Universidade de São Paulo

Estudo Empírico de Modelos de Regressão para Dados de Alta Dimensão

Bolsista:
João Victor Alcantara Pimenta
nº 11820812

Orientador:
Mário de Castro

Número do Processo:
105938/2020-3

São Carlos
Período: 01/04/2020 a 01/08/2021

Conteúdo

1 Introdução **2**
1.1 Atividades Desenvolvidas 2

2 Detalhamento do conteúdo estudado **5**
2.1 Replicando os resultados do Artigo 5
2.2 Ranqueamento 7
2.3 Predição 9
2.4 Classificação 11
2.5 Robustez dos resultados 15

1 Introdução

O projeto revolve em torno do artigo publicado por *Wang, F., Mukherjee, S., Richardson, S., Hill, S. M. (2019)* [1] que trata de métodos de regressão para dados de alta dimensão, buscando o escopo de validade e performance de cada um em determinadas condições. Métodos 'de penalidade' são amplamente usados para regressões em dados de alta dimensão (quando o número de 'variáveis' é maior que o número de 'observações'), e, apesar de sua teoria estar sendo estabelecida, sua aplicação em situações com dados finitos, como é de praxe em situações reais, não é bem fundada.

O objetivo do artigo é percorrer diferentes cenários e configurações destes dados finitos, testando empiricamente, para cada situação, o desempenho dos diferentes método selecionados (Lasso, Adaptive Lasso, Elastic Net, Ridge Regression, SCAD, the Dantzig Selector and Stability Selection). Isso é feito com uma implementação em R de cada um dos métodos e rodando uma simulação nos dados gerados (dados independentes e semi correlacionados) 64 vezes, da qual é tirada a média, a qual consideramos o resultado do método naquela condição. O artigo vai ainda tratar de três parâmetros, Ranqueamento (medido por pAUC), Classificação (medido por TPR, PPV, TPR e PPV) e Predição (medido por RMSE). Em todos, varia-se algumas condições, dentre elas: p (dimensionalidade), n (tamanho amostral), s_0 (esparcidade), SNR (razão sinal-barulho). Para os *'Synthetic (pairwise) correlation design'* varia-se ainda p^b (tamanho do bloco de correlação), p (correlação em bloco), ${s^b}_0$ (sinais por bloco). Finalmente para *'Semisynthetic ("low"/"high") correlation designs'* varia-se também p^b (tamanho do bloco de correlação) e ${s^b}_0$ (sinais por bloco).

O objetivo deste projeto de pesquisa é portando compreender os métodos empregados, junto com o algoritmo e resultados. Podendo assim replicá-los e validar os achados do artigo. Em suma, o artigo encontra uma alternância na superioridade dos métodos. Com diferentes métodos performando melhor para diferentes condições e métricas. Sem nenhum 'vencedor' absoluto. Os resultados encontrados ao replicar os métodos foram de extrema concordância aos observados na dissertação dos autores. Uma nota de importância é que, assim como os autores, não considerou-se o método Dantzig depois de perceber que seu desempenho é em geral, sempre inferior a outros métodos, e seu tempo de execução incrivelmente alto em comparação ao tempo médio dos outros métodos. Além disso foi observado um bom desempenho geral, não sempre superior, mas consistente, do método Lasso.

Além disso, buscou-se uma extensão para as ideias discutidas pelos autores. Simulando para isso, cenários onde a robustez dos resultados não foi testada originalmente. Mais especificamente para diferentes graus de liberdade na distribuição dos erros. Para o estudo deste, primeiro, foi necessário, fundamentar conhecimentos de estatística e, mais especificamente, de modelos de regressão.

1.1 Atividades Desenvolvidas

Neste primeiro momento, anterior ao artigo, foram desenvolvidas as seguintes atividades:

- Estudo direcionado pelo livro *Introduction to Statistical Thought* [2]

- Estudo inicial e introdutório à linguagem R, a partir de atividades ligadas aos conceitos de estatística discutidos no Livro;
- Estudo ligado aos capítulos e temas:
  * 1.1 - 1.8 (Probabilities) → Basic Probabilities; Probability Density; Parametric Families of Distributions (Binomial, Poisson, Exponential and Normal); Centers, Spreads, Means and Moments; Joint, Marginal and Conditional Probability; Association, Dependence, Independence; Simulation (Calculating Probabilities and Evaluating Statistical Procedures); R; Results for Large Samples.
  * 2.1 - 2.4 (Modes of Inference) → Data; Data description (Summary statistics, Displaying Distributions and Exploring Relationships); Likelihood (The likelihood function, Central Limit Theorem and Likelihood for several Parameters); Estimation (The Maximum Likelihood Estimate, Accuracy of Estimation and The Sampling Distribution of an Estimator.
  * 3.1 - 3.2 (Regression) →Introduction to concept; Normal Linear Models (Inference for Linear Models).

- Estudo Dirigido pelo livro *An Introduction to Statistical Learning* [3]
  - Aprofundamento na Linguagem em R em conceitos ligados à estatística.
    * 2 (Statistical Learning) → 2.1 - What is Statistical Learning (Why Estimate f, How Do We Estimate f, The trade-Off Between Prediction Accuracy, Supervised Versus Unsupervised Learning, Regression Versus Classification Problems); 2.2 - Assessing Model Accuracy (Measuring the Quality of the Fit, The Bias-Variance Trade-Off, The Classification Setting); 2.3 - Lab: Introduction to R; 2.4 - Exercises.
    * 3 (Linear Regression) → 3.1 Simple Linear Regression (Estimating the coefficients, Assessing the Accuracy of the coefficient Estimates, Assessing the Accuracy of the Model); 3.2 Multiple Linear Regression (Estimating the Regression Coefficients); 3.3 Other (Qualitative Predictors, Extensions of the Linear Model, Potential Problems); 3.5 Comparison of the Linear Regression with the K-Nearest Neighbors; 3.6 Lab: Linear Regression. 3.7 Exercises.
    * 4 (Classification) → 4.1 An overview of Classification; 4.2 Why Not Linear Regression?; 4.3 Logistic Regression (The Logistic Model, Estimating the Regression Coefficients, Making Predictions, Multiple Logistic Regression, Logistic Regression for ¿ 2 Response Class); 4.4 Linear Discriminant Analysis (Using Bayes' Theorem for Classification, LDA for p = 1, LDA for p ¿ 1, Quadratic Discriminant Analysis); 4.5 A Comparison of Classification Methods; 4.6 Lab: Logistic Regression; 4.7 Exercises.
    * 5 (Resampling Models) → 5.1 Cross-Validation (The Validation Set Approach, Leave-One-Out Cross-Validation, k-Fold Cross-Validation, Bias-Variance Trade-off for k-Fold Cross-Validation); 5.2 The Bootstrap; 5.3 Lab: Cross-Validation and the Bootstrap; 5.4 Exercises.
    * 6 (Linear Model Selection and Regularization) → 6.1 Subset Selection (Best Subset Selection, Stepwise Selection, Choosing the Optimal Model);

6.2 Shrinkage Methods (Ridge Regression, The Lasso, Selecting the Tuning Parameter); 6.3 Dimension Reduction Methods (Principal Components Regression, Partial Least Squares); 6.4 Considerations in High Dimensions (High-Dimensional Data, What goes wrong in High Dimensions?, Regression in High Dimensions, Interpreting Results in High Dimensions); 6.5 Lab 1: Subset Selection Methods; 6.6 Lab 2: Ridge Regression and Lasso; 6.7 Lab 3: PCR and PLS Regression; 6.8: Exercises.

- Estudo por Cursos Disponibilizados Pela USP no Coursera:
  - **(1)** *"Statistics with R"* Uma especialização composta de 4 cursos de estatística fornecido pela Duke University. Cada um deles culmina em um projeto de conclusão prático sobre o tema. São eles e seus respectivos conteúdos:
  - Introduction to Probability and Data with R
    - Introdução à dados e manipulação em R assim como probabilidade básica;
    - EDA e introdução à inferência;
    - Mais sobre probabilidades e probabilidades condicionais, Introdução ao Teorema de Bayes;
    - Distribuições Binomial e Normal.
  - Inferential Statistics
    - CLT e intervalos de confiança;
    - Inferência e Significância;
    - Inferência para comparar médias;
    - Inferência para proporções.
  - Linear Regression and Modeling
    - Regressões lineares;
    - Outliers, inferência em Regressão Linear;
    - Regressão Múltipla (numérica e categórica)
  - Bayesian Statistics
    - Conceitos relacionados à base teórica da Teoria de Bayes; - Inferência Bayesiana;
    - Tomada de decisão e teste de hipótese Bayesiana;
    - Regressões Bayesianas e model averaging.
  - **(2)** *"Data Science: Foundations Using R"* : Especialização da John Hopkins University em Data Science. Composto dos cursos:
  - The data Science Toolbox
    - Tópicos de introdução e Git, R, Rstudio e Big Data;
    - Noções gerais de Data Science.
  - R Programming
    - História e noções em R;
    - Funções e controle de estruturas em R;
    - Funções Loop e Debugging;

    - Simulações em R e Profiling.
- Getting and Cleaning Data
    - Achar e ler diferentes formatos de Data;
    - Formas de armazenamento de Data e ferramentas de extração;
    - Noções e funções em organização de Data;
- **(3)** *"Grafing with GGplot2"* Pequeno curso de produtor independente nas noções de criação de gráfico com a biblioteca do R: GGplot2

Desenvolvidas estas atividades iniciais, partiu-se para o entendimento do artigo. Lido e bem estabelecido, passou-se para o código que suplementa o artigo, onde parte da implementação dos métodos é feita. Importante notar que no repositório dos autores o código não é funcional, apesar de ter fornecido a base para o desenvolvimento do código local, além de ter funções já feitas de visualização dos dados que muito foram úteis. Foi necessário adaptar e complementar o código para poder fazer a reprodução dos resultados, que por sua vez, foram de alta concordância com os do artigo, como será indicado. Replicado os resultados, estendeu-se o escopo do artigo procurando por robustez em outras configurações dos erros dos dados. Implementou-se o erro com diferentes graus de liberdade da distribuição e observou-se razoável robustez nos principais métodos.

2 Detalhamento do conteúdo estudado

2.1 Replicando os resultados do Artigo

Desenvolvido o algoritmo e implementado os métodos de regressão tratados no artigo, pôde-se gerar imagens comparáveis às do artigo, com outros dados, locais, para replicar seus resultados. O desempenho dos métodos em cada um dos cenários propostos sugerido pelo artigo e sua versão local é a seguinte:

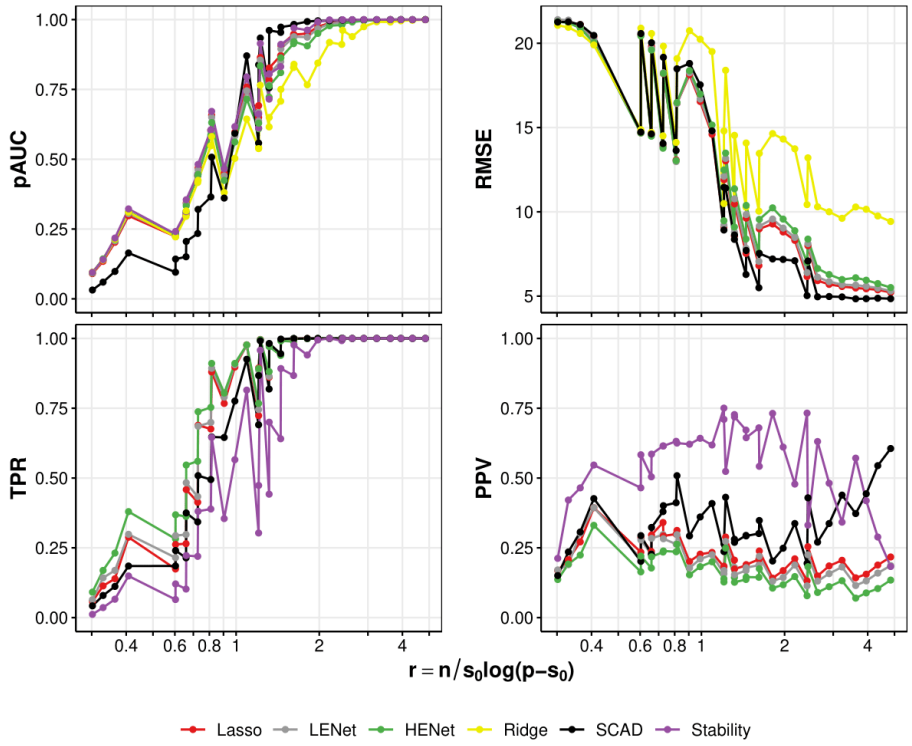


Figura 1: Versão do Artigo de cada um dos métodos para diferentes parâmetros e condições de dados. Visão geral.

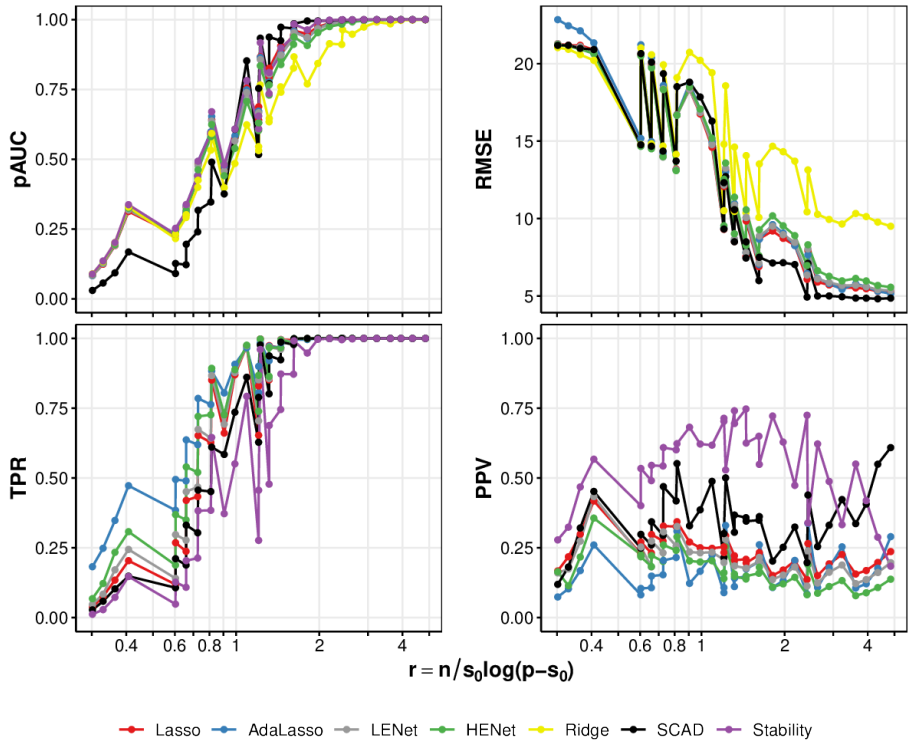
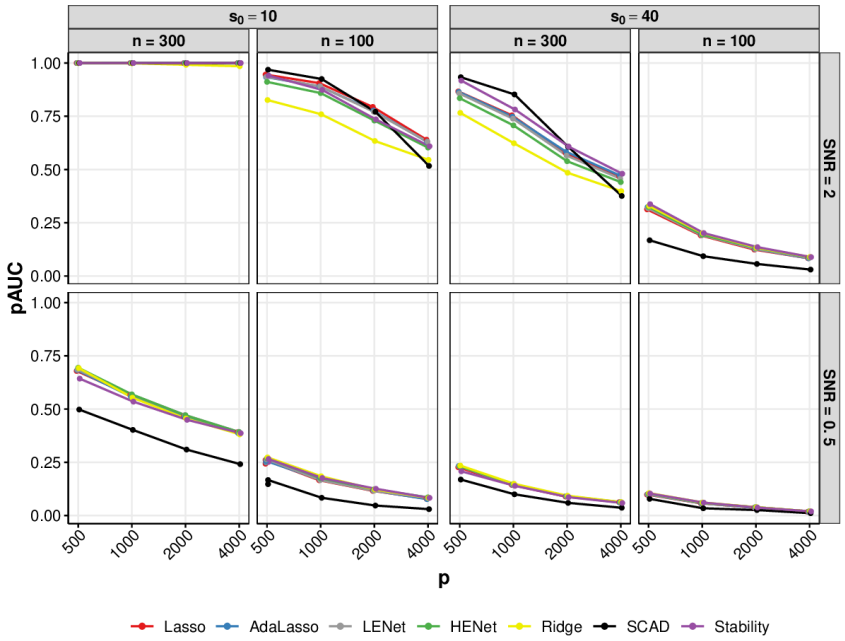


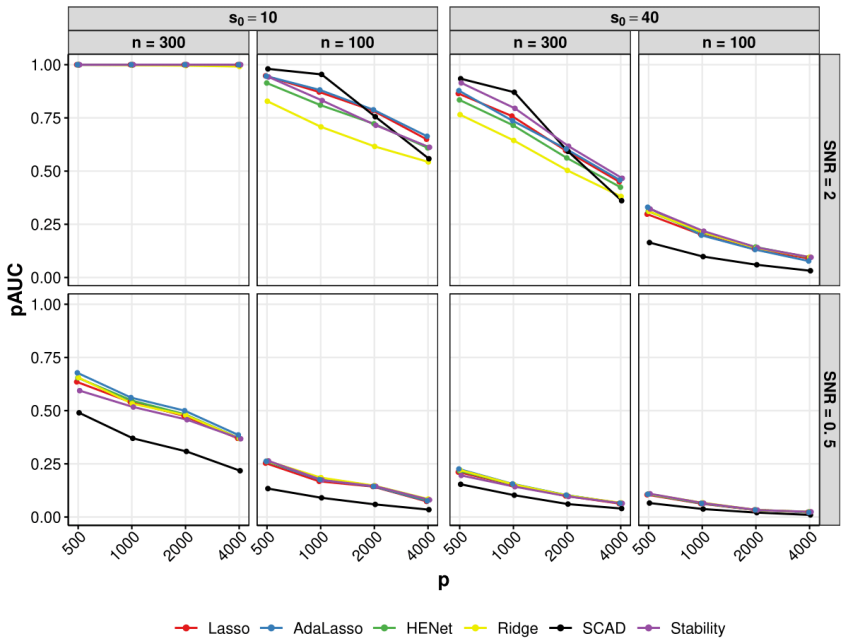
Figura 2: Versão Local de cada um dos métodos para diferentes parâmetros e condições de dados. Visão geral.

Que dá uma boa ideia de que os resultados foram devidamente processados, ou pelo menos que são possíveis de serem replicados de forma coerente em diferentes máquinas. Também é observável o comportamento geral dos métodos, apesar de que cada um será explorado nos outros resultados obtidos em:

2.2 Ranqueamento

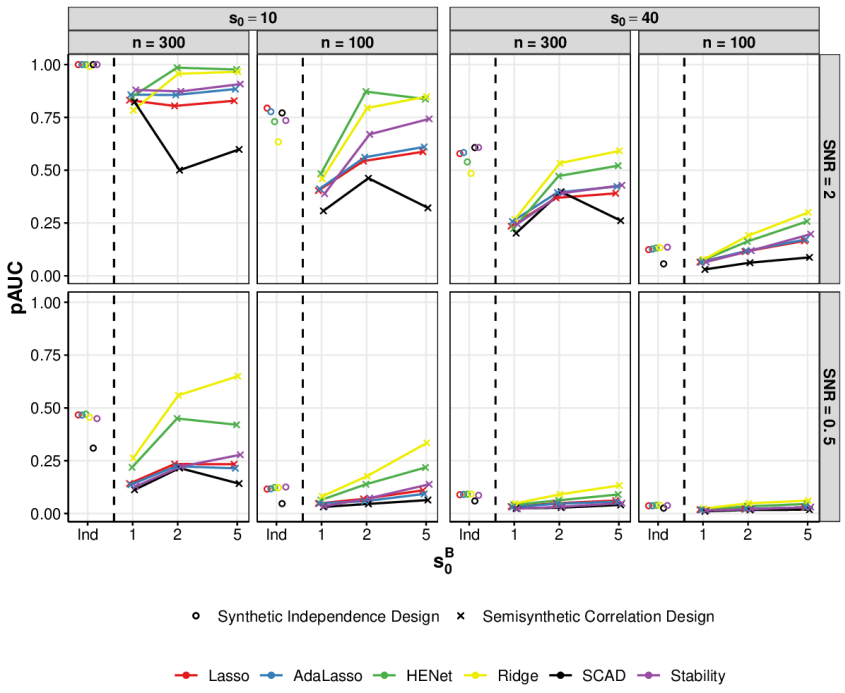


(a)

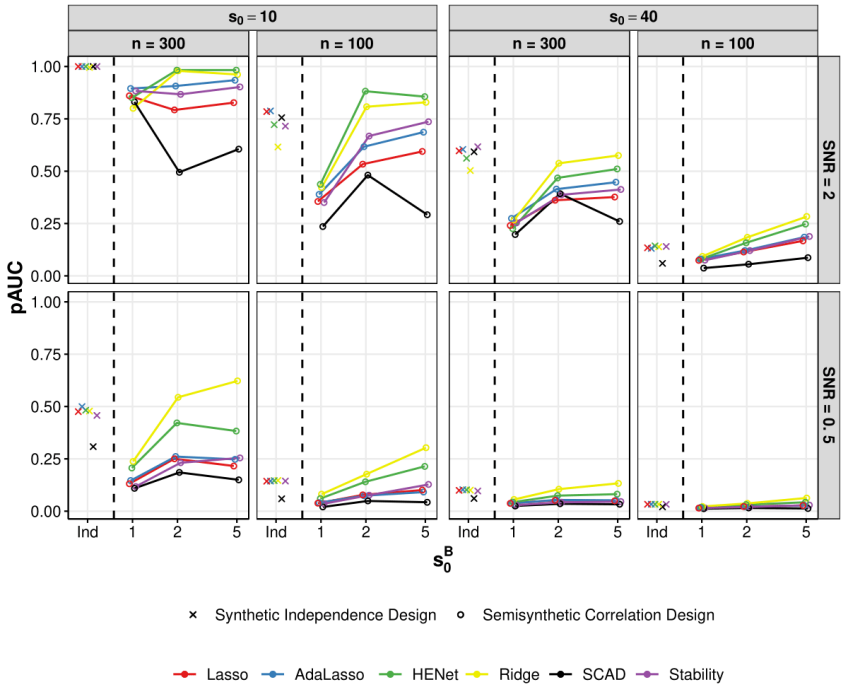


(b)

Figura 3: Versões (a) Local e (b) Artigo de pAUC em dados independentes para um set de configuração apontado



(a)



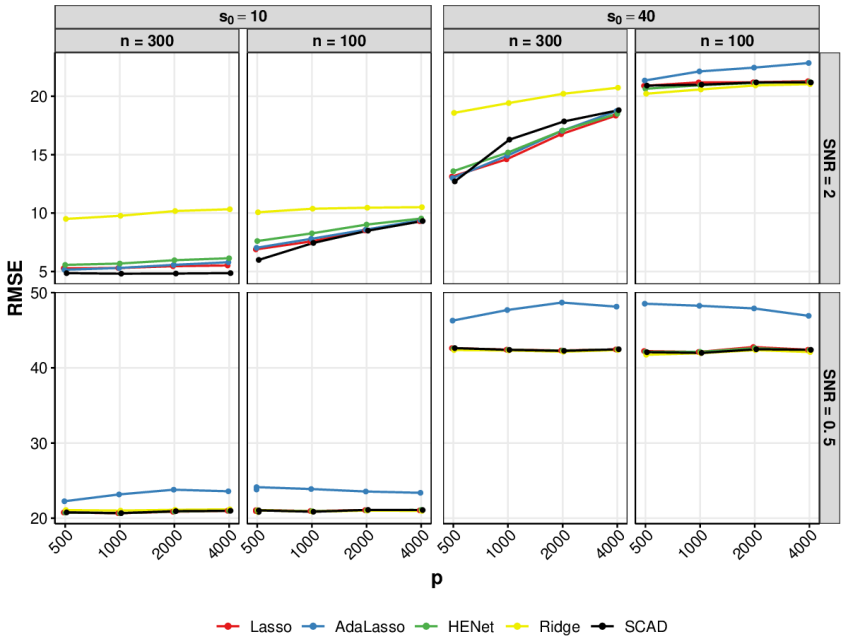
(b)

Figura 4: Versões (a) Local e (b) Artigo de pAUC em dados semi-sintéticos para um set de configuração apontado

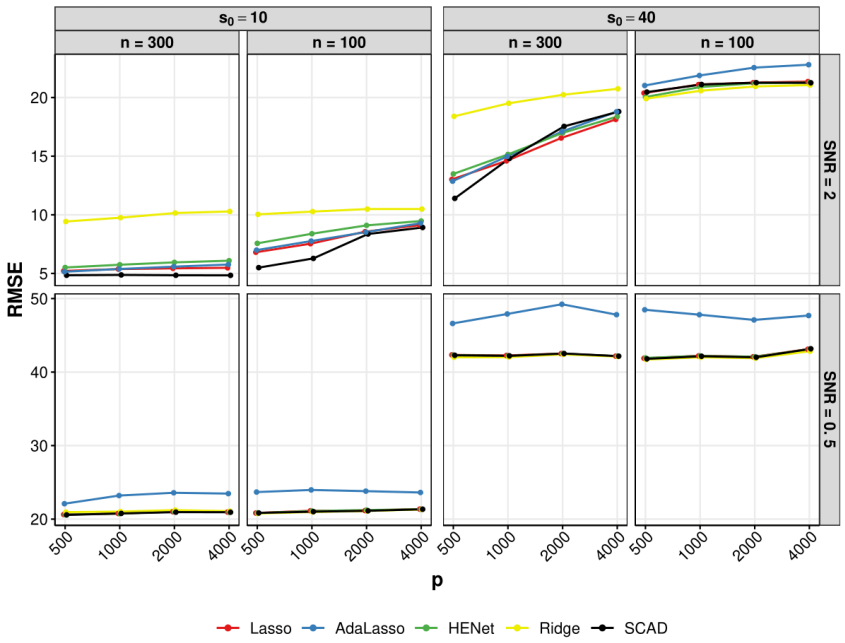
Assim como os autores, observa-se nesse parâmetro uma superioridade do Lasso e Ada-Lasso em contextos de não correlação ou baixa correlação entre os dados. E em contextos de alta correlação, uma superioridade do método Ridge. O método Scad, apesar de ser superior em instâncias específicas, cenários 'fáceis' (grande r e SNR), sofre uma transição de fase, como os autores chamam atenção, e deve ser usado para obter superioridade

somente se houver certeza das condições serem favoráveis.

2.3 Predição

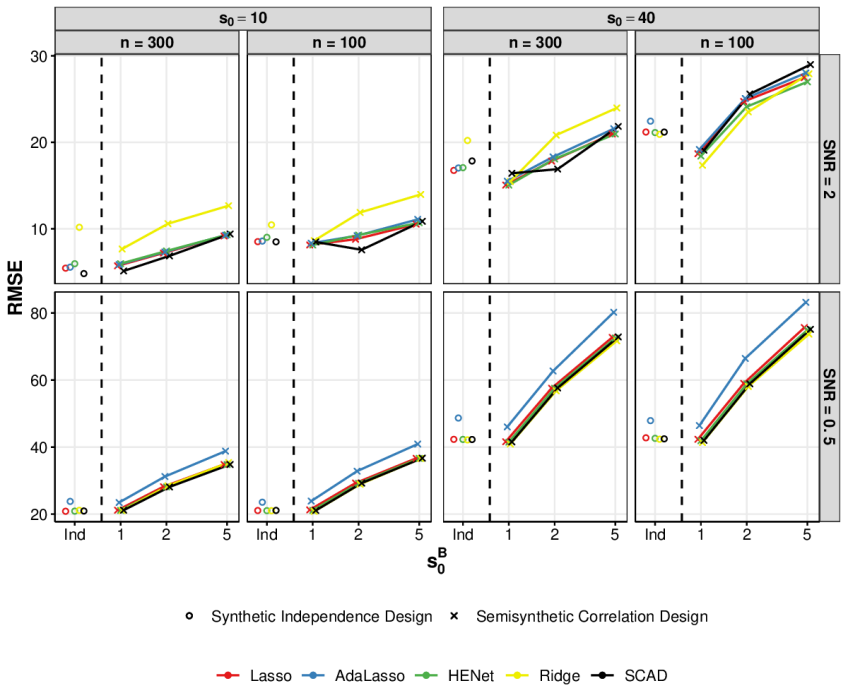


(a)

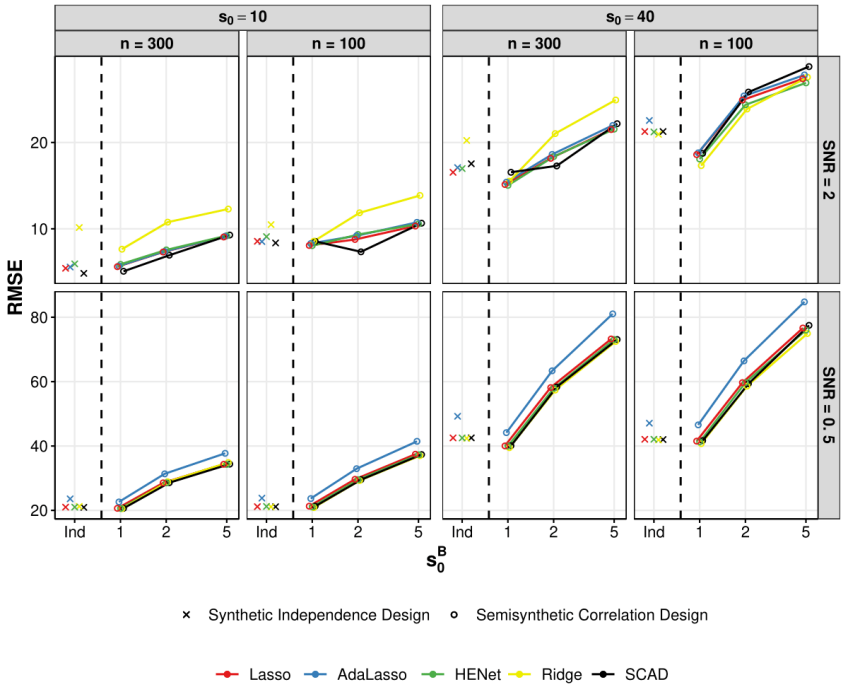


(b)

Figura 5: Versões (a) Local e (b) Artigo de RMSE em dados independentes para um set de configuração apontado



(a)



(b)

Figura 6: Versões (a) Local e (b) Artigo de RMSE em dados semi-sintéticos para um set de configuração apontado

Neste parâmetro, Lasso se mostra ainda uma boa opção para ambos cenários fracamente e fortemente correlacionados. No caso de usar o SCAD, é novamente necessário ter certeza das condições serem ideais, já que existe sua 'transição de fase'. Adalasso não é superior dessa vez, como foi na última avaliação. Além disso Ridge pode oferecer pequenos ganhos, somente em cenários difíceis. Diferente do HEnet, que é mais consistente em

cenários difíceis e pode oferecer pequenos ganhos em relação ao Lasso.

2.4 Classificação

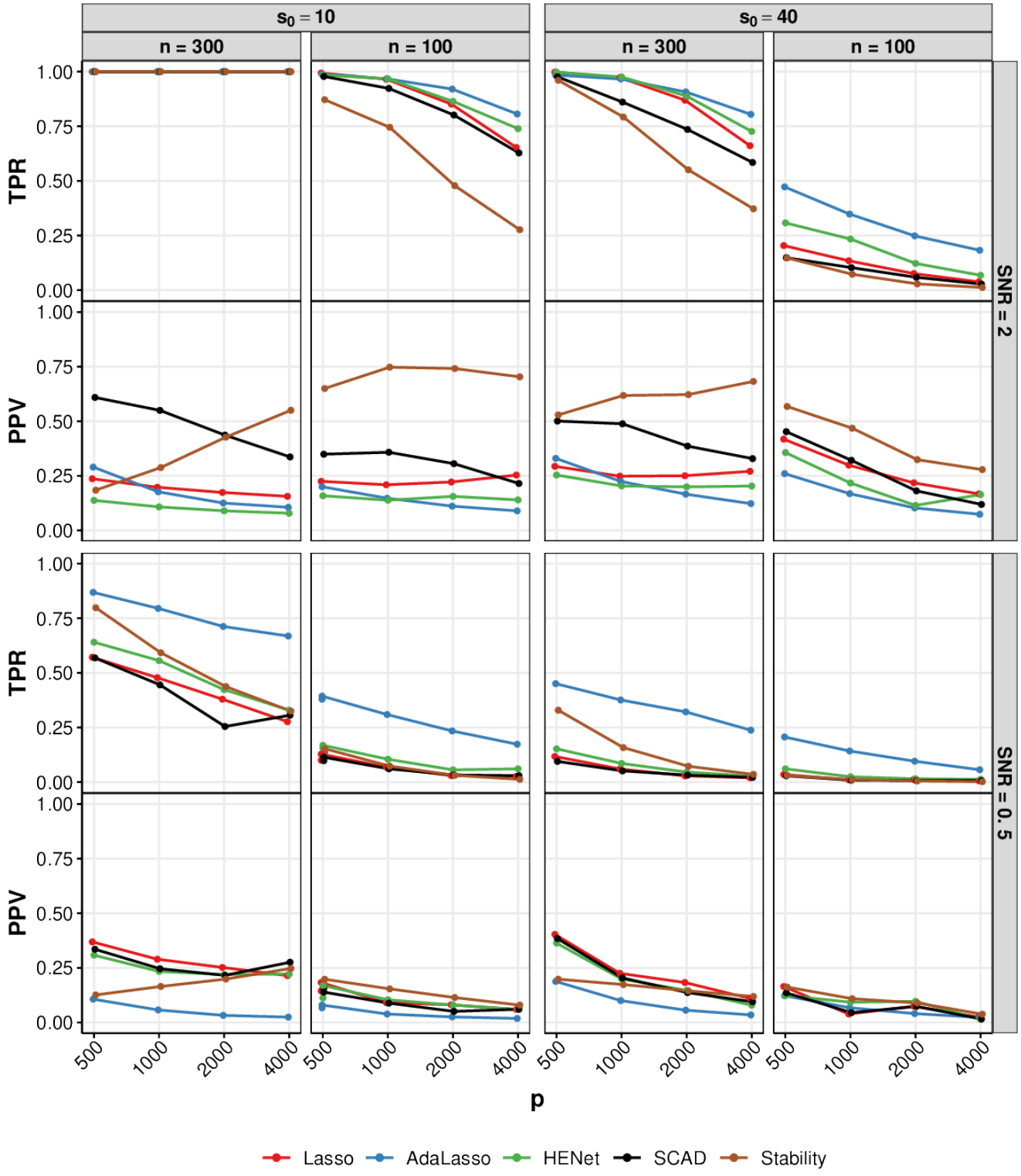


Figura 7: Versão Local de TPR, PPV, TPR, PPV em dados independentes para um set de configuração apontado

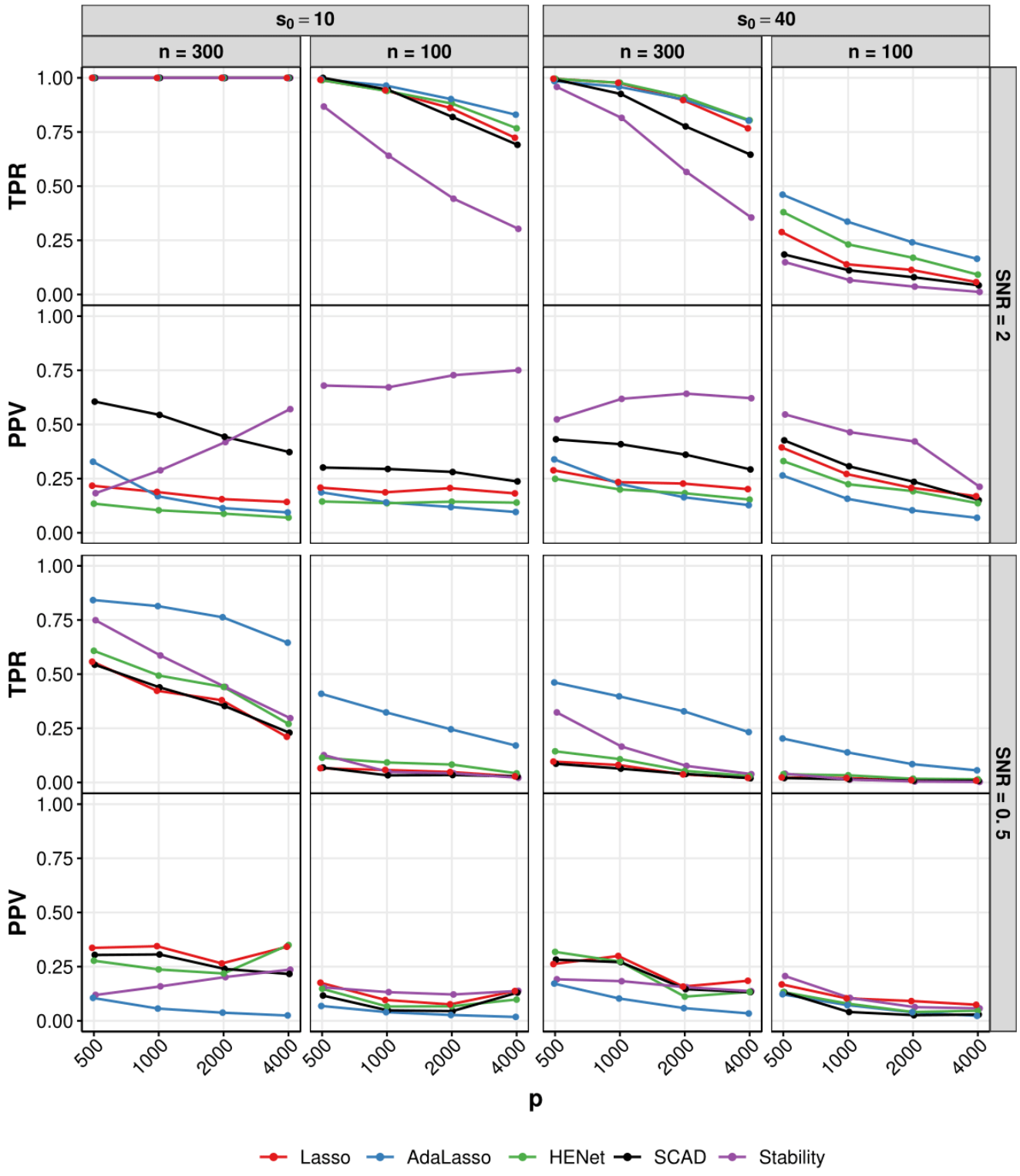


Figura 8: Versão Artigo de TPR, PPV, TPR, PPV em dados independentes para um set de configuração apontado

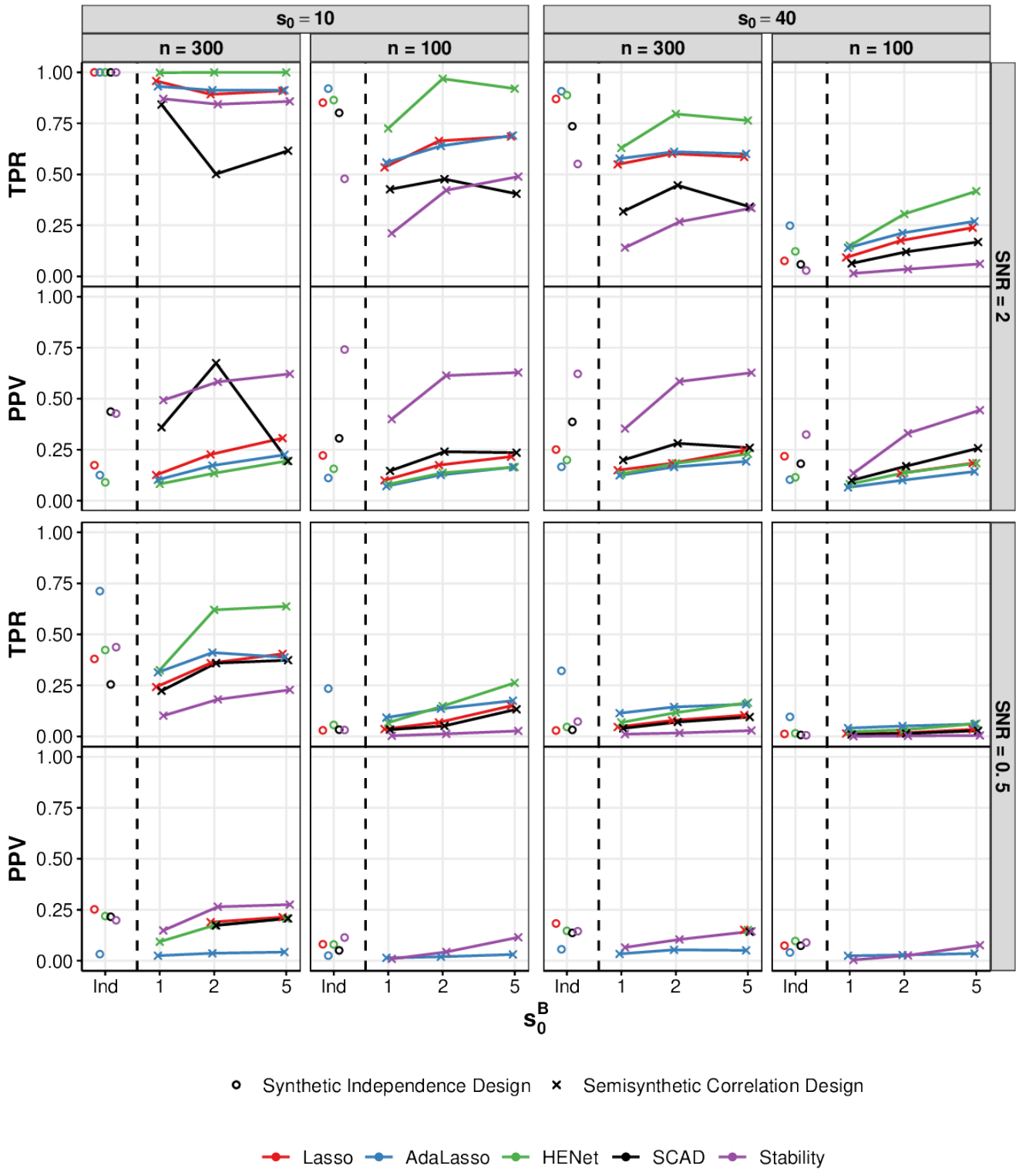


Figura 9: Versão Local de TPR, PPV, TPR, PPV em dados semi-sintéticos para um set de configuração apontado

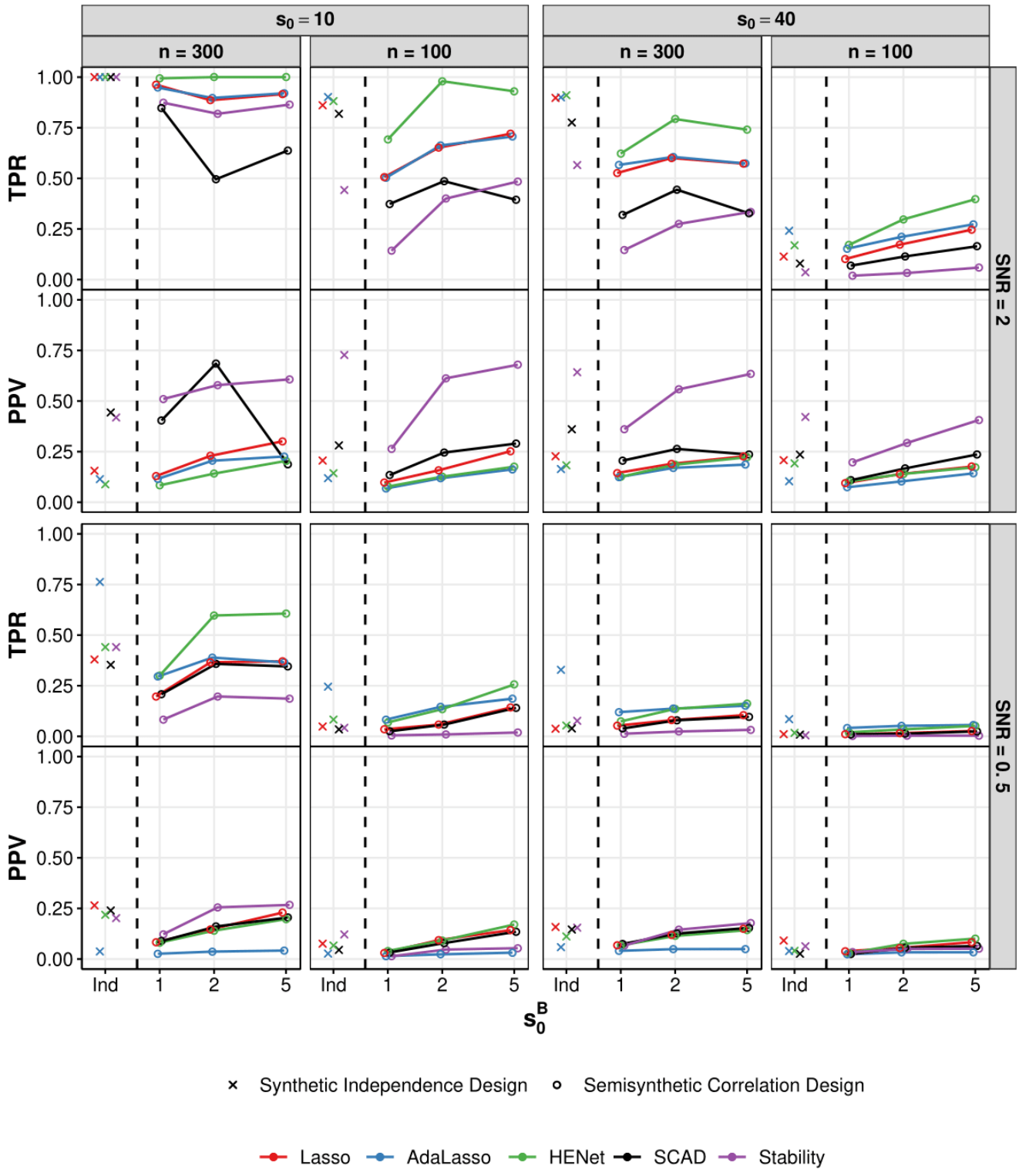


Figura 10: Versão Artigo de TPR, PPV, TPR, PPV em dados semi-sintéticos para um set de configuração apontado

Aqui, a escolha depende do contexto. Em busca da melhor proporção de falsos positivos, Stability é uma escolha segura para um escopo amplo. Já se o objetivo for aumentar o número de sinais reais selecionados, Adalasso tem superioridade aos outros métodos, com exceção de situações com alta correlação. Onde HEnet assume o protagonismo. Lasso ainda é uma opção generalista que performa bem para um largo escopo.

2.5 Robustez dos resultados

É dito no artigo que todas as simulações feitas, o são usando erros normalmente distribuídos. Sabendo que em situações reais isso pode fugir bastante da regra, busca-se testar a robustez destes métodos com erros distribuídos em diferentes graus de liberdade. Para comparação, utilizou-se a imagem mais generalista do artigo. Obtemos para os principais métodos em performance do artigo:

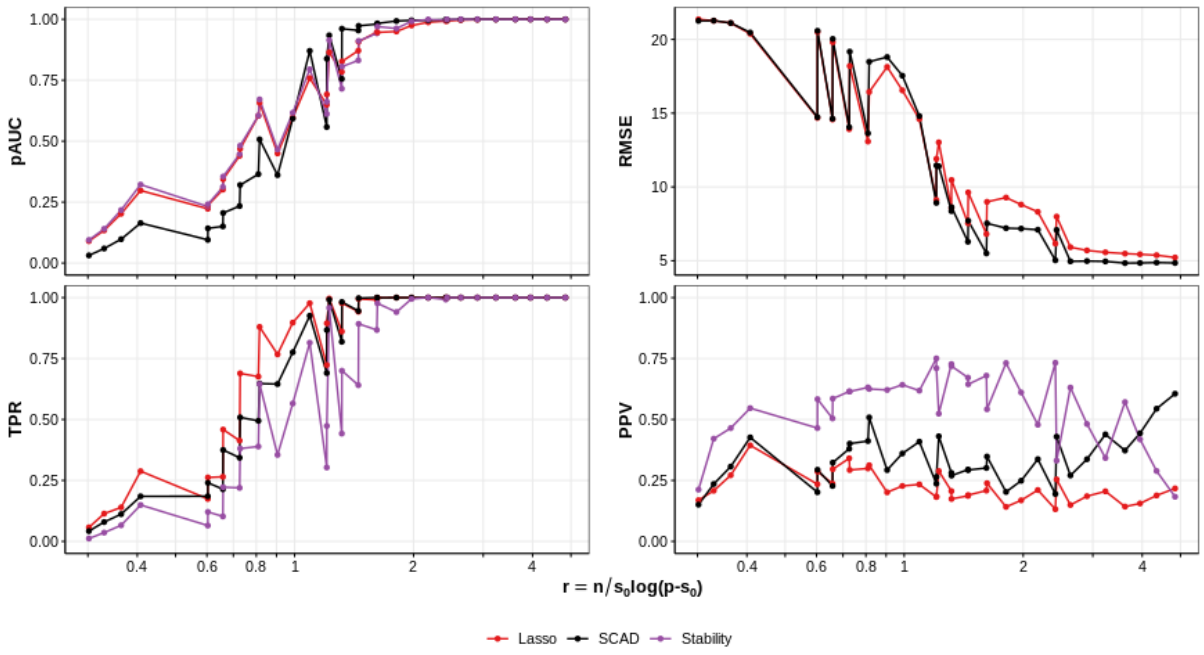
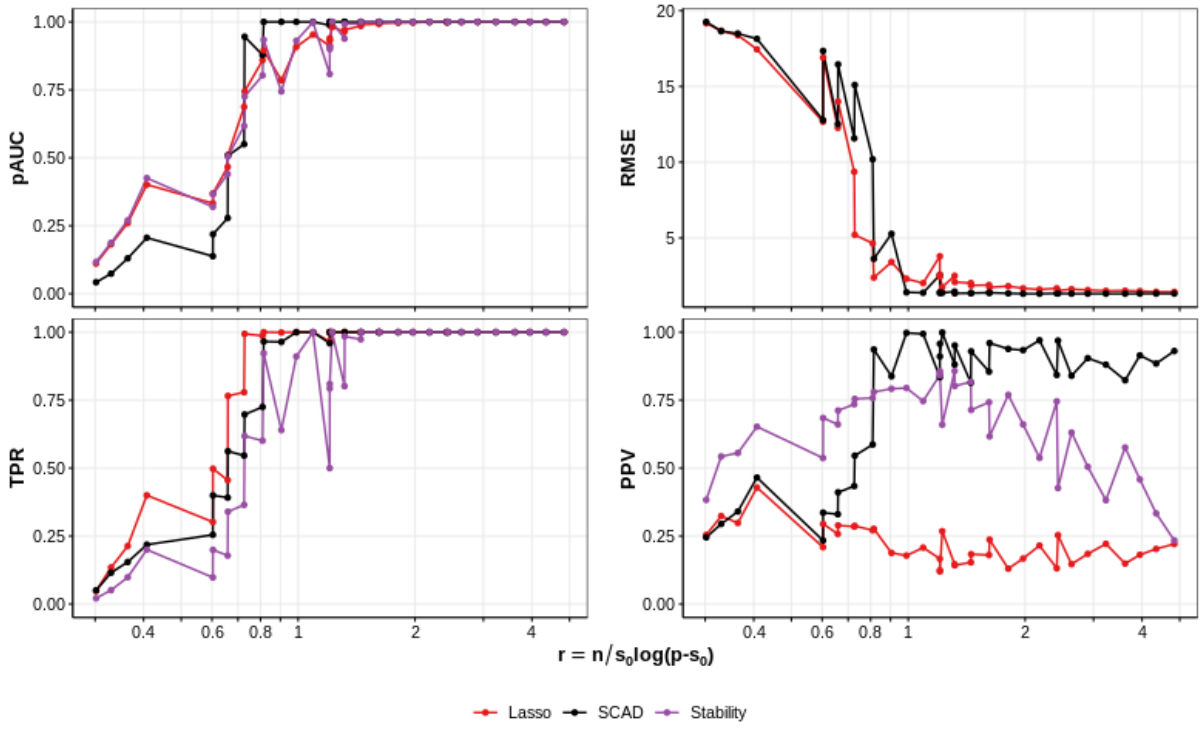
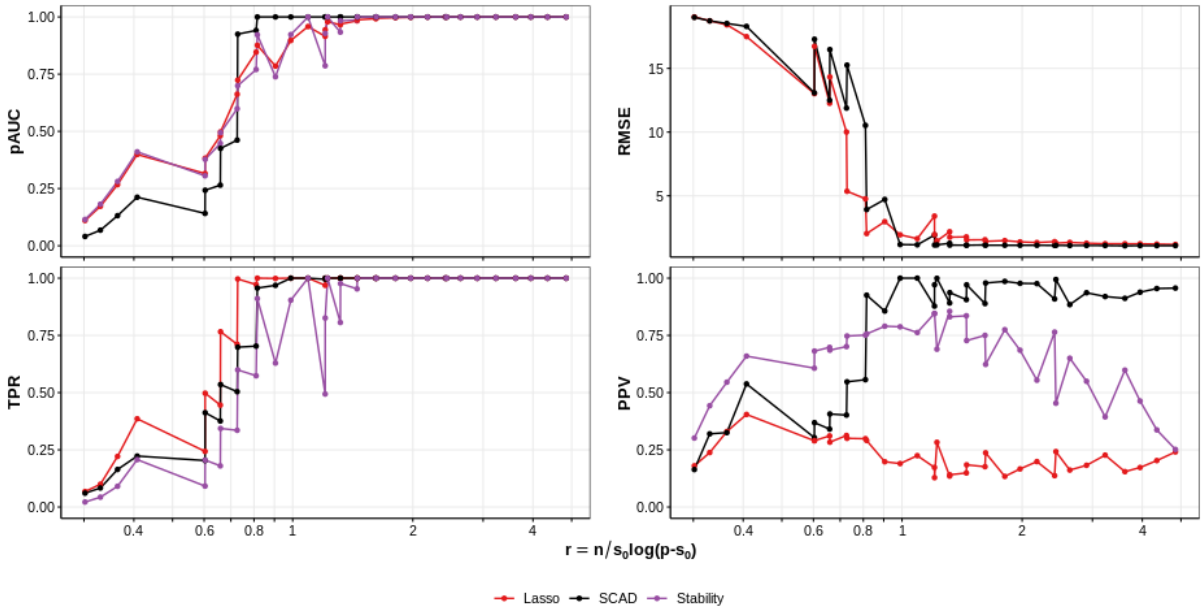


Figura 11: Artigo - Gaussiana



(a)



(b)

Figura 12: Versões (a) Local - Grau de liberdade 5 (b) Local - Grau de liberdade 15

Onde pode-se afirmar alguma robustez na maioria dos métodos sem grandes perdas de detalhe. Com uma exceção notável do parâmetro de PPV, onde o método SCAD parece aumentar o índice de forma muito mais intensa que o observado nos resultados do artigo. Como o SCAD era um método já de muita variação com as condições dadas, era de alguma forma esperado tal resultado vindo dele.

Referências

[1] Wang, F., Mukherjee, S., Richardson, S., Hill, S. M. (2019). High-dimensional regression in practice: an empirical study of finite-sample prediction, variable selection and ranking. Statistics and Computing, 30(3), 697–719. https://doi.org/10.1007/s11222-019-09914-9

[2] LAVINE, Michael. Introduction to Statistical Thought. [S.I.]: University Press of Florida, 2009. 434 p. ISBN 9781616100483. Disponível em: Introduction to Statistical Thought.. Acesso em: 01 set. 2020.

[3] JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. An Introduction to Statistical Learning: with Applications in R. 1. ed. New York: Springer-Verlag, 2013. 426 p. ISBN 1431-875X. DOI 10.1007/978-1-4614-7138-7. Springer - 2013. Disponível em: An Introduction to Statistical Learning: with Applications in R. Acesso em: 01 set. 2020.

São Carlos, 01/08/2021