Índices de Qualidade de Agrupamento baseados no Grafo de Gabriel

José Geraldo Fernandes
Escola de Engenharia
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
josegeraldof@ufmg.br

***Resumo*—Este trabalho analisa a relação entre índice de qualidade de agrupamentos e o desempenho de classificadores em problemas supervisionados. Propõe-se indicadores baseados no Grafo de Gabriel e testa-se sua validade a partir de um modelo de regressão linear e um teste de correlação. Todo o código desenvolvido está disponível em repositório Git.**

I. Introdução

A. Grafo de Gabriel

O Grafo de Gabriel [1] é uma construção a partir de um conjunto de pontos $\mathcal{S}$, que define os vértices do grafo, para outro conjunto de arestas $\mathcal{E}$ tal que dois pontos $\boldsymbol{x_i}, \boldsymbol{x_j}$ são adjacentes, definem uma aresta, se não há outro ponto de $\mathcal{S}$ dentro da hiperesfera definida com a distância entre os dois pontos $\boldsymbol{x_i}, \boldsymbol{x_j}$ como diâmetro, como na Equação 1.

$$\begin{gathered}(\boldsymbol{x_i}, \boldsymbol{x_j}) \in \boldsymbol{\mathcal{E}} \leftrightarrow \\ \delta^2(\boldsymbol{x_i}, \boldsymbol{x_j}) \leq [\delta^2(\boldsymbol{x_i}, \boldsymbol{x_k}) + \delta^2(\boldsymbol{x_j}, \boldsymbol{x_k})] \ \forall \boldsymbol{x_k} \in \mathcal{S}\end{gathered} \tag{1}$$

Onde δ é a métrica de distância na construção e comumente definida como a distância euclidiana.

A construção desse grafo é uma técnica da Geometria Computacional e emprestada para os problemas de aprendizado como uma forma de expressar as características de vizinhança da estrutura do conjunto de dados. Note como a definição da métrica de distância ótima pode representar um ganho nessa representação quando a distribuição de classes, em um problema de classificação, é melhor mapeada e discriminada em uma métrica que outra [2].

B. Métricas de Distância

As métricas de distâncias são parte essencial de muitas técnicas de aprendizado de máquina. É proveitoso representar os dados em um espaço tal que amostras similares conjugam uma distância menor que para outras amostras semanticamente díspares. A própria noção de semelhança e disparidade depende também da aplicação, em um mesmo espaço algumas características podem ter relevância superior a depender do contexto.

Considere, por exemplo, um conjunto de dados composto por amostras de fala, dada uma extração de características acústicas. Nesse cenário, dependendo se o interesse é classificar as amostras pela idade do falante, seu gênero ou até identificar aspectos mais culturais como o sotaque ou a emoção do discurso, os atributos extraídos terão relevância diferente no problema.

Uma abordagem para contornar esse obstáculo é determinar a priori o conjunto apropriado de características da extração a partir de conhecimentos específicos do problema. Contudo, esse processo é custoso e perde em generalização, pode-se, portanto, considerar o aprendizado de métrica uma forma de automatizar essa etapa para qualquer problema.

Uma formulação geral para esse modelo, que explicita as relações de interesse no espaço, é como a função de custo que, comumente [3], avalia as seguintes restrições:

$$\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) : \boldsymbol{x}_i \, e \, \boldsymbol{x}_j \text{ devem ser próximos}\}$$

$$\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) : \boldsymbol{x}_i \, e \, \boldsymbol{x}_j \text{ devem ser distantes}\}$$

$$\mathcal{R} = \{(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k) : \boldsymbol{x}_i \text{ deve ser mais próximo de } \boldsymbol{x}_j \text{ que de } \boldsymbol{x}_k\}$$

Assim, o problema de otimização é descrito como na Equação 2.

$$\min_M \ l(M, \mathcal{S}, \mathcal{D}, \mathcal{R}) + \lambda R(M) \tag{2}$$

Onde M é o parâmetro a ser otimizado, l é uma função de custo, R um regularizador e $\lambda \geq 0$ seu parâmetro.

C. Filtro de Sobreposição

Os filtros de sobreposição baseados no Grafo de Gabriel foram adotados como uma forma de realizar regularização no Classificador por Arestas de Suporte (CLAS) [4], [5].

Baseado na diferença de classes entre o vértice e as amostras conectadas por suas arestas, define-se uma grandeza Q que representa a qualidade dessa amostra como na Equação 3, onde V representa o número de arestas e V_{eq} o número dessas para amostras de classe coincidente.

$$Q(\boldsymbol{x}_i) = \frac{V_{eq}(\boldsymbol{x}_i)}{V(\boldsymbol{x}_i)} \tag{3}$$

Amostras na região de fronteira e sobreposição terão sua qualidade afetada quão maior for a mistura. Segue, portanto, um filtro simples de amostras com qualidade inferior a um limiar. Adota-se um limiar dinâmico TC que representa a qualidade média de amostras de classe específica.

Neste trabalho, contudo, é interesse utilizar esses indicadores, qualidade do vértice (Q) e proporção apontada pelo filtro (Di) como uma forma de avaliar o espaço de um conjunto de dados e um novo espaço aprendido com método de aprendizado de métrica.

II. Revisão Bibliográfica

O aprendizado de métrica é um processo importante para a utilização de classificadores baseados em distância. Além disso, considerando essa abordagem como um forma de mapeamento otimizado dos dados, é considerável também sua utilidade para redução de dimensionalidade e *clustering*.

Em *Local Fisher Discriminant Analysis* [6] os autores combinam duas funções de agrupamento, considerando aspectos globais e locais, na função de custo, adaptando duas técnicas de aprendizado de métrica, *Fisher Discriminant Analysis* e *lLcal-Preserving Projection*.

Já em *Large Margin Nearest Neighbors* (LMNN) [7] utiliza-se uma abordagem de k-vizinhos mais próximos para combinar um incentivo de aproximação de agrupamentos coincidentes e outro de afastamento para dissidentes, modelado, também, na função de custo do problema de otimização da matriz de transformação.

Para avaliação dessas técnicas, é comum a análise comparativa em função do desempenho na classificação [8]–[10]. Apesar dessa abordagem ser direta ao interesse do problema, é importante considerar um número grande e diverso de conjunto de dados para eliminar um viés de objetivo.

Muitas vezes, a higiene do espaço de representação de dados é tão importante quanto o resultado final do preditor. Nesse sentido, sugere-se índices de qualidade de *clustering* para avaliar essa perspectiva.

III. Metodologia

A. Base de Dados

Para avaliação dos índices de qualidade dos agrupamentos seleciona-se um conjunto de 15 *datasets* padrão do repositório UCI [11] em todos os testes. Esses são separados em 10 partições determinadas para validação cruzada *k-fold* [12].

Para confiabilidade do resultado, as bases de dados selecionadas são amplamente utilizadas em aplicações semelhantes. Como pré-processamento, fez-se uma normalização de todos os atributos e conversão dos categóricos em númericos, necessário para o algoritmo CLAS. Todos são de problemas de classificação binária. Segue a descrição da seleção: *Statlog Australian Credit Approval* (australian); *Banknote Authentication* (banknote); *Breast Cancer Wisconsin* (breastcancer); *Breast Cancer Hess Probes* (breastHess); *Liver Disorders* (bupa); *Climate Model Simulation Crashes* (climate); *Pima Indian Diabetes* (diabetes); *Fertility* (fertility); *Statlog German Credit Data* (german); *Gene Expression* (golub); *Haberman's Survival* (haberman); *Statlog Heart Dicease* (heart); *Indian Liver Patient* (ILPD); *Parkinsons* (parkinsons); *Connectionist Bench Sonar, Mines vs. Rocks* (sonar).

A Tabela I mostra as principais características dessa seleção, η representa a proporção entre as classes. Note a alta diversidade dos problemas para atestar a generalização do método.

Tabela I
Características das bases de dados selecionadas.

Dataset	**Amostras**	**Atributos**	$\boldsymbol{\eta}$
australian	690	14	0.44
banknote	1372	4	0.44
breastcancer	683	6	0.65
breastHess	133	30	0.74
bupa	345	6	0.42
climate	540	18	0.91
diabetes	768	8	0.65
fertility	100	9	0.12
german	1000	24	0.70
golub	72	50	0.65
haberman	306	3	0.74
heart	270	13	0.56
ILPD	579	10	0.72
parkinsons	195	22	0.75
sonar	208	60	0.47

B. Classificação

O problema de classificação segue o padronizado. Para cada *dataset* separou-se 10 *folds* fixos, desses aplicou-se os métodos de aprendizado de métrica, LMNN e LFDA, no conjunto de treinamento para conseguir o *dataset* no novo espaço, também avaliou-se o *dataset* sem modificação.

Aplicou-se o algoritmo CLAS e uma SVM com *kernel* RBF para classificação. Avaliou-se a performance com a área sob a curva ROC (AUC) com validação cruzada nos *folds*.

C. Avaliação dos Índices

Para avaliar o espaço aprendido pelos métodos de *metric learning*, utilizou-se os índices de qualidade de agrupamentos propostos: a qualidade média dos vértices (q); e, a taxa de filtragem de dados por sobreposição (Di). Esses foram utilizados sobre o conjunto de dados completo.

Em seguida, para validação, aplicou-se um modelo de regressão linear para mensurar a capacidade de previsão que os índices têm da acurácia do classificador. Ademais, um teste de correlação foi realizado com o mesmo objetivo.

IV. Resultados

A. Classificação

As Tabelas II e III mostram os resultados obtidos com a AUC média de todos os conjunto de dados e suas modificações aprendidas por ambos classificadores.

B. Índices de Qualidade

As Tabelas IV e V mostram os índices de qualidade calculados com todo o conjunto de dados no formato tradicional e suas modificações.

Dos índices propostos, as Tabelas IV e V mostram o resultado encontrado análogo.

Tabela II
DESEMPENHO OBTIDO PELO CLAS, AUC MÉDIA, PARA CADA BASE DE DADOS.

Dataset	Base	LMNN	LFDA
australian	0.85 ± 0.04	0.86 ± 0.04	0.86 ± 0.03
banknote	0.99 ± 0.01	1.00 ± 0.00	0.97 ± 0.03
breastcancer	0.96 ± 0.03	0.95 ± 0.04	0.97 ± 0.02
breastHess	0.81 ± 0.12	0.79 ± 0.14	0.74 ± 0.11
bupa	0.63 ± 0.10	0.58 ± 0.05	0.72 ± 0.08
climate	0.84 ± 0.07	0.87 ± 0.08	0.88 ± 0.11
diabetes	0.72 ± 0.04	0.72 ± 0.04	0.73 ± 0.06
fertility	0.59 ± 0.26	0.58 ± 0.23	0.59 ± 0.19
german	0.67 ± 0.04	0.69 ± 0.04	0.70 ± 0.05
golub	0.77 ± 0.17	0.68 ± 0.23	0.59 ± 0.17
haberman	0.56 ± 0.09	0.58 ± 0.09	0.57 ± 0.07
heart	0.80 ± 0.08	0.79 ± 0.08	0.84 ± 0.06
ILPD	0.57 ± 0.09	0.57 ± 0.10	0.65 ± 0.07
parkinsons	0.90 ± 0.15	0.91 ± 0.13	0.79 ± 0.11
sonar	0.88 ± 0.08	0.85 ± 0.08	0.76 ± 0.13

Tabela III
DESEMPENHO OBTIDO PELA SVM, AUC MÉDIA, PARA CADA BASE DE DADOS.

Dataset	Base	LMNN	LFDA
australian	0.86 ± 0.04	0.86 ± 0.04	0.86 ± 0.04
banknote	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01
breastcancer	0.97 ± 0.01	0.97 ± 0.02	0.97 ± 0.01
breastHess	0.78 ± 0.10	0.78 ± 0.12	0.74 ± 0.12
bupa	0.67 ± 0.07	0.67 ± 0.07	0.70 ± 0.04
climate	0.58 ± 0.09	0.68 ± 0.11	0.73 ± 0.11
diabetes	0.71 ± 0.05	0.70 ± 0.05	0.72 ± 0.04
fertility	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00
german	0.67 ± 0.05	0.68 ± 0.04	0.69 ± 0.05
golub	0.82 ± 0.17	0.54 ± 0.12	0.48 ± 0.16
haberman	0.51 ± 0.04	0.50 ± 0.03	0.55 ± 0.07
heart	0.80 ± 0.06	0.82 ± 0.07	0.81 ± 0.06
ILPD	0.50 ± 0.01	0.50 ± 0.01	0.50 ± 0.01
parkinsons	0.81 ± 0.11	0.83 ± 0.12	0.77 ± 0.14
sonar	0.84 ± 0.09	0.87 ± 0.06	0.81 ± 0.09

Tabela IV
ÍNDICE DE QUALIDADE MÉDIA DOS VÉRTICES DO CONJUNTO DE DADOS COMPLETO.

Dataset	Base	LMNN	LFDA
australian	0.55	0.55	0.55
banknote	0.56	0.56	0.56
breastcancer	0.65	0.65	0.65
breastHess	0.75	0.75	0.72
bupa	0.42	0.43	0.41
climate	0.08	0.07	0.07
diabetes	0.35	0.36	0.35
fertility	0.89	0.86	0.89
german	0.70	0.70	0.70
golub	0.64	0.63	0.67
haberman	0.73	0.74	0.73
heart	0.56	0.56	0.55
ILPD	0.70	0.71	0.71
parkinsons	0.76	0.76	0.76
sonar	0.47	0.48	0.43

Tabela V
PROPORÇÃO DE AMOSTRAS AVALIADAS COMO SOBREPOSIÇÃO DO CONJUNTO DE DADOS COMPLETO.

Dataset	Base	LMNN	LFDA
australian	0.48	0.49	0.46
banknote	0.45	0.45	0.45
breastcancer	0.29	0.29	0.29
breastHess	0.36	0.36	0.34
bupa	0.48	0.53	0.50
climate	0.55	0.62	0.62
diabetes	0.52	0.52	0.53
fertility	0.40	0.28	0.43
german	0.47	0.46	0.47
golub	0.47	0.42	0.40
haberman	0.42	0.47	0.52
heart	0.50	0.49	0.48
ILPD	0.45	0.42	0.47
parkinsons	0.36	0.30	0.39
sonar	0.49	0.52	0.52

C. Validação

Para cada classificador ajustou-se um modelo de regressão linear dos índices de qualidade para o resultado do desempenho desse. Registrou-se o tamanho do coeficiente, β, e o valor-p, P, da estimativa para cada atributo. As Tabelas VI e VII mostram os resultados obtidos e as Figuras 1 e 2 mostram as curvas ajustadas.

Tabela VI
REGRESSÃO LINEAR A PARTIR DA PROPORÇÃO DO FILTRO (DI).

CLAS						
	Base		**LMNN**		**LFDA**	
Atributo	β	P	β	P	β	P
intercept	0.94	$2e-03$	0.81	$6e-04$	1.01	$4e-04$
discard	-0.37	$5e-01$	-0.10	$8e-01$	-0.53	$3e-01$
SVM						
	Base		**LMNN**		**LFDA**	
Atributo	β	P	β	P	β	P
intercept	0.99	$4e-03$	0.84	$1e-03$	1.18	$2e-04$
discard	-0.56	$4e-01$	-0.24	$6e-01$	-0.96	$7e-02$

Tabela VII
REGRESSÃO LINEAR A PARTIR DO ÍNDICE DE QUALIDADE MÉDIO (q).

CLAS						
	Base		**LMNN**		**LFDA**	
Atributo	β	P	β	P	β	P
intercept	0.87	$4e-06$	0.88	$5e-06$	0.88	$3e-06$
q	-0.18	$4e-01$	-0.18	$4e-01$	-0.19	$3e-01$
SVM						
	Base		**LMNN**		**LFDA**	
Atributo	β	P	β	P	β	P
intercept	0.76	$1e-04$	0.76	$1e-04$	0.77	$8e-05$
q	-0.05	$8e-01$	-0.05	$8e-01$	-0.05	$8e-01$

Também, calculou-se a associação dos índices com o desempenho a partir dos testes de correlação de Pearson [13] e Spearman [14]. As Tabelas VIII e IX mostram os resultados obtidos para a estimativa, ρ, e seu valor-p associado, P.

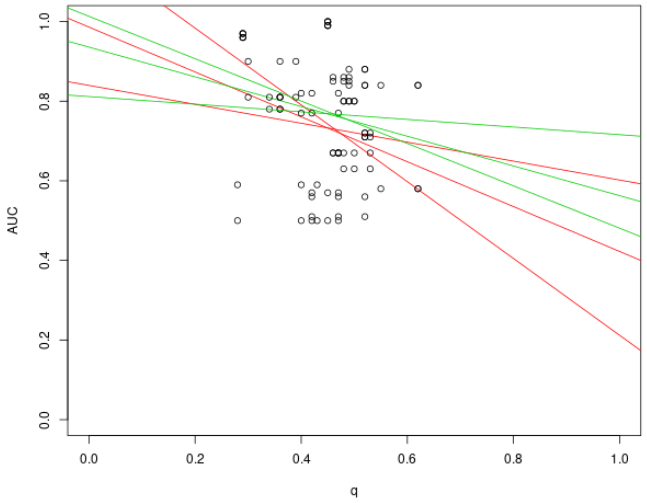

Figura 1. Curvas ajustadas da regressão linear a partir do desempenho dos diferentes classificadores e espaços contra a taxa do filtro (Di), em vermelho SVM e em verde CLAS.

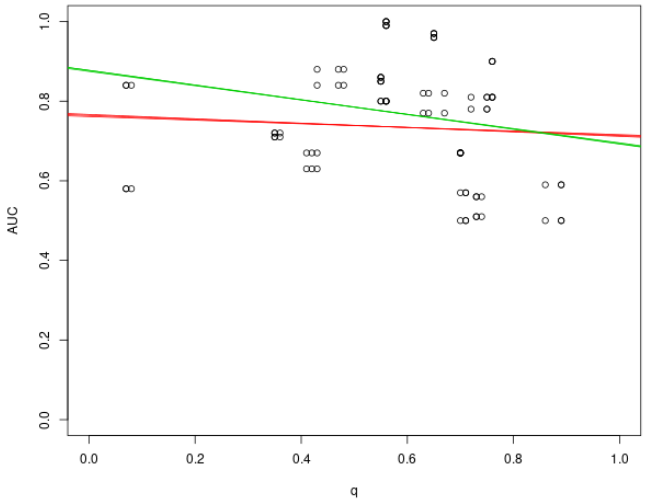

Figura 2. Curvas ajustadas da regressão linear a partir do desempenho dos diferentes classificadores e espaços contra o índice de qualidade médio (q), em vermelho SVM e em verde CLAS.

Tabela VIII
TESTE DE CORRELAÇÃO A PARTIR DA PROPORÇÃO DO FILTRO (DI).

CLAS						
	Base		**LMNN**		**LFDA**	
Método	ρ	P	ρ	P	ρ	P
Pearson	-0.18	$5e-01$	-0.07	$8e-01$	-0.31	$3e-01$
Spearman	-0.08	$8e-01$	-0.07	$8e-01$	-0.35	$2e-01$
SVM						
	Base		**LMNN**		**LFDA**	
Método	ρ	P	ρ	P	ρ	P
Pearson	-0.24	$4e-01$	-0.14	$6e-01$	-0.49	$7e-02$
Spearman	-0.07	$8e-01$	-0.07	$8e-01$	-0.40	$1e-01$

Tabela IX
TESTE DE CORRELAÇÃO A PARTIR DO ÍNDICE DE QUALIDADE MÉDIA (q).

CLAS						
	Base		**LMNN**		**LFDA**	
Método	ρ	P	ρ	P	ρ	P
Pearson	-0.26	$4e-01$	-0.26	$4e-01$	-0.27	$3e-01$
Spearman	-0.22	$4e-01$	-0.23	$4e-01$	-0.26	$3e-01$
SVM						
	Base		**LMNN**		**LFDA**	
Método	ρ	P	ρ	P	ρ	P
Pearson	-0.06	$8e-01$	-0.06	$8e-01$	-0.07	$8e-01$
Spearman	-0.24	$4e-01$	-0.25	$4e-01$	-0.26	$4e-01$

V. DISCUSSÕES

Como esperado, os métodos de aprendizado de métrica geraram resultados de desempenho ligeiramente melhores. Esperava-se, portanto, uma conclusão mais definitiva a partir dos índices de qualidade de agrupamentos. Seguindo o raciocínio que, apesar da pequena diferença em performance, um espaço mais representativo e com menor sobreposição é benéfico para o tratamento do problema.

De fato, o diferencial desses indicadores é mais ilativo, o que valida os métodos aplicados de *metric learning*. Resta, contudo, uma relação quantitativa desses índices com o interesse do problema, e a validação dos índices propostos.

Para validar essa hipótese aplicou-se a regressão e o teste de correlação, infelizmente, no entanto, o resultado obtido ainda é frágil para sustentar a hipótese. Os altos valores de valor-p não invalidam a hipótese nula em uma estimativa razoável para a grande maioria das observações. Há apenas uma tendência das curvas no ajuste de regressão, mais estável para o índice de qualidade médio.

VI. CONCLUSÕES

Este trabalho apresentou uma tentativa de aproximação entre o desempenho dos classificadores e índices de qualidade de agrupamentos a partir de um ajuste de regressão linear e um teste de correlação. Também, propôs dois índices de qualidade de agrupamentos baseados na estrutura do grafo construído no espaço.

Mostrou-se que técnicas de aprendizado de métrica respondem bem aos indicadores mas têm performance tímida no resultado dos classificadores contra o espaço tradicional. A tentativa de relacionar essas grandezas foi frustrada.

Uma continuação deste trabalho é o aumento do *benchmark* de base de dados para tentar aumentar a robustez dos modelos de validação.

AGRADECIMENTO

Este trabalho foi possível pela disponibilização das bases de dados pelo repositório *UCI Machine Learning* [11].

REFERÊNCIAS

[1] K. R. Gabriel and R. R. Sokal, "A new statistical approach to geographic variation analysis," *Systematic zoology*, vol. 18, no. 3, pp. 259–278, 1969.

[2] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," *Michigan State Universiy*, vol. 2, no. 2, p. 4, 2006.

[3] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," *arXiv preprint arXiv:1306.6709*, 2013.

[4] L. Torres, C. Castro, F. Coelho, F. Sill Torres, and A. Braga, "Distance-based large margin classifier suitable for integrated circuit implementation," *Electronics Letters*, vol. 51, no. 24, pp. 1967–1969, 2015.

[5] L. C. B. Torres, "Classificador por arestas de suporte (clas): Métodos de aprendizado baseados em grafos de gabriel," 2016.

[6] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis.," *Journal of machine learning research*, vol. 8, no. 5, 2007.

[7] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification.," *Journal of machine learning research*, vol. 10, no. 2, 2009.

[8] I. Fehervari, A. Ravichandran, and S. Appalaraju, "Unbiased evaluation of deep metric learning algorithms," *arXiv preprint arXiv:1911.12528*, 2019.

[9] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3279–3286, 2015.

[10] P. Moutafis, M. Leng, and I. A. Kakadiaris, "An overview and empirical comparison of distance metric learning methods," *IEEE transactions on cybernetics*, vol. 47, no. 3, pp. 612–625, 2016.

[11] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[12] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, pp. 1137–1145, Montreal, Canada, 1995.

[13] D. Best and D. Roberts, "Algorithm as 89: the upper tail probabilities of spearman's rho," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 24, no. 3, pp. 377–379, 1975.

[14] M. Hollander, D. A. Wolfe, and E. Chicken, *Nonparametric statistical methods*. John Wiley & Sons, 2013.