 primeiras amostras geradas. O mesmo aplica-se para a base de quimiometria, com a diferença de terem sido coletadas 10.000 amostras

MCMC com *thin* igual a 10 e de terem sido desconsideradas as 1.000 primeiras para a estimativa da média a posteriori.

As Figuras 3.3(a) e 3.3(b) comparam as curvas de validação cruzada do LASSO em cada base de dados com as curvas provenientes dos outros métodos. Como o LASSO Bayesiano utiliza uma distribuição para o parâmetro de regularização, a seleção do modelo é realizada automaticamente e só é possível obter um EQMcv para o método.

Em ambos os casos o LASSO Bayesiano não alcançou uma boa performance, enquanto que o LASSO foi o método com o melhor desempenho, com um EQMcv de 0.571 pra $\alpha = 2.250$ nos dados de próstata e um EQMcv de 0.454 para $\alpha = 3.025$ nos dados de quimiometria.

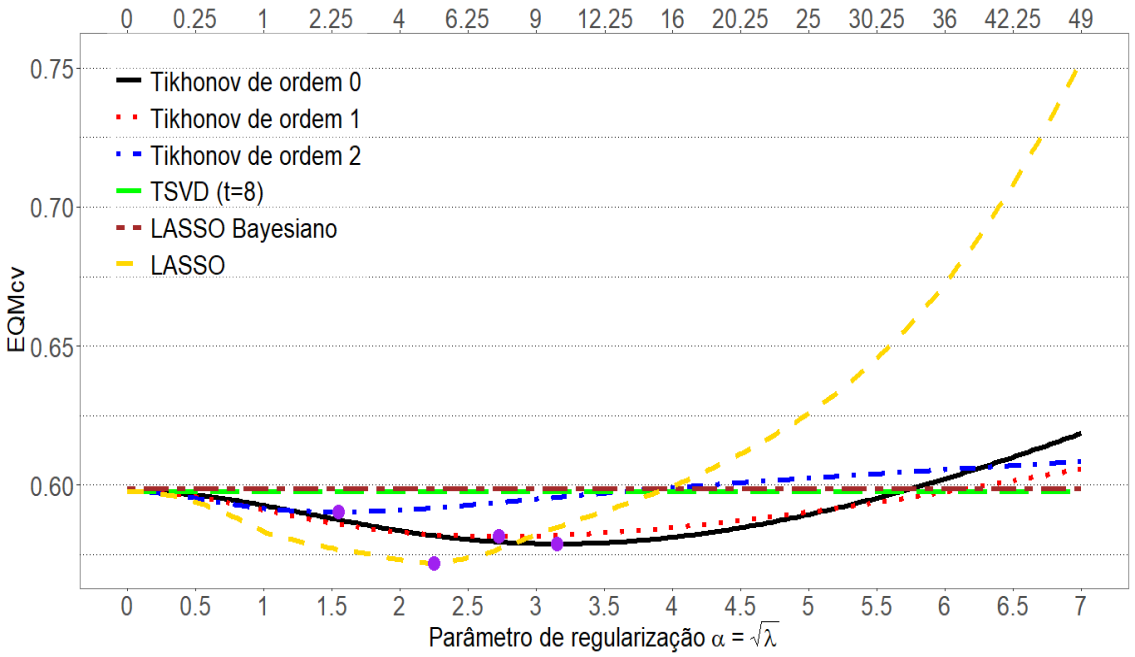
Na Figura 3.3(b), diferentemente da Figura 3.3(a), existe uma diferença numérica quando $\alpha = 0$ entre os EQMcv's do LASSO e dos métodos Tikhonov.

A Tabela 3.1 apresenta os coeficientes dos dados de próstata gerados por cada método, com os melhores parâmetros selecionados pela validação cruzada, além de apresentar os valores gerados pela função *lm* do *software* R, que calcula os coeficientes da regressão linear. No caso do LASSO Bayesiano, apenas aplicou-se o método com os mesmo hiperparâmetros, fornecidos pelo *default* da função *blasso*.

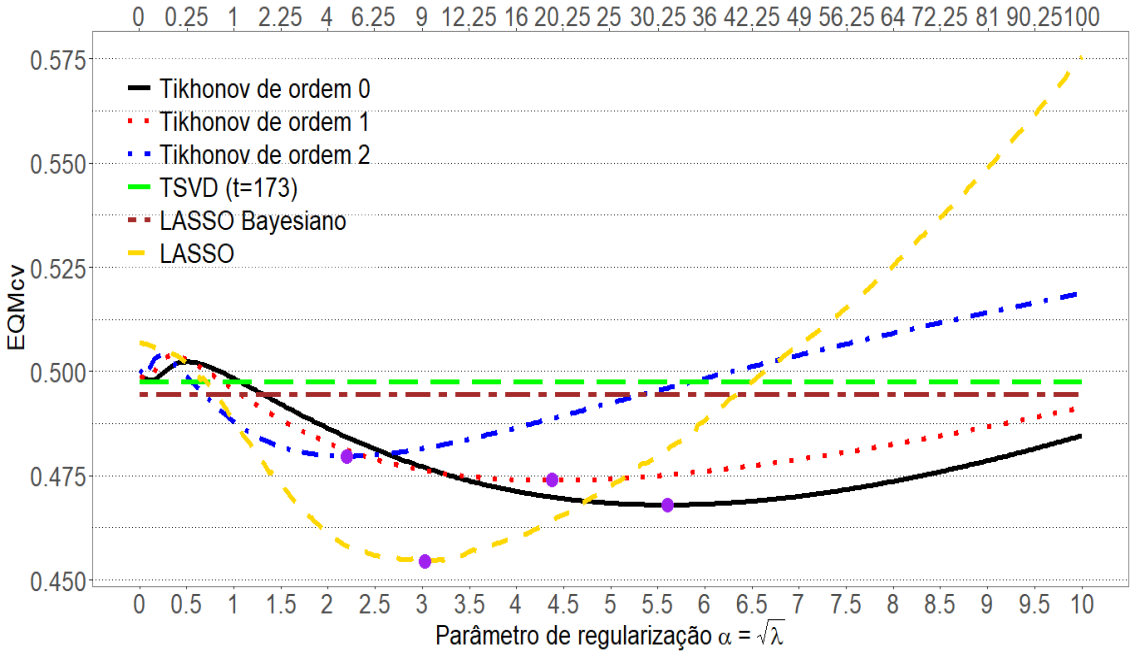
Para a aplicação dos métodos, utilizou-se todo o banco de dados, após o mesmo ter sido padronizado. Portanto, a estimativa do intercepto é a mesma para todos.

Abreviação da variável	Função lm	TSVD ($t = 8$)	Tikhonov de ordem 0 ($\alpha = 3.150$)	Tikhonov de ordem 1 ($\alpha = 2.725$)	Tikhonov de ordem 2 ($\alpha = 1.550$)	LASSO ($\alpha = 2.250$)	LASSO Bayesiano
Intercepto	2.478	2.478	2.478	2.478	2.478	2.478	2.478
lcavol	0.665	0.665	0.554	0.587	0.642	0.601	0.453
lweight	0.266	0.266	0.255	0.267	0.248	0.242	0.204
age	-0.158	-0.158	-0.114	-0.085	-0.102	-0.084	-0.049
lbph	0.140	0.140	0.120	0.110	0.118	0.102	0.087
svi	0.315	0.315	0.275	0.241	0.238	0.250	0.212
lcp	-0.148	-0.148	-0.031	-0.005	-0.035	0.000	0.059
gleason	0.036	0.036	0.048	0.036	0.009	0.009	0.049
pgg45	0.126	0.126	0.090	0.082	0.106	0.068	0.072

Tabela 3.1: Coeficientes calculados segundo cada método de regularização (base `PROS`).



(a) `Base PROS.`



(b) `Base QUIM.`

Figura 3.3: Sensibilidade da métrica de avaliação de desempenho segundo os valores de α para cada regularização.

Como o método TSVD utilizou todos os valores singulares, os resultados são iguais aos da solução pseudoinversa, que, para estes dados, equivale à solução de mínimos quadrados do modelo de regressão linear.

Para compreender a extensão do viés agregado à cada coeficiente, é interessante utilizar os coeficientes da regressão linear como referência. Pela distância euclidiana, o LASSO Bayesiano foi o método que mais se distanciou da solução de mínimos quadrados.

Embora exista uma variação razoável no valor de cada coeficiente dentre os métodos, é possível observar uma diferença mais extrema: o LASSO excluiu do modelo a variável *lcp*, enquanto os outros métodos apenas reduziram satisfatoriamente o efeito da mesma.

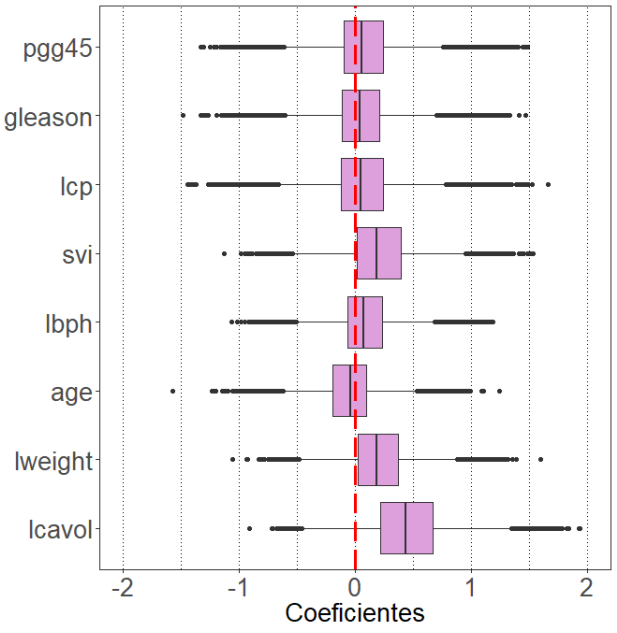


Figura 3.4: Boxplot das amostras a posteriori para cada coeficiente do LASSO Bayesiano (base `PROS`).

Mesmo o LASSO Bayesiano não excluindo diretamente variáveis do modelo, a Figura 3.4, que resume as informações da distribuição a posteriori de cada coeficiente, sugere que apenas três dentre os oito coeficientes estão bem distanciados de zero: *lcavol*, *lweight* e *svi*.

Capítulo 4

Aplicação em Dados Genéticos

Redução de custos, reprodutibilidade e redução do tempo necessário para alcançar um objetivo são pilares para quem trabalha com melhoramento de plantas ou animais. Tais profissionais buscam balancear as vantagens e desvantagens desses métodos de melhoramento, escolhendo aquele que melhor atende às suas necessidades.

O melhoramento tradicional de plantas ou animais envolve uma seleção baseada na observação e avaliação das características fenotípicas de interesse (característica externas, morfológicas, fisiológicas e comportamentais dos indivíduos). Realizada a seleção, o melhoramento se dá a partir do cruzamento dos indivíduos selecionados.

Posteriormente, técnicas mais avançadas que levavam em consideração as relações de parentesco e o efeito do ambiente sobre o fenótipo foram implementadas tendo em vista que uma boa compreensão do delineamento de um experimento e dos dados gerados era e ainda é um pré-requisito fundamental para a análise dos valores quantitativos expressos nos fenótipos.

Embora o controle dos fatores ambientais seja fundamental para a seleção dos principais indivíduos a serem utilizados no melhoramento, sabe-se que o desempenho de um animal ou planta também é influenciado pela genética e sua interação com o ambiente. Assim, pode se utilizar essa informação genética a fim de se definir, através de previsões do fenótipo, os candidatos a indivíduos elite.

Considerando-se que os dados provenientes dos fenótipos (ou traços) foram coletados corretamente no campo, a seleção fenotípica, que é realizada com base na avaliação direta do fenótipo, é eficaz e a identificação dos indivíduos superiores torna-se uma tarefa possível. Apesar de eficaz, esse procedimento está longe de ser eficiente quando o melhoramento é realizado em espécies com longos ciclos de desenvolvimento. A espera excessiva para a avaliação dos fenótipos, que é uma consequência desses longos ciclos, retarda o procedimento para a obtenção de indivíduos superiores. A existência de ciclos com períodos longos é bastante comum principalmente no melhoramento de plantas e trata-se de um problema sério por fazer do melhoramento com base na seleção fenotípica um procedimento custoso.

Nesse contexto, o melhoramento genético apresenta-se como uma ferramenta alternativa com grande potencial de aplicação para a definição dos cruzamentos a serem realizados, visando-se a criação de novos genótipos e a indicação mais ágil dos indivíduos superiores a serem utilizados comercialmente (Resende et al., 2008). A pesquisa genética molecular passou a ganhar destaque, sustentada no argumento de que a utilização da informação em um nível molecular traria um ganho genético mais rápido que os melhoramentos baseados somente nas informações fenotípicas (Meuwissen et al., 2001).

A Seleção Auxiliada por Marcadores Moleculares (MAS) foi uma das primeiras técnicas de melhoramento genético propostas e trouxe consigo a possibilidade de atenuar problemas envolvendo os longos ciclos de melhoramento e os altos custos provenientes da manutenção e fenotipagem dos ensaios experimentais. Sendo marcadores moleculares estruturas que fornecem as informações genéticas necessárias para a implementação da metodologia, a MAS requer o conhecimento prévio das associações desses marcadores com regiões específicas do genoma chamadas *Quantitative Trait Loci* (QTL), que são regiões reguladoras de características quantitativas.

Apesar dos avanços nas tecnologias utilizadas para a identificação dos QTL's e de suas respectivas associações com os marcadores moleculares, a MAS não é tão eficaz em problemas reais. A dificuldade em se trabalhar com QTL's motivou uma mudança de paradigma, minimizando-se o foco em efeitos individuais dos traços

e passando-se a trabalhar com a informação genética conjunta para predições dos fenótipos (Grattapaglia, 2014). À utilização conjunta e direta das informações presentes na fita de DNA, deu-se o nome de Seleção Genômica (GS), também conhecida como Seleção Genômica Ampla (GWS).

Um dos principais pressupostos da Seleção Genômica é que, dada a enorme quantidade de marcadores utilizados no estudo, todos os QTL's relacionados à característica de interesse estão correlacionados com pelo menos um dos marcadores utilizados no melhoramento, fazendo da detecção dos QTL's e do estudo das respectivas associações um passo desnecessário (Lin et al., 2014).

Por um determinado período, a nova metodologia permaneceu pouco utilizada, em grande parte pela inviabilidade da utilização dos marcadores moleculares até então disponíveis. Com o surgimento dos marcadores SNP (*Single Nucleotide Polymorphism*), que são variações na sequência de DNA em um único nucleotídeo e existentes em pelo menos 1% da população estudada, tal metodologia voltou a ganhar espaço tendo em vista que a genotipagem de SNP's é de fácil obtenção e relativamente barata, viabilizando a construção de bases genéticas com uma quantidade numerosa de marcadores. Essas amplas bases de dados com informações genéticas no nível dos indivíduos são, então, juntamente com os respectivos fenótipos, utilizadas para o treinamento e validação de modelos estatísticos de previsão ou classificação, posteriormente utilizados não somente para aprimorar a compreensão da conexão existente entre elementos genômicos e fenótipos como também para auxiliar na definição dos cruzamentos a serem realizados.

A Seleção Genômica limita-se à construção de modelos estatísticos que irão realizar predições dos chamados GEBV's (*Genomic Estimated Breeding Values*), que podem ser obtidos de diversas formas e em diversos contextos. No contexto apresentado neste trabalho, os GEBV's serão obtidos através da retirada dos efeitos de delineamento dos fenótipos.

A seleção dos novos indivíduos é, então, realizada com base em seus genótipos, utilizando-se o modelo estatístico já construído para a predição dos GEBV's. Assim sendo, a seleção dos melhores será diretamente baseada nos valores preditos dos

GEBV's, observando-se o *rank*.

Dada a viabilidade econômica da utilização de SNP's e de sua cobertura na fita de DNA, a seleção com base nesses marcadores pode fornecer, desde que possua poder preditivo satisfatório, uma vantagem relativa à seleção fenotípica devido à redução dos custos e à classificação precoce daqueles que serão considerados candidatos a indivíduos elite.

Devido ao número de variáveis (marcadores) ser frequentemente muito maior que o número de observações (indivíduos – plantas ou animais), a estimação dos efeitos dos marcadores e a consequente previsão dos GEBV's não é uma tarefa tão simples. A falta de graus de liberdade para o cálculo das estimativas através de métodos como Mínimos Quadrados Ordinários juntamente com a existência de muitos marcadores em LD (*Linkage Disequilibrium*) — marcadores associados — são alguns dos desafios a serem enfrentados.

O método BLUP (*Best Linear Unbiased Predictor*) proposto por Henderson (1984) procura contornar este problema de *big data*, embora possua uma suposição pouco observada na prática que é a exigência de igualdade entre as variâncias dos marcadores. Métodos bayesianos, ao propor a utilização de distribuições a priori para essas variâncias, tornam-se mais realistas (Meuwissen et al., 2001).

4.1 Ciclo de Melhoramento de Plantas Via GS

A Figura 4.1 apresenta de forma sucinta um esquema do melhoramento de plantas através da Seleção Genômica.

Dentre as diversas fases da Seleção Genômica, a Figura 4.1 destaca duas: o desenvolvimento do modelo preditivo e a seleção das melhores mudas, que foram obtidas através do cruzamento das árvores do grupo utilizado no treinamento do modelo.

Eventualmente o melhorista pode coletar uma amostra aleatória das mudas com as melhores previsões de GEBV's e realizar o respectivo plantio sob um delineamento experimental. Depois de alguns anos tais indivíduos podem ter as suas

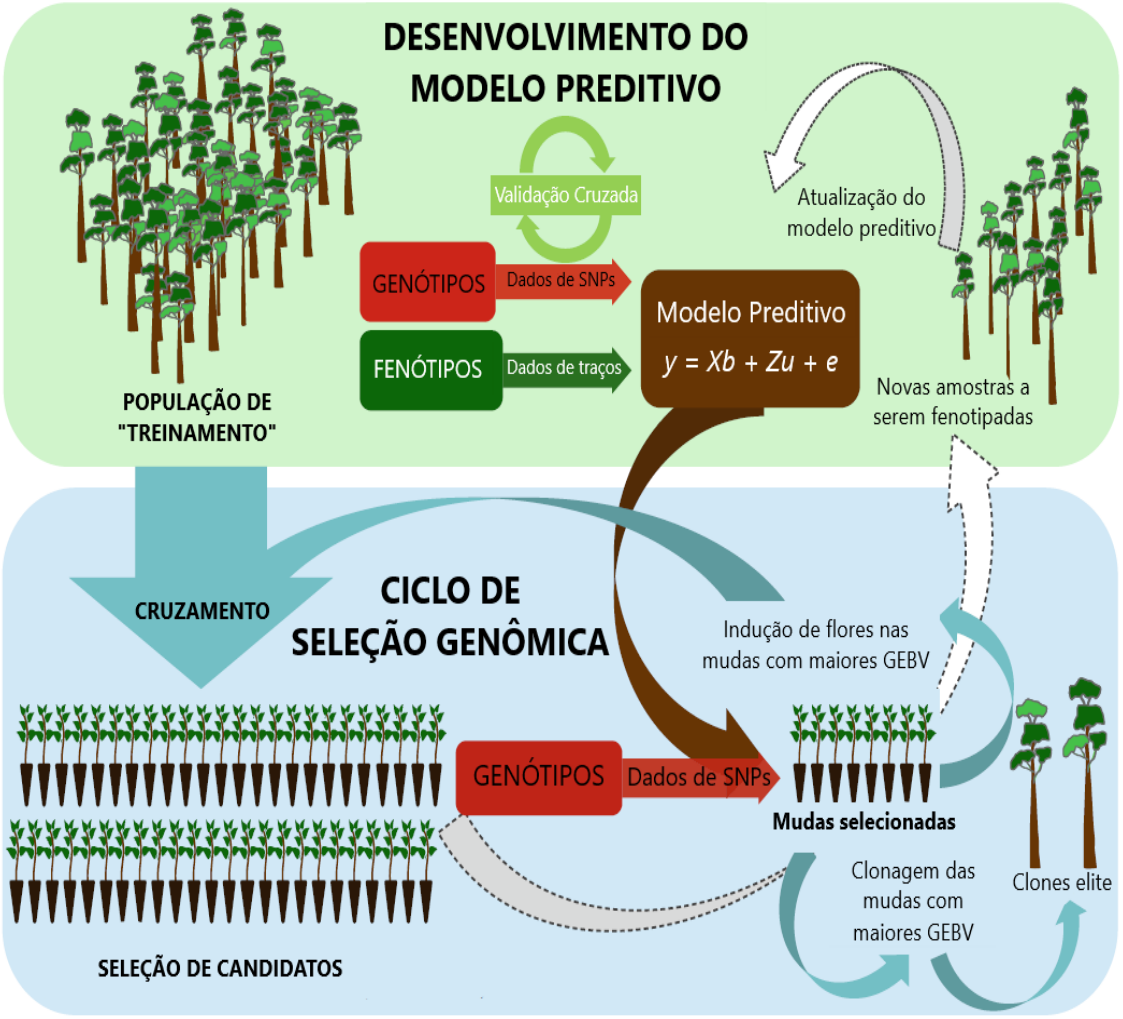


Figura 4.1: Ciclo GS (Figura retirada de Grattapaglia (2014), com adaptações).

informações fenotípicas registradas e serem utilizados para atualizar o modelo preditivo (Grattapaglia, 2014). Enquanto a maturação necessária para a fenotipagem não acontecer, o melhorista induz o florescimento e realiza o cruzamento das mudas para a realização de novas seleções.

Em estudos genômicos existe uma associação negativa entre o tamanho efetivo da população (conceito que está relacionado à compatibilidade genética existente entre indivíduos) e a extensão dos desequilíbrios de ligação entre os marcadores e os QTL's (Grattapaglia, 2014). Na prática, conforme os ciclos de seleção vão sendo realizados e novas populações vão sendo geradas através dos cruzamentos dos ancestrais elite, tem-se um maior compartilhamento de material genético em comum

entre os indivíduos e isso caracteriza uma redução do tamanho efetivo da população. Com essa redução, a extensão dos desequilíbrios de ligação entre marcadores e QTL's aumenta, ocasionando um aumento do poder preditivo da GS.

4.2 Base de Dados

Para a implementação dos modelos e avaliação da Seleção Genômica será utilizada uma base de dados reais com informações do genoma de árvores de eucaliptos. A base de dados envolve informações de 33.398 marcadores do tipo SNP de 999 plantas de eucalipto plantadas em 2006 na cidade de Brotas, São Paulo, Brasil. As árvores de eucalipto utilizadas no estudo foram obtidas através de uma amostra, com base em variáveis de crescimento, de 1.000 indivíduos dentre os 2.784 eucaliptos plantados sob um delineamento de blocos completos com parcela. As 2.784 plantas se subdividem em 45 famílias. Apesar de serem amostradas 1.000 plantas, não foi possível coletar a informação genética de uma delas, sobrando assim 999 para estudo (Lima, 2014).

Uma vez que os SNP's representam pares de bases nitrogenadas do DNA, as variáveis do banco de dados de eucalipto são codificadas originalmente através de três combinações de pares de letras. A critério de exemplo, se uma variável apresentar as combinações "AA", "AT" e "TT", o par "AT" é o genótipo heterozigoto e os demais pares são genótipos homozigotos. Para a realização dos estudos relacionados a essas bases, realizou-se uma codificação diferente de tal forma que o genótipo homozigoto de maior frequência fosse identificado pelo número "1", o genótipo homozigoto de menor frequência pelo número "-1" e o genótipo heterozigoto pelo número "0".

Foram coletados 15 fenótipos de relevância comercial, que podem ser agrupados em químicos, físicos e de crescimento. A Tabela 4.1 apresenta as estatísticas descritivas acerca dos fenótipos das plantas de eucalipto que estão disponíveis para estudos.

Os quatro primeiros fenótipos da Tabela 4.1 são de crescimento. Em seguida, com exceção da Relação SG, Celulose, Hemicelulose e dos traços relativos à lignina,

Fenótipo	Mínimo	Máximo	Desvio Padrão	Média	Mediana	Frequência de NA's
Diâmetro à Altura do Peito (DAP) (cm)	12.35	23.24	1.83	16.60	16.55	0
Altura (m)	19.60	27.60	1.36	24.19	24.20	0
Volume da Madeira (m^3)	0.12	0.50	0.06	0.24	0.23	0
Incremento médio anual (IMA) ($m^3 \cdot ha^{-1} \cdot ano$)	28.87	121.64	14.92	58.63	56.66	0
Celulose (%)	41.80	55.40	1.77	48.88	48.80	0
Hemicelulose (%)	13.90	20.20	0.88	17.27	17.30	0
Relação SG (Siringil/Guaiacil)	1.83	4.20	0.41	2.93	2.90	0
Lignina Insolúvel (%)	20.70	28.80	1.11	25.24	25.30	0
Lignina Solúvel (%)	2.20	4.90	0.41	3.55	3.50	0
Lignina Total (%)	24.40	32.10	1.05	28.78	28.80	0
Densidade da Madeira ($kg \cdot m^{-3}$)	407.10	646.50	35.90	512.51	511.80	0
Ângulo Microfibrilar	10.50	17.50	1.22	12.94	12.90	651
Comprimento da Fibra (mm)	0.59	0.93	0.06	0.75	0.75	649
Largura da Fibra (μm)	17.20	24.70	1.14	19.84	19.80	649
Rigidez ($mg \cdot 100m^{-1}$)	4.40	11.00	1.02	7.11	7.00	650

Tabela 4.1: Estatísticas resumo dos fenótipos relativos aos dados reais.

todos os demais podem ser considerados fenótipos físicos.

A avaliação do mínimo, máximo e desvio padrão, juntamente com a informação da média, permite concluir que não existem observações muito discrepantes das demais em nenhum dos traços em questão.

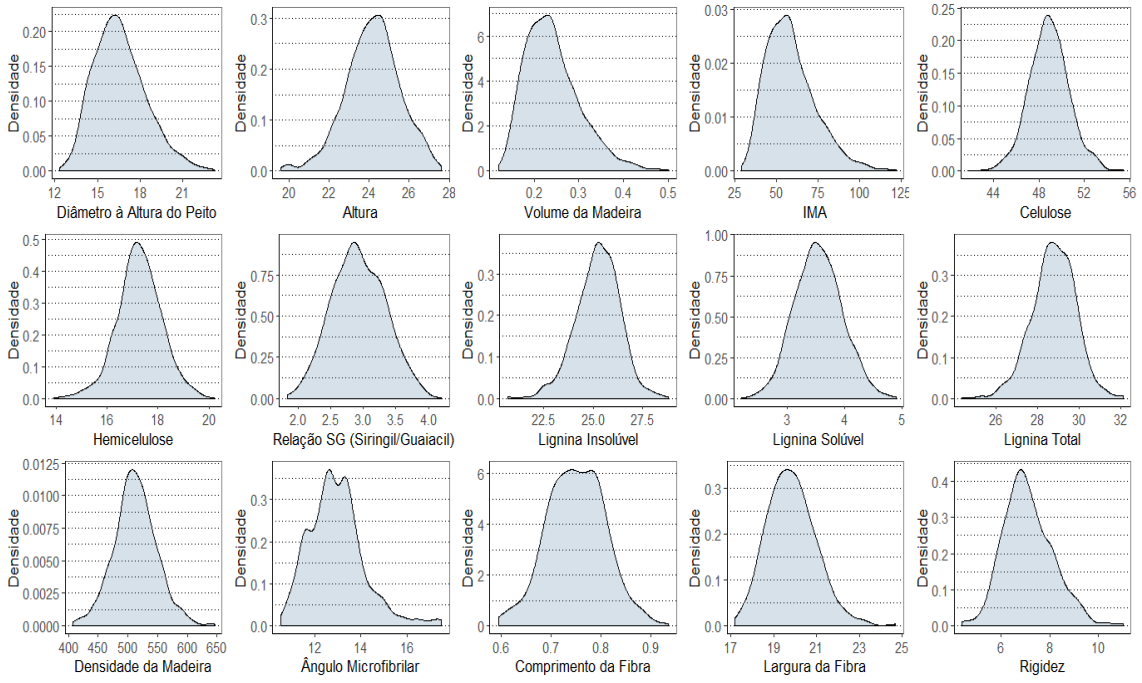


Figura 4.2: Densidades estimadas.

A Figura 4.2 apresenta as estimativas das respectivas densidades por meio da utilização do kernel gaussiano e fornece suporte às conclusões sobre o comportamento

desses fenótipos. Embora não seja muito acentuada, três traços possuem densidades com assimetria à direita: Volume da Madeira, IMA e Ângulo Microfibrilar. Para os demais traços, os gráficos apontam para um padrão de simetria.

Embora os custos de genotipagem tenham diminuído drasticamente nos últimos anos, os custos de fenotipagem permanecem altos, limitando o aumento do número de observações para a Seleção Genômica (Lin et al., 2014).

Devido às dificuldades inerentes ao processo de coleta dos fenótipos, apenas quatro, dentre os quinze apresentados na Tabela 4.1, tiveram seus valores devidamente registrados para todas as 999 plantas genotipadas: DAP, Altura, Volume da Madeira e IMA. Para os demais fenótipos, registrou-se os respectivos valores para cerca de 350 plantas, escolhidas de modo a maximizar a variabilidade e representatividade amostral (Lima, 2014). Posteriormente, foi aplicado o modelo de espectroscopia de infravermelho próximo (*Near Infrared Reflectance Spectroscopy* - NIRS) para se realizar as imputações das observações faltantes (Lima, 2014). Todavia, o modelo NIRS não foi capaz de realizar boas imputações para os quatro últimos fenótipos da Tabela 4.1.

4.3 Limpeza e Imputação da Matriz de marcadores

Para a análise dos dados genéticos, serão avaliados três métodos de regularização: Ridge, LASSO e LASSO Bayesiano. Mas antes, é preciso realizar um pré-processamento da matriz de marcadores e da variável resposta.

Para a obtenção de modelos com uma boa qualidade preditiva, foram adotados alguns procedimentos comumente utilizados na literatura de Seleção Genômica:

1. Foram excluídos 94 marcadores constantes, ou seja, marcadores que estavam codificados com apenas um genótipo.

2. Foram excluídos 5.723 marcadores por possuírem a frequência do alelo menos comum (*Minor Allele Frequency* - *MAF*) inferior a 1%.

Como erros de genotipagem induzem à existência de dados faltantes na matriz de marcadores, um outro procedimento comum é a exclusão do marcador cuja frequência dos genótipos faltantes exceda 10%, ou equivalentemente, cuja taxa de genotipagem (*call rate*) seja inferior a 90%. Embora existam genótipos faltantes nos dados em estudo, tal procedimento não foi necessário já que nenhum marcador apresentou *call rate* inferior a 10%.

Realizados os procedimentos de limpeza da base de dados, procedeu-se com a imputação dos marcadores restantes que possuíam algum genótipo faltante, utilizando o valor médio dos genótipos de cada marcador.

4.4 As Etapas da Modelagem

A abordagem de modelagem utilizada nesta aplicação caracteriza-se por duas etapas bem definidas:

1. ajuste por efeitos de delineamento;

2. aplicação dos modelos de regularização utilizando os resíduos obtidos na primeira etapa como variável resposta.

Como foi realizada uma seleção de 1.000 sem que fosse levado em consideração o delineamento experimental aplicado nas 2.784 plantas de origem, optou-se apenas pelo ajuste por efeito de bloco. Assim, para a retirada do efeito de delineamento, aplicou-se um modelo com intercepto e blocos aleatórios.

Os resíduos obtidos, também chamados de GEBV's, tornam-se, então, a variável resposta dos modelos de regularização, sobre os quais aplicou-se um procedimento de validação cruzada 5-fold para encontrar os melhores modelos segundo um critério: a média das habilidades preditivas obtidas para cada grupo da validação.

A habilidade preditiva pode ser calculada pela correlação de Pearson entre os GEBV's (resíduos) e os valores preditos. Tal medida é importante para a Seleção Genômica devido à sua relação linear com o ganho genético (Lin et al., 2014).

4.5 Herdabilidade e Escolha dos Fenótipos

Sabe-se que a habilidade preditiva de um determinado modelo em um determinado fenótipo é influenciada por diversos fatores, sendo a herdabilidade um deles (Lin et al., 2014). A herdabilidade no sentido restrito é a proporção da variabilidade fenotípica que pode ser explicada pelos efeitos genéticos aditivos, representando, assim, o quão herdável estima-se que um determinado fenótipo é (los Campos et al., 2013). Tal medida considera apenas efeitos aditivos, desconsiderando-se efeitos de dominância, epistasia ou interação entre genótipo e ambiente (Lin et al., 2014). Existem diversas formas de se calcular a herdabilidade. No presente estudo, utilizou-se o modelo misto rrBLUP (Endelman, 2011), a partir do qual calcula-se a herdabilidade como sendo

$$h^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2}, \tag{4.1}$$

em que $\hat{\sigma}_g^2$ é a variância estimada genética e $\hat{\sigma}_e^2$ é a variância estimada dos resíduos. Para a estimação de σ_g^2, calculou-se $\hat{\sigma}_g^2 = \sum_{j=1}^{M} 2p_j(1-p_j)\hat{\sigma}_u^2$ (Gianola et al., 2009), em que p_j é a frequência do alelo mais comum no j-ésimo marcador e $\hat{\sigma}_u^2$ corresponde à variância do vetor de efeitos aleatórios.

Neste trabalho o modelo misto rrBLUP não será explorado em detalhes, mas será utilizado para fornecer a herdabilidade como uma ferramenta descritiva que informa a influência da genética no produto final, que é o fenótipo. Para o cálculo das mesmas, utilizou-se a matriz de marcadores após a sua limpeza e imputação, embora as imputações não tenham sido utilizadas para o cálculo de p_j. Da mesma forma, como variável resposta do modelo misto rrBLUP, optou-se por utilizar os resíduos obtidos através do ajuste pelos efeitos de delineamento.

A Tabela 4.2 apresenta as herdabilidades calculadas pelo método rrBLUP para cada traço.

Implementou-se os modelos de regularização utilizando-se o fenótipo Relação SG, que possui alta herdabilidade, e o fenótipo DAP, que possui herdabilidade mo-

Fenótipo	Herdabilidade
Diâmetro à Altura do Peito (DAP) (cm)	0.432
Altura (m)	0.330
Volume da Madeira (m^3)	0.417
Incremento médio anual (IMA) ($m^3 \cdot ha^{-1} \cdot ano$)	0.418
Celulose (%)	0.554
Hemicelulose (%)	0.642
Relação SG (Siringil/Guaiacil)	0.845
Lignina Insolúvel (%)	0.660
Lignina Solúvel (%)	0.695
Lignina Total (%)	0.661
Densidade da Madeira ($kg \cdot m^{-3}$)	0.575
Ângulo Microfibrilar	0.127
Comprimento da Fibra (mm)	0.623
Largura da Fibra (μm)	0.103
Rigidez ($mg \cdot 100m^{-1}$)	0.281

Tabela 4.2: Herdabilidades dos fenótipos.

derada. Com isso, será possível verificar o impacto da herdabilidade de um fenótipo na habilidade preditiva obtida.

4.6 Aplicação dos Modelos e Resultados

Para a aplicação dos modelos de regularização LASSO e Ridge, fez-se uso do mesmo pacote utilizado anteriormente nos dados de próstata e quimiometria (*glmnet*). Já para a aplicação do LASSO Bayesiano, optou-se por utilizar o pacote BLR do *software* R, amplamente utilizado em Seleção Genômica. Diferentemente da metodologia aplicada nos exemplos de próstata e quimiometria, não foi realizado nenhum tipo de padronização na matriz de marcadores e o intercepto foi calculado diretamente, através da opção *intercept=TRUE*, presente nas funções dos respectivos pacotes.

A primeira versão do BLR foi lançada em 2010 e, desde então, o pacote tem sido citado em diversas publicações envolvendo análises genéticas para programas de melhoramento de plantas e animais (los Campos et al., 2013).

Relativamente ao pacote *monomvn*, utilizado nos exemplos com as bases de próstata e quimiometria, o pacote BLR demonstra superioridade no tempo de processamento, mas não fornece uma matriz com as amostras dos coeficientes que foram

coletadas no MCMC, retornando apenas a média para cada coeficiente. Contudo, as amostras coletadas da variância e dos parâmetros de regularização são salvos em arquivos de texto.

Apesar do BLR possuir um algoritmo que reproduz a proposta de implementação computacional de Park and Casella (2008), existe uma pequena diferença relativa à priori para σ^2: enquanto Park and Casella (2008) sugerem a utilização de uma priori gama inversa, o pacote BLR opera com uma priori qui-quadrado inversa escalonada. Comparada à distribuição gama inversa, a qui-quadrado inversa escalonada descreve a mesma distribuição de dados, mas utiliza uma parametrização diferente. A parametrização para a priori de σ^2 no BLR é descrita como segue:

$$f(\sigma^2|v,S) = \frac{\left(\frac{S}{2}\right)^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right)} \frac{\exp\left(-\frac{S}{2\sigma^2}\right)}{(\sigma^2)^{1+\frac{v}{2}}}, \tag{4.2}$$

cuja esperança é dada por $\frac{S}{v-2}$ (para $v > 2$) e cuja moda é dada por $\frac{S}{v+2}$ (los Campos et al., 2013). Assim, a densidade (4.2) é equivalente à distribuição gama inversa (3.26) com o primeiro parâmetro igual a $\frac{v}{2}$ e o segundo igual a $\frac{S}{2}$.

A parametrização do BLR para a priori do σ^2 mostra-se mais conveniente no contexto de Seleção Genômica. los Campos et al. (2013) fornecem algumas recomendações a respeito dos hiperparâmetros a serem utilizados no LASSO Bayesiano. Para a definição dos hiperparâmetros da priori do σ^2, recomenda-se utilizar um valor pequeno para v a fim de se reduzir a influência da priori, mas que esse seja maior do que 4 para evitar problemas com o valor esperado de σ^2. Já para o parâmetro S, recomenda-se utilizar

$$S = \text{Var}(\boldsymbol{y})(1-h^2)(v-2), \tag{4.3}$$

em que h^2 é a herdabilidade.

No tocante aos hiperparâmetros da distribuição a priori gama de λ^2, com parametrização tal que

$$\pi(\lambda^2) = \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}e^{-\delta\lambda^2}, \tag{4.4}$$

sendo r o parâmetro de forma e δ o parâmetro de taxa, recomenda-se defini-los de modo que a densidade tenha uma moda e seja relativamente não informativa em sua vizinhança. Uma vez que a moda de uma distribuição gama pode ser dada por $\frac{(r-1)}{\delta}$ (para $r > 1$), ao se definir o parâmetro de taxa como

$$\delta = (r-1)\frac{h^2}{2(1-h^2)MS_x}, \tag{4.5}$$

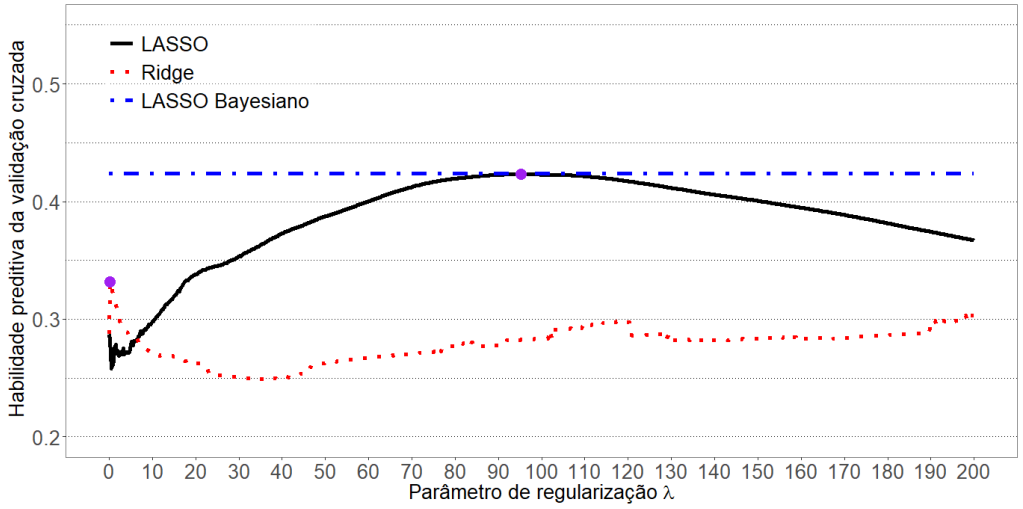
em que $MS_x = \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(x_ij-\bar{x_j})^2}{n}$, tem-se que a referida moda está bem definida para todo $r > 1$ e é igual a $2\frac{(1-h^2)}{h^2}MS_x$. Essa moda nada mais é do que um ponto alvo, definido a partir da informação a priori da herdabilidade. Com relação ao parâmetro de forma é interessante que este seja ligeiramente maior que um para garantir que a moda esteja bem definida e que a priori seja não informativa em torno dessa, já que o coeficiente de variação será alto (los Campos et al., 2013).

Seguindo as recomendações para a implementação do LASSO Bayesiano, definiu-se, tanto para o fenótipo DAP como para a Relação SG, $r = 1.01$ como parâmetro de forma da priori de λ^2, enquanto o parâmetro de taxa δ foi definido pela Equação (4.5), considerando-se h^2 igual a 0.432 para a DAP e igual a 0.845 para a Relação SG conforme os valores apresentados na Tabela 4.2. Do mesmo modo, definiu-se S (um dos parâmetros da priori de σ^2) pela Equação (4.3). Por fim, o parâmetro v, referente aos graus de liberdade da distribuição qui-quadrado inversa escalonada, foi definido como sendo igual a 5 independente do fenótipo em questão.

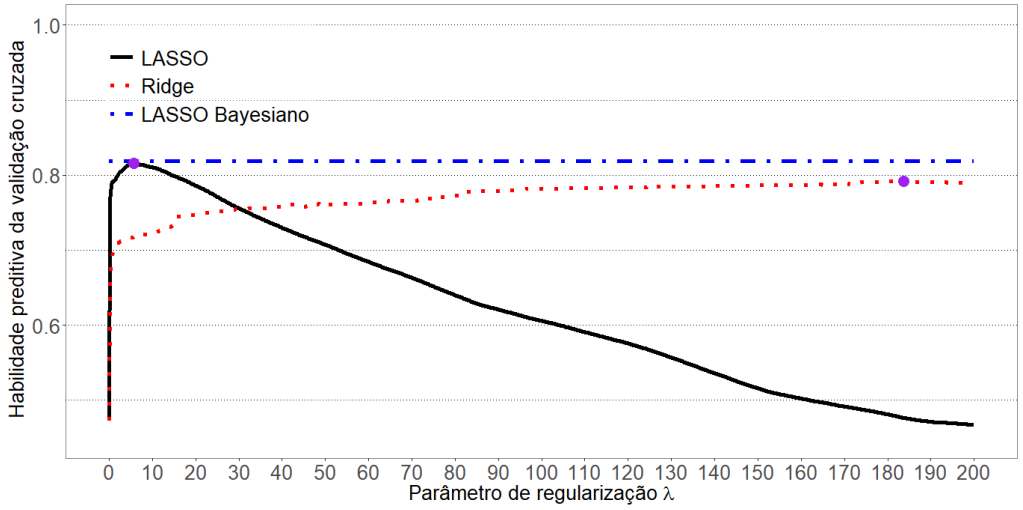
Dessa forma, o LASSO Bayesiano foi implementado e, com 100.000 iterações e um *thin* igual a 10, foram coletadas 10.000 amostras MCMC. Para a estimação das médias das distribuições a posteriori dos coeficientes, desconsiderou-se as 1.000 primeiras amostras geradas.

Para a avaliação da performance dos modelos em estudo, calculou-se a média das habilidades preditivas obtidas pelo procedimento 5-fold, que será chamada de habilidade preditiva da validação cruzada (HP_{cv}). Tanto para o método LASSO

quanto para o Ridge, a HP_{cv} foi calculada para cada λ pertencente a um *grid* que abrange valores de 0 a 200, variando 0.25 de elemento a elemento. Também levou-se em consideração as transformações de λ para λ_{glmnet} (*input* para a função *glmnet*) apresentadas na seção 3.4, que são necessárias para se ter equivalência com a formulação dos modelos apresentada pela Equação (3.4). No caso do LASSO Bayesiano, a validação cruzada foi aplicada apenas uma vez, obtendo-se uma única HP_{cv}, já que a calibração do λ é estabelecida automaticamente através da sua hiperpriori.



(a) `DAP`



(b) `Relação SG`

Figura 4.3: Sensibilidade da métrica de avaliação de desempenho segundo os valores de λ (curva de validação) para os métodos LASSO e Ridge.

As Figuras 4.3(a) e 4.3(b) apresentam as habilidades preditivas de validação cruzada para cada parâmetro do LASSO e Ridge, assim como apresenta a obtida pelo LASSO Bayesiano. Os pontos destacados sobre as curvas de validação do LASSO e Ridge representam as regiões de melhor desempenho destes métodos.

Ao se observar os pontos de melhor desempenho de cada método, percebe-se que os métodos que realizam seleção de variáveis (LASSO e LASSO Bayesiano) se destacaram em relação ao Ridge na tarefa de prever os GEBV's de ambos os fenótipos, sendo que, no caso do fenótipo de baixa herdabilidade, essa vantagem foi satisfatoriamente grande. Em ambos os casos o LASSO Bayesiano alcançou um desempenho similar ao obtido pelo LASSO com seu melhor λ, segundo a métrica HP_{cv}.

Na aplicação do modelo Ridge, com o fenótipo Relação SG, a respectiva curva de validação se aproxima da melhor HP_{cv} obtida pelo LASSO, assim como se aproxima da HP_{cv} obtida pelo LASSO Bayesiano. Todavia, a aproximação é assintótica e, sabendo-se que os modelos são completamente dominados pelo ruído quando a regularização é demasiadamente intensa, conclui-se que modelos com um regularização maior que as avaliadas não irão obter resultados muito melhores, esperando-se, inclusive, que o desempenho diminua a partir de um determinado ponto.

A Tabela 4.3 apresenta as melhores habilidades preditivas de validação cruzada juntamente com os respectivos parâmetros de regularização. No caso do LASSO Bayesiano, apresenta-se a única habilidade preditiva de validação cruzada calculada para comparação com as demais.

Fenótipo	LASSO	Ridge	LASSO Bayesiano
DAP (baixa herdabilidade)	0.423 ($\lambda = 95.25$)	0.331 ($\lambda = 0.25$)	0.423
Relação SG (alta herdabilidade)	0.815 ($\lambda = 5.75$)	0.791 ($\lambda = 183.75$)	0.818

Tabela 4.3: Habilidades preditivas de validação cruzada.

Percebe-se que o LASSO Bayesiano, com os hiperparâmetros definidos segundo as recomendações de los Campos et al. (2013), mostrou-se um método extremamente competitivo na aplicação.

Para uma análise direcionada aos coeficientes gerados por cada modelo, os

três métodos de regularização foram aplicados levando-se em consideração todas as plantas do banco de dados. No caso do LASSO e do Ridge, foram utilizados os parâmetros de regularização presentes na Tabela 4.3, enquanto que para o LASSO Bayesiano apenas aplicou-se o método com os mesmos hiperparâmetros em toda a base. A Figura 4.4 apresenta os gráficos que informam a magnitude dos coeficientes gerados em cada método.

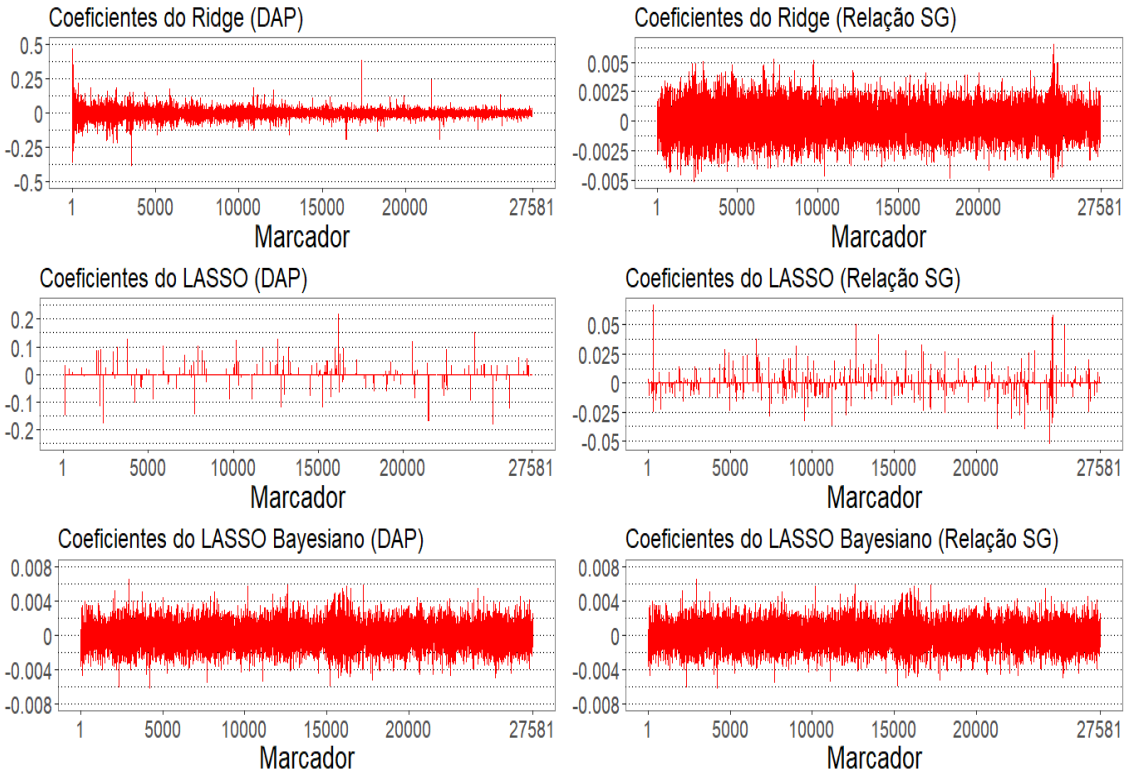


Figura 4.4: Valores dos coeficientes de todos os 27.581 marcadores SNP.

A informação mais impactante apresentada na Figura 4.4 é a esparsidade existente nos gráficos dos coeficientes gerados pelo LASSO. Devido à sua propriedade de seleção de variáveis, existem apenas 164 coeficientes diferentes de zero quando o fenótipo em questão é a DAP e 484 quando o fenótipo em questão é a Relação SG.

Embora o Ridge se caracterize por manter os valores dos coeficientes em um mesmo patamar, o primeiro gráfico mostra que o método permitiu que alguns poucos coeficientes fossem significantemente maiores que a maioria. O mesmo não acontece para a aplicação do Ridge com a Relação SG.

Como para o cálculo dos coeficientes do LASSO Bayesiano utilizou-se a média a posteriori, e não a moda, a seleção de variáveis passa a não ser direta, justificando a ausência de esparsidade. Diferentemente do esperado, o LASSO Bayesiano não permitiu, nesta aplicação, que alguns coeficientes se destacassem em relação aos demais, ou seja, que alguns coeficientes sofressem uma regularização menor do que a maioria.

Capítulo 5

Conclusão

A obtenção de soluções regularizadas para um determinado sistema linear pode ser dada de diferentes maneiras, podendo ser viabilizada através de procedimentos que variam desde a simples exclusão dos valores singulares substancialmente pequenos, provenientes da matriz $\boldsymbol{X}$, até procedimentos mais complexos envolvendo otimizações numéricas ou distribuições de probabilidades. Tais métodos são úteis para o desenvolvimento de modelos com uma boa qualidade preditiva uma vez que permitem a obtenção de coeficientes que, embora levemente enviesados, possuem uma variância menor em relação à obtida pelo modelo linear sem regularização. O *trade-off* entre viés e variância é estabelecido por um parâmetro que regula a intensidade da regularização e saber escolher o melhor valor para esse parâmetro é um passo importante da modelagem.

Dentre as medidas que podem ser utilizadas para se avaliar o desempenho do modelo, o Erro Quadrático Médio mostra-se bastante prático já que possui um comportamento conhecido ao se variar o parâmetro de regularização λ. Todavia, a escolha da medida vai depender dos interesses práticos em questão.

Com o estudo dos modelos de regularização, percebe-se que o método novo e emergente chamado LASSO Bayesiano possui algumas vantagens em relação aos demais aqui avaliados. Embora tenha um custo computacional maior, o método traz a opção de se utilizar uma priori para λ, permitindo uma automatização da calibração deste parâmetro, além de fornecer intervalos de credibilidade. Diferentemente

do LASSO Bayesiano, os demais métodos estudados não fornecem nenhum tipo de estimativa intervalar, sendo preciso recorrer a métodos de reamostragem como o *bootstrap* caso se deseje uma.

Na aplicação realizada em dados genéticos, três dentre os diversos métodos de regularização foram avaliados e aqueles que permitem efeitos maiores apenas para alguns poucos marcadores (modelos que induzem a seleção de variáveis) obtiveram melhor performance. O LASSO, especificamente, mostrou-se muito efetivo em todos as três bases de dados estudadas, mesmo sendo extremamente simples.

Os resultados da aplicação em Seleção Genômica, em conformidade com outros estudos realizados em outras bases de dados da área, evidenciaram a associação existente entre a herdabilidade e a habilidade preditiva, mostrando que a Seleção Genômica é uma boa alternativa para o melhoramento genético quando fenótipos de alta herdabilidade estão em questão.

As linhas de pesquisa envolvendo Seleção Genômica e regularização são diversas e ainda há muito a ser explorado. Os métodos bayesianos são os mais citados em publicações sobre Seleção Genômica. A escolha da priori dos coeficientes para estes modelos é ponto chave para definir se será realizada seleção de variáveis ou apenas regularização. Além das versões bayesianas do LASSO e Ridge, profissionais que lidam com métodos para Seleção Genômica frequentemente recorrem a outras versões bayesianas para regularização como os métodos Bayes A, Bayes B, Bayes C, dentre outros. A principal diferença entre esses métodos encontra-se nas definições das prioris dos coeficientes e de σ^2, que podem ter maior ou menor massa na vizinhança do zero e caudas mais ou menos pesadas.

Apêndice A

Decomposição SVD

A obtenção de uma solução para o problema inverso $\boldsymbol{X\beta} = \boldsymbol{y}$, assim como o estudo das propriedades dessa solução gerada, podem ser desenvolvidos utilizando a decomposição em valores singulares (SVD), que é baseada na fatoração da matriz $\boldsymbol{X}$ tal que

$$\boldsymbol{X} = \boldsymbol{USV'}, \tag{A.1}$$

em que $\boldsymbol{U_{n\times n}}$ e $\boldsymbol{V_{p\times p}}$ são matrizes ortogonais e $\boldsymbol{S_{n\times p}}$ é uma matriz diagonal cujos elementos são chamados de valores singulares.

Os valores singulares contidos em $\boldsymbol{S}$ correspondem à raiz quadrada dos autovalores de $\boldsymbol{X}$. Como o posto de $\boldsymbol{X}$ — aqui codificado por r — é equivalente ao número de autovalores não nulos e como o posto máximo que $\boldsymbol{X}$ pode ter é o mínimo entre o número de linhas e colunas, a matriz $\boldsymbol{S}$ tem no máximo $\min(n, p)$ valores singulares não nulos, sendo os demais necessariamente iguais a zero. Se os valores singulares forem organizados tais que $s_1 \geq s_2 \geq \cdots \geq s_{\min(n,p)} \geq 0$ e $s_{\min(n,p)+1} = s_{\min(n,p)+2} = \cdots = s_{\max(n,p)} = 0$, considerando que apenas os r primeiros valores singulares são estritamente maiores do que zero, então a matriz $\boldsymbol{S}$ passa a ter o seguinte aspecto:

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S_r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{A.2}$$

em que $\boldsymbol{S_r}$ é uma matriz quadrada r por r. Assim,

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{U_r} & \boldsymbol{U_0} \end{pmatrix} \begin{pmatrix} \boldsymbol{S_r} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{V_r} & \boldsymbol{V_0} \end{pmatrix}' = \boldsymbol{U_r S_r V_r'}, \tag{A.3}$$

em que $\boldsymbol{U_r}$ e $\boldsymbol{V_r}$ denotam as primeiras r colunas de $\boldsymbol{U}$ e $\boldsymbol{V}$ respectivamente. Portanto, a decomposição de $\boldsymbol{X}$ em $\boldsymbol{U_r S_r V_r'}$ é uma versão compacta da decomposição SVD haja vista a anulação de $\boldsymbol{U_0}$ e $\boldsymbol{V_0}$ pelos elementos nulos de $\boldsymbol{S}$.

Apêndice B

Desigualdade Triangular

Seja $\boldsymbol{\beta_{mq}} = \boldsymbol{\beta_+} + \boldsymbol{\beta_{Null}}$, tem-se que

$$\| \boldsymbol{\beta_{mq}} \|_2 = \sqrt{\langle \boldsymbol{\beta_+} + \boldsymbol{\beta_{Null}}, \boldsymbol{\beta_+} + \boldsymbol{\beta_{Null}} \rangle} \tag{B.1}$$

$$= \sqrt{\langle \boldsymbol{\beta_+}, \boldsymbol{\beta_+} \rangle + 2 \langle \boldsymbol{\beta_+}, \boldsymbol{\beta_{Null}} \rangle + \langle \boldsymbol{\beta_{Null}}, \boldsymbol{\beta_{Null}} \rangle} \tag{B.2}$$

$$= \| \boldsymbol{\beta_+} \|_2 + \| \boldsymbol{\beta_{Null}} \|_2 + \sqrt{2 \langle \boldsymbol{\beta_+}, \boldsymbol{\beta_{Null}} \rangle}. \tag{B.3}$$

Então, para provar que $\| \boldsymbol{\beta_{mq}} \|_2 = \| \boldsymbol{\beta_+} \|_2 + \| \boldsymbol{\beta_{Null}} \|_2$, basta provar que $\boldsymbol{\beta_+}$ e $\boldsymbol{\beta_{Null}}$ são ortogonais, ou seja,

$$\langle \boldsymbol{\beta_+}, \boldsymbol{\beta_{Null}} \rangle = \left\langle \boldsymbol{V_r S_r^{-1} U_r' y}, \sum_{i=r+1}^{p} \alpha_i \boldsymbol{V_{.i}} \right\rangle = \sum_{i=r+1}^{p} \alpha_i \left\langle \boldsymbol{V_r S_r^{-1} U_r' y}, \boldsymbol{V_{.i}} \right\rangle = 0. \tag{B.4}$$

As soluções $\boldsymbol{\beta_+}$ e $\boldsymbol{\beta_{Null}}$ só irão ser ortogonais se $\left\langle \boldsymbol{V_r S_r^{-1} U_r' y}, \boldsymbol{V_{.i}} \right\rangle$ for igual a zero para todo i.

Levando em consideração que $\boldsymbol{V_r S_r^{-1} U_r' y} = \sum_{j=1}^{r} \frac{\boldsymbol{U_{.j}' y}}{\boldsymbol{s_j}} \boldsymbol{V_{.j}}$, tem-se

$$\left\langle \boldsymbol{V_r S_r^{-1} U_r' y}, \boldsymbol{V_{.i}} \right\rangle = \left\langle \sum_{j=1}^{r} \frac{\boldsymbol{U_{.j}' y}}{\boldsymbol{s_j}} \boldsymbol{V_{.j}}, \boldsymbol{V_{.i}} \right\rangle = \sum_{j=1}^{r} \frac{\boldsymbol{U_{.j}' y}}{\boldsymbol{s_j}} \left\langle \boldsymbol{V_{.j}}, \boldsymbol{V_{.i}} \right\rangle. \tag{B.5}$$

Como $j = 1, 2, \ldots, r$, $i = r+1, r+2, \ldots, p$ e $\boldsymbol{V}$ possui colunas ortogonais, o

produto interno $\langle \boldsymbol{V_{.j}}, \boldsymbol{V_{.i}} \rangle$ será nulo para todas as configurações possíveis de i e j, fazendo com que $\left\langle \boldsymbol{V_r S_r^{-1} U_r' y}, \boldsymbol{V_{.i}} \right\rangle$ seja nulo para todo i.

O mesmo argumento é utilizado para provar que $\| \boldsymbol{\beta_{mq}} \|_2^2 = \| \boldsymbol{\beta_+} \|_2^2 + \| \boldsymbol{\beta_{Null}} \|_2^2$, sendo que a única diferença na demostração é a ausência da raiz quadrada na Equação (B.1).

Apêndice C

Variância do Estimador Ridge

Relação entre a variância do estimador de mínimos quadrados e a variância do estimador Ridge.

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) - \mathrm{Var}(\hat{\boldsymbol{\beta}}_{\boldsymbol{2}}) = \sigma^2 \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} - \boldsymbol{W}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{W}' \right]$$

$$= \sigma^2 \left[\boldsymbol{W}\boldsymbol{W}^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{W}^{-1})'\boldsymbol{W}' - \boldsymbol{W}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{W}' \right]$$

$$= \sigma^2 \boldsymbol{W} \left[\boldsymbol{W}^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{W}^{-1})' - (\boldsymbol{X}'\boldsymbol{X})^{-1} \right] \boldsymbol{W}'$$

$$= \sigma^2 \boldsymbol{W} \left\{ \left[\boldsymbol{I} + \lambda(\boldsymbol{X}'\boldsymbol{X})^{-1} \right] (\boldsymbol{X}'\boldsymbol{X})^{-1} \left[\boldsymbol{I} + \lambda(\boldsymbol{X}'\boldsymbol{X})^{-1} \right]' - (\boldsymbol{X}'\boldsymbol{X})^{-1} \right\} \boldsymbol{W}'$$

e, como $(\boldsymbol{I} + \lambda(\boldsymbol{X}'\boldsymbol{X})^{-1})$ é simétrica,

$$\sigma^2 \boldsymbol{W} \left[(\boldsymbol{I} + \lambda(\boldsymbol{X}'\boldsymbol{X})^{-1})(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{I} + \lambda(\boldsymbol{X}'\boldsymbol{X})^{-1}) - (\boldsymbol{X}'\boldsymbol{X})^{-1} \right] \boldsymbol{W}'$$

$$= \sigma^2 \boldsymbol{W} \left\{ (\boldsymbol{X}'\boldsymbol{X})^{-1} + 2\lambda \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} \right]^2 + \lambda^2 \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} \right]^3 - (\boldsymbol{X}'\boldsymbol{X})^{-1} \right\} \boldsymbol{W}'$$

$$= \sigma^2 \boldsymbol{W} \left\{ 2\lambda \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} \right]^2 + \lambda^2 \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} \right]^3 \right\} \boldsymbol{W}'$$

$$= \sigma^2 \left[\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I} \right]^{-1} \boldsymbol{X}'\boldsymbol{X} \left\{ 2\lambda \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} \right]^2 + \lambda^2 \left[(\boldsymbol{X}'\boldsymbol{X})^{-1} \right]^3 \right\} \left[(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X} \right]'$$

$$= \sigma^2 \left[\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I} \right]^{-1} \left[2\lambda\boldsymbol{I} + \lambda^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \right] \left[(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1} \right]'. \qquad \text{(C.1)}$$

Referências Bibliográficas

Andrews, D. F. & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society.*, 36:99–102. 55

Aster, R. C., Borchers, B., & Thurber, C. H. (2011). *Parameter Estimation and Inverse Problems*, volume 90. Academic Press. 28, 31, 36

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with r package rrblup. *Plant Genome*, 4:250–255. 76

Farebrother, R. W. (1976). Further results on the mean square error of ridge regression. *Journal of the Royal Statistical Society.*, 38:248–250. 50

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22. 61

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., & Fernando, R. (2009). Additive genetic variability and the bayesian alphabet. *Genetics Society of America*, 183:347–363. 76

Gramacy, R. B. (2018). *Estimation for Multivariate Normal and Student-t Data with Monotone Missingness*. CRAN. R package version 1.9. 61

Grattapaglia, D. (2014). Breeding forest trees by genomic selection: current progress and the way forward. in "genomics of plant genetic resources vol 1 pp 651-682. eds r. tuberosa, a. graner & e. frison. Technical report, DOI 10.1007/978-94-007-7572-5_26, Springer Science+ Business Media Dordrecht. 69, 71

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer New York. 36, 46, 48, 52

Henderson, C. (1984). *Applications of Linear Models in Animal Breeding.* University of Guelph. 70

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67. 50

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R.* Springer Texts in Statistics. Springer New York.

Kuhn, M. & Johnson, K. (2013). *Applied Predictive Modeling.* Springer. 36

Lavrentiev, M. M. (1967). *Some Improperly Posed Problems of Mathematical Physics*, volume 11. Springer.

Lima, B. M. (2014). *Bridging genomics and quantitative genetics of Eucalyptus: genome-wide prediction and genetic parameter estimation for growth and wood properties using high-density SNP data.* PhD thesis, Escola Superior de Agricultura "Luiz de Queiroz". 72, 74

Lin, Z., Hayes, B. J., & Daetwyler, H. D. (2014). Genomic selection in crops, trees and forages: a review. *Crop & Pasture Science.* 69, 74, 75, 76

los Campos, G., Pérez, P., Vazquez, A. I., & Crossa, J. (2013). Genome-enabled prediction using the blr (bayesian linear regression) r-package. In: *Genome-Wide Association Studies and Genomic Prediction*, C. Gondro, J. van der Werf, & B. Hayes, ed., chapter 12, pages 299–320. Springer Science+Business Media. 76, 77, 78, 79, 81

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics Society of America.* 68, 70

Park, T. & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103:681–686. 45, 52, 53, 55, 58, 59, 60, 61, 62, 78

Resende, M. D. V., Lopes, P. S., da Silva, R. L., & Pires, I. E. (2008). Seleção genômica ampla (gws) e maximização da eficiência do melhoramento genético. *Pesquisa Florestal Brasileira*, 56:63–77. 68

Souto, G. (2000). Decomposição em valores singulares. Trabalho de conclusão de curso, Universidade Federal de Santa Catarina.

Souza, G. S. (1998). *Introdução aos Modelos de Regressão Linear e Não-Linear.* Embrapa. 21

Theobald, C. M. (1974). Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society.*, 36:103–106. 50