Rank-based Molecular Prognosis and Network-guided Biomarker Discovery for Breast Cancer

Yunlong Jiao

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à MINES ParisTech

Rank-based Molecular Prognosis and Network-guided Biomarker Discovery for Breast Cancer

Pronostic moléculaire basé sur l'ordre des gènes et découverte de biomarqueurs guidé par des réseaux pour le cancer du sein

École doctorale n°432

SCIENCES DES MÉTIERS DE L'INGÉNIEUR

Spécialité BIO-INFORMATIQUE

COMPOSITION DU JURY :

M. Francis BACH
INRIA, Président

M. Stéphan CLEMENCON
Télécom ParisTech, Rapporteur

M. Risi KONDOR
University of Chicago, Rapporteur

Mme Chloé-Agathe AZENCOTT
MINES ParisTech, Examinateur

M. Joaquin DOPAZO
Fundación Progreso y Salud, Examinateur

M. Jean-Philippe VERT
MINES ParisTech, Examinateur

Soutenue par **Yunlong JIAO**
le 11 septembre 2017

Dirigée par **Jean-Philippe VERT**

MINES ParisTech | PSL RESEARCH UNIVERSITY PARIS

Abstract

Breast cancer is the second most common cancer worldwide and the leading cause of women's death from cancer. Improving cancer prognosis has been one of the problems of primary interest towards better clinical management and treatment decision making for cancer patients. With the rapid advancement of genomic profiling technologies in the past decades, easy availability of a substantial amount of genomic data for medical research has been motivating the currently popular trend of using computational tools, especially machine learning in the era of data science, to discover molecular biomarkers regarding prognosis improvement. This thesis is conceived following two lines of approaches intended to address two major challenges arising in genomic data analysis for breast cancer prognosis from a methodological standpoint of machine learning: rank-based approaches for improved molecular prognosis and network-guided approaches for enhanced biomarker discovery. Furthermore, the methodologies developed and investigated in this thesis, pertaining respectively to learning with rank data and learning on graphs, have a significant contribution to several branches of machine learning, concerning applications across but not limited to cancer biology and social choice theory.

Résumé

Le cancer du sein est le deuxième cancer le plus répandu dans le monde et la principale cause de décès due à un cancer chez les femmes. L'amélioration du pronostic du cancer a été l'une des principales préoccupations afin de permettre une meilleure gestion et un meilleur traitement clinique des patients. Avec l'avancement rapide des technologies de profilage génomique durant ces dernières décennies, la disponibilité aisée d'une grande quantité de données génomiques pour la recherche médicale a motivé la tendance actuelle qui consiste à utiliser des outils informatiques tels que l'apprentissage statistique dans le domaine de la science des données afin de découvrir les biomarqueurs moléculaires en lien avec l'amélioration du pronostic. Cette thèse est conçue suivant deux directions d'approches destinées à répondre à deux défis majeurs dans l'analyse de données génomiques pour le pronostic du cancer du sein d'un point de vue méthodologique de l'apprentissage statistique : les approches basées sur le classement pour améliorer le pronostic moléculaire et les approches guidées par un réseau donné pour améliorer la découverte de biomarqueurs. D'autre part, les méthodologies développées et étudiées dans cette thèse, qui concernent respectivement l'apprentissage à partir de données de classements et l'apprentissage sur un graphe, apportent une contribution significative à plusieurs branches de l'apprentissage statistique, concernant au moins les applications à la biologie du cancer et la théorie du choix social.

Acknowledgments

First and foremost, I would like to thank Jean-Philippe Vert for being an inspiring advisor and a supportive supervisor, for having welcomed me into CBIO with an amazing funding opportunity, for sharing his experience and ideas with me that have both consciously and unconsciously shaped my academic and communication skills, for setting an excellent example of a researcher with contagious enthusiasm and a group leader with motivational leadership, without whom I would simply never have completed this thesis.

During my PhD, I was fortunately offered the opportunity to work with Joaquin Dopazo, who proposed and led one of my doctoral projects and mentored me with encouragement and inspiration during my stay at CIPF, and Stefan Kobel, who patiently trained my presentation skills and enriched my background knowledge in biochemistry during my stay at Roche; both mentors have been huge influences to me, for which I cannot express enough gratitude.

Many other people have contributed, directly or indirectly, to the work presented in this thesis, and I would like to thank: Elsa Bernard, Erwan Scornet, Véronique Stoven and Thomas Walter, for participating in the DREAM challenge as a team; Fabian Heinemann and Sven Dahlmanns, for the discussion on the project of analyzer failure prediction; Eric Sibony and Anna Korba, for suggesting and collaborating on the project of rank aggregation; Marta Hidalgo, Cankut Çubuk, Alicia Amadoz, José Carbonell-Caballero and Rubén Sánchez, for the comments and help on the project of network analysis; last but not least, Vincent Brunet, for always being so responsive and helpful whenever I had an embarrassingly trivial problem with the server.

I would also like to thank my thesis reviewers, Risi Kondor and Stéphan Clémençon, for their time, interest and helpful comments, and other members on my defense jury, Chloé-Agathe Azencott, Francis Bach, Joaquin Dopazo and Jean-Philippe Vert, for their time and insightful questions.

The few people who have been an immensely significant part of my professional and personal life during my PhD must be specially mentioned, in that my unexpected encounter with them and their involvement in my life afterwards can only be described by no better words than *kizuna* (a special bond of friendship).

MeiMei channn, thank you for being the first person who ever talked to me at our first ITN summer school in Tübingen and then becoming one of my closest friends two years later at the fourth time we met, for those countless times of selflessly helping and teaching me with programming, biology and everything you know, for always being there for me caring every little thing happening around me, for listening to my joys and misery and also sharing happiness and frustration in life, for having never complained about my constant complaints and never been bored of my tedious stories, for every moment

during the very few times we could meet that I cherish for the rest of my life, or simply for agreeing instantaneously to have more than two dinners until we got totally bloated every time we hang out.

Puppy Peeter, thank you for showing up in CBIO since when the lab just seemed to me a much different place to be in, for helping Google translate all the abstracts in this thesis into French together with lovely Lucile, for having the best taste in food, except for cheese, and sharing as much interest in burgers as I do, for being the first one and the only one in the lab for a long time who would think calling me by a different name was not an inappropriate thing, or simply for being so adorable to talk to, to be around with or just to look at.

Cankut, thank you for being such a great labmate, flatmate and frrriend during my stay in Valencia, for showing me around so many times that had made me fall in love with every bit of the city, for being one of the most truly selfless and genuinely sympathetic people I know, for having the cutest beagle in the world, MoMo, who would lick me every morning until I woke up, or simply for being one of the most important reasons that my stay in Spain was such an unforgettable experience that I keep going around telling everyone how much it means to me.

MI LOBE SEÑORITO, shank you for being Shpanish first of all, then for teaching me sho much matsch, including eigenvaluesh in particular, and influenshing me with your shrewd wishdom in life, for trushting me blindfolded and opening up to me sho easily that makesh me feel sho very shpecial, for making me shunny in a gloomy placshe without even having to try, for putting up with my shilliness and grumpiness and even being shilly and grumpy together, or shimply for running down with me to my favorite reshtaurant in Parish every Tueshday, but the one shing I am not at all shankful for is how late you came in my life when I will have to leave shoon.

jacoPoo, thank you for having made my last year in Paris so wOndErfUl, for bringing up the Italian soul in me by giving me an Italian name now everyone knows me by and teaching me how to speak with a hand, for showing me the aesthetic side of you that enlightens my capacity for art, for always understanding me and sticking with me under any circumstances, or simply for being my brother from another mother who made me leaving Paris so much harder than it should have been.

Ana! Thank you too for having made my last year in Paris so special, for having the unique personality of being the meanest on the outside and sweetest on the inside, for bringing the competition of being shameless to another level for me, for not only enduring but treasuring the superficiality and stupidity of me, or simply for being my sister from another mother who too made me leaving Paris so much harder than it should have been.

I would also like to thank all the former and current members of CBIO:

Nelle, for having helped me a lot during and especially at the beginning of my PhD as an admirable *senpai* (a senior colleague) to me, for giving me plenty of valuable advice on building a professional career, and for co-founding the CBOG (CBIO Beer Organizing Group); Véronique, for being one of the most optimistic and delightful people I know who tells the funniest stories non-stop while being a respectable professor; Thomas, for being a determined researcher and a motivating character to me; Chloé, for giving a lot of helpful comments and advice on learning with networks, and for setting up an outstanding example for me as a researcher with a successful career established at my age; Victor, for being the only other person in the lab who did not speak French for three years; Nino, for being so kind and encouraging all the time with whom I could talk about science, even comfortably in French; Marine, for having to sit in the same office with me with whom I could "professionally" and casually talk to from time to time at work; Beyrem, for being a funny guy; Xiwei, for so many pieces of important information that I managed to not have myself evicted by the French prefectures; Benoît, for accomplishing the mission impossible that you had single-handedly changed what Paris and France meant to me; Joe, for being such an adorable human being I like to hang out a lot with but at the same time such an annoying yet weirdly charming one I can never really get mad at; and many others from CBIO I will apologetically skip naming, for the enjoyable moments and pleasant conversations over a cup of coffee or a pint of beer occasionally. Besides the regulars, Ramona, Ilaria and several other visitors brought appreciable dynamic to CBIO, for which I am very grateful.

During my secondments in Germany and in Spain, many people came across that made my short stay abroad much less lonely, and I would deeply thank: Kathrin, Miaolin and others colleagues from Roche, for their amiable company in Penzberg; Edgar, Carol, Pau, Kinza, Sema, Julen and other colleagues from CIPF, for their delightful friendship in Valencia, especially outside of the lab, and Javi, for hosting me in his apartment with enormous generosity and warm-heartedness when I went to Madrid for visa affairs.

Finally, this thesis is dedicated to the most important people in my life even though they would never have read this, my parents, for always believing in me since I was born, for having supported every decision I made, for raising me up and providing everything I needed but never asking anything in return.

Funding-wise, my PhD was supported by the European Union 7th Framework Program through the Marie Curie Initial Training Network (ITN) Machine Learning for Personalized Medicine (MLPM) grant No. 316861, and by the European Research Council grant ERC-SMAC-280032.

Paris, July 2017

Yunlong

Contents

List of Figures

List of Tables

List of Symbols

Learning Setup

$\mathcal{X}$	Input Space
$\mathbf{x}$	Input Vector or Uncertain Ranking
$\mathcal{Y}$	Output Space
y	Output Response of Interest
$\mathcal{D}$	Dataset
m, N	Number of Observations
n	Dimensionality

Permutation and Ranking

$[\![n]\!]$	Item Set or $\{1, 2, \ldots, n\}$
$\mathbb{S}_n$	Symmetric Group on $[\![n]\!]$
n_c	Number of Concordant Pairs
n_d, d	Number of Discordant Pairs or Kendall Tau Distance
σ, π, τ	Permutation or Total Ranking
R	Partial Ranking
$\mathbf{R}$	Multivariate Ranking

Kernel Learning

$\mathcal{F}$	Feature Space
K	Positive Definite Kernel
K_τ	Kendall Kernel
K_M	Mallows Kernel
Φ, ϕ	Kendall Embedding

Learning on Graphs

$\mathcal{G}$	Graph
$\mathcal{V}$	Vertex Set
$\mathcal{E}$	Edge Set
L	Graph Laplacian
Ψ	Graph Wavelets
Ω	Graph Dual Wavelets
P	Regularization or Penalty Function
β	Linear Prediction Coefficients

Kemeny Aggregation

σ^*	Kemeny Consensus
$\mathcal{K}$	Set of Kemeny Consensuses
θ	Euclidean Angle
r	Approximate Voting Rule

Other Notations

$\mathbb{R}$	Set of Real Numbers
$\mathbb{P}$	Probability
$\mathbb{E}$	Expectation

List of Thesis Deliverables

Working Papers and Preprints

- Y. Jiao and J.-P. Vert. *Network-based Wavelet Smoothing for Analysis of Genomic Data*. Technical report, École nationale supérieure des mines de Paris, 2017.
- Y. Jiao, M. R. Hidalgo, C. Çubuk, A. Amadoz, J. Carbonell-Caballero, J.-P. Vert and J. Dopazo. *Signaling Pathway Activities Improve Prognosis for Breast Cancer*. 2017. Submitted. bioRxiv preprint bioRxiv-132357.
- Y. Jiao and J.-P. Vert. *The Kendall and Mallows Kernels for Permutations*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2017. In press. HAL preprint HAL-01279273.
- E. Bernard, Y. Jiao, E. Scornet, V. Stoven, T. Walter and J.-P. Vert. *Kernel Multitask Regression for Toxicogenetics*. Molecular Informatics, 2017. In press. bioRxiv preprint bioRxiv-171298.

Published Papers

- Y. Jiao, A. Korba and E. Sibony. *Controlling the Distance to a Kemeny Consensus without Computing It*. In Proceedings of the 33rd International Conference on Machine Learning (ICML-16), pages 2971–2980, 2016.
- F. Eduati, L. Mangravite, T. Wang, H. Tang, J. Bare *et al. Prediction of human population responses to toxic compounds by a collaborative competition*. Nature Biotechnology, vol. 33, no. 9, pages 933–940, 2015.
- Y. Jiao and J.-P. Vert. *The Kendall and Mallows Kernels for Permutations*. In D. Blei and F. Bach, editors, Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pages 1935–1944, 2015.

Patents and Patent Applications

- Y. Jiao, J.-P. Vert, F. Heinemann, S. Dahlmanns and S. Kobel. *Failure State Prediction for Automated Analyzers for Analyzing a Biological Sample*, 2016. Pending European patent filed by Roche Diagnostics GmbH, F. Hoffmann–La Roche AG, December 2016.

Software

- Y. Jiao. ***kernrank**, version 1.0.2*. `https://github.com/YunlongJiao/kernrank`, 2016. Online; accessed April 2016. Open-source `R` package publicly available on GitHub.

Chapter 1

Introduction

Abstract: *Breast cancer is the second most common cancer worldwide and the leading cause of women's death from cancer. Improving cancer prognosis has been one of the problems of primary interest towards better clinical management and treatment decision making for cancer patients. With the rapid advancement of genomic profiling technologies in the past decades, easy availability of a substantial amount of genomic data for medical research has been motivating the currently popular trend of using computational tools, especially machine learning in the era of data science, to discover molecular biomarkers regarding prognosis improvement. This chapter briefly summarizes the general background of breast cancer with a particular focus on breast cancer prognosis, reviews the prospects and challenges in genomic data analysis, and overviews the methodologies and contribution of the thesis work in this research area.*

Résumé : *Le cancer du sein est le deuxième cancer le plus répandu dans le monde et la principale cause de décès due à un cancer chez les femmes. L'amélioration du pronostic du cancer a été l'une des principales préoccupations afin de permettre une meilleure gestion et un meilleur traitement clinique des patients. Avec l'avancement rapide des technologies de profilage génomique durant ces dernières décennies, la disponibilité aisée d'une grande quantité de données génomiques pour la recherche médicale a motivé la tendance actuelle qui consiste à utiliser des outils informatiques tels que l'apprentissage statistique dans le domaine de la science des données afin de découvrir les biomarqueurs moléculaires en lien avec l'amélioration du pronostic. Ce chapitre résume brièvement le contexte général du cancer du sein avec un point particulier sur son pronostic, détaille les perspectives et les défis dans l'analyse des données génomiques, et présente les méthodologies et contributions de la thèse dans ce domaine de recherche.*

1.1 General Background of Breast Cancer

Breast cancer refers to a malignant tumor that has developed from cells in the breast. Uncontrolled growth of cancer cells can invade nearby healthy breast tissue over time, and if cancer cells get into the lymph nodes that are small organs that filter out foreign substances in the body, they could then have a system of spreading further into other parts of the body and form new tumors in distant organs or tissues, a process called distant metastasis that aggravates the situation to a significant extent. Breast cancer is the most common cancer in women worldwide and second most common cancer overall for both genders in terms of incidence rates (following lung cancer), and it is the leading cause of cancer death among women in developing countries and the second leading cause of cancer death (following lung cancer) among women in developed countries [Torre 2015].[1] Over 521,900 women worldwide were estimated to have died in 2012 due to breast cancer [Ferlay 2013].[2] Survival rates have in general been improving over the past decades, as a result of increased awareness, earlier detection through mammographic screening, adequate medical care and cancer treatment advances, with the caveat that rates vary greatly worldwide and still remain quite low in less developed countries.

Diagnosis of cancer, determination of the presence (or extent) of the disease, is performed by means of (incisional) *biopsy*, a medical test in which surgeons extract sample cells or tissues for pathologists to examine under microscope or further analyze chemically. If diagnosed early, the initial treatment for breast cancer is usually accomplished by complete removal of tumor by surgery or radiation (mastectomy or less-extensive breast-conserving surgery) without damage to the rest of the body. After the initial treatment (or in case that the initial treatment should not be applicable), many patients receive additional treatment, including adjuvant chemotherapy, hormone therapy and targeted therapy, to lower the risk of relapse, that is the recurrence risk of cancer-related conditions, and/or to prevent metastasis. However, as the most common type of adjuvant therapy, chemotherapy usually involves cytotoxic drugs and has strong deleterious side effects, and the intake of such aggressive treatment should hence be minimized for those that will not necessarily need it. Therefore, to identify those patients who should receive adjuvant chemotherapy is of chief importance in improving the feasibility of treatment deployment in routine clinical management of cancer. The decision of whether to receive such treatment or not is made based on prognosis of the cancer patient, that is the estimation of the risk of relapse or likely course of outcome if no additional treatment is given after the initial treatment, and further treatments are considered most beneficial for patients with poor prognosis and some cases of good prognosis can even choose the option to forgo chemotherapy.[3] In order to quantify prognosis results, a patient is

[1] See more cancer facts and statistics summary at https://www.cancer.org/research/cancer-facts-statistics.html.

[2] See more of contemporary estimates of the incidence of, mortality and prevalence from major types of cancer at http://globocan.iarc.fr/.

[3] In fact, two questions need to be addressed in decision making for cancer treatment: prognosis

usually categorized into prognostic groups of high or low risk corresponding to one of the four common types of survival risk: distant metastasis-free survival, (local or distant) recurrence-free survival, disease-free survival, overall survival. Note that the following discussion applies to any specific survival unless specified otherwise.

Conventionally, breast cancer prognosis is based solely on clinico-pathological information collected from patients and tumors. Several commonly used clinico-pathological parameters have been well established to be indicative of likely prognosis of patients, and thus widely adopted in the clinical management of breast cancer. For example, it is known that breast cancer with cancer cells detected in lymph nodes has a higher risk of relapse than breast cancer *in situ*, and thus requires to be treated with certain adjuvant chemotherapies that are usually more aggressive [Moffat 2014]. In fact, doctors most often evaluate the severity of breast cancer based on the Nottingham grading system, a score-based grading system using clinico-pathological parameters such as the size and shape of the nucleus in the tumor cells and how many dividing cells are present [on Cancer 2010]. High-grade tumors look the most abnormal from normal cells and tend to be the most invasive, and are thus classified with poor prognosis. As another example, hormone receptors in breast cancer, estrogen-receptor (ER) and progesterone-receptor (PR), play an important role in normal glandular development and in breast cancer progression, and their status is therefore highly prognostic (as well as predictive to the responsiveness of hormone and endocrine therapies) [Moffat 2014]. Some online tools exist to perform prognosis of cancer patients and aid physicians weigh against the risks and benefits of adjuvant treatments, among which stands out the renowned *Adjuvant! Online*[4]. Notably, the six predictors that are shown highly prognostic and used by *Adjuvant! Online* to predict cancer-related mortality and relapse are: patient age, tumor size, grade, hormone receptor status, number of positive lymph nodes and comorbidity level.

Due to the intrinsic heterogeneity across breast cancer tumors, patients of similar clinico-pathological type can have remarkably different survival outcome. An example constituted by [van 't Veer 2008] will be quoted here. Large meta-analyses show that recurrence is likely in 20–30% of young women with early-stage (lymph node-negative) breast cancer, but in the United States 85–90% of women with this type of cancer receive adjuvant chemotherapy, among whom 55–75% therefore undergo a toxic therapy that they would very likely not benefit from but may experience the undesirable side effects. Since cancer is a inherently complex disease, the unwanted situation is mostly due to the fact that clinico-pathological information alone is far from sufficient to reliably identify those patients who are likely to

that is the estimation of the course of outcome if no additional treatment is given and hence the identification of those patients who are most likely to need additional treatment; prediction that is the identification of patients who are most likely to benefit from a specific treatment and hence the determination on which treatment should be most responsive and effective for a patient. While prognosis and prediction are equally important and usually discussed together in literature, prediction will be mostly omitted from discussion for ease of the presentation of this thesis.

[4] https://www.adjuvantonline.com/.

relapse, let alone to accurately characterize the outcome of each particular case in order to personalize the best therapeutic option. It is recognized as an important yet challenging task to improve prognosis for each diseased individual and identify more efficient prognostic features, burgeoning the research of interest in interrogating breast cancer at the molecular level.

1.2 Towards Molecular Prognosis

As [Vogelstein 2004] put it, who are pioneers in cancer molecular biology research:

> *"The revolution in cancer research can be summed up in a single sentence: cancer is, in essence, a genetic disease."*

Among many explanations on cancer biology, a widely accepted one states that cancer is caused by genomic abnormalities, such as the accumulation of mutations or the dysregulation of gene expression involving tumor suppressor genes and oncogenes in cancer cells. For decades, the number of genes with established involvement in cancer development has been increasing significantly, and it has been appreciated that their biological functions are organized by a few principles, named *the hallmarks of cancer*, which rationalize the complexities of cancer and are all underlaid by genome instability generating genetic diversity [Hanahan 2000, Hanahan 2011]. It is now common knowledge that genomic features contain unique characteristics of each individual being and offer the opportunity of scrutinizing the individuality of each breast tumor. Often termed by *biomarkers* are such molecular features, typically genes, whose abnormal presence or dysfunctional behavior characterizes the biological heterogeneity of tumours, leading to molecular subtyping of cancer, and can thus be indicative of prognosis. While biomarkers can be associated to any phenotype of interest in general, the discussion will particularly focus on biomarkers related to breast cancer prognosis in accordance with the objective of the present thesis.

Many biomarkers related to breast cancer survival have been reported in the literature. For example, somatic mutations in gene TP53 show association with worse survival, independent of other risk predictors, see for instance a meta-analysis by [Pharoah 1999]. Worse breast cancer survival of gene BRCA mutation carriers versus non-carriers have been confirmed by several meta-analyses [Zhong 2015, Zhu 2016]. Over-expression of gene HER2, pathologically termed as *HER2-positive*, is linked to poorer outcome of node-negative breast cancers [Chia 2008], a widely-observed association that has led to the advent of several HER2-directed therapies [Arteaga 2012]. Notably, major molecular subtypes of breast cancer are determined by the gene expression status, over- or under-expression, of hormone receptors and HER2, based on which physicians usually perform prognosis and plan treatments [Schnitt 2010]. For a review on currently established and emerging biomarkers for breast cancer prognosis, see [Weigel 2010].

From the foundation and completion of Human Genome Project (HGP) to the foundation of The Cancer Genome Atlas Research Network (TCGA), the rapid

advancement of genomic profiling technologies in the past decades have paved way to the advent of the current "omics" revolution. Nowadays, thousands up to millions of genomic features can be efficiently collected from biological samples available for medical research. Taking gene expression profiling as an example, DNA microarray, a hybridization-based technology, measures the relative expression activity of a large number of predetermined list of target genes in a single experiment (Figure 1.1) [Lockhart 1996]. RNA-seq, a next-generation sequencing-based technology, was later invented to provide expression measurements of gene sequences at lower cost and higher throughput (or larger genome coverage) with many advantages benchmarked against previous technologies [Wang 2009].

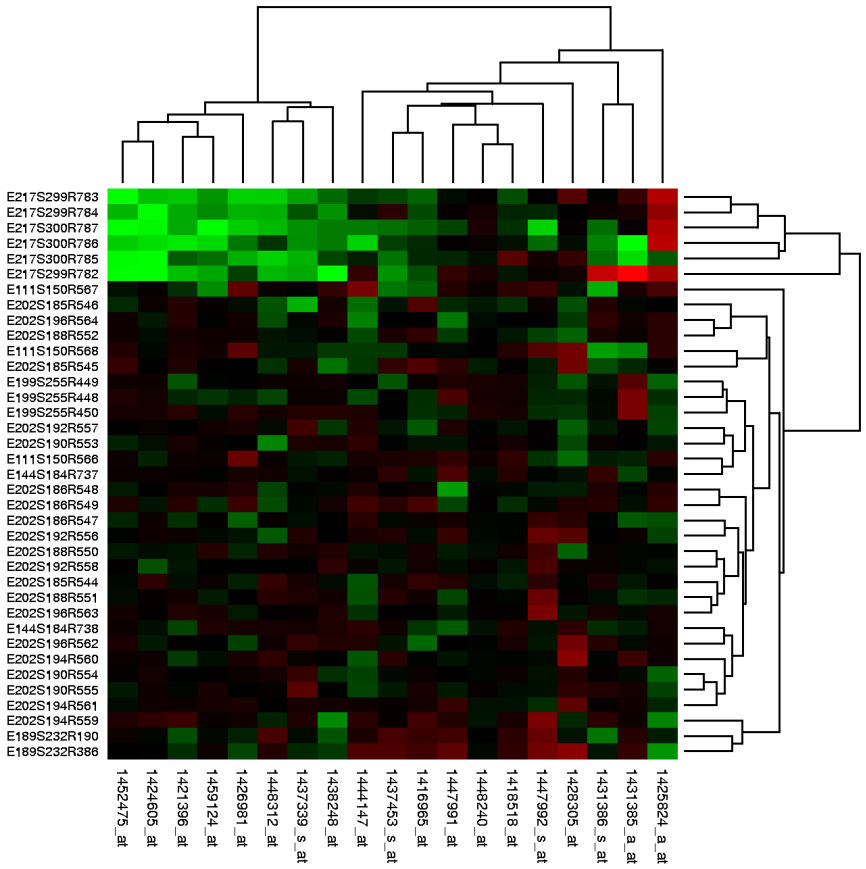


Figure 1.1: This image from [Commons 2017] illustrates an example of gene expression values from microarray experiments represented as a heatmap of two color dyes, with patients in rows and probes in columns, to visualize results of data analysis.

The revolution of gene expression profiling technologies fostered the development of multigene expression signatures for breast cancer prognosis, a group of biomarker genes whose combined expression pattern refines prognosis (usually with incremen-

tal value added to the use of standard clinico-pathological parameters). The research of prognostic signatures has resulted in many success stories [Sotiriou 2009]. Notably, as of today there exist at least six different prognostic multigene expression signatures commercially available to aid clinical decision making of breast cancer:[5]

- MammaPrint® (Agendia, Amsterdam, The Netherlands) [van 't Veer 2002] is a 70-gene microarray-based expression profile for stratifying breast cancer into high- or low-risk prognostic groups. As one of the earliest success stories, it was the first test approved by the Food and Drug Administration (FDA) in the United States and by regulators in the European Union as an adjunct prognostic assay for women patients satisfying criteria[6] including stage I/II, invasive infiltrating carcinoma, tumor size less than 5.0 cm, lymph node negative (or up to three lymph nodes positive).

- Prosigna® Breast Cancer Prognostic Gene Signature Assay or PAM50 (Nanostring Technologies, Seattle, WA, USA) [Parker 2009] is a 50-gene assay for classifying breast tumors into five intrinsic subtypes (luminal A, luminal B, HER2-enriched, basal-like, normal-like) that are prognostic independent of standard clinico-pathological parameters. It is the second FDA-approved test in the United States to estimate distant recurrence risk for stage I/II (including one to three positive nodes), ER-positive breast cancer in postmenopausal women treated with adjuvant endocrine therapy, and it also received clearance in the European Union.

- Oncotype DX® (Genomic Health, Redwood City, CA, USA) [Paik 2004] is a 21-gene signature for categorizing tamoxifen-treated breast cancer patients into groups of low-, intermediate- or high-risk recurrence. It is the most widely used prognostic assay for ER-positive cancers in the United States.

- MapQuant Dx™ Genomic Grade Index (Ipsogen, France) [Sotiriou 2006] is a microarray-based 97-gene assay for reclassifying histologically intermediate-grade ER-positive cancers into high or low molecular grade with significantly different prognosis.

- Breast Cancer Index℠ (BioTheranostics, San Diego, CA, USA) [Ma 2008] is comprised of two signatures, a 5-gene molecular grade index and the ratio of two independent biomarkers HOXB13:IL17BR, and can assess the risk of distant recurrence in ER-positive, lymph node-negative breast cancers.

- EndoPredict® (Sividon Diagnostics GmbH, Koln, Germany) [Filipits 2011] is a 11-gene signature for stratifying patients with ER-positive cancer into high or low risk of recurrence if treated with adjuvant endocrine therapy alone.

[5] See for reference `http://www.breastcancer.org/symptoms/testing/types`.

[6] Indications for ordering an assay can vary in accordance with the clearance issued by the country of application.

More details about these signatures are found in [Győrffy 2015]. Notably, another rather famous 76-gene signature (Veridex LLC, a Johnson & Johnson company, San Diego, CA, USA) [Wang 2005a] could be used to predict the development of distant metastases within 5 years in lymph node-negative primary breast cancer patients (irrespective of age and tumor size) who did not receive systemic treatment, which was later confirmed in multiple independent studies on patient data obtained from different institutions [Foekens 2006, Desmedt 2007, Zhang 2009].

1.3 Genomic Data Analysis: Topics, Prospects and Challenges

In order to study the substantial amount of genomic data available for medical research, the use of computational tools such as machine learning has become a popular trend [Barillot 2012]. In fact, machine learning is particularly suitable for analyzing genomic data by developing algorithms or building models to discover unseen patterns, identify complex relationships and predict for phenotypic phenomenon of interest. While genomic data analysis of cancer is a research field encompassing a broad range of topics, the present thesis is specifically devoted to breast cancer prognosis and related biomarker discovery.

Molecular Prognosis

In the language of machine learning, *cancer prognosis* is usually formulated as *predictive modeling* (or discriminative modeling succeeding supervised learning). In fact, an extensive body of findings in the genre of genomic data analysis are inferred from empirical evidence of relationship between the genomic features and the survival information collected over large population of patients. Given a set of m observations $\mathcal{D} := \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$, where $\mathbf{x}_i \in \mathcal{X}$ denotes the feature vector of the i-th sample, typically the expression measurements of n genes (or i-th row in Figure 1.1[7]) in gene expression data analysis when $\mathcal{X} = \mathbb{R}^n$, and $y_i \in \mathcal{Y}$ denotes the outcome of the i-th sample, typically the survival time when $\mathcal{Y} = \mathbb{R} \times \{0, 1\}$ of (positive) survival observation with a right-censoring flag, or the prognostic group when $\mathcal{Y} = \{1, \ldots, K\}$ of $K \geq 2$ groups categorized by thresholding the observed survival time, the objective is then to infer a predictive function $h : \mathcal{X} \to \mathcal{Y}$ which can then be used to predict survival risk or classify prognostic group for any new sample. These two learning tasks are termed respectively as *survival analysis* and *classification* in machine learning literature.

Survival analysis is generally referred to a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest, hereby referring to the survival time when $\mathcal{Y} = \mathbb{R} \times \{0, 1\}$. In clinical management of cancer, patients are usually followed for a specified time period and the focus is

[7] Probes are hybridization fragments of DNA, therefore probe-specific measurements in microarray data usually need post-processing to estimate gene-specific measurements.

on the time at which the event of interest occurs such as metastasis, recurrence or death. If the event had occurred during the follow-up, the survival time is documented by the observed time to event; if the event had not occurred by the end of the follow-up (or the patient dropped out of the study), the event had not yet been observed and the survival time is documented by the follow-up (or drop-out) time with a flag, meaning that survival time can only be considered at least as long as the duration of follow-up. A survival observation is called right-censored if it is incomplete as in the latter case. Survival time is therefore a variable consisting of two components: the documented survival time (usually measured in days) and a right-censoring flag indicating whether the survival is exact or lower-bounded, leading to $\mathcal{Y} = \mathbb{R} \times \{0, 1\}$ in survival analysis. A number of methods are available in literature to analyze the relationship of the feature vector with the survival time, among which two are worth special mention. The Kaplan-Meier method [Kaplan 1958] is a nonparametric estimator and graphical method of depicting survival probabilities as a function of time. It is widely used to obtain descriptive statistics for survival observations that can be further combined with statistical tests to compare the survival experience for two or more groups of patients[8]. The Cox proportional hazards model [Cox 1972] is a popular regression model for analyzing survival data that builds an easily interpretable model associating the relationship of the survival hazards to predictive features in order to describe the likely course of outcome. For a textbook-oriented overview of survival analysis, see [Hosmer 1999].

Classification is another classical topic in machine learning and statistics where the outcome variable belongs to one of a few predetermined categories, specifically $\mathcal{Y} == \{1, \ldots, K\}$ representing $K \geq 2$ prognostic groups. Based on their clinical records of survival time, cancer patients can be categorized into high-risk and low-risk (and sometimes a third intermediate-risk) groups typically by binarizing the continuous survival time at a 5-year threshold. In fact, deployment of cancer treatment usually relies on such manageable categorization of patients into prognostic groups. Compared to survival analysis, classification bypasses the difficulty in accurately depicting the course of survival outcome but instead seeks a coarse yet clinically meaningful description of survival outcome. Popular classification methods include Fisher's linear discriminant [Fisher 1936], logistic regression [Cox 1958], decision trees [Breiman 1984], Support Vector Machines [Cortes 1995], Random Forests [Breiman 2001], Gradient Boosting Machines [Friedman 2001], see [Hastie 2009] for details and many other algorithms for classification.

Biomarker Discovery

The predictive modeling framework discussed above assumes that a representation of all sample vectors consisting of n genomic features is already determined and will be included in building a predictive model. In the era where we have easy access to thousands up to millions of genomic features for a biological sample albeit most

[8] Patients are usually grouped by molecular subtypes typically by clustering approaches based on their genomic features.

of which can be irrelevant or redundant for the inference task under consideration, it is crucial to determine which features to be incorporated in the model, a question usually termed as *feature selection* in machine learning or *biomarker discovery* in computational biology. On one hand, inferring a predictive model with a large number of features from a relatively small number of samples, which is usually the case in biomedical applications, is essentially difficult from the viewpoint of statistical inference, a phenomenon referred to as *the curse of dimensionality*, which often leads to unreliable models that overfit the observed samples and generalize poorly when used to predict for future samples. Reducing the number of features representing each sample by selecting only a few important features has proven an efficient way to limit this difficulty.[9] On another hand, the identification of a few informative genomic features helps suggest discerning interpretation and key insights into molecular cancer biology. Further, a few identified biomarkers can facilitate the design of more affordable prognostic gene signatures as it is still cheaper and faster to measure the activity of a few targeted genes nowadays.

Many feature selection techniques exist and are organized into three categories, depending on how they are combined with the construction of the predictive model: filter methods, wrapper methods and embedded methods. (Univariate) filter methods select a list of relevant features from the entire feature set independent from the predictive models used, by assessing the relevance of each feature to the response of interest with an importance score, typically by applying some statistical test such as χ^2-test or calculating some information measure univariately such as Information Gain [Xing 2001], and removing those low-scoring ones. Being the computationally fastest methods, filter methods can easily scale to a large number of features and accommodate any predictive model, whereas they usually ignore the interaction between features and special attributes of the predictive model considered. Taking into account the dependencies between features and the hypothesis of the predictive model, wrapper methods aim to directly find the best combination of features by evaluating all possible feature subsets as input to the model and picking the one with which the resulting model performs the best. Due to the fact that the space of feature subsets grows exponentially with the number of features, exhaustive search over the full space of feature subsets is in general computationally impossible, and hence heuristic or greedy algorithms are often adopted to guide the search for a satisfactory candidate of feature subset. Popular wrapper methods include simulated annealing [Kirkpatrick 1983] and sequential elimination such as stepwise regression [Hocking 1976]. Embedded methods enable feature selection during the process of constructing a predictive model, and as these methods are usually tailored to each specific model utilized, they are therefore far less computationally intensive than wrapper methods. Popular embedded methods include a wealth of regularization methods such as the lasso [Hastie 2015] and recursive feature elimination

[9] Besides feature selection, another efficient approach of dimensionality reduction is via feature extraction such as principal component analysis. While feature selection finds a subset of informative features *as is* without altering the original representation of data, feature extraction transforms the data in the high-dimensional feature space to a space of lower dimension.

embedded in Support Vector Machines [Guyon 2002]. For an overview of feature selection methods, see [Guyon 2003, Li 2016] for an introductory review from the methodological viewpoint of machine learning and [Saeys 2007, Hira 2015] with a particular emphasis on applications in bioinformatics.

Prospects and Challenges

While survival analysis, classification and feature selection are themselves extensively studied and still active research areas of machine learning research, their applications in genomic data analysis are a particularly demanding task. In fact, it has been widely recognized as a challenging problem to extract potentially valuable information from genomic data for reasons of multiple folds. To start with, cancer is intrinsically a highly complex disease and consequently the heterogeneity underlying cancer patients renders inevitable obstacle in analyzing cancer data, in other words, high-throughput experimental data are noisy by nature leading to a decline in the informativeness of such data. In addition, from the viewpoint of machine learning, a relatively small number of clinical samples (typically at the scale of $10^2 \sim 10^3$) versus a large number of genomic features (typically at the scale of $10^3 \sim 10^6$) adds difficulty in making reliable inference from analyzing observed samples that could generalize well to future samples and in identifying prognostic biomarkers reusable for future patients. Another major concern specially regarding biomarker discovery is the *a posteriori* interpretation of the computational findings in terms of biological relevance to the mechanism of cancer. To address the challenges in genomic data analysis, there is a pressing need for bioinformatics-oriented methods built upon state-of-the-art machine learning algorithms as a stepping stone.

Despite the computational challenges confronted by machine learning applications in cancer prognosis, many success stories are prominent. For example, the above-mentioned PAM50 test, the 50-gene classifier for subtyping breast cancer, is constructed upon a learning algorithm called the nearest shrunken centroid method [Tibshirani 2002]. Another example comes from the *DREAM 7 — Sage Bionetworks–DREAM Breast Cancer Prognosis Challenge* [Margolin 2013], a competition-based crowd-source effort that systematically assessed and confirmed the potential of computational models designed to predict breast cancer survival by combining various types of molecular features with standard clinico-pathological parameters to improve prognosis performance (Figure 1.2) [Bilal 2013]. Notably, the best-performing model of the competition [Cheng 2013b] was built upon, in addition to clinico-pathological features, such molecular features called *attractor metagenes* that are pan-cancer signatures of coexpressed genes previously identified in rich gene expression datasets by an iterative attractor-finding algorithm [Cheng 2013a]. For a recent survey on machine learning applications in cancer prognosis, see [Kourou 2015]. Worth special mention are two lines of ideas to address the difficulty in cancer prognosis and biomarker discovery, which have primarily motivated the work presented in this thesis.

Since high-throughput high-dimensional gene expression data are often subject

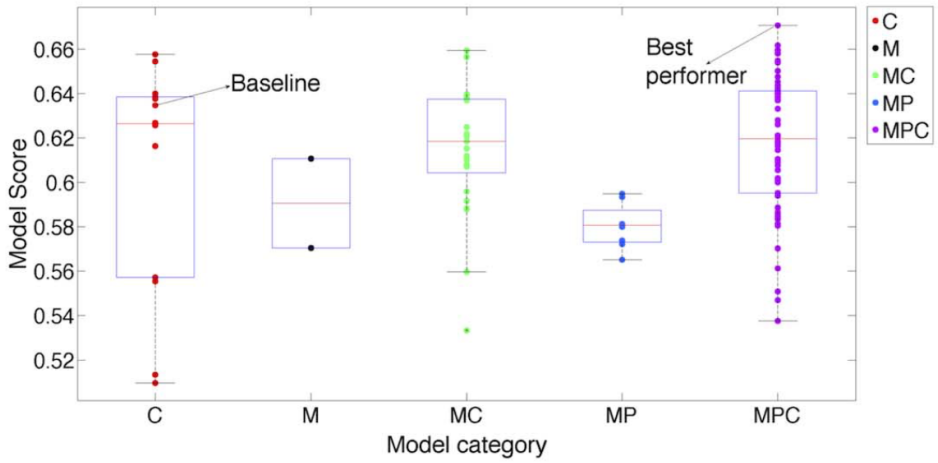


Figure 1.2: This figure from [Bilal 2013, Figure 2] illustrates that the best performer among submissions to the pilot competition uses a combination of clinical and molecular features that are deliberately selected subject to prior knowledge (the MPC category). Models submitted are categorized by the type of features they use: only clinical features (C), only molecular features (M), molecular and clinical features (MC), molecular features selected using prior knowledge (MP), molecular features selected using prior knowledge and clinical features (MPC).

to high measurement noise, the ranking of the expression levels of multiple genes are presumably more robust predictors, in the sense that they can be less sensitive to noise, than their real-valued measurements. This can be particularly beneficial in many biomedical applications when the informativeness (or signal-to-noise ratio) in data is low. Pioneering the exploration of these ideas is the top scoring pairs (TSP) [Geman 2004], an algorithm for classifying gene expression profiles by pairwise microarray comparison, together with successive extensions and further investigations by [Tan 2005, Xu 2005, Lin 2009]. These methods generate simple and accurate decision rules to discriminate cancer samples from normal ones based on the relative reversals of pairwise ordering comparing the expression of a few genes. However, when it comes to biomedical classification on difficult tasks such as cancer prognosis that usually involves the collaborative functional activities of a relatively large number of gene, the performance of TSP-family classifiers degrades drastically, probably due to the naively simple majority voting scheme adopted by those classifiers. In order to improve cancer outcome prediction, many studies employed TSP algorithm as a feature selection technique that is further embedded into more complex classification methods such as Support Vector Machines [Shi 2011] or decision trees [Czajkowski 2011] in microarry data analysis.

Cancer is a "network disease". In fact, it has already been quoted above that cancer is a genetic disease. As more and more cancer-related genes were identified and arranged into signaling pathways through which they act, it became apparent that these pathways are interconnected and present crossroads at differ-

ent levels [Vogelstein 2004], indicating that tumor progression is the consequence of network-level dysregulation of interactions between genes, RNAs, proteins and other molecules that control at least the hallmarks of cancer [Hanahan 2011]. Moreover, biological networks, including protein-protein interaction, coexpression and regulatory networks, or metabolic and signaling pathways, are a common way of depicting functional relationships between genes that have been accumulated from decades of biomedical research, and they can be potentially valuable when incorporated as domain-specific knowledge during the process of the computational analysis of genomic data so as to, for instance, improve stability and interpretability of biomarker discovery (Figure 1.3). Approaches to pathway and network analysis techniques range broadly, including gene set enrichment analysis that identifies genes of interest appearing in pathways more frequently than expected by chance [Subramanian 2005], network modeling that infers the activities and interactions of various genetic components in pathway or networks, see for instance [Tarca 2008, Drier 2013, Vandin 2011, Hidalgo 2017], network-guided predictive modeling that consults the structure of *a priori* known network and constrains the predictive modeling procedures discussed above so that the "ideal" model or biomarkers selected should be coherent with the network, see for instance [Li 2010, Rapaport 2007, Jacob 2009]. For a recent review of pathway and network analysis of cancer genomes, see [Creixell 2015] with a focus on approaches applied to somatic single nucleotide variants (SNVs) and altered RNA expression and [Azencott 2016] with a particular emphasis on biomarker discovery.

1.4 Contribution of the Thesis

The thesis work is conceived following the two lines of ideas intended to address two major questions from the methodological standpoint of machine learning: rank-based approaches for improved molecular prognosis and network-guided approaches for enhanced biomarker discovery. Furthermore, despite their biomedical application in cancer prognosis to which this thesis is largely devoted, the methodologies developed and investigated in this thesis, pertaining respectively to learning with rank data and learning on graphs, have a significant contribution to several branches of machine learning, concerning applications across but not limited to cancer biology and social choice theory. This thesis will be organized by projects, each presented in one chapter.

Rank-based Approaches for Improved Molecular Prognosis

The first line of ideas is to perform gene expression data analysis based on exploiting exclusively the ranking of the expression levels of multiple genes while their real-valued measurements are disregarded, which integrates the idea of relative reversals of pairwise ordering inherited from TSP-family classifiers in the paradigm of kernel learning. From the point view of machine learning, the problem reduces to the study of a particular type of structured data, specifically rankings. It is well-known that

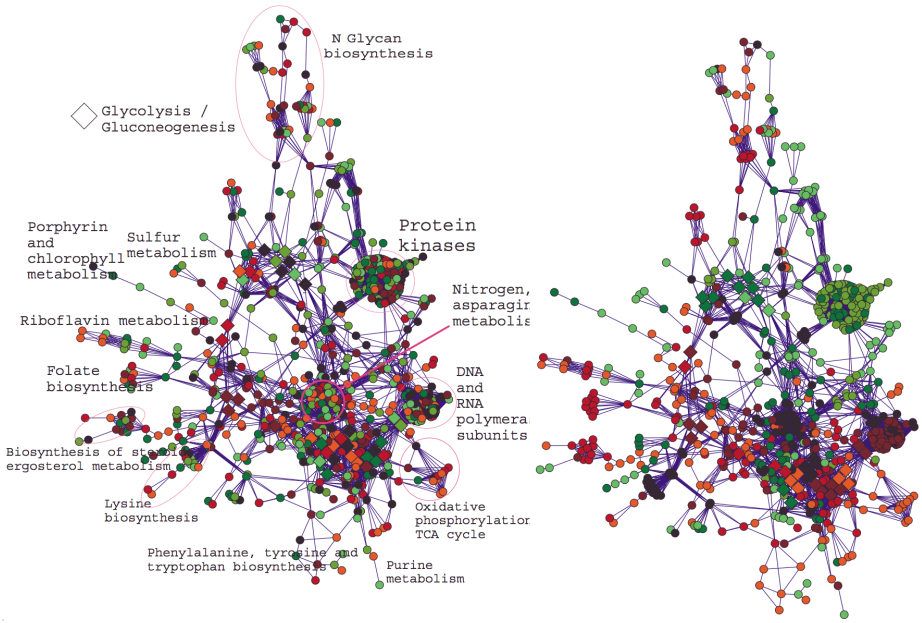


Figure 1.3: This figure from [Rapaport 2007, Figure 3] illustrates an example of metabolic pathways, mapped by coefficients of the decision function obtained by applying a network-free model (left) and a network-guided model (right) in color, positive in red and negative in green with intensities reflecting absolute values, where some large highly connected functional parts of the network with annotations such as proteinkinases and DNA and RNA polymerase subunits were identified by the network-guided model, rendering readily available interpretability of the involvement of the selected genes in cancer.

kernel methods have found many successful applications where the input data are discrete or structured including strings and graphs [Gärtner 2004]. The first project of my doctoral studies was focused on proposing computationally attractive kernels for rank data and applying kernel methods to problems involving rankings. Central to this work was the observation that the widely used Kendall tau correlation and the Mallows similarity measure are indeed positive definite kernels for total rankings. These kernels were further tailored to more complex types of rank data that prevail in real-world applications, especially uncertain rankings which are converted from real-valued vectors by keeping simply the relative ordering of the values of multiple features thereof. Thanks to these kernels, many off-the-shelf kernel machines are available to solve various problems at hand [Shawe-Taylor 2004, Schölkopf 2004]. It is worth special mention that, despite that the project was initially motivated by biomedical applications, the prospective contribution of this work concerns applications from many fields of machine learning pertaining to learning from rankings, or learning to rank. This study will be presented in Chapter 2.

The study of the Kendall kernel for rankings has paved an unprecedented way towards a deeper understanding of a classical problem called Kemeny aggregation [Kemeny 1959] from the field of social choice theory. Kemeny aggregation searches for a consensus ranking that best represents a collection of individual rankings in the sense that the sum of the Kendall tau distance between each ranking and the consensus is minimized. Although Kemeny aggregation is often considered to provide the "golden" solution among all ranking aggregation criteria, the Kemeny consensus is known to be NP-hard to find [Bartholdi III 1989]. Many tractable approximations to the Kemeny consensus have therefore been proposed and extensively studied, see for instance [Ali 2012]. Since the Kendall kernel derives from an inner product of a Euclidean space, the Kendall tau distance derives from a squared Euclidean distance. As a result, the combinatorial problem of Kemeny aggregation is endowed with an intuitive interpretation from a geometric point of view. Based on this observation, a tractable upper bound of the estimation error in terms of the distance between the exact Kemeny consensus and an approximate solution is established. This upper bound requires little assumption on the approximation procedure or the collection of rankings to aggregate. Due to its remote connection to cancer prognosis or the primary objective of this thesis, this study will be presented in Appendix A.

Network-guided Approaches for Enhanced Biomarker Discovery

The second line of ideas of performing genomic data analysis for cancer prognosis is to consult biological networks as prior knowledge in order to improve the selection efficacy of molecular features. Two projects were initiated on network-guided analysis of genomic data for suggesting candidate biomarkers related to cancer prognosis.

In one project, we focused on the study of structured regularization in generalized linear models [McCullagh 1989] and the Cox proportional hazards model [Cox 1972] where the regularization method is designed so that genes closer on

the biological network are encouraged to be selected simultaneously as candidate biomarkers. In fact, in order to achieve simultaneous modularity and sparsity coherent with the presumed network structure, a popular method called network-based wavelet smoothing has been successfully applied in many applications from the field of signal processing [Shuman 2013]. Therefore, we were intrigued to investigate the potential of this method in survival analysis of breast cancer with a gene expression dataset guided by a protein-protein interaction network, albeit the methodology is generally applicable to various types of genomic data and biological networks. In particular, the method allows to designate genes as candidates for biomarkers in form of gene modules with intra-collaborative functionality rendering readily interpretable insights related to cancer survival. Numerical results demonstrated that, compared to several network-free and some established network-based regularization methods, network-based wavelet smoothing was able to improve the selection efficacy of genes related to cancer survival in terms of stability, connectivity and interpretability, while achieving competitive performance of survival risk prediction. This study will be presented in Chapter 3.

In another project, we focused on a particular type of biological network namely signaling pathway network. Based on a modeling framework of cell signaling proposed by [Hidalgo 2017], gene expression profiles can be translated into personalized profiles of signaling pathway activities by integrating known signaling pathways. When gene-level profiles are replaced by these derived pathway-level profiles as input to many off-the-shelf computational tools, a simple scheme emerges where gene-level analysis is easily promoted to pathway-level analysis of gene expression data. The advantage is remarkable in that, when combined with feature selection methods, the proposed scheme enables direct identification of pathway-level mechanistic signatures as an alternative to conventional gene-based signatures, which provides more informative insights into the cellular functions and biological processes involved in cancer. This study will be presented in Chapter 4.

Other Contributions

During the course of my doctoral studies, I have undertaken some other projects as well. In 2013, Elsa Bernard, Erwan Scornet, Véronique Stoven, Thomas Walter, Jean-Philippe Vert from our laboratory and I participated in the *DREAM 8 NIEHS–NCATS–UNC Toxicogenetics Challenge*, an international bioinformatics competition where participants were asked to predict the response of human cell lines exposed to various toxic chemical compounds based on the molecular characterization of chemicals and the transcriptome of cell lines. Finally our team won second place with a kernel bilinear regression model. Oral presentation was accepted to *NIPS Workshop on Machine Learning in Computational Biology (MLCB)* and later invited to *RECOMB Conference on Regulatory and Systems Genomics*. This work has been accepted for publication in [Bernard 2017] and it has also been published as part of the crowd-source collaboration as a result of the competition in [Eduati 2015], whereas this work will be excluded from this thesis due to the fact

that it was not well polished by the time of drafting the manuscript.

During my internship at Roche Diagnostics GmbH, Penzberg, Germany, I worked on failure state prediction for automated analyzers for analyzing biological samples in collaboration with Jean-Philippe Vert, Fabian Heinemann, Sven Dahlmanns and Stefan Kobel, and a European patent regarding the application was filed by Roche Diagnostics GmbH, F. Hoffmann–La Roche AG in December 2016 and is currently pending approval [Jiao 2016c]. Due to corporate confidentiality policies, this study will not be included in this thesis.

Chapter 2

The Kendall and Mallows Kernels for Permutations

Publication and Dissemination: *The work in this chapter has been published as joint work with Jean-Philippe Vert in [Jiao 2015], orally presented at ICML 2015 and accepted for publication in [Jiao 2017b].*

Abstract: *We show that the widely used Kendall tau correlation coefficient and the related Mallows kernel are positive definite kernels for permutations. They offer computationally attractive alternatives to more complex kernels on the symmetric group to learn from rankings, or learn to rank. We show how to extend these kernels to partial rankings, multivariate rankings and uncertain rankings. Examples are presented on how to formulate typical problems of learning from rankings such that they can be solved with state-of-the-art kernel algorithms. We demonstrate promising results on clustering heterogeneous rank data and high-dimensional classification problems in biomedical applications.*

Résumé : *Nous prouvons ici que le tau de Kendall, un coefficient de corrélation populaire, et le noyau de Mallows sont deux noyaux définis positifs pour les permutations. Ils offrent des alternatives computationnellement intéressantes comparés à d'autres noyaux plus complexes sur le groupe symétrique pour apprendre à partir de données de classements, ou apprendre à classer. Nous montrons comment étendre ces noyaux à des classements partiels, des classements multivariés et des classements incertains. Nous présentons des exemples sur comment formuler des problèmes classiques pour apprendre à partir de données de classements afin qu'ils puissent être résolus par les algorithmes à noyaux de l'état de l'art. Nous obtenons des résultats prometteurs sur le regroupement de données hétérogènes de classements et sur les problèmes de classification en grande dimension dans les applications biomédicales.*

2.1 Introduction

A permutation is a 1-to-1 mapping from a finite set into itself. Assuming the finite set is ordered, a permutation can equivalently be represented by a total ranking of the elements of the set. Permutations are ubiquitous in many applications involving preferences, rankings or matching, such as modeling and analyzing data describing the preferences or votes of a population [Diaconis 1988, Marden 1996], learning or tracking correspondences between sets of objects [Huang 2009], or estimating a consensus ranking that best represents a collection of individual rankings [Dwork 2001, Ailon 2008, Arrow 2012]. Another potentially rich source of rank data comes from real-valued vectors in which the relative ordering of the values of multiple features is more important than their absolute magnitude. For example, in the case of high-dimensional gene expression data, [Geman 2004] showed that simple classifiers based on binary comparisons between the expression of different genes in a sample show competitive prediction accuracy with much more complex classifiers built on quantitative gene expression levels, a line of thoughts that have been further investigated by [Tan 2005, Xu 2005, Lin 2009]. In these approaches, an n-dimensional feature vector is first transformed into a vector of ranks by sorting its entries, which are presented as input to training a classifier.

Working with permutations is, however, computationally challenging. There are $n!$ permutations over n items, suggesting that various simplifications or approximations are necessary in pursuit of efficient algorithms to analyze or learn permutations. Such simplifications include for example, reducing ranks to a series of binary decisions [Ailon 2008, Balcan 2008], or estimating a parametric distribution over permutations [Lebanon 2008, Helmbold 2009, Huang 2009].

Kernel algorithms form a class of methods that have been proved successful in numerous applications and enjoy great popularity in the machine learning community [Cortes 1995, Vapnik 1998, Schölkopf 2002, Shawe-Taylor 2004]. The essential idea behind these methods is to define a symmetric positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ over an input space $\mathcal{X}$, which expresses our belief of similarities between pairs of points in the input space, and which implicitly defines an embedding $\Phi : \mathcal{X} \to \mathcal{F}$ of the input space $\mathcal{X}$ to a Hilbert space $\mathcal{F}$ in which the kernel becomes an inner product:

$$\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \quad K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\mathcal{F}} .$$

Key to kernel methods is the fact that kernel algorithms only manipulate data through evaluation of the kernel function, allowing to work implicitly in the potentially high- or even infinite-dimensional space $\mathcal{F}$. This *kernel trick* is particularly interesting when $K(\mathbf{x}, \mathbf{x}')$ is inexpensive to evaluate, compared to $\Phi(\mathbf{x})$ and $\Phi(\mathbf{x}')$. In particular, kernel methods have found many applications where the input data are discrete or structured, such as strings or graphs, thanks to the development of numerous kernels for these data [Haussler 1999, Kashima 2003, Gärtner 2004, Shawe-Taylor 2004, Schölkopf 2004, Vishwanathan 2009].

In this context, it is surprising that relatively little attention has been paid to the problem of defining positive definite kernels between permutations, which could pave the way to benefiting from computationally efficient kernel methods in problems involving permutations. A notable exception is the work of [Kondor 2008, Kondor 2010], who exploit the fact that the right-invariant positive definite kernels on the symmetric group are fully characterized by Bochner's theorem [Kondor 2008, Fukumizu 2008]. They derive interesting kernels, such as a diffusion kernel for rankings or partial rankings, and demonstrate that kernel methods are flexible to handle rank data of diverse types. However, the kernels proposed in their papers have typically a computational complexity that grows exponentially with the number of items to rank, and remain prohibitive to compute for more than a few items.

Here we study new computationally attractive positive definite kernels for permutations and rankings. Our main contribution is to show that two widely-used and computationally efficient measures of similarity between permutations, the Kendall tau correlation coefficient and the Mallows kernel, are positive definite. Although these measures compare two permutations of n items in terms of $\binom{n}{2}$ pairwise comparisons, they can be computed in $O(n \log n)$, which allows us to use kernel methods for problems involving rank data over a large number of items. We show how these kernels can be extended to partial rankings, multivariate rankings, and uncertain rankings which are particularly relevant when the rankings are obtained by sorting a real-valued vector where ties or almost-ties occur. We illustrate the benefit of kernel learning with the new kernels on two applications, one concerning the unsupervised clustering of rank data with kernel k-means, one focusing on the supervised classification of genomic data with Support Vector Machines (SVMs), reaching in both cases state-of-the-art performances.

The chapter is organized as follows. In Section 2.2, we prove our main theorem showing that the Kendall and Mallows kernels are positive definite. We extend them to partial, multivariate and uncertain rankings respectively in Section 2.3.1, 2.3.2 and 2.3.3. We highlight the relation to the diffusion kernel of [Kondor 2010] in Section 2.4. Finally we illustrate the relevance of kernel methods for unsupervised (Section 2.5) and supervised (Section 2.6) tasks. Data and `R` code for reproducing the experiments in this chapter are available via `https://github.com/YunlongJiao/kendallkernel_demo`. I have also developed `kernrank`, an `R` package implementing kernel functions and kernel methods for analyzing rank data [Jiao 2016a].

2.2 The Kendall and Mallows Kernels for Permutations

Let us first fix some notations. Given a list of n items $\{x_1, x_2, \ldots, x_n\}$, a *total ranking* is a strict ordering on the n items of the form

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_n} , \tag{2.1}$$

where $\{i_1, \ldots, i_n\}$ are distinct indices in $\{1, 2, \ldots, n\} =: [\![n]\!]$. A *permutation* is a 1-to-1 mapping from a finite set into itself, i.e., $\sigma : [\![n]\!] \to [\![n]\!]$ such that $\sigma(i) \neq \sigma(j)$

for $i \neq j$. Each total ranking can be equivalently represented by a permutation σ in the sense that $\sigma(i) = j$ indicates that a ranker assigns rank j to item i where higher rank coincides higher preference. For example, the ranking $x_2 \succ x_4 \succ x_3 \succ x_1$ is associated to the permutation $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 2 & 3 \end{pmatrix}$, meaning $\sigma(1) = 1$, $\sigma(2) = 4$, $\sigma(3) = 2$ and $\sigma(4) = 3$. There are $n!$ different total rankings, and we denote by $\mathbb{S}_n$ the set of all permutations over n items. Endowed with the composition operation $(\sigma_1\sigma_2)(i) = \sigma_1(\sigma_2(i))$, $\mathbb{S}_n$ is a group called the *symmetric group*.

Given two permutations $\sigma, \sigma' \in \mathbb{S}_n$, the number of concordant and discordant pairs between σ and σ' are respectively

$$n_c(\sigma, \sigma') = \sum_{i<j} \left[\mathbb{1}_{\{\sigma(i)<\sigma(j)\}} \mathbb{1}_{\{\sigma'(i)<\sigma'(j)\}} + \mathbb{1}_{\{\sigma(i)>\sigma(j)\}} \mathbb{1}_{\{\sigma'(i)>\sigma'(j)\}} \right],$$
$$n_d(\sigma, \sigma') = \sum_{i<j} \left[\mathbb{1}_{\{\sigma(i)<\sigma(j)\}} \mathbb{1}_{\{\sigma'(i)>\sigma'(j)\}} + \mathbb{1}_{\{\sigma(i)>\sigma(j)\}} \mathbb{1}_{\{\sigma'(i)<\sigma'(j)\}} \right].$$

As their names suggest, $n_c(\sigma, \sigma')$ and $n_d(\sigma, \sigma')$ count how many pairs of items are respectively in the same or opposite order in the two rankings σ and σ'. n_d is frequently used as a distance between permutations, often under the name *Kendall tau distance*, and underlies two popular similarity measures between permutations:

- The *Mallows kernel* defined for any $\lambda \geq 0$ by

$$K_M^\lambda(\sigma, \sigma') = e^{-\lambda n_d(\sigma, \sigma')}, \tag{2.2}$$

- The *Kendall kernel* defined as

$$K_\tau(\sigma, \sigma') = \frac{n_c(\sigma, \sigma') - n_d(\sigma, \sigma')}{\binom{n}{2}}. \tag{2.3}$$

The Mallows kernel plays a role on the symmetric group similar to the Gaussian kernel on Euclidean space, for example for statistical modeling of permutations [Mallows 1957, Critchlow 1985, Fligner 1986, Meilă 2007] or nonparametric smoothing [Lebanon 2008], and the Kendall kernel [Kendall 1938, Kendall 1948] is probably the most widely-used measure of rank correlation coefficient. In spite of their pervasiveness, to the best of our knowledge the following property has been overlooked:

Theorem 2.1. *The Mallows kernel K_M^λ, for any $\lambda \geq 0$, and the Kendall kernel K_τ are positive definite.*

Proof. Consider the Kendall mapping $\Phi : \mathbb{S}_n \to \mathbb{R}^{\binom{n}{2}}$ defined by

$$\Phi(\sigma) = \left(\frac{1}{\sqrt{\binom{n}{2}}} (\mathbb{1}_{\{\sigma(i)>\sigma(j)\}} - \mathbb{1}_{\{\sigma(i)<\sigma(j)\}}) \right)_{1 \leq i < j \leq n}.$$

Then one immediately sees that, for any $\sigma, \sigma' \in \mathbb{S}_n$,

$$K_\tau(\sigma, \sigma') = \Phi(\sigma)^\top \Phi(\sigma'),$$

showing that K_τ is positive definite, and that

$$\begin{aligned}\|\Phi(\sigma) - \Phi(\sigma')\|^2 &= K_\tau(\sigma,\sigma) + K_\tau(\sigma',\sigma') - 2K_\tau(\sigma,\sigma') \\ &= 1 + 1 - 2\Big(\frac{n_c(\sigma,\sigma') - n_d(\sigma,\sigma')}{\binom{n}{2}}\Big) \\ &= \frac{4}{\binom{n}{2}} n_d(\sigma,\sigma')\,,\end{aligned} \tag{2.4}$$

showing that $-n_d$ is conditionally positive definite and therefore that K_M^λ is positive definite for all $\lambda \geq 0$ [Schoenberg 1938]. □

Although the Kendall and Mallows kernels correspond respectively to a linear and Gaussian kernel on an $\binom{n}{2}$-dimensional embedding of $\mathbb{S}_n$ such that they can in particular be computed in $O(n^2)$ time by a naive implementation of pair-by-pair comparison, it is interesting to notice that more efficient algorithms based on divide-and-conquer strategy can significantly speed up the computation, up to $O(n \log n)$ using a technique based on Merge Sort algorithm [Knight 1966]. Computing in $O(n \log n)$ a kernel corresponding to an $O(n^2)$-dimensional embedding of $\mathbb{S}_n$ is a typical example of the kernel trick, which allows to scale kernel methods to larger values of n than what would be possible for methods working with the explicit embedding.

2.3 Extensions of the Kendall Kernel to Rank Data

2.3.1 Extension to Partial Rankings

In this section we show how the Kendall and Mallows kernels can efficiently be adapted to partial rankings, a situation frequently encountered in practice. For example, in a movie recommender system, each user only grades a few movies that he has watched based on personal interest. As another example, in a chess tournament, each game results in a relative ordering between two contestants, and one would typically like to find a single ranking of all players that globally best represents the large collection of binary outcomes.

As opposed to a total ranking (2.1), *partial rankings* arise in diverse form which can be generally described by

$$X_1 \succ X_2 \succ \cdots \succ X_k,$$

where $X_1, \ldots, X_k$ are k disjoint subsets of n items $\{x_1, \ldots, x_n\}$. For example, $\{x_2, x_4\} \succ x_6 \succ \{x_3, x_8\}$ in a social survey could represent the fact that items 2 and 4 are ranked higher by an interviewee than item 6, which itself is ranked higher than items 3 and 8. Note that it is uninformative of the relative order between items 2 and 4, and of how item 1 is rated. For ease of analysis, a partial ranking is often associated with a subset $R \subset \mathbb{S}_n$ of permutations which are compatible with all partial orders described by the partial ranking. In this study, two particularly interesting types are:

(i) Interleaving partial rankings. Such a partial ranking is of the form

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k}, \quad k \leq n,$$

where we have a total ranking for k out of n items. This type of partial ranking is frequently encountered in real life, for example in a social survey an interviewer is inexperienced to rank all items listed so that there exist interleaved inaccessible values. The interleaving partial ranking corresponds to the set of permutations compatible with it:

$$A_{i_1,\ldots,i_k} = \{\sigma \in \mathbb{S}_n | \sigma(i_a) > \sigma(i_b) \text{ if } a < b, a, b \in [1, k]\}. \tag{2.5}$$

(ii) Top-k partial rankings. Such a partial ranking is of the form

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k} \succ X_{\text{rest}}, \quad k \leq n,$$

where we have a total ranking for k out of n items and also know that these k items are ranked higher than all the other items. For example, the top k hits returned by a search engine leads to a top k partial ranking; under a voting system in election, voters express their vote by ranking some (or all) of the candidates in order of preference. The top-k partial ranking corresponds to the set of compatible permutations:

$$B_{i_1,\ldots,i_k} = \{\sigma \in \mathbb{S}_n | \sigma(i_a) = n + 1 - a, a \in [1, k]\}. \tag{2.6}$$

To extend any kernel K over $\mathbb{S}_n$ to a kernel over the set of partial rankings, we propose to represent a partial ranking by its compatible subset $R \subset \mathbb{S}_n$ of permutations, and define a kernel between two partial rankings R and R' by adopting the *convolution kernel*, written with a slight abuse of notations as

$$K(R, R') = \frac{1}{|R||R'|} \sum_{\sigma \in R} \sum_{\sigma' \in R'} K(\sigma, \sigma'). \tag{2.7}$$

As a convolution kernel, it is positive definite as long as K is positive definite [Haussler 1999]. However, a naive implementation to compute (2.7) typically requires $O((n-k)!(n-k')!)$ operations when the number of observed items in partial rankings R, R' is respectively $k, k' < n$, which can quickly become prohibitive. Fortunately Theorem 2.2 guarantees that we can circumvent the computational burden of naively implementing (2.7) with the Kendall kernel K_τ on at least the two particular cases of partial rankings (2.5) or (2.6).

Theorem 2.2. *The Kendall kernel K_τ between two interleaving partial rankings of respectively k and m observed items, or between a top-k partial ranking and a top-m partial ranking, of form (2.7) can be computed in $O(k \log k + m \log m)$ operations.*

Proof. The proof is constructive. We show here explicitly how to compute the Kendall kernel between two interleaving partial rankings while the idea remains similar for the case of top-k partial rankings. Denote by $[\![n]\!]$ the item set to be

ranked and by $A_{i_1,\ldots,i_k}, A_{j_1,\ldots,j_m} \subset \mathbb{S}_n$ two interleaving partial rankings of size k, m respectively, whose subsets of item indices are denoted by $I := \{i_1, \ldots, i_k\}$ and $J := \{j_1, \ldots, j_m\}$. We will lighten the notation by writing $A_I := A_{i_1,\ldots,i_k}$ and $A_J := A_{j_1,\ldots,j_m}$ and recall that by definition,

$$A_I = \{\pi \in \mathbb{S}_n | \pi(i_a) > \pi(i_b) \text{ if } a < b, a, b \in [1, k]\},$$
$$A_J = \{\pi' \in \mathbb{S}_n | \pi'(j_a) > \pi'(j_b) \text{ if } a < b, a, b \in [1, m]\}$$

are subsets of $\mathbb{S}_n$ compatible with the two partial rankings respectively. In particular, $|A_I| = n!/k!$ and $|A_J| = n!/m!$. Note that every item that does not appear in the partial ranking corresponding to A_I (or A_J) can be interleaved at any possible order with the other items for some permutation in that set.

Key observation to our proof is the "symmetry" of A_I (or A_J) in the sense that (i) for every item pair $\{i, j\}$ such that $i, j \in I$, all permutations in A_I are identical on the relative order of items i and j; (ii) for every item pair $\{i, j\}$ such that $i, j \in I^{\complement}$, there exists a unique permutation $\rho = (i, j) \circ \pi \in A_I$ for each $\pi \in A_I$ by swapping the ranks of items i, j in π such that $(\pi(i) - \pi(j))(\rho(i) - \rho(j)) < 0$ and ρ is identical with π on the absolute ranks of all the other items.

By the definition of convolution kernel and Theorem 2.1, we have

$$\begin{aligned} K_\tau(A_I, A_J) &= \frac{1}{|A_I||A_J|} \sum_{\substack{\pi \in A_I \\ \pi' \in A_J}} \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \operatorname{sgn}(\pi(i) - \pi(j)) \operatorname{sgn}(\pi'(i) - \pi'(j)) \\ &= \sum_{1 \le i < j \le n} \frac{k!m!}{(n!)^2 \binom{n}{2}} \sum_{\substack{\pi \in A_I \\ \pi' \in A_J}} \operatorname{sgn}(\pi(i) - \pi(j)) \operatorname{sgn}(\pi'(i) - \pi'(j)) . \quad (2.8) \end{aligned}$$

As we will always regard the item set $[\![n]\!]$ as the universe, we will write the complement of set $S \subset [\![n]\!]$ as $S^{\complement} := [\![n]\!] \setminus S$. Since the item set can be divided into four disjoint subsets that are $[\![n]\!] = (I \cap J) \sqcup (I \setminus J) \sqcup (J \setminus I) \sqcup (I \cup J)^{\complement}$, any (unordered) item pair $\{i, j\}$ can be categorized uniquely into one out of ten cases:

1. both items in $I \cap J$.
2. one item in $I \cap J$, the other in $I \setminus J$.
3. one item in $I \cap J$, the other in $J \setminus I$.
4. one item in $I \cap J$, the other in $(I \cup J)^{\complement}$.
5. one item in $I \setminus J$, the other in $J \setminus I$.
6. both items in $I \setminus J$.
7. both items in $J \setminus I$.
8. both items in $(I \cup J)^{\complement}$.
9. one item in $I \setminus J$, the other in $(I \cup J)^{\complement}$.

10. one item in $J \setminus I$, the other in $(I \cup J)^{\complement}$.

Now we can split and case-by-case regroup the additive terms in (2.8) into ten parts. We denote by s_1 to s_{10} the subtotal corresponding to cases 1 to 10, i.e.,

$$K_\tau(A_I, A_J) = \sum_{l=1}^{10} s_l := \sum_{l=1}^{10} \Big\{ \sum_{\substack{\{i,j\} \text{ in} \\ \text{case } l}} \frac{k!m!}{(n!)^2 \binom{n}{2}} \times \sum_{\substack{\pi \in A_I \\ \pi' \in A_J}} \operatorname{sgn}(\pi(i) - \pi(j)) \operatorname{sgn}(\pi'(i) - \pi'(j)) \Big\} .$$

It is straightforward to see that s_6 to s_{10} are all equal to 0 due to the symmetry of A_I and/or A_J. For example for every item pair $\{i, j\}$ in case 6, since both items i and j appear in I, their relative order is fixed in the sense that $\operatorname{sgn}(\pi(i) - \pi(j))$ remains constant for all $\pi \in A_I$; since both items are absent from J, we can pair up permutations $\pi', \rho' \in A_J$ such that $\operatorname{sgn}(\pi'(i) - \pi'(j)) = -\operatorname{sgn}(\rho'(i) - \rho'(j))$. As a result all additive terms in s_6 cancel out each other and thus $s_6 = 0$.

Now we will take efforts to compute s_1 to s_5. For every item pair $\{i, j\}$ in case 1 such that $i, j \in I \cap J$, since $i, j \in I$, their relative order remains unchanged for all $\pi \in A_I$ and let us denote by $\tau \in \mathbb{S}_{|I \cap J|}$ the total ranking of the observed items indexed by $I \cap J$ with respect to A_I. Since also $i, j \in J$, we can denote by $\tau' \in \mathbb{S}_{|I \cap J|}$ the total ranking of the observed items indexed by $I \cap J$ with respect to A_J. Therefore we have

$$\begin{aligned} s_1 &= \sum_{\substack{1 \le i < j \le n \\ i,j \in I \cap J}} \frac{k!m!}{(n!)^2 \binom{n}{2}} \sum_{\substack{\pi \in A_I \\ \pi' \in A_J}} \operatorname{sgn}(\pi(i) - \pi(j)) \operatorname{sgn}(\pi'(i) - \pi'(j)) \\ &= \frac{1}{\binom{n}{2}} \sum_{\substack{1 \le i < j \le n \\ i,j \in I \cap J}} \operatorname{sgn}(\tau(i) - \tau(j)) \operatorname{sgn}(\tau'(i) - \tau'(j)) \\ &= \frac{\binom{|I \cap J|}{2}}{\binom{n}{2}} K_\tau(\tau, \tau') , \end{aligned}$$

where the last line is by the definition of Kendall kernel between τ and τ' on the common items in $I \cap J$.

For every item pair $\{i, j\}$ in case 2, we may assume without loss of generality that $i \in I \cap J, j \in I \setminus J$, or equivalently $i, j \in I$ and $i \in J, j \notin J$. The relative order of i, j in $\pi \in A_I$ is thus determined by τ but not fixed for all $\pi' \in A_J$. Let us denote by $\sigma \in \mathbb{S}_k$ the total ranking corresponding to the k observed items in A_I and by $\sigma' \in \mathbb{S}_m$ the total ranking of the m observed items in A_J. In fact, there are $(m + 1)$ possible positions for j to interleave in some $\pi' \in A_J$ and the number of

positions with a lower relative order of j to i is $\sigma'(i)$. Therefore we have

$$\begin{aligned}
s_2 =& \sum_{\substack{i\in I\cap J\\ j\in I\setminus J}} \frac{k!m!}{(n!)^2\binom{n}{2}} \sum_{\substack{\pi\in A_I\\ \pi'\in A_J}} \operatorname{sgn}(\pi(i)-\pi(j))\operatorname{sgn}(\pi'(i)-\pi'(j)) \\
=& \frac{1}{\binom{n}{2}} \sum_{\substack{i\in I\cap J\\ j\in I\setminus J}} \operatorname{sgn}(\tau(i)-\tau(j)) \frac{m!}{n!} \sum_{\pi'\in A_J} \operatorname{sgn}(\pi'(i)-\pi'(j)) \\
=& \frac{1}{\binom{n}{2}} \sum_{i\in I\cap J} \sum_{j\in I\setminus J} \Big\{ \operatorname{sgn}(\tau(i)-\tau(j)) \frac{m!}{n!} \\
& \qquad \times \frac{n!}{(m+1)!} \left[\sigma'(i) - ((m+1)-\sigma'(i))\right] \Big\} \\
=& \frac{1}{\binom{n}{2}(m+1)} \sum_{i\in I\cap J} \left[2\sigma'(i)-m-1\right] \sum_{j\in I\setminus J} \operatorname{sgn}(\tau(i)-\tau(j)) \\
=& \frac{1}{\binom{n}{2}(m+1)} \sum_{i\in I\cap J} \Big\{ \left[2\sigma'(i)-m-1\right] \\
& \qquad \times \left[2(\sigma(i)-\tau(i)) - k + |I\cap J|\right] \Big\},
\end{aligned}$$

where the last line concludes from basic deductive calculation. Similarly we have for case 3,

$$\begin{aligned}
s_3 = \frac{1}{\binom{n}{2}(k+1)} \sum_{i\in I\cap J} \Big\{ & \left[2\sigma(i)-k-1\right] \\
& \times \left[2(\sigma'(i)-\tau'(i)) - m + |I\cap J|\right] \Big\}.
\end{aligned}$$

For every item pair $\{i, j\}$ in case 4, we may assume without loss of generality that $i \in I \cap J, j \in (I \cup J)^{\complement}$. As j is absent from I (or J respectively), there are $(k+1)$ (or $(m+1)$ resp.) possible positions for j to interleave in some $\pi \in A_I$ (or $\pi' \in A_J$ resp.) and the number of positions with a lower relative order of j to i is $\sigma(i)$ (or $\sigma'(i)$ resp.). The times we get $(\pi(i)-\pi(j))(\pi'(i)-\pi'(j)) > 0$ for all possible interleaved positions of j in some $\pi \in A_I, \pi' \in A_J$ is in total $[\sigma(i)\sigma'(i) + (k+1-\sigma(i))(m+1-\sigma'(i))]$, and the times we get $(\pi(i)-\pi(j))(\pi'(i)-\pi'(j)) < 0$

is in total $[\sigma(i)(m+1-\sigma'(i)) + \sigma'(i)(k+1-\sigma(i))]$. Therefore we have

$$\begin{aligned}
s_4 =& \sum_{\substack{i\in I\cap J \\ j\in (I\cup J)^{\complement}}} \frac{k!m!}{(n!)^2\binom{n}{2}} \sum_{\substack{\pi\in A_I \\ \pi'\in A_J}} \operatorname{sgn}(\pi(i)-\pi(j))\operatorname{sgn}(\pi'(i)-\pi'(j)) \\
=& \sum_{i\in I\cap J} \frac{k!m!}{(n!)^2\binom{n}{2}} \frac{(n!)^2}{(k+1)!(m+1)!} |(I\cup J)^{\complement}| \\
&\times \Big\{ \big[\sigma(i)\sigma'(i) + (k+1-\sigma(i))(m+1-\sigma'(i))\big] \\
&\quad - \big[\sigma(i)(m+1-\sigma'(i)) + \sigma'(i)(k+1-\sigma(i))\big] \Big\} \\
=& \frac{|(I\cup J)^{\complement}|}{\binom{n}{2}(k+1)(m+1)} \\
&\times \sum_{i\in I\cap J} \big[2\sigma(i)-k-1\big]\big[2\sigma'(i)-m-1\big] .
\end{aligned}$$

For case 5, similar derivation (as case 4) with interleaving i in A_J and interleaving j in A_I leads to

$$\begin{aligned}
s_5 =& \sum_{\substack{i\in I\setminus J \\ j\in J\setminus I}} \frac{k!m!}{(n!)^2\binom{n}{2}} \sum_{\substack{\pi\in A_I \\ \pi'\in A_J}} \operatorname{sgn}(\pi(i)-\pi(j))\operatorname{sgn}(\pi'(i)-\pi'(j)) \\
=& \sum_{i\in I\setminus J} \sum_{j\in J\setminus I} \frac{k!m!}{(n!)^2\binom{n}{2}} \frac{(n!)^2}{(k+1)!(m+1)!} \\
&\times \Big\{ \big[\sigma(i)(m+1-\sigma'(j)) + \sigma'(j)(k+1-\sigma(i))\big] \\
&\quad - \big[\sigma(i)\sigma'(j) + (k+1-\sigma(i))(m+1-\sigma'(j))\big] \Big\} \\
=& \frac{-1}{\binom{n}{2}(k+1)(m+1)} \\
&\times \sum_{i\in I\setminus J} \big[2\sigma(i)-k-1\big] \sum_{j\in J\setminus I} \big[2\sigma'(j)-m-1\big] .
\end{aligned}$$

Finally $K_\tau(A_{i_1,\ldots,i_k}, A_{j_1,\ldots,j_m}) = s_1 + s_2 + s_3 + s_4 + s_5$ concludes the proof. The algorithms are summarized in Algorithm 2.1 for interleaving partial rankings and Algorithm 2.2 for top-k rankings. Note that in both algorithms, the first step is the computationally most intensive one, where we need to identify the total ranking restricted to the items present in the partial rankings. This can be achieved by any sorting algorithm, leading the algorithms to a time complexity $O(k\log k + m\log m)$. □

Note that the convolution kernel (2.7) taking the Mallows kernel K_M^λ is not straightforward to evaluate, which will be further discussed in Section 2.4. However, since we have extended the Kendall kernel to partial rankings, an exponential kernel can be constructed trivially following (2.4), for which the computation remains just as simple as the extended Kendall kernel. Since this technique also applies in following sections, we focus mainly on extending Kendall kernel henceforth.

Algorithm 2.1 Kendall kernel for two interleaving partial rankings.

Input: two partial rankings $A_{i_1,\ldots,i_k}, A_{j_1,\ldots,j_m} \subset \mathbb{S}_n$, corresponding to subsets of item indices $I := \{i_1,\ldots,i_k\}$ and $J := \{j_1,\ldots,j_m\}$.

1: Let $\sigma \in \mathbb{S}_k$ be the total ranking corresponding to the k observed items in $A_{i_1,\ldots,i_k}$, and $\sigma' \in \mathbb{S}_m$ be the total ranking corresponding to the m observed items in $A_{j_1,\ldots,j_m}$.

2: Let $\tau \in \mathbb{S}_{|I\cap J|}$ be the total ranking of the observed items indexed by $I \cap J$ in $A_{i_1,\ldots,i_k}$, and $\tau' \in \mathbb{S}_{|I\cap J|}$ the total ranking of the observed items indexed by $I \cap J$ in partial ranking $A_{j_1,\ldots,j_m}$.

3: Initialize $s_1 = s_2 = s_3 = s_4 = s_5 = 0$.

4: If $|I \cap J| \geq 2$, update

$$s_1 = \frac{\binom{|I\cap J|}{2}}{\binom{n}{2}} K_\tau(\tau, \tau').$$

5: If $|I \cap J| \geq 1$ and $|I \setminus J| \geq 1$, update

$$s_2 = \frac{1}{\binom{n}{2}(m+1)} \sum_{i\in I\cap J} \Big\{ \big[2\sigma'(i) - m - 1\big] \times \big[2(\sigma(i) - \tau(i)) - k + |I \cap J|\big] \Big\}.$$

6: If $|I \cap J| \geq 1$ and $|J \setminus I| \geq 1$, update

$$s_3 = \frac{1}{\binom{n}{2}(k+1)} \sum_{i\in I\cap J} \Big\{ \big[2\sigma(i) - k - 1\big] \times \big[2(\sigma'(i) - \tau'(i)) - m + |I \cap J|\big] \Big\}.$$

7: If $|I \cap J| \geq 1$ and $|(I \cup J)^{\complement}| \geq 1$, update

$$s_4 = \frac{|(I \cup J)^{\complement}|}{\binom{n}{2}(k+1)(m+1)} \times \sum_{i\in I\cap J} \big[2\sigma(i) - k - 1\big]\big[2\sigma'(i) - m - 1\big].$$

8: If $|I \setminus J| \geq 1$ and $|J \setminus I| \geq 1$, update

$$s_5 = \frac{-1}{\binom{n}{2}(k+1)(m+1)} \times \sum_{i\in I\setminus J} \big[2\sigma(i) - k - 1\big] \sum_{j\in J\setminus I} \big[2\sigma'(j) - m - 1\big].$$

Output: $K_\tau(A_{i_1,\ldots,i_k}, A_{j_1,\ldots,j_m}) = s_1 + s_2 + s_3 + s_4 + s_5$.

Algorithm 2.2 Kendall kernel for a top-k partial ranking and a top-m partial ranking.

Input: a top-k partial ranking and a top-m partial ranking $B_{i_1,\ldots,i_k}, B_{j_1,\ldots,j_m} \subset \mathbb{S}_n$, corresponding to subsets of item indices $I := \{i_1,\ldots,i_k\}$ and $J := \{j_1,\ldots,j_m\}$.

1: Let $\sigma \in \mathbb{S}_k$ be the total ranking corresponding to the k observed items in $B_{i_1,\ldots,i_k}$, and $\sigma' \in \mathbb{S}_m$ be the total ranking corresponding to the m observed items in $B_{j_1,\ldots,j_m}$.

2: Let $\tau \in \mathbb{S}_{|I\cap J|}$ be the total ranking of the observed items indexed by $I \cap J$ in $B_{i_1,\ldots,i_k}$, and $\tau' \in \mathbb{S}_{|I\cap J|}$ the total ranking of the observed items indexed by $I \cap J$ in partial ranking $B_{j_1,\ldots,j_m}$.

3: Initialize $s_1 = s_2 = s_3 = s_4 = s_5 = 0$.

4: If $|I \cap J| \geq 2$, update

$$s_1 = \frac{\binom{|I\cap J|}{2}}{\binom{n}{2}} K_\tau(\tau, \tau') \,.$$

5: If $|I \cap J| \geq 1$ and $|I \setminus J| \geq 1$, update

$$s_2 = \frac{1}{\binom{n}{2}} \sum_{i \in I\cap J} \left[2(\sigma(i) - \tau(i)) - k + |I \cap J|\right] .$$

6: If $|I \cap J| \geq 1$ and $|J \setminus I| \geq 1$, update

$$s_3 = \frac{1}{\binom{n}{2}} \sum_{i \in I\cap J} \left[2(\sigma'(i) - \tau'(i)) - m + |I \cap J|\right] .$$

7: If $|I \cap J| \geq 1$ and $|(I \cup J)^{\complement}| \geq 1$, update

$$s_4 = \frac{|I \cap J| \cdot |(I \cup J)^{\complement}|}{\binom{n}{2}} \,.$$

8: If $|I \setminus J| \geq 1$ and $|J \setminus I| \geq 1$, update

$$s_5 = \frac{-|I \setminus J| \cdot |J \setminus I|}{\binom{n}{2}} \,.$$

Output: $K_\tau(B_{i_1,\ldots,i_k}, B_{j_1,\ldots,j_m}) = s_1 + s_2 + s_3 + s_4 + s_5$.

2.3.2 Extension to Multivariate Rankings

In contrast to the rankings defined in previous sections, a *multivariate ranking* can be seen as a collection of multiple (univariate) partial/total rankings from the same ranker based on different sources. For example, a commercial survey is designed to analyze the preference routines of a customer based on various categories such as music, movies and novels, where the item sets are generally incomparable in crossing categories; an electoral system asks a voter to express his opinion in consecutive sessions across years, where the candidates are usually different across elections. In that case, it is desirable to process and integrate the rank data from different sources when extensively comparing the similarity between two rankers. Known as "data fusion", this problem is well studied in the kernel learning literature [Lanckriet 2004b, Schölkopf 2004].

Let us now denote that a ranker is represented by a multivariate ranking $\mathbf{R} = (R_1, \ldots, R_p)$, in which each component R_j for $1 \leq j \leq p$ is a partial ranking over n_j items, i.e., a subset of permutations (or exactly one permutation when all n_j items are totally ranked) in $\mathbb{S}_{n_j}$. Suppose K is any kernel for univariate rankings, a kernel for multivariate rankings that integrates information from several variates can be constructed by a weighted average of the kernels evaluated individually for each variate, written with a slight abuse of notations as

$$K(\mathbf{R}, \mathbf{R}') = \sum_{j=1}^{p} \mu_j K(R_j, R'_j) \quad \text{s.t.} \quad \sum_{j=1}^{p} \mu_j = 1\,, \tag{2.9}$$

where a kernel K for partial rankings has been defined in (2.7). A practically simple approach would be to set the weights $\mu_j = 1/p$ for $1 \leq j \leq p$ in (2.9), but the weights can be learned as well through multiple kernel learning under appropriate setting [Lanckriet 2004a, Bach 2004, Sonnenburg 2006, Gönen 2011].

2.3.3 Extension to Uncertain Rankings

When data to analyze are n-dimensional real-valued quantitative vectors, converting them to permutations in $\mathbb{S}_n$ by ranking their entries can be beneficial in cases where we trust more the relative ordering of the values than their absolute magnitudes. For example in social surveys or recommender systems, users are sometimes asked to rate a score for each item individually instead of providing a preference order on the item set. The scale of ratings usually varies according to personal preference of each user and it can therefore be safer to adopt ranking-based methods to analyze such score-based rating data [Kamishima 2003]. As another example, an interesting line of work in the analysis of gene expression data promotes the development of classifiers based on relative gene expression within a sample, based on the observations that gene expression measurements are subject to various measurement errors such as technological biases and normalization issues, while assessing whether a gene is more expressed than another gene is generally a more robust task [Geman 2004, Tan 2005, Xu 2005, Lin 2009]. This suggests that the Kendall kernel can be relevant for analyzing quantitative vectors.

The Kendall kernel for quantitative vectors now takes exactly the same form as for permutations, i.e.,

$$K_\tau(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^\top \Phi(\mathbf{x}')\,, \tag{2.10}$$

where $\Phi : \mathbb{R}^n \to \mathbb{R}^{\binom{n}{2}}$ is defined for $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ by

$$\Phi(\mathbf{x}) = \left(\frac{1}{\sqrt{\binom{n}{2}}} (\mathbb{1}_{\{x_i > x_j\}} - \mathbb{1}_{\{x_i < x_j\}}) \right)_{1 \le i < j \le n}. \tag{2.11}$$

In this case, the interpretation of the Kendall kernel in terms of concordant and discordant pairs (2.3) is still valid, with the caveats that in the presence of ties between entries of $\mathbf{x}$, say two coordinates i and j such that $x_i = x_j$, the tied pair $\{x_i, x_j\}$ will be neither concordant nor discordant. This implies in particular that if $\mathbf{x}$ has ties or so does $\mathbf{x}'$, then $|K_\tau(\mathbf{x}, \mathbf{x}')| < 1$ strictly. Notably in the presence of ties, the fast implementation of Kendall kernel still applies to quantitative vectors in $O(n \log n)$ time [Knight 1966]. However, feature mapping (2.11) is by construction very sensitive to the presence of entry pairs that are ties or almost-ties. In fact, each entry of $\Phi(\mathbf{x})$ is, up to a normalization constant, the Heaviside step function which takes discrete values in $\{-1, 0, +1\}$, and thus can change abruptly even when $\mathbf{x}$ changes slightly but reverses the ordering of two entries whose values are close. In addition to pairwise relative ordering as defined in (2.11), it can be wise to also exploit the information given by pairwise absolute difference in the feature values.

We propose to make the mapping more robust by assuming a random noise $\varepsilon \sim \mathcal{P}$ added to the feature vector $\mathbf{x}$ and checking where $\Phi(\mathbf{x}+\varepsilon)$ is on average (similarly to, e.g., [Muandet 2012]). In other words, we consider a smoother mapping $\Psi : \mathbb{R}^n \to \mathbb{R}^{\binom{n}{2}}$ defined by

$$\Psi(\mathbf{x}) = \mathbb{E}\Phi(\mathbf{x} + \varepsilon) =: \mathbb{E}\Phi(\tilde{\mathbf{x}}), \tag{2.12}$$

where ε is an n-dimensional random noise and $\tilde{\mathbf{x}} := \mathbf{x} + \varepsilon$ denotes the random-jittered vector of $\mathbf{x}$. The corresponding kernel is the underlying dot product as usual:

$$G(\mathbf{x}, \mathbf{x}') = \Psi(\mathbf{x})^\top \Psi(\mathbf{x}') = \mathbb{E}\Phi(\tilde{\mathbf{x}})^\top \mathbb{E}\Phi(\tilde{\mathbf{x}}') = \mathbb{E}K_\tau(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')\,, \tag{2.13}$$

where $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}'$ are independently noise-perturbed versions of $\mathbf{x}$ and $\mathbf{x}'$. In fact, we can deduce from (2.11) that Ψ is equivalently written as

$$\Psi(\mathbf{x}) = \left(\frac{1}{\sqrt{\binom{n}{2}}} (\mathbb{P}(\tilde{x}_i > \tilde{x}_j) - \mathbb{P}(\tilde{x}_i < \tilde{x}_j)) \right)_{1 \le i < j \le n}.$$

Depending on the noise distribution, various kernels are thus obtained. For example, assuming specifically that $\varepsilon \sim (\mathcal{U}[-\frac{a}{2}, \frac{a}{2}])^n$ the n-dimensional uniform noise of window size a centered at 0, the (i, j)-th entry of $\Psi(\mathbf{x})$ for all $i < j$ becomes

$$\Psi_{ij}(\mathbf{x}) = \frac{1}{\sqrt{\binom{n}{2}}} g_a(x_i - x_j)\,, \tag{2.14}$$

where

$$g_a(t) := \begin{cases} 1 & t \geq a \\ 2(\frac{t}{a}) - (\frac{t}{a})^2 & 0 \leq t \leq a \\ 2(\frac{t}{a}) + (\frac{t}{a})^2 & -a \leq t \leq 0 \\ -1 & t \leq -a \end{cases} .$$

Note that g_a is odd, continuous, piecewise quadratic between $[-a, a]$ and constant elsewhere at ± 1, and thus can be viewed as smoothed version of the Heaviside step function to compare any two entries x_i and x_j from their difference $x_i - x_j$ (Figure 2.1).

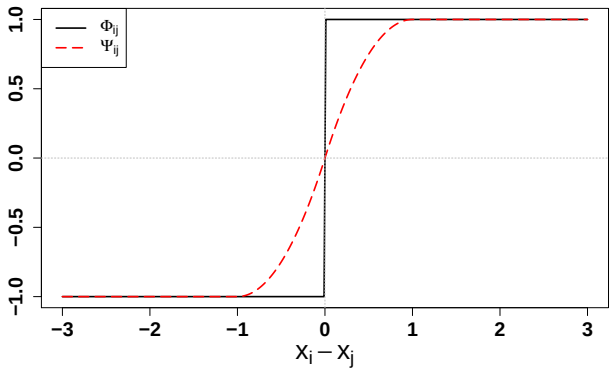


Figure 2.1: Smooth approximation (in red) of the Heaviside function (in black) used to define the mapping (2.14) for $a = 1$.

Although the smoothed kernel (2.13) can be an interesting alternative to the Kendall kernel (2.10), we unfortunately lose for G the computational trick that allows to compute K_τ in $O(n \log n)$. Specifically, we have two ways to compute G:
(i) Exact evaluation. The first alternative is to compute explicitly the $\binom{n}{2}$-vector representation Ψ in the feature space, and then take the dot product to obtain G. While the kernel evaluation is exact, an analytic form of the smoothed mapping (2.12) is required and the computational cost is linear with the dimension of the feature space, i.e., $O(n^2)$.
(ii) Monte Carlo approximation. The second alternative requires the observation that the smoothed mapping $\Psi(\mathbf{x}) = \mathbb{E}\Phi(\tilde{\mathbf{x}})$ appears in the form of expectation and can thus be approximated by a D-sample mean of jittered points mapped by Φ into the feature space:

$$\Psi_D(\mathbf{x}) = \frac{1}{D} \sum_{j=1}^{D} \Phi(\tilde{\mathbf{x}}^j) ,$$

where $\tilde{\mathbf{x}}^1, \ldots, \tilde{\mathbf{x}}^D$ are i.i.d. noisy versions of $\mathbf{x}$. The dot product induces a kernel:

$$G_D(\mathbf{x}, \mathbf{x}') = \Psi_D(\mathbf{x})^\top \Psi_D(\mathbf{x}') = \frac{1}{D^2} \sum_{i=1}^{D} \sum_{j=1}^{D} K_\tau \left(\tilde{\mathbf{x}}^i, \tilde{\mathbf{x}}'^j \right) , \tag{2.15}$$

which is a D^2-sample empirical estimate of $G(\mathbf{x}, \mathbf{x}') = \mathbb{E}K_\tau(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$ when $\mathbf{x}, \mathbf{x}'$ are independently jittered with identically distributed noise. Since K_τ is of computational complexity $O(n \log n)$, computing G_D requires $O(D^2 n \log n)$.

Note that the second alternative is faster to compute than the first one as long as, up to constants, $D^2 < n/\log n$, and small values of D are thus favored on account of computational consideration. In that case, however, the approximation performance can be unappealing. To better understand the trade-off between the two alternatives, the question should be addressed upon how large D should be so that the approximation error is not detrimental to the performance of a learning algorithm if we use the approximate kernel G_D instead of G. Lemma 2.1 provides a first answer to this question, showing that the approximation error of the kernel is upper bounded by $O(1/\sqrt{D})$ with high probability:

Lemma 2.1. *For any $0 < \delta < 1$, the following holds:*
(a) For any $\mathbf{x} \in \mathbb{R}^n$, with probability greater than $1 - \delta$,

$$\|\Psi_D(\mathbf{x}) - \Psi(\mathbf{x})\| \leq \frac{1}{\sqrt{D}} \left(2 + \sqrt{8 \log \frac{1}{\delta}}\right) .$$

(b) For any $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$, with probability greater than $1 - \delta$,

$$\sup_{i=1,\ldots,m} \|\Psi_D(\mathbf{x}_i) - \Psi(\mathbf{x}_i)\| \leq \frac{1}{\sqrt{D}} \left(2 + \sqrt{8 \log \frac{m}{\delta}}\right) .$$

Proof. For any $\mathbf{x} \in \mathbb{R}^n$, note that $\|\Phi(\mathbf{x})\| \leq 1$. We can therefore apply [Boucheron 2013, Example 6.3] to the random vector $X_j = \Phi(\tilde{\mathbf{x}}^j) - \Psi(\mathbf{x})$ that satisfies $\mathbb{E}X_j = 0$ and $\|X_j\| \leq 2$ a.s. to get, for any $u \geq 2/\sqrt{D}$,

$$\mathbb{P}\left(\|\Psi_D(\mathbf{x}) - \Psi(\mathbf{x})\| \geq u\right) \leq \exp\left(-\frac{\left(u\sqrt{D} - 2\right)^2}{8}\right) .$$

We recover (a) by setting the right-hand side equal to δ and solving for u. (b) then follows by a simple union bound. □

The uniform approximation bound of Lemma 2.1 in turn implies that learning with the approximate kernel G_D can be almost as good with the kernel G, as we now discuss. For that purpose, let us consider for example the case where the smoothed kernel G is used to train a Support Vector Machine (SVM) from a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \subset (\mathbb{R}^n \times \{-1, +1\})^m$, specifically to estimate a function $h(\mathbf{x}) = \mathbf{w}^\top \Psi(\mathbf{x})$ by solving

$$\min_{\mathbf{w}} F(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \widehat{R}(\mathbf{w}), \tag{2.16}$$

where $\widehat{R}(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^m \ell(y_i \mathbf{w}^\top \Psi(\mathbf{x}_i))$ is the empirical loss, with $\ell(y_i \mathbf{w}^\top \Psi(\mathbf{x}_i)) = \max(0, 1 - y_i \mathbf{w}^\top \Psi(\mathbf{x}_i))$ the hinge loss associated to the i-th point, λ the regularization parameter. Now suppose that instead of training the SVM with smoothed

feature mapping on the original points $\{\Psi(\mathbf{x}_i)\}_{i=1,\dots,m}$, we first randomly jitter $\{\mathbf{x}_i\}_{i=1,\dots,m}$ D times at each point, resulting in $\{\tilde{\mathbf{x}}_i^j\}_{i=1,\dots,m;j=1,\dots,D}$, and then replace each $\Psi(\mathbf{x}_i)$ by the D-sample empirical average of jittered points mapped by Φ into the feature space, that is

$$\Psi_D(\mathbf{x}_i) := \frac{1}{D}\sum_{j=1}^{D}\Phi(\tilde{\mathbf{x}}_i^j)\,.$$

Note that $\Psi_D(\mathbf{x}_i)^\top\Psi_D(\mathbf{x}_j) = G_D(\mathbf{x}_i,\mathbf{x}_j)$, hence training an SVM with the Monte Carlo approximate G_D instead of exact version G is equivalent to solving (2.16) with $\{\Psi_D(\mathbf{x}_i)\}_{i=1,\dots,m}$ in the hinge loss instead of $\{\Psi(\mathbf{x}_i)\}_{i=1,\dots,m}$. Theorem 2.3 quantifies the approximation performance in terms of objective function F which helps to answer the question on the trade-off between G and G_D in computational complexity and learning accuracy.

Theorem 2.3. *For any $0 \le \delta \le 1$, the solution $\widehat{\mathbf{w}}_D$ of the SVM trained with the Monte Carlo approximation (2.15) with D random-jittered samples for each training point satisfies, with probability greater than $1-\delta$,*

$$F(\widehat{\mathbf{w}}_D) \le \min_{\mathbf{w}} F(\mathbf{w}) + \sqrt{\frac{8}{\lambda D}}\left(2+\sqrt{8\log\frac{m}{\delta}}\right).$$

Proof. Let $\widehat{\mathbf{w}}$ be a solution to the original SVM optimization problem, and $\widehat{\mathbf{w}}_D$ a solution to the perturbed SVM, i.e., a solution of

$$\min_{\mathbf{w}} F_D(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \widehat{R}_D(\mathbf{w}), \tag{2.17}$$

with $\widehat{R}_D(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^m \ell(y_i\mathbf{w}^\top\Psi_D(\mathbf{x}_i))$. Since the hinge loss is 1-Lipschitz, i.e., $|\ell(a)-\ell(b)| \le |a-b|$ for any $a,b\in\mathbb{R}$, we obtain that for any $\mathbf{u}\in\mathbb{R}^{\binom{n}{2}}$:

$$\begin{aligned}\left|\widehat{R}(\mathbf{u}) - \widehat{R}_D(\mathbf{u})\right| &\le \frac{1}{m}\sum_{i=1}^m\left|\mathbf{u}^\top\left(\Psi(\mathbf{x}_i)-\Psi_D(\mathbf{x}_i)\right)\right| \\ &\le \|\mathbf{u}\|\sup_{i=1,\dots,m}\|\Psi_D(\mathbf{x}_i)-\Psi(\mathbf{x}_i)\|\,.\end{aligned} \tag{2.18}$$

Now, since $\widehat{\mathbf{w}}_D$ is a solution of (2.17), it satisfies

$$\|\widehat{\mathbf{w}}_D\| \le \sqrt{\frac{2F_D(\widehat{\mathbf{w}}_D)}{\lambda}} \le \sqrt{\frac{2F_D(0)}{\lambda}} = \sqrt{\frac{2}{\lambda}},$$

and similarly $\|\widehat{\mathbf{w}}\| \le \sqrt{2/\lambda}$ because $\widehat{\mathbf{w}}$ is a solution of the original SVM optimization

problem. Using (2.18) and these bounds on $\|\widehat{\mathbf{w}}_D\|$ and $\|\widehat{\mathbf{w}}\|$, we get

$$\begin{aligned}
&F(\widehat{\mathbf{w}}_D) - F(\widehat{\mathbf{w}}) \\
&\quad = F(\widehat{\mathbf{w}}_D) - F_D(\widehat{\mathbf{w}}_D) + F_D(\widehat{\mathbf{w}}_D) - F(\widehat{\mathbf{w}}) \\
&\quad \leq F(\widehat{\mathbf{w}}_D) - F_D(\widehat{\mathbf{w}}_D) + F_D(\widehat{\mathbf{w}}) - F(\widehat{\mathbf{w}}) \\
&\quad = \widehat{R}(\widehat{\mathbf{w}}_D) - \widehat{R}_D(\widehat{\mathbf{w}}_D) + \widehat{R}_D(\widehat{\mathbf{w}}) - \widehat{R}(\widehat{\mathbf{w}}) \\
&\quad \leq (\|\widehat{\mathbf{w}}_D\| + \|\widehat{\mathbf{w}}\|) \sup_{i=1,\ldots,m} \|\Psi_D(\mathbf{x}_i) - \Psi(\mathbf{x}_i)\| \\
&\quad \leq \sqrt{\frac{8}{\lambda}} \sup_{i=1,\ldots,m} \|\Psi_D(\mathbf{x}_i) - \Psi(\mathbf{x}_i)\| \ .
\end{aligned}$$

Theorem 2.3 then follows from Lemma 2.1. □

It is known that compared to the exact solution of (2.16), an $O(m^{-1/2})$-approximate solution is sufficient to reach the optimal statistical accuracy [Bottou 2008]. This accuracy can be attained in our analysis when $D = O(m/\lambda)$, and since typically $\lambda \sim m^{-1/2}$ [Steinwart 2005], this suggests that it is sufficient to take D of order $m^{3/2}$. Going back to the comparison strategy of the two alternatives G and G_D, we see that the computational cost of computing the full $m \times m$ Gram matrix with the exact evaluation is $O(m^2 n^2)$, while the cost of computing the approximate Gram matrix with $D = O(m^{3/2})$ random samples is $O(m^2 D^2 n \log n) = O(m^5 n \log n)$. This shows that, up to constants and logarithmic terms, the Monte Carlo approximation is interesting when $m = o(n^{1/3})$, otherwise the exact evaluation using explicit computation in the feature space is preferable.

Interestingly we can look at the extended Kendall kernel (2.13) to uncertain rankings from the perspective of Hilbert space embeddings of probability distributions [Smola 2007]. In fact, for x fixed, the smoothed mapping $\Psi(\mathbf{x}) = \mathbb{E}\Phi(\mathbf{x}+\varepsilon)$ is exactly an embedding for the distribution $\mathcal{P}$ of an additive noise ε in the reproducing kernel Hilbert space (RKHS) associated with Kendall kernel. As a consequence, the idea of smoothed kernel $G(\mathbf{x}, \mathbf{x}')$ for $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ is essentially equivalent to how [Muandet 2012, Lemma 4] defines kernels on two probability distributions from $\{\mathcal{P} + \mathbf{x} | \mathbf{x} \in \mathcal{X}\}$ using the Kendall kernel as the level-1 embedding kernel and linear inner product as the level-2 kernel in the feature space. As a result, given a fixed training set $\mathcal{D}$, training an SVM with G in place of K_τ is equivalent to training a Flex-SVM instead of an ordinary SVM with K_τ [Muandet 2012]. In this case, Theorem 2.3 provides an error bound in terms of the optimal accuracy for cases when training a Flex-SVM if exact evaluation of G is intractable and its Monte Carlo approximate G_D is employed. This serves to obtain a trade-off between computation complexity and approximation accuracy which is particularly interesting when we are working in high dimensions.

2.4 Relation of the Mallows Kernel and the Diffusion Kernel on $\mathbb{S}_n$

It is interesting to relate the Mallows kernel (2.2) to the diffusion kernel on the symmetric group proposed by [Kondor 2010], which is the diffusion kernel [Kondor 2002] on the Cayley graph of $\mathbb{S}_n$ generated by adjacent transpositions with left-multiplication. This graph, illustrated for a specific case of $n = 4$ in Figure 2.2, is defined by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \mathbb{S}_n$ as vertices, and undirected edge set $\mathcal{E} = \{\{\sigma, \pi\sigma\} : \sigma \in \mathbb{S}_n, \pi \in Q\}$, where $Q = \{(i, i+1) | i = 1, \dots, n-1\}$ the set of all adjacent transpositions. Note Q is symmetric in the sense that $\pi \in Q \Leftrightarrow \pi^{-1} \in Q$, and the graph adjacency relation is a right-invariant relation, that is $\sigma \sim \sigma' \Leftrightarrow \sigma'\sigma^{-1} \in Q$. The corresponding graph Laplacian is the matrix Δ with

$$\Delta_{\sigma,\sigma'} = \begin{cases} 1 & \text{if } \sigma \sim \sigma' \\ -(n-1) & \text{if } \sigma = \sigma' \\ 0 & \text{otherwise} \end{cases},$$

where $n-1$ is the degree of vertex σ (number of edges connected with vertex σ), and the *diffusion kernel* on $\mathbb{S}_n$ is finally defined as

$$K^{\beta}_{\text{dif}}(\sigma, \sigma') = [e^{\beta\Delta}]_{\sigma,\sigma'} \tag{2.19}$$

for some diffusion parameter $\beta \in \mathbb{R}$, where $e^{\beta\Delta}$ is the matrix exponential. K^{β}_{dif} is a right-invariant kernel on the symmetric group [Kondor 2010, Proposition 2], and we denote by $\kappa^{\beta}_{\text{dif}}$ the positive definite function induced by K^{β}_{dif} such that $K^{\beta}_{\text{dif}}(\sigma, \sigma') = \kappa^{\beta}_{\text{dif}}(\sigma'\sigma^{-1})$. Since the Mallows kernel K^{λ}_{M} is straightforwardly right-invariant, we denote by κ^{λ}_{M} the positive definite function induced by the Mallows kernel K^{λ}_{M} such that $K^{\lambda}_{M}(\sigma, \sigma') = \kappa^{\lambda}_{M}(\sigma'\sigma^{-1})$. One way to interpret the diffusion kernel (2.19) is by the heat equation on the Cayley graph

$$\frac{d}{d\beta} K^{\beta}_{\text{dif}} = \Delta K^{\beta}_{\text{dif}} \quad \text{s.t. } K^{\beta}_{\text{dif}}|_{\beta=0} = I.$$

K^{β}_{dif} is thus the product of a continuous process, expressed by the graph Laplacian Δ, gradually transforming local structure $K^{\beta}_{\text{dif}}|_{\beta=0} = I$ to a kernel with stronger and stronger off-diagonal effects as β increases.

Interestingly, the Mallows kernel can also be interpreted with the help of the Cayley graph. Indeed, it is well-known that the Kendall tau distance $n_d(\sigma, \sigma')$ is the minimum number of adjacent swaps required to bring σ to σ', i.e. $n_d(\sigma, \sigma')$ equals to the shortest path distance on the Cayley graph [Drutu 2017, Exercise 7.73], or simply written

$$n_d(\sigma, \sigma') = d_{\mathcal{G}}(\sigma, \sigma').$$

Different from the diffusion kernel for which communication between permutations is a diffusion process over the graph, the Mallows kernel

$$K^{\lambda}_{M}(\sigma, \sigma') = e^{-\lambda n_d(\sigma,\sigma')} = e^{-\lambda d_{\mathcal{G}}(\sigma,\sigma')}$$

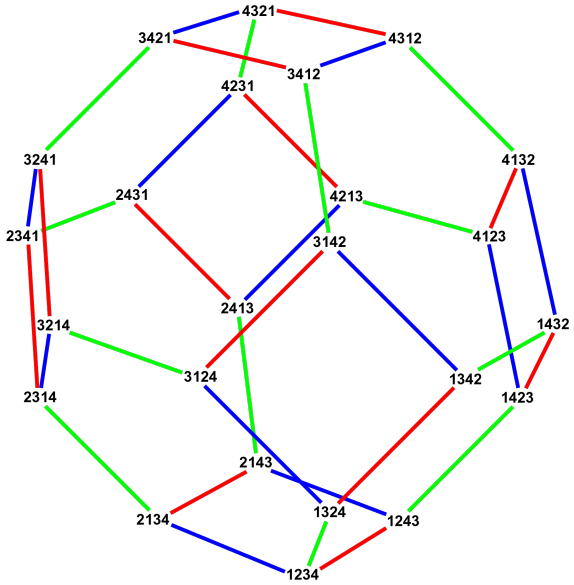


Figure 2.2: Cayley graph of $\mathbb{S}_4$, generated by the transpositions (1 2) in blue, (2 3) in green, and (3 4) in red.

considers exclusively the shortest path over the graph when expressing the similarity between permutations σ, σ'.

A notable advantage of the Mallows kernel over the diffusion kernel is that the Mallows kernel enjoys faster evaluation. On one hand if data instances are total rankings, i.e. $\sigma, \sigma' \in \mathbb{S}_n$, evaluating $K^{\beta}_{\mathrm{dif}}(\sigma, \sigma')$ would require exponentiating an $n!$-dimensional Laplacian matrix by naive implementation, and can reduce to exponentiating matrices of smaller sizes by careful analysis in the Fourier space, which still remains problematic if working dimension n is large [Kondor 2010]. However, evaluating $K^{\lambda}_{M}(\sigma, \sigma')$ only takes $O(n \log n)$ time. On the other hand if data instances are partial ranking of size $k \ll n$, i.e. $R, R' \subset \mathbb{S}_n$, and we take convolution kernel (2.7) to extend the two kernels, the analysis of exploring the sparsity of the Fourier coefficients of the group algebra of partial rankings R, R' of size k reduces the evaluation of both the diffusion kernel and the Mallows kernel to $O((2k)^{2k+3})$ time, provided that the exponential kernel Fourier matrices $[\widehat{\kappa}(\mu)]_{\geq[\dots]_{n-k}}$ are precomputed before any kernel evaluations take place [Kondor 2010, Theorem 13].

2.5 Application: Clustering and Modeling Rank Data

In this section we illustrate the potential benefit of kernel-based algorithms using the Kendall and Mallows kernels for the purpose of unsupervised cluster analysis, i.e., partitioning a collection of rank data into sub-groups and/or estimating densities

of a collection of rank data. This is in particular of great practical interest in social choice theory in order to explore the heterogeneity and identify typical sub-groups of voters with a common behavior to understand, for example, their political support for various parties [Gormley 2006, Gormley 2008, Marden 1996].

2.5.1 Clustering with Kernel k-means

Let $\{\sigma_i\}_{i=1}^m \subset \mathbb{S}_n$ be a collection of m permutations representing, say, the preferences of customers over n products or the votes of electorate over n candidates. We aim at partitioning these permutations into $c \leq m$ clusters $\{S_j\}_{j=1}^c$. One approach to cluster rank data is to follow a method similar to k-means in the symmetric group. Assuming that each cluster S_j has a "center" $\pi_j \in \mathbb{S}_n$ serving as a prototype permutation of that cluster, the classic k-means clustering attempts to put each point in the cluster with the nearest center so as to minimize the sum of Kendall tau distance of each permutation to the corresponding center of its cluster. Specifically, when the number of clusters c is fixed, the objective is to find:

$$\underset{\{S_j,\pi_j\in\mathbb{S}_n\}}{\arg\min} \sum_{j=1}^{c} \sum_{i:\sigma_i\in S_j} n_d(\sigma_i,\pi_j)\,. \tag{2.20}$$

Note that (2.20) reduces to a single-ranking aggregation problem when $c = 1$, where the center π is commonly known as Kemeny consensus [Kemeny 1962] which is NP-hard to find [Bartholdi III 1989]. With the objective in (2.20) being non convex, Lloyd's algorithm is usually employed to find local minima in an iterative manner consisting of two steps: the *assignment step* assigns each point to its closest cluster, and the *update step* updates each of the c cluster centers using the points assigned to that cluster; the algorithm repeats until all the cluster centers remain unchanged in an iteration. While the assignment step is usually fast, the update step is indeed equivalent to solving a Kemeny consensus problem for each cluster, i.e., $\arg\min_{\pi_j\in\mathbb{S}_n} \sum_{i:\sigma_i\in S_j} n_d(\sigma_i,\pi_j)$. Since the exact Kemeny-optimal ranking is difficult to find, approximate techniques are usually employed in practice such as Borda Count [de Borda 1781] or Copeland's method [Copeland 1951].

As the Kendall tau distance is conditionally positive definite, we can propose as an alternative to use the kernel k-means approach [Girolami 2002, Zhang 2002] that relaxes the assumption that the cluster center are permutations, and instead works implicitly in the feature space where cluster centers can be any vector in $\mathbb{R}^{\binom{n}{2}}$ by considering the problem:

$$\underset{\left\{S_j,\mu_j\in\mathbb{R}^{\binom{n}{2}}\right\}}{\arg\min} \sum_{j=1}^{c} \sum_{i:\sigma_i\in S_j} \|\Phi(\sigma_i)-\mu_j\|^2\,,$$

for which local minima can be found efficiently by Algorithm 2.3. Note that μ_j does not match a true permutation $\pi_j \in \mathbb{S}_n$ in general, and the Kemeny consensus problem in the update step is thus bypassed. It is worthwhile to note that the algorithm is not exclusive for clustering permutations, kernel k-means clustering

can be applied respectively to total/partial/multivariate/uncertain rankings with appropriate kernels defined.

Algorithm 2.3 Kernel k-means for clustering heterogeneous rank data.

Input: a collection of permutations $\{\sigma_i\}_{i=1}^m$ and a kernel function K over $\mathbb{S}_n$, or a kernel matrix evaluated between pairwise data points $\mathbf{K} = (K(\sigma_i, \sigma_j))_{1 \le i,j \le m}$; the number of clusters $c \le m$.

1: Randomly initialize cluster assignment for each data points and form c clusters $S_1, \ldots, S_c$.

2: For each data point, find its new cluster assignment, i.e., for $i = 1, \ldots, m$,

$$j^*(\sigma_i) = \arg\min_j d_{ij} \,,$$

where

$$\begin{aligned} d_{ij} &:= \left\| \Phi(\sigma_i) - \frac{1}{|S_j|} \sum_{\sigma_\ell \in S_j} \Phi(\sigma_\ell) \right\|^2 \\ &= K(\sigma_i, \sigma_i) - \frac{2}{|S_j|} \sum_{\sigma_\ell \in S_j} K(\sigma_i, \sigma_\ell) + \frac{1}{|S_j|^2} \sum_{\sigma_v, \sigma_\ell \in S_j} K(\sigma_v, \sigma_\ell) \,. \end{aligned}$$

3: Form updated clusters, i.e., for $j = 1, \ldots, c$,

$$S_j = \{\sigma_i : j = j^*(\sigma_i), i = 1, \ldots, m\} \,.$$

4: Repeat 2-3 until all cluster assignments remain unchanged in an iteration.

Output: Cluster assignments $\{S_j\}_{j=1}^c$.

2.5.2 Mallows Mixture Model with Kernel Trick

An alternative to k-means clustering is to consider mixture models, which provide a method for modeling heterogeneous population in data by assuming a mixture of standard models for rankings in each homogeneous sub-population. Mixture models not only allow to cluster data, but more generally to estimate a distribution on the space of permutation that can then be used for other purposes, such as combining evidences. One popular choice of probabilistic distribution over $\mathbb{S}_n$ is the Mallows model [Mallows 1957], which takes the form in expressing the occurring probability of σ by

$$f(\sigma|\pi, \lambda) = C(\lambda) \exp[-\lambda n_d(\sigma, \pi)] \,, \tag{2.21}$$

where the central ranking $\pi \in \mathbb{S}_n$ and the precision $\lambda \ge 0$ are model parameters, and the normalization constant $C(\lambda) = 1/\sum_{\sigma' \in \mathbb{S}_n} \exp[-\lambda n_d(\sigma', \pi)]$ is chosen so

that $f(\cdot|\pi,\lambda)$ is a valid probability mass function over $\mathbb{S}_n$. Notably, $C(\lambda)$ does not depend on the center π due to the symmetry of $\mathbb{S}_n$.

We follow the mixture modeling setup in [Murphy 2003]. Now suppose that a population consists of c sub-populations, a Mallows mixture model assumes that an observation comes from group j with probability $p_j \geq 0$ for $j = 1,\ldots,c$ and, given that the observation belongs to sub-population j, it is generated from a Mallows model with central ranking π_j and precision λ_j, i.e., the occurring probability of σ in the Mallows mixture model is written as

$$f(\sigma) = \sum_{j=1}^{c} p_j f(\sigma|\pi_j,\lambda_j) = \sum_{j=1}^{c} p_j C(\lambda_j)\exp[-\lambda_j n_d(\sigma,\pi_j)]\,. \tag{2.22}$$

Denoting $\underline{\pi} = \{\pi_j\}_{j=1}^c$, $\underline{\lambda} = \{\lambda_j\}_{j=1}^c$, $\underline{p} = \{p_j\}_{j=1}^c$ such that $\sum_{j=1}^c p_j = 1$, the log-likelihood of a collection of m i.i.d. permutations $\underline{\sigma} = \{\sigma_i\}_{i=1}^m$ is therefore:

$$L(\underline{\pi},\underline{\lambda},\underline{p}|\underline{\sigma}) = \sum_{i=1}^{m}\log f(\sigma_i) = \sum_{i=1}^{m}\log\left\{\sum_{j=1}^{c} p_j C(\lambda_j)\exp[-\lambda_j n_d(\sigma_i,\pi_j)]\right\}. \tag{2.23}$$

The Mallows mixture model is usually fitted by maximum likelihood using the EM algorithm. Specifically, by introducing latent (membership) variables $\underline{z} = \{z_{ij} : i = 1,\ldots,m, j = 1,\ldots,c\}$ where $z_{ij} = 1$ if σ_i belongs to group j and 0 otherwise, the complete log-likelihood of data is

$$L_C(\underline{\pi},\underline{\lambda},\underline{p}|\underline{\sigma},\underline{z}) = \sum_{i=1}^{m}\sum_{j=1}^{c} z_{ij}[\log p_j + \log C(\lambda_j) - \lambda_j n_d(\sigma_i,\pi_j)]\,.$$

The EM algorithm can be implemented to find local maximum likelihood estimates following two steps iteratively until convergence: the *E-step* calculates the expected value of membership variables $\hat{\underline{z}}$ conditioned on the current estimates of the model parameters $\underline{\pi},\underline{\lambda},\underline{p}$, and the *M-step* updates the model parameters $\underline{\pi},\underline{\lambda},\underline{p}$ by maximizing the expected complete log-likelihood $\hat{L}_C = L_C(\underline{\pi},\underline{\lambda},\underline{p}|\underline{\sigma},\hat{\underline{z}})$ where membership variables are replaced by their expected values. The final estimate $\hat{z}_{ij}$ amounts to our belief of σ_i belonging to group j, and can thus be used to form clusters $\{S_j\}_{j=1}^c$ serving a partition of data where

$$S_j = \left\{\sigma_i : \hat{z}_{ij} = \max_{\ell}\hat{z}_{i\ell}, i = 1,\ldots,m\right\}. \tag{2.24}$$

A closer look at the EM algorithm reveals that optimizing $\hat{L}_C$ with respect to $\underline{\pi}$ alone in the M-step is indeed equivalent to finding a (weighted) Kemeny consensus for each group, i.e., solving $\arg\min_{\pi_j\in\mathbb{S}_n}\sum_{i=1}^m \hat{z}_{ij} n_d(\sigma_i,\pi_j)$, for which exact solution is difficult as above-mentioned in the context of k-means clustering. Similarly to the idea of kernel k-means in contrast to classic k-means, we propose to seek ways to bypass the Kemeny consensus problem by working in the feature space instead. Note that the Mallows probability mass function (2.21) is equivalently written as $f(\sigma|\pi,\lambda) \propto \exp\left[-\lambda\left\|\Phi(\sigma)-\Phi(\pi)\right\|^2\right]$ up to a constant scaling on λ by using (2.4),

we propose to relax the constraint that the center has to match a true permutation $\pi \in \mathbb{S}_n$ and consider the following two alternatives in place of f following the mixture modeling approach stated above:

(i) Kernel Mallows. The Mallows probability mass function over $\mathbb{S}_n$ (2.21) is generalized to admit any point in the feature space $\mu \in \mathbb{R}^{\binom{n}{2}}$ to be the population center, i.e.,

$$g(\sigma|\mu,\lambda) = C(\mu,\lambda)\exp\left[-\lambda\left\|\Phi(\sigma)-\mu\right\|^2\right], \tag{2.25}$$

where the normalization constant $C(\mu,\lambda) = 1/\sum_{\sigma'\in\mathbb{S}_n}\exp\left[-\lambda\left\|\Phi(\sigma')-\mu\right\|^2\right]$ is chosen so that $g(\cdot|\mu,\lambda)$ is a valid probability mass function over $\mathbb{S}_n$. Notably, $C(\mu,\lambda)$ now depends on the center μ as well.

If we replace the probability mass function of classic Mallows f in (2.23) by that of kernel Mallows g, the Kemeny consensus problem is averted when the EM algorithm is used to fit a local maximum likelihood estimate. However, another computational setback arises that the expected complete log-likelihood $\hat{L}_C$ to maximize in the M-step of the EM algorithm is separately concave with respect to μ or λ, but not jointly concave. Hence alternating optimization is often used in practice with the caveats of intensive computation and no guarantee to attain global optima for the M-step optimization at each iteration.

(ii) Kernel Gaussian. Note that (2.25) has a similar form to the Gaussian density, therefore we consider for $\sigma \in \mathbb{S}_n$,

$$g^\dagger(\sigma|\mu,\lambda) = \sqrt{\left(\frac{\lambda}{\pi}\right)^{\binom{n}{2}}}\exp\left[-\lambda\left\|\Phi(\sigma)-\mu\right\|^2\right], \tag{2.26}$$

which is exactly $\mathcal{N}(\Phi(\sigma)|\mu,(2\lambda)^{-1}I)$, i.e., the $\binom{n}{2}$-dimensional Gaussian distribution with mean μ and isotropic covariance matrix $(2\lambda)^{-1}I$ injected by $\Phi(\sigma)$. Notably, $g^\dagger(\cdot|\mu,\lambda)$ is not a valid probability mass function over $\mathbb{S}_n$.

The mixture modeling approach stated above using $g^\dagger$ instead of f is in fact equivalently stated in Algorithm 2.4. It is worthwhile to note that the algorithm also applies to total/partial/multivariate/uncertain rankings with appropriate kernels defined as [Wang 2003, Table 2] provides the counterpart of Algorithm 2.4 in case that a kernel matrix evaluated between data points is given instead. However, since $g^\dagger$ itself is not a valid probability mass function over $\mathbb{S}_n$, an evident drawback is that we now lose the probabilistic interpretation of the mixture distribution as in (2.22).

2.5.3 Experiments

Clustering 1980 APA election data. In the 1980 American Psychological Association (APA) presidential election, voters were asked to rank 5 candidates in order of preference, and 5738 votes in form of total rankings were reported and thus used in our experiment. The dataset was thoroughly studied by [Diaconis 1988].

We first use k-means approaches to cluster the data. We compare the proposed kernel k-means algorithm (Algorithm 2.3 with Kendall kernel K_τ) to the classic

Algorithm 2.4 Kernel trick embedded Gaussian mixture model for clustering heterogeneous rank data.

Input: a collection of permutations $\{\sigma_i\}_{i=1}^m$ and a kernel function K over $\mathbb{S}_n$; the number of clusters $c \leq m$.

1: Compute feature points $\Phi(\sigma_i) \in \mathbb{R}^{\binom{n}{2}}$ mapped by the Kendall embedding.
2: Fit a Gaussian mixture model for $\{\Phi(\sigma_i)\}_{i=1}^m$ in $\mathbb{R}^{\binom{n}{2}}$ using maximum likelihood with the EM algorithm under the constraint of isotropic covariance matrix, i.e., $\Sigma = (2\lambda)^{-1} I$.
3: Use the membership estimates $\hat{\underline{z}}$ to form clusters by (2.24).

Output: Cluster assignments $\{S_j\}_{j=1}^c$.

k-means algorithm formulated as (2.20). For the classic k-means where cluster centers are required to be a prototype permutation, three methods are employed in the center-update step for each iteration: brute-force search of Kemeny-optimal ranking, approximate ranking induced by Borda Count and Copeland's method. In each case, we vary the number of clusters ranging from 2 to 10 and the algorithm is repeated 50 times with randomly initialized configurations for each fixed number of clusters. We observe from Figure 2.3 that the kernel k-means or classic k-means with approximate centers runs much faster than optimal k-means for which the Kemeny-optimal ranking is time-consuming to find by a brute-force search. Further, Figure 2.4 shows that the kernel k-means outperforms all three methods based on classic k-means in terms of the average silhouette scores of the clustering results, which justifies that the kernel k-means splits the data into more consistent sub-groups in the sense that instances, measured by Kendall tau distance on average, are more similar in the same cluster and more dissimilar in different clusters.

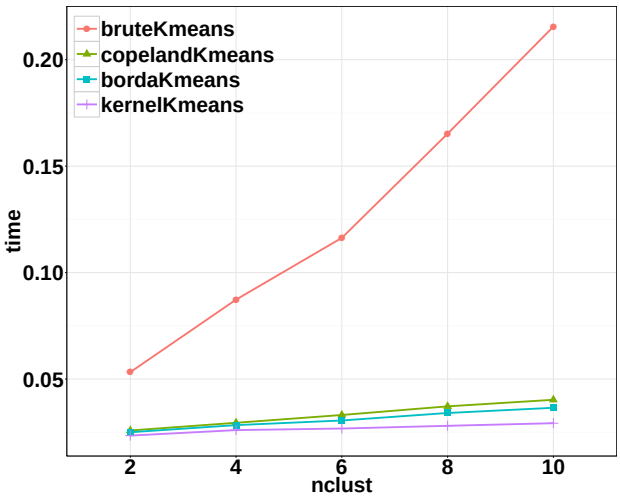


Figure 2.3: Computational time (in seconds) of k-means algorithms per run across different number of clusters.

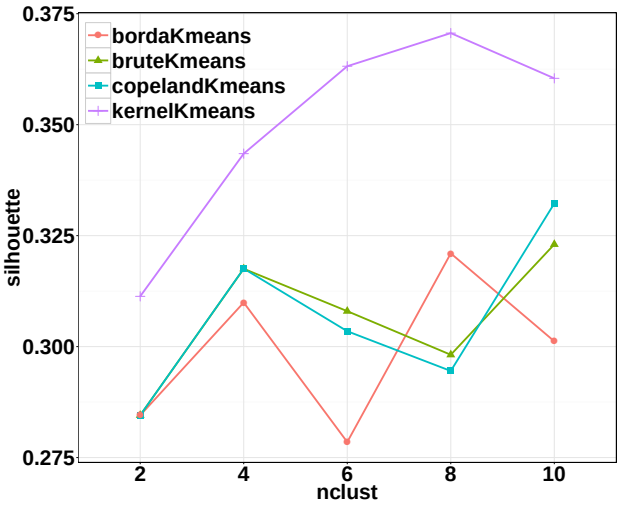


Figure 2.4: Average silhouette scores of k-means methods across different number of clusters.

Further, good clustering algorithms are supposed be robust to "perturbation" in data, in the sense that clusters formed by running an algorithm on bootstrap replicas of the original data should be similar. In other words, if we bootstrap the complete dataset twice and form a clustering with respect to each, the two clustering assignments should be close to each other. Note that in order to measure the similarity of two clustering assignments, we use the (adjusted) Rand index defined by the percentage of instance pairs falling in the same or in different clusters by the two assignments [Hubert 1985]. We now compare the stability performance of the proposed kernel k-means and other k-means algorithms. Specifically, for each fixed number of clusters, we repeatedly use a bootstrap replica of the dataset to search for centroids returned by running k-means algorithms, and partition the original dataset with these identified centroids. The Rand index for two such clustering assignments is computed and the computation is repeated for 100 times accounting for the random process of bootstrapping. Results are shown in Figure 2.5. We observe that, for each fixed number of clusters, kernel k-means has higher stability scores than the classic k-means algorithms in general. Notably, the discrepancy between kernel k-means and the others in terms of their stability performance is even sharper when the number of clusters becomes large. In conclusion, evidence advocates again the use of kernel k-means over classic k-means algorithms in clustering rank data.

Mixture modeling is then used to fit the data and a partition of the votes is converted from the fitted models forming a clustering result. Baseline models are the Mallows mixture models fitted by the EM algorithm [Murphy 2003] using three different center-update algorithms at each iteration: brute-force search for Kemeny-optimal ranking, approximate ranking induced by Borda Count and Copeland's method. As proposed in this chapter, we embed the kernel trick in Mallows mixture

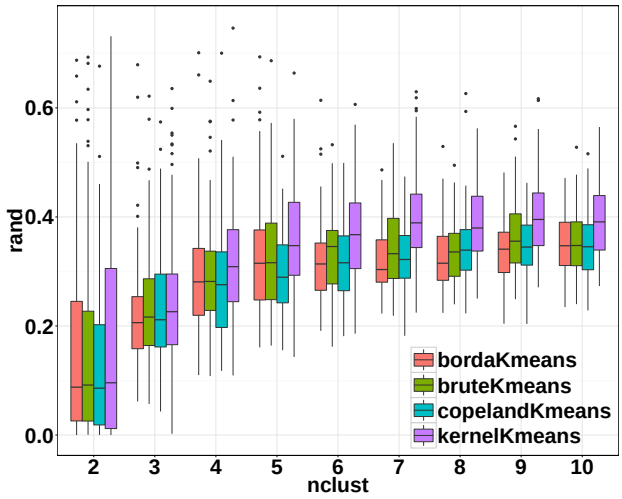


Figure 2.5: Across different number of clusters, Rand index between clustering assignments by running k-means algorithm on bootstrap replicas of the 1980 APA election data. For each fixed number of clusters, the boxplot represents the variance over 100 repeated runs.

modeling with two alternatives g (2.25) and $g^{\dagger}$ (2.26) in place of f (2.21). In each case, we vary the number of clusters ranging from 2 to 10 and the algorithm is repeated 50 times with randomly initialized configurations for each fixed number of clusters. As shown in Figure 2.6, modeling a Gaussian mixture to data in the feature space, or equivalently using $g^{\dagger}$ instead of f, provides a preferable split of the data into sub-groups with higher average silhouette scores across different number of clusters selected *a priori*.

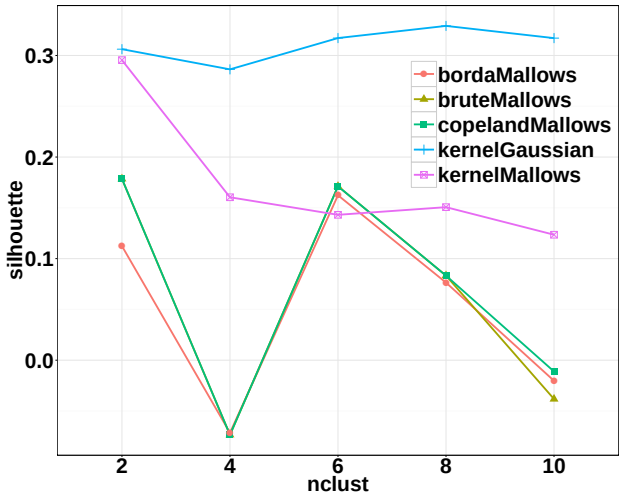


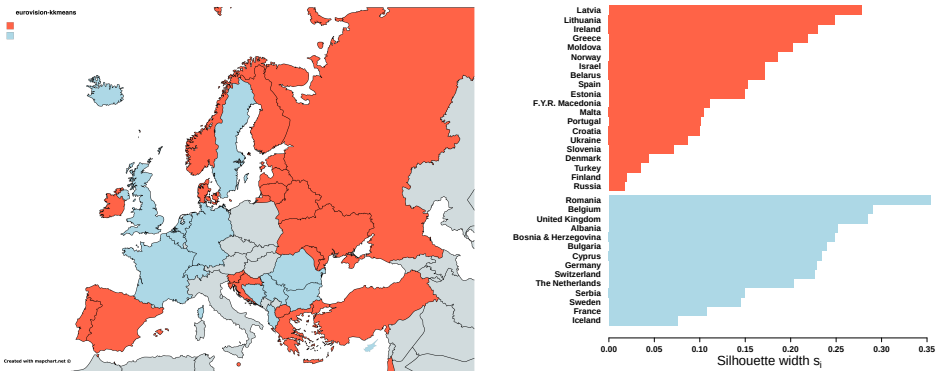
Figure 2.6: Average silhouette scores of Mallows mixture modeling methods across different number of clusters.

Clustering ESC voting data. We finally perform clustering on a dataset of multivariate partial rankings. In the finale of Eurovision Song Contest (ESC), each participating country casts one top-k vote over the finalists who represent their home country. Taken from [Jacques 2014], the dataset consists of 34 multivariate ranking instances, each being a series of 6 partial votes over top 8 finalists from 2007 to 2012 respectively.

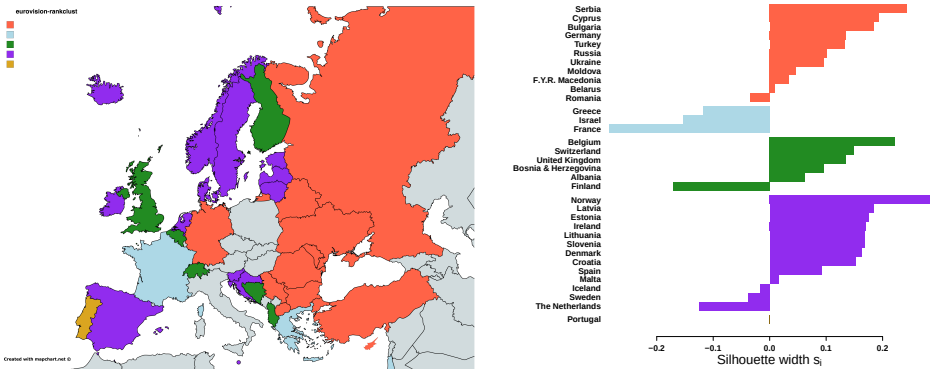
In comparison with the mixture of Insertion Sorting Rank (ISR) model for clustering multivariate partial rank data proposed by [Jacques 2014], we implement the kernel k-means algorithm (Algorithm 2.3) with the extended Kendall kernel to multivariate rankings (2.9) and equal weights $\mu_j = 1/p$ where $p = 6$ corresponding to the six contests across years. For each fixed number of clusters, the kernel k-means algorithm is repeated 100 times with randomly initialized configurations while 10 times for the ISR mixture modeling approach. We vary the number of clusters ranging from 2 to 6, and the optimal number is selected to be 2 for kernel k-means with respect to highest average silhouette score while 5 for the ISR mixture model with respect to highest BIC value. It consumes 2 hours in total to fit the ISR mixture model in order for clustering while it only takes less than 10 seconds to form the partition of data with kernel k-means. Although it is beyond the scope of this study to further explore the meaningful voting blocs, the colored map of Asia-Europe in terms of clustering results of participating countries to the ESC according to their voting behavior (Figure 2.7, Left) depicts that there exists interesting geographical alliances between countries in the voting data. For example, country-clusters returned by both algorithms present a sign of strong amity within Eastern Europe. Silhouette plots for both algorithms are shown in Figure 2.7 (Right). Despite a relatively small number of clusters selected for the kernel k-means, the silhouette plot (Figure 2.7a, Right) attests that the underlying clusters are well formed. Note that both silhouette plots opt for the same distance used by kernel k-means, which may show bias against a clustering scheme based on probabilistic modeling with ISR mixtures. However, the two approaches behave distinctly in identifying subgroups. For example, the ISR mixture model distinguishes Portugal as a singleton among all countries, while interpreting such clustering results remains to be studied. On the other hand, the k-means based approach tends to find more evenly distributed subgroups, in the sense that the number of individuals in each subgroup is more consistent. Therefore kernel k-means clustering is favored if the study of interest lies in populous behaviors in voting despite of potential outlier individuals. Notably the detection of outliers can be done by other kernel algorithms (Section 2.7).

2.6 Application: Supervised Classification of Biomedical Data

In this section we illustrate the relevance of supervised classification of rank data with an SVM using the Kendall kernel, when the ranking are derived from a high-dimensional real-valued vector. More precisely, we investigate the performance of



(a) Country-clusters returned by kernel k-means, where the number of clusters 2 is selected with respect to highest silhouette score averaged over all countries.



(b) Country-clusters returned by ISR mixture modeling, where the number of clusters 5 (including in particular "Portugal" as a singleton) is selected with respect to highest fitted BIC value.

Figure 2.7: Clustering results of participating countries to the ESC according to their voting behavior illustrated by geographic map and silhouette plot.

classifying high-dimensional biomedical data, motivated by previous work demonstrating the relevance of replacing numerical features by pairwise comparisons in this context [Geman 2004, Tan 2005, Xu 2005, Lin 2009].

For that purpose, we collected 10 datasets related to human cancer research publicly available online [Li 2003, Schroeder 2011, Shi 2011], as summarized in Table 2.1. The features are proteomic spectra relative intensities for the *Ovarian Cancer* dataset and gene expression levels for all the others. The contrasting classes are typically "Non-relapse v.s. Relapse" in terms of cancer prognosis, or "Normal v.s. Tumor" in terms of cancer identification. The datasets have no missing values, except the *Breast Cancer 1* dataset for which we performed additional preprocessing to remove missing values as follows: first we removed two samples (both labeled "relapse") from the training set that have around 10% and 45% of missing gene values; next we discarded any gene whose value was missing in at least one sample, amounting to a total of 3.5% of all genes.

Table 2.1: Summary of biomedial datasets.

Dataset	No. of features	No. of samples (training/test)		Reference
		C_1	C_2	
Breast Cancer 1	23624	44/7 (Non-relapse)	32/12 (Relapse)	[van 't Veer 2002]
Breast Cancer 2	22283	142 (Non-relapse)	56 (Relapse)	[Desmedt 2007]
Breast Cancer 3	22283	71 (Poor Prognosis)	138 (Good Prognosis)	[Wang 2005b]
Colon Tumor	2000	40 (Tumor)	22 (Normal)	[Alon 1999]
Lung Adenocarcinoma 1	7129	24 (Poor Prognosis)	62 (Good Prognosis)	[Beer 2002]
Lung Cancer 2	12533	16/134 (ADCA)	16/15 (MPM)	[Gordon 2002]
Medulloblastoma	7129	39 (Failure)	21 (Survivor)	[Pomeroy 2002]
Ovarian Cancer	15154	162 (Cancer)	91 (Normal)	[Petricoin 2002]
Prostate Cancer 1	12600	50/9 (Normal)	52/25 (Tumor)	[Singh 2002]
Prostate Cancer 2	12600	13 (Non-relapse)	8 (Relapse)	[Singh 2002]

We compare the Kendall kernel to three standard kernels, namely linear kernel, homogeneous 2nd-order polynomial kernel and Gaussian RBF kernel with bandwidth set with "median trick", using SVM (with regularization parameter C) and Kernel Fisher Discriminant (KFD, without tuning parameter) as classifiers. In addition, we include in the benchmark classifiers based on Top Scoring Pairs (TSP) [Geman 2004], namely (1-)TSP, k-TSP [Tan 2005][1] and APMV (all-pairs majority votes, i.e. $\binom{n}{2}$-TSP). Finally we also test SVM with various kernels using as input only top features selected by TSP [Shi 2011].

In all experiments, each kernel is centered (on the training set) and scaled to unit norm in the feature space. For KFD-based models, we add 10^{-3} on the diagonal of the centered and scaled kernel matrix, as suggested by [Mika 1999]. The Kendall kernel we use in practice is a soft version to (2.10) in the sense that the extremes

[1] While the original k-TSP algorithm selects only top k *disjoint* pairs with the constraint that k is *less than* 10, we do not restrict ourselves to any of these two conditions since we consider k-TSP in this study essentially a feature pair scoring algorithm.

± 1 can still be attained in the presence of ties, specifically we use

$$K_\tau(\mathbf{x}, \mathbf{x}') = \frac{n_c(\mathbf{x}, \mathbf{x}') - n_d(\mathbf{x}, \mathbf{x}')}{\sqrt{(n_0 - n_1)(n_0 - n_2)}},$$

where $n_0 = \binom{n}{2}$ and n_1, n_2 are the number of tied pairs in $\mathbf{x}, \mathbf{x}'$ respectively.

For the three datasets that are split into training and test sets, we report the performance on the test set; otherwise we perform a 5-fold cross-validation repeated 10 times and report the mean performance over the $5 \times 10 = 50$ splits to evaluate the performance of the different methods. In addition, on each training set, an internal 5-fold cross-validation is performed to tune parameters, namely the C parameter of SVM-based models (optimized over a grid ranging from 10^{-2} to 10^3 in log scale), and the number k of k-TSP in case of feature selection (ranging from 1 to 5000 in log scale).

Table 2.2 and Figure 2.8 summarize the performance of each model across the datasets. An SVM with the Kendall kernel achieves the highest average prediction accuracy overall (79.39%), followed by a linear SVM trained on a subset of features selected from the top scoring pairs (77.16%) and a standard linear SVM (76.09%). The SVM with Kendall kernel outperforms all the other methods at a p-value of 0.07 according to a Wilcoxon rank test. We note that even though models based on KFD generally are less accurate than those based on SVMs, the relative order of the different kernels is consistent between KFD and SVM, adding evidence that the Kendall kernel provides an interesting alternative to other kernels in this context. The performance of TSP and k-TSP, based on majority vote rules, are comparatively worse than SVMs using the same features, as already observed by [Shi 2011].

We further studied how the performance of different kernels depends on the choice of the C parameter or the SVM (Figure 2.9), and on the number of features used (Figure 2.10), on some representative datasets. We observe that compared to other kernels, an SVM with the Kendall kernel is relatively insensitive to hyper-parameter C especially when C is large, which corresponds to a hard-margin SVM. This may explain in part the success of SVMs in this setting, since the risk of choosing a bad C during training is reduced. Regarding the number of features used in case of feature selection, we notice that it does not seem to be beneficial to perform feature selection in this problem, explaining why the Kendall kernel which uses all pairwise comparisons between features outperforms other kernels restricted to a subset of these pairs. In particular, the feature space of the Kendall and Mallows kernels is precisely the space of binary pairwise comparisons defined by [Geman 2004], and the results show that instead of selecting a few features in this space as the Top Scoring Pairs (TSP)-family classifiers do [Geman 2004, Tan 2005, Xu 2005, Lin 2009], one can simply work with *all* pairs with the kernel trick.

Finally, as a proof of concept we empirically compare on one dataset the smooth alternative (2.13) and its Monte Carlo approximate (2.15) with the original Kendall kernel. We studied how the performance varies with the amount of noise added to the samples (Figure 2.11), and how the performance varies with the number of

Table 2.2: Prediction accuracy (%) of different methods across biomedical datasets (ordered by decreasing average accuracy across datasets). Models are named after candiate methods (SVM or KFD) and candiate kernels, namely linear kernel (linear), 2nd-order homogeneous polynomial kernel (poly), Gaussian RBF kernel (rbf) or Kendall kernel (kdt), and whether feature selection is combined (TOP) or not (ALL). Prediction accuracy of the best-performing models for each dataset is in boldface.

	Average	BC1	BC2	BC3	CT	LA1	LC2	MB	OC	PC1	PC2
SVMkdtALL	79.39	78.95	71.31	67.34	85.78	**70.98**	97.99	63.67	99.48	**100.00**	58.40
SVMlinearTOP	77.16	**84.21**	69.29	67.11	84.19	63.92	97.32	**65.17**	99.41	85.29	55.70
SVMlinearALL	76.09	78.95	**71.67**	64.27	86.73	70.23	97.99	62.67	99.64	73.53	55.17
SVMkdtTOP	75.50	52.63	70.61	65.81	85.46	67.70	97.99	58.33	99.92	97.06	59.47
SVMpolyALL	74.54	68.42	71.62	63.66	78.43	70.53	98.66	61.17	99.28	79.41	54.23
KFDkdtALL	74.33	63.16	59.41	67.22	85.46	59.08	**99.33**	59.33	98.73	97.06	54.57
kTSP	74.03	57.89	58.22	64.47	**87.23**	61.70	97.99	56.00	99.92	**100.00**	56.83
SVMpolyTOP	73.99	63.16	69.44	66.26	79.14	65.98	**99.33**	60.00	99.21	88.24	49.10
KFDlinearALL	71.81	63.16	60.43	**67.52**	77.26	57.24	97.99	59.50	**100.00**	73.53	**61.43**
KFDpolyALL	71.39	63.16	60.48	67.38	75.10	58.52	97.99	60.33	**100.00**	73.53	57.43
TSP	69.71	68.42	49.58	57.80	85.61	58.96	95.97	52.67	99.80	76.47	51.83
SVMrbfALL	69.31	63.16	71.41	65.87	81.18	70.84	93.96	63.83	98.85	26.47	57.50
KFDrbfALL	66.50	63.16	60.38	66.17	84.33	58.62	97.32	60.17	98.34	26.47	50.00
APMV	61.91	**84.21**	65.98	33.96	64.49	33.60	89.93	42.17	85.19	73.53	46.00

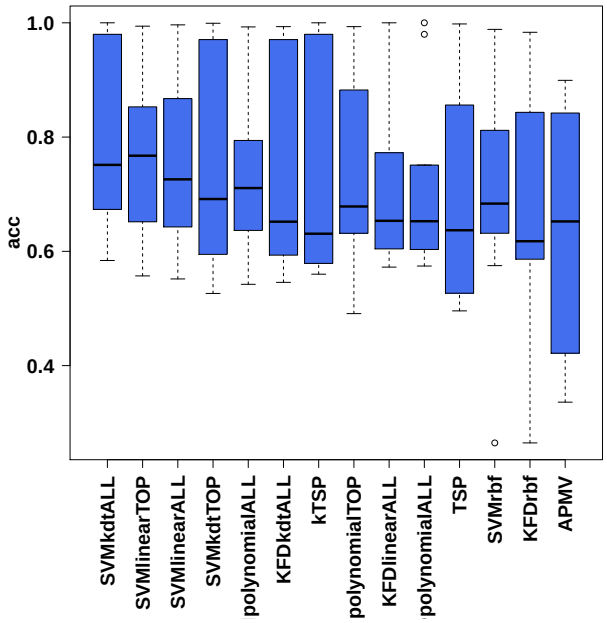


Figure 2.8: Model performance comparison (ordered by decreasing average accuracy across datasets).

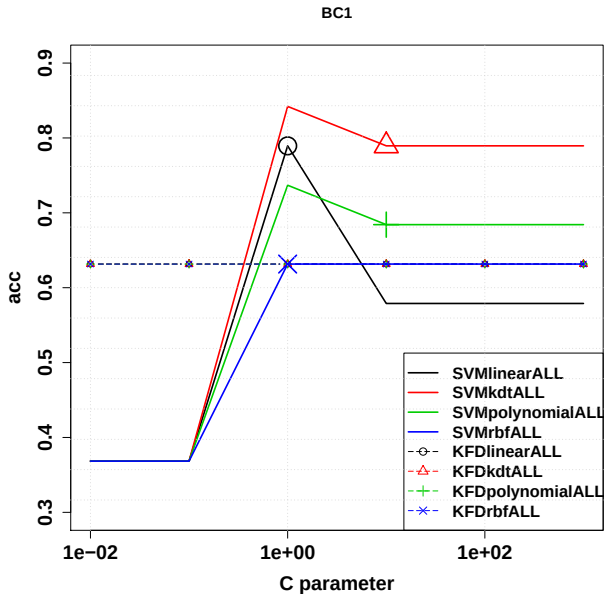


Figure 2.9: Sensitivity of kernel SVMs to C parameter on the *Breast Cancer 1* dataset. (Special marks on SVM lines denote the parameter returned by cross-validation.)

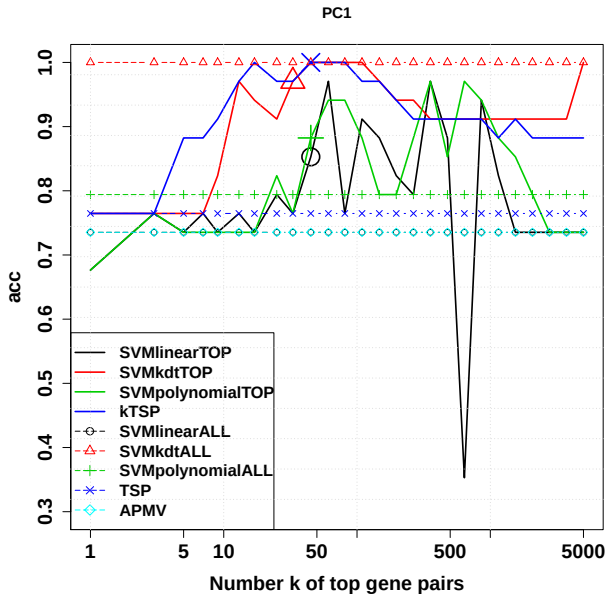


Figure 2.10: Impact of TSP feature selection on the *Prostate Cancer 1* dataset. (Special marks on SVM lines denote the parameter returned by cross-validation.)

samples in the Monte Carlo scheme for a given amount of noise (Figure 2.12). It confirms that the smooth alternative (2.13) can improve the performance of the Kendall kernel, and that the amount of noise (window size) should be considered as a parameter of the kernel to be optimized. Although the D^2-sample Monte Carlo approximate kernel (2.15) mainly serves as a fast estimate to the exact evaluation of (2.13), it shows that the idea of jittered input with specific noise can also bring a tempting benefit for data analysis with Kendall kernel, even when D is small. This also justifies the motivation of our proposed smooth alternative (2.13). Last but not least, despite the fact that the convergence rate of D^2-sample Monte Carlo approximate to the exact kernel evaluation is guaranteed by Theorem 2.3, experiments show that the convergence in practice is typically faster than the theoretical bound, and even faster in case that the window size a is small. This is due to the fact that the convergence rate is also dependent of the observed data distribution in the input space, for which we have not made any specific assumption in our analysis.

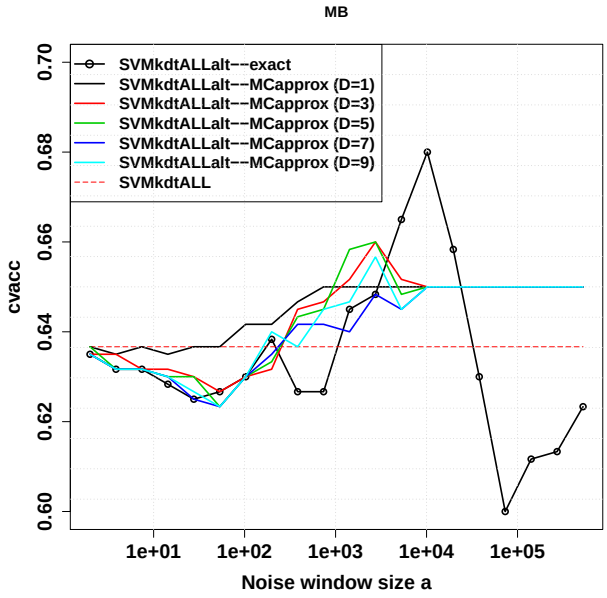


Figure 2.11: Empirical performance of smoothed alternative to Kendall kernel on the *Medulloblastoma* dataset.

2.7 Discussion

Based on the observation that the popular Kendall tau correlation between total rankings is a positive definite kernel, we presented some extensions and applications pertaining to learning with the Kendall kernel and the related Mallows kernel. We showed that both kernels can be evaluated efficiently in $O(n \log n)$ time, and that

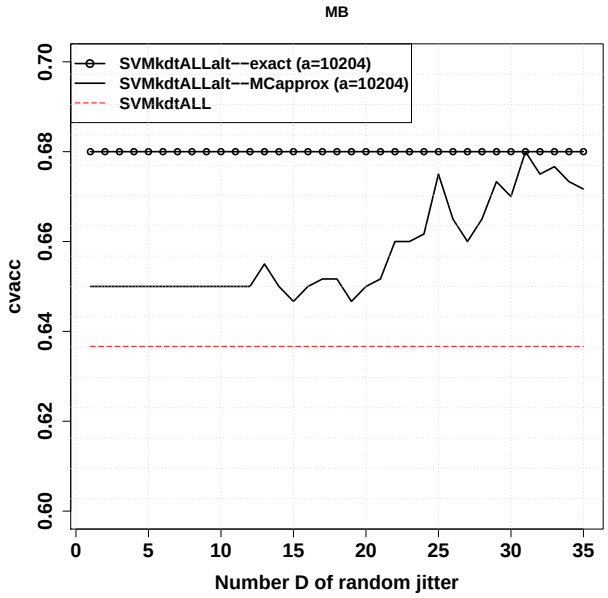


Figure 2.12: Empirical convergence of Monte Carlo approximate at the fixed window size attaining maximum underlying accuracy from the left plot.

the Kendall kernel can be extended to partial rankings containing k items out of n in $O(k \log k)$ time as well as to multivariate rankings. When permutations are obtained by sorting real-valued vectors, we proposed an extension of the Kendall kernel based on random perturbations of the input vector to increase its robustness to small variations, and discussed two possible algorithms to compute it. We further highlighted a connection between the fast Mallow kernel and the diffusion kernel of [Kondor 2010]. We also reported promising experimental results on clustering of heterogeneous rank data and classifying biomedical data demonstrating that for highly noisy data, the Kendall kernel is competitive or even outperforms other state-of-the-art kernels.

We believe that computationally efficient kernels over the symmetric group pave the way to numerous applications beyond the ones we pursued in this chapter. In unsupervised data mining, kernel density estimation for example can be applied to modeling the distribution over a collection of rankings, and by the representer theorem the resulting distribution depends solely on the observed data points circumventing the exponentially large cardinality of the symmetric group, from which a consensus ranking that best represents the data is the one with the highest probability. As more complicated cases, there is much interest beyond finding a single consensus ranking typically in the context of political votes or social choices: groups of homogeneous sub-populations in data can be clustered by algorithms such as kernel k-means or spectral clustering [Filippone 2008]; dependencies or principle structural factors in data can be found by kernel canonical correlation analysis [Lai 2000]

or kernel principle component analysis [Schölkopf 1999a]; outliers in a collection of rank data can be detected with one-class SVMs [Schölkopf 1999b, Tax 2004]. In a more predictive setting, Support Vector Machines and kernel ridge regression are representative delegates for solving classification and regression problems amongst many other kernel algorithms [Schölkopf 2002]. Notably, the input/output kernels formalism allows us to predict rankings as well as learn from rankings where a wealth of algorithms such as multi-class SVMs or structural SVMs [Crammer 2002, Tsochantaridis 2005, Bakir 2007] are ready to suit the problem at hand.

Deeper understanding of the Kendall and Mallows kernels calls for more theoretical work of the proposed kernels. In particular, a detailed analysis of the Fourier spectra of the Kendall and Mallows kernels is provided in [Mania 2016]. Those authors also introduced a tractable family of normalized polynomial kernels of degree p that interpolates between Kendall (degree one) and Mallows (infinite degree) kernels.

There are many interesting extensions of the current work. One direction would be to include high-order comparisons in measuring the similarity between permutations. Since the fast computation of the Kendall and Mallows kernels is balanced by the fact that they only rely on pairwise statistics between the ranks, computationally tractable extension to higher-order statistics, such as three-way comparisons, could potentially enhance the discriminative power of the proposed kernels. Another interesting direction would be to extend the proposed kernels to rankings on partially ordered set. In fact, the current work lies on the assumption that a (strict) total order can be associated with the (finite) set of items given to rank $\{x_1, \ldots, x_n\}$, which is implicitly presumed when we label the items by the subscripts $[\![n]\!]$ and then define the Kendall and Mallows kernels by comparing all item pairs (i, j) for $i < j$ (Section 2.2). However, there are cases when the item set is intrinsically associated with a (strict) partial order such that some item pairs are conceptually incomparable. In that case, we can collect all comparable item pairs into a set denoted by E and define the kernels by comparing only those item pairs (i, j) in E. Notably evaluating the extended kernels is still fast as we can simply replace the Merge Sort algorithm for total orders (Section 2.2) by a topological sort algorithm for partial orders [Cormen 2009, Section 22.4]. We leave further investigations of this generalization to future work.

Chapter 3

Network-based Wavelet Smoothing for Analysis of Genomic Data

Publication and Dissemination: *The work in this chapter is under preparation for submission as joint work with Jean-Philippe Vert in [Jiao 2017c].*

Abstract: *Biological networks are a common way of describing information on relationships between genes that are accumulated from many years of biomedical research, and they are thus potentially valuable when incorporated as prior knowledge to guide biomarker discovery in genomic data analysis. In this chapter, we focus on network-based regularization methods through a predictive framework with linear models, and propose to use a class of methods based on wavelet smoothing over undirected graphs that directly detect subnetworks composing of collaboratively functional gene modules. We perform breast cancer survival analysis using a large gene expression dataset and a protein-protein interaction network obtained from a public database, and demonstrate that the proposed methods are able to improve gene selection in terms of stability, connectivity and interpretability while achieving competitive performance of survival risk prediction. Our results also serve a comparative study benchmarking several network-free and network-based regularization methods for gene selection related to breast cancer survival.*

Résumé : *Les réseaux biologiques sont un moyen classique de représenter l'information sur les relations entre les gènes qui sont accumulées depuis de nombreuses années en recherche biomédicale. Il est donc intéressant de les incorporés comme connaissances préalables pour aider à la découverte de biomarqueurs dans l'analyse des données génomiques. Dans ce chapitre, nous nous concentrons sur les méthodes de régularisation par de tels réseaux dans un cadre prédictif avec des modèles linéaires, dans de tels cas nous proposons d'utiliser une classe de méthodes basées sur le*

débruitage par ondelettes sur un graphe non orienté qui détectent directement les sous-réseaux composés de modules de gènes qui sont collaborativement fonctionnels. Nous effectuons une analyse de survie du cancer du sein à l'aide d'un grand ensemble de données d'expression génétique et d'un réseau d'interactions protéine-protéine obtenu à l'aide d'une base de données publique. Nous démontrons que les méthodes proposées sont capables d'améliorer la sélection de gènes en termes de robustesse, de connectivité et d'interprétation à performance égale en prédiction du risque de survie. Nos résultats fournissent également une étude comparative de plusieurs méthodes de régularisation sans réseau et basées sur un réseau donné pour la sélection de gènes liés à la survie du cancer du sein.

3.1 Introduction

Genomic data analysis is a rapidly developing research area that receives increasing attention, thanks to the recent advancement of technologies in gene expression profiling that monitors the activity of a large number of genes in a single experiment. Recall from Section 1.3, identifying genes related to a clinical phenotype of interest, such as drug resistance or disease progression, is a central yet challenging topic in genomic research commonly known as *biomarker discovery*. A typical approach follows a predictive framework that builds a predictive model linking genomic data to a clinical outcome further combined with a regularization method for feature selection and to address the high-dimensionality of high-throughput genomic data. For example, linear regression can be used to model the relationship between the quantitative measurements of toxicity response to a drug and the expression levels of all genes, and many regularization methods have been proposed in literature for identifying a few genes that are potentially related to the targets of the drug, including the well-studied and widely-used lasso [Tibshirani 1996] and elastic net [Zou 2005].

Despite the usefulness of the lasso-type regularization methods, the genes selected purely by such algorithmic approaches often lack proper mechanistic interpretation in terms of biological relevance and it remains a demanding task to determine *a posteriori* whether and how the selected genes cooperate in some biological process. In fact, plentiful information about the interaction between gene products and the underlying biological functions is accumulated from extensive biomedical research over the years and can be obtained through many publicly available databases, including Human Protein Reference Database (HPRD) [Keshava Prasad 2009], Gene Ontology (GO) [Ashburner 2000] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kanehisa 2000]. Although these databases can help verify the collaborative functionality of a list of selected genes, the interpretation is usually non-trivial as many selected genes may seem unrelated or even

unreliable.

To overcome the difficulty of *a posteriori* interpretation of selected genes, there is therefore a need to develop methods that integrate prior knowledge in the process of biomarker detection in genomic data analysis in order to promote biological relevance. Although various databases register different aspects of interaction between gene products, this information is usually provided in the form of biological networks, which can be represented by graphs where vertices are genes and edges indicate some notion of interaction between the gene products of connected vertices. It has been demonstrated that genes closer on the network are more likely to be involved in similar biological functions and vice versa (see, e.g., [Stuart 2003]). The incorporation of biological networks to enhance biomarker discovery consequently follows the principle that genes closer on the network should tend to be selected simultaneously. Note that this research topic is commonly phrased as *network-guided biomarker discovery*, for which we refer to a tutorial-oriented survey by [Azencott 2016].

In particular, we are interested in network-guided feature selection under a predictive framework via regularization methods. This network-based regularization should encourage that genes closer on the network contribute similarly to the predictive model built for some clinical phenotype and then, if selected, tend to be selected simultaneously. Notably, Laplacian regularization (Section 3.2.2) is a classic method that attempts to encourage smoothness between the coefficients of a linear model corresponding to neighboring features on the given network [Belkin 2004] but the regularization itself does not enforce sparsity nor enable feature selection. [Li 2008] proposed to combine sparsity-inducing penalties such as lasso with the Laplacian regularization for network-constrained feature selection with an application in genomic data analysis, and is further generalized in an adaptive fashion by [Li 2010]. Besides Laplacian-based methods, many others stem from extending the standard lasso to structured regularization for identifying genes by groups. If meaningful gene groups can be defined by known functional gene modules on the network for instance, Group lasso [Jacob 2009, Yuan 2006] can be used to force that genes belonging to the same groups are selected or disregarded simultaneously. Graph-fused lasso [Tibshirani 2005] encourages that the direct neighboring genes on the network share the same coefficients in a linear model, leading to the a partition of the network into subnetworks as functional gene modules. Pairwise L_γ penalty [Pan 2010] aims to model the intuition that direct neighboring genes in a network should be more likely to participate in the same biological process and thus tend to be selected simultaneously. Note that Laplacian regularization, which is intended for global smoothness, and group lasso, which is designed for group-wise gene signature selection, can also be combined [Tian 2013].

We propose in this study to use a class of regularization methods for network-guided feature selection under a predictive modeling framework that simultaneously enjoy global smoothness over the network and directly identify subnetworks consisting of a few connected features. In fact, since the Laplacian regularization is known to be equivalent to a quadratic penalty with respect to the graph spectral

domain [Belkin 2004], the method essentially performs a graph Fourier transform on the coefficients of the linear models and then attenuates the high-frequency components thereof, therefore inducing global smoothness for the predictive models. A wealth of studies in the field of signal processing have been devoted to extending Fourier smoothing in order to simultaneously achieve spatial-temporal sparse coding, a topic that has been well established for data of regular structure such as time series or images [Mallat 1999] and has attracted much attention for data residing on irregular structure such as general graphs or manifolds [Shuman 2013]. Following this trend, we propose to study network-based wavelet smoothing with an application to gemonic data analysis. Notably, when applied to biological networks, the global smoothness as well as localization properties of the network-based wavelet smoothing estimates a predictive model that directly enables the detection of subnetworks readily translated into functional gene modules, rendering interpretable biological insights concerning the particular phenotype of interest.

The chapter is organized as follows. In Section 3.2 we first elaborate the predictive modeling framework with regularization, with a particular emphasis on reviewing the network-based methods in literature, and then propose to use a class of novel methods based on wavelet smoothing on graphs. In Section 3.3, we perform simulated experiments and breast cancer survival analysis with gene expression data guided by a protein-protein interaction (PPI) network obtained from HPRD database. Promising results demonstrate the usefulness of the proposed methods for biomarker discovery related to breast cancer survival, while they serve a comparative study benchmarking several methods of network-free and network-guided biomarker discovery. Finally we conclude and discuss in Section 3.4.

3.2 Methods

3.2.1 Feature Selection Under Predictive Modeling Framework

In supervised learning, let $\mathcal{D} := \{(\mathbf{x}_i, y_i) : i = 1, \ldots, m\}$ denote a dataset of m observations where each $\mathbf{x}_i \in \mathbb{R}^n$ is an n-dimensional feature vector that is paired with a quantitative measurement of some clinical phenotype y_i to predict depending on the particular application. For instance, the feature values in $\mathbf{x}_i = (x_{i1}, \ldots, x_{in})^\top$ can denote the expression levels of n genes of sample i, while the quantity of the clinical phenotype can be a response $y_i \in \mathbb{R}$ measuring the resistance to a drug of the sample, or a binary label $y_i \in \{-1, +1\}$ denoting whether a specific treatment is applied to the cancer patient, or a right-censored survival time $y_i = (T_i, \Delta_i) \in \mathbb{R} \times \{0, 1\}$ where, for some predetermined censoring time C, $(T_i|\Delta_i = 1)$ denotes the observed survival time of the diseased patient prior to C or $(T_i|\Delta_i = 0)$ is equal to the censoring time C meaning that the information is censored. Typically in biomedical applications, the number of genes is usually larger than the number of observations, i.e., $n > m$ or even $n \gg m$. We further assume that the feature data are standardized to have zero mean and unit variance, i.e., $\sum_{i=1}^m x_{ij}/m = 0$ and $\sum_{i=1}^m x_{ij}^2/(m-1) = 1$ for $j = 1, \ldots, n$.

We consider linear models in this study where the quantity y is linked with the underlying feature vector $\mathbf{x}$ via a linear combination $\beta^\top \mathbf{x}$ for some coefficient vector $\beta \in \mathbb{R}^n$. For simplicity of notations, we do not consider intercepts explicitly in the discussion, but it can be easily included in the model by augmenting the feature vector with a dummy variable that takes constant value 1. Given the dataset $\mathcal{D}$, an empirical procedure to estimate the coefficients β is by solving optimization problems of the form:

$$\min_{\beta \in \mathbb{R}^n} \ell(y_1, \ldots, y_m, \beta^\top \mathbf{x}_1, \ldots, \beta^\top \mathbf{x}_m)\,, \tag{3.1}$$

where ℓ is a loss function measuring the empirical cost on the training data and should be carefully designed for specific application. For example, when $y_i \in \mathbb{R}$ and

$$\ell(y_1, \ldots, y_m, \beta^\top \mathbf{x}_1, \ldots, \beta^\top \mathbf{x}_m) = \frac{1}{m} \sum_{i=1}^{m} (y_i - \beta^\top \mathbf{x}_i)^2\,, \tag{3.2}$$

it recovers linear regression and returns the least squares solution; when $y_i \in \{-1, +1\}$ and

$$\ell(y_1, \ldots, y_m, \beta^\top \mathbf{x}_1, \ldots, \beta^\top \mathbf{x}_m) = \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-y_i \beta^\top \mathbf{x}_i))\,, \tag{3.3}$$

it recovers logistic regression for binary classification; when $y_i = (T_i, \Delta_i) \in \mathbb{R} \times \{0, 1\}$ and

$$\ell(y_1, \ldots, y_m, \beta^\top \mathbf{x}_1, \ldots, \beta^\top \mathbf{x}_m) = -\frac{1}{m} \sum_{i=1}^{m} \Delta_i \{\beta^\top \mathbf{x}_i - \log(\sum_{j: T_j > T_i} \exp(\beta^\top \mathbf{x}_j))\}\,, \tag{3.4}$$

it recovers the Cox proportional hazards model for survival analysis [Cox 1972]. We would like to point out that the choice of the loss function is mostly application-oriented and is primarily not the concern of this study. In particular, the following discussion and our proposed methods will not depend on the choice of the loss function.

In order to avoid overfitting or to enable feature selection, it is usually suggested to incorporate appropriate regularization in (3.1). Specifically, we aim to solve optimization problems of the form:

$$\min_{\beta \in \mathbb{R}^n} \ell(y_1, \ldots, y_m, \beta^\top \mathbf{x}_1, \ldots, \beta^\top \mathbf{x}_m) + \lambda P(\beta)\,, \tag{3.5}$$

where P is a penalty term, especially one that induces sparsity for feature selection, and $\lambda \geq 0$ is a regularization parameter that trades off between the loss term and the penalty term. Classic penalty terms include the *ridge* [Hoerl 1970] which penalizes the squared L_2 norm of the coefficient vector, i.e.,

$$P^{\text{ridge}}(\beta) = \|\beta\|_2^2\,. \tag{3.6}$$

It is well-known that the ridge regularization is a shrinkage method that usually yields more robust estimate but does not enable feature selection. The *lasso* [Tibshirani 1996] is probably the simplest and most widely used sparsity-inducing regularization method which penalizes the L_1 norm of the coefficient vector, i.e.,

$$P^{\text{lasso}}(\beta) = \|\beta\|_1 \,. \tag{3.7}$$

The lasso enables feature selection by allowing a few non-zero coefficients in the estimate of β. Another widely used penalty extends the lasso by penalizing a weighted sum of the L_1 norm and squared L_2 norm on the coefficient vector that leads to a regularization method called *elastic net* [Zou 2005], i.e.,

$$P^{\text{e-net}}(\beta;\nu) = \nu\|\beta\|_1 + (1-\nu)\frac{1}{2}\|\beta\|_2^2 \,, \tag{3.8}$$

where $0 \leq \nu \leq 1$ is a regularization parameter balacing between the two norms. In particular, the elastic net regularization reduces to the ridge when $\nu = 0$, and enables feature selection for $0 < \nu \leq 1$ including as a special case the lasso when $\nu = 1$. Despite the fact that the elastic net regularization usually helps produce more reliable solutions when applied to biomedical data, the selected genes corresponding to non-zero coefficients often give no clear biological meaning in terms of their collaborative functionality. One possible solution is to devise a $L_{2,1}$ mixed norm, often termed as *group lasso* [Yuan 2006, Jacob 2009], to force that eventually certain genes that belong to the same group, if selected, will be selected simultaneously. The groups of genes are defined *a priori* by external knowledge. For example, genes that belong to the same pathway or contribute to the same biological process can be grouped together. However, as most biomedical databases provide domain-specific knowledge on gene interactions in the form of biological networks, we are interested in directly exploiting the network structure in order to guide biomarker discovery.

3.2.2 Network-guided Feature Selection: A Review of Related Work

Suppose a network that specifies the relationships between features is represented by an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$, where $\mathcal{V}$ denotes the set of vertices representing the n features indexed by $\{1, \ldots, n\}$, $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ denotes the set of edges with $(u, v) \in \mathcal{E}$ representing a link between vertices u and v, $A \in \mathbb{R}^{n\times n}$ is the weighted adjacency matrix such that $A_{uv} = A_{vu} =: a(u, v)$ with $a(u, v) > 0$ the weight assigned to an edge $(u, v) \in \mathcal{E}$ and $a(u, v) = 0$ if $(u, v) \notin \mathcal{E}$. In particular, we assume that no self-loop exists, i.e., $(u, u) \notin \mathcal{E}$ for any vertex u. Depending on the network under consideration, the weighted edges can be used to register the uncertainty of the existence of a link or the strength of the interaction between connected vertices.

Recall that a common assumption for network-guided feature selection under predictive framework is that we would like to devise penalty terms that encourage the coefficients corresponding to those features closer on the network to be similar so that, if selected, they tend to be selected together. To this end, let us first

define the graph Laplacian $L = D - A$ where D is a diagonal matrix with $D_{uu} = \sum_{v \neq u} a(u, v) =: d(u)$ the degree of the vertex u, i.e.,

$$L_{uv} = \begin{cases} d(u) & \text{if } u = v, \\ -a(u, v) & \text{if } (u, v) \in \mathcal{E}, \\ 0 & \text{otherwise.} \end{cases}$$

The graph Laplacian is a key concept in spectral graph theory that shares many properties with the Laplace operator on compact Riemannian manifolds and reflects many properties of the graph structure [Chung 1997]. For example, L is a symmetric, positive semi-definite matrix whose number of zero eigenvalues is equal to the number of maximally connected components of the graph. Note that some authors prefer to use alternatively the *normalized* graph Laplacian defined as $\mathcal{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$, particularly accounting for the degrees of different vertices. While the discussion in this study does not depend on which version of Laplacian (normalized or non-normalized) is used, they usually give very different results in practice as we will empirically demonstrate in Section 3.3. An important observation regarding the graph Laplacian is that it can be used to define measure of smoothness with respect to the graph structure for any vector whose covariates naturally reside on the vertices of the graph. We define the penalty term for *Laplacian regularization* as

$$P^{\text{lap}}(\beta) = \beta^\top L \beta = \sum_{(u,v) \in \mathcal{E}} (\beta_u - \beta_v)^2 a(u, v)\,. \tag{3.9}$$

In words, the Laplacian regularization method "shrinks" the pairwise difference between neighboring features to be small taking into account the edge weights and hence encourages solutions to be smooth over the graph.

It is very interesting to understand how the Laplacian regularization (3.9) achieves global smoothness from a spectral perspective. As $L \in \mathbb{R}^{n \times n}$ is symmetric and semi-positive, we have the eigendecomposition

$$L = X \Lambda X^\top$$

for an orthogonal matrix $X = (\chi_1 | \dots | \chi_n)$ where χ_i denotes the i-th column of X and a diagonal matrix $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ with $0 = \lambda_1 \leq \dots \leq \lambda_n$. By analogy to the Laplace operator on Riemannian manifolds, the eigenbasis of L, namely $\chi_1, \dots, \chi_n$, forms the Fourier basis of the graph spectral domain with "frequencies" $\lambda_1, \dots, \lambda_n$ respectively. $\hat{\beta} := X^\top \beta$ with coordinates $\hat{\beta}_i = \chi_i^\top \beta$ is called the *Fourier transform* of β, and $\beta = X\hat{\beta} = \sum_{i=1}^n \hat{\beta}_i \chi_i$ gives the *inverse Fourier transform*. Since we have

$$\beta^\top L \beta = \| X \Lambda^{\frac{1}{2}} X^\top \beta \|_2^2 = \hat{\beta}^\top \Lambda \hat{\beta} = \sum_{i=1}^n \lambda_i \hat{\beta}_i^2\,, \tag{3.10}$$

the penalty term of the Laplacian regularization is essentially a weighted squared L_2 norm of the Fourier transform of β in the graph spectral domain with frequencies acting as the weights. In other words, the Laplacian regularization attenuates the high-frequency components, thereby inducing global smoothness of β over the

graph. Following this direction, [Rapaport 2007] studied a spectral method generalizing the Laplacian regularization by considering functions of the frequencies as weights in the norm that allow finer control over the estimated coefficients and applied the method to classify microarray data.

However, due to the non-singularity of its quadratic form, the Laplacian regularization (3.9) alone does not enable feature selection. [Li 2008] suggested to combine it with the lasso leading to a penalty that enables network-constrained feature selection which we call *Laplacian lasso*, i.e.,

$$P^{\text{laplasso}}(\beta;\nu) = \nu\|\beta\|_1 + (1-\nu)\beta^\top L\beta\,, \tag{3.11}$$

where $0 \leq \nu \leq 1$ is a regularization parameter balancing between the lasso term for sparsity and the Laplacian term for smoothness. Specifically, the Laplacian term achieves global smoothness by attenuating high-frequency components in β and the lasso term allows selection of a few relevant features potentially connected on the network. Detailed analysis on the grouping effect and asymptotic properties of the penalty is found in [Li 2010]. Note that the penalty proposed by the authors appears with the normalized Laplacian instead. A possible extension of the Laplacian lasso regularization is to replace the lasso term by a group lasso term in (3.11) so that features are forced to be selected effectively by predetermined groups [Tian 2013]. However, to define such meaningful groups requires extra effort and domain expertise concerning specific application.

Another strategy, often termed as *graph-fused lasso* [Tibshirani 2005], directly extends the Laplacian regularization (3.9) by replacing the squared 2-norm by 1-norm on the pairwise difference of connected features, i.e.,

$$P^{\text{gflasso}}(\beta) = \sum_{(u,v)\in\mathcal{E}} |\beta_u - \beta_v| a(u,v)\,. \tag{3.12}$$

This regularization method results in a piece-wise constant estimate of the coefficient vector that achieves smoothness and structured sparsity simultaneously. A general class of penalties that induce structured sparsity, often termed as *generalized lasso* [Tibshirani 2011], is written in the form of

$$P^{\text{genlasso}}(\beta) = \|D^\top\beta\|_1\,, \tag{3.13}$$

where $D \in \mathbb{R}^{n\times d}$ is predefined and reflects the structure of desired sparsity in β by d linear constraints. In particular, the generalized lasso (3.13) reduces to the ordinary lasso (3.7) when D take the identity matrix, and encapsulates the graph-fused lasso (3.12) as a special case when D takes the oriented incidence matrix of the undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$ with any orientation, i.e., $D \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{E}|}$ such that, for each column indexed by edge e connecting vertices u and v, $a(u,v)$ appears in the row corresponding to one vertex of e, $-a(u,v)$ appears in the row corresponding to the other vertex of e, 0 appears in all other rows. As we will see shortly, another interesting choice of D takes in columns an orthogonal system of wavelet basis in order to promote structured spatial smoothness, in which case the generalized lasso recovers special cases of wavelet smoothing. This last observation motivates us to study two wavelet-based regularization methods.

3.2.3 Network-based Wavelet Smoothing for Feature Selection

All the above-mentioned network-based regularization methods share the objective of obtaining an estimate of the coefficient vector that enjoys global smoothness and selects features that preferably form subnetworks over the graph. Towards the same goal, our idea is to consult graph wavelets and wavelet smoothing that are well-known in the field of signal processing to achieve simultaneous localization in both frequency and space, former attempting global smoothness over the graph and latter granting the ability to detect subnetwork directly.

Suppose $\mathcal{G}$ is a graph with vertex set $\mathcal{V} = \{1, \ldots, n\}$ and we call a *graph vector* an n-dimensional real-valued vector whose covariates reside on the vertices of the graph. Intuitively, a *graph wavelet* is a graph vector such that is purposefully crafted to reflect the information regarding some local structure underlying the graph and, when combined with any graph vector, to extract its locally irregular behavior.[1] Before delving into the technical details of how to construct wavelets on general graphs, let us denote by $\Psi \in \mathbb{R}^{n\times d}$ whose columns $\psi_1, \ldots, \psi_d$ form a set of graph wavelets. We assume that $d \geq n$ and Ψ has full row rank, and we call the set of wavelets complete if $d = n$ or overcomplete if $d > n$. Any graph vector $\mathbf{f} \in \mathbb{R}^n$ can thus be represented by a linear combination of the building-block wavelets such that $\mathbf{f} = \Psi\mathbf{w}$ where $\mathbf{w} \in \mathbb{R}^d$ is a (possibly non-uniquely) representation of $\mathbf{f}$ that reflects details of its locally irregular behavior. Notably, wavelets should be determined exclusively by the graph $\mathcal{G}$ regardless of any graph vectors considered. Let us now write $\Omega \in \mathbb{R}^{n\times d}$ whose columns $\omega_1, \ldots, \omega_d$ are another set of graph vectors lying on $\mathcal{G}$, provided that $d \geq n$ and Ω has full row rank. Given the graph wavelets in $\Psi \in \mathbb{R}^{n\times d}$, we say $\Omega \in \mathbb{R}^{n\times d}$ record the *dual wavelets* of the graph if $\Omega^\top$ is the Moore-Penrose pseudoinverse of Ψ, i.e.

$$\Omega^\top = \Psi^+ = \Psi^\top(\Psi\Psi^\top)^{-1}\,.$$

Now, for any graph vector $\mathbf{f} \in \mathbb{R}^n$, we define the (unique) *wavelet representation* of $\mathbf{f}$ by applying the *wavelet transform*

$$\mathbf{w} = \Omega^\top\mathbf{f} \in \mathbb{R}^d\,,$$

and the *inverse wavelet transform*

$$\mathbf{f} = \Psi\mathbf{w} \in \mathbb{R}^n$$

reconstructs $\mathbf{f}$. It is worth noting that, in this definition the set of dual wavelets are defined according to a set of (primal) wavelets. In fact, equivalently we can first define a set of dual wavelets and then obtain the set of corresponding (primal) wavelets, due to the full rank assumption and the fact that $(\Omega^+)^+ = \Omega$ always holds.

[1] See [Hammond 2011] for an example of rigorous definition of graph wavelets in terms of frequency-spatial localization in small-scale limit.

Now we introduce two regularization methods based on graph wavelet smoothing. The first method is largely motivated by sparse basis pursuit methods [Chen 2001]. We assume that the idealized coefficients β in the predictive model should have a sparse wavelet representation θ implying that only a few wavelets are involved in building the prediction. Under the regularization framework of this study, we propose to solve (3.5) with a penalty that reads:

$$P^{\text{w-synthesis}}(\beta) = \min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \quad \text{s.t. } \beta = \Psi\theta\,. \tag{3.14}$$

Specifically, we aim to solve the following optimization problem:

$$\min_{\theta \in \mathbb{R}^d} \ell(y_1, \ldots, y_m, (\Psi\theta)^\top \mathbf{x}_1, \ldots, (\Psi\theta)^\top \mathbf{x}_m) + \lambda\|\theta\|_1\,, \tag{3.15}$$

in which $\beta = \Psi\theta$ reconstructs the coefficient vector in the underlying linear model from its wavelet representation that is sought to be sparse. We term (3.15) as a synthesis approach to wavelet smoothing or simply *wavelet-synthesis* method, and hence we call (3.14) the wavelet-synthesis penalty. By analogy to the lasso penalty (3.7) which is defined as the L_1 norm of the coefficient vector with respect to the Euclidean basis, wavelet-synthesis penalty (3.14) is the L_1 "norm" of the coefficient vector with respect to the wavelet "basis", indeed an (over)complete system of wavelets. By the definition of wavelets, the estimated coefficient vector β should be globally smooth and localized on the graph, such that after thresholding small values in β, the remaining coordinates of β should result in a few subnetworks identified. In particular, the location, size and shape of the potential subnetworks are inherently specified by the wavelets which in turn rely solely on the underlying graph.

The second wavelet-based method exploits the advantages of using wavelet representation for the feature vectors in data. So far, we have been focusing on directly seeking for a coefficient vector $\beta \in \mathbb{R}^n$ in the predictive model with structured sparsity, both in the overview of previous work and in the first wavelet-based method we introduced. In fact, the wavelet representation can also be applied to the feature vector $\mathbf{x} \in \mathbb{R}^n$ to obtain a relatively compact representation of data that "decorrelates" the feature vector concerning local behaviors with regard to the graph structure, a trick that has shown advantages in many applications [Kim 2014, Tremblay 2014]. To this end, we propose to first transform all feature vectors in data to their wavelet representation and build regularized linear models in the wavelet domain. Specifically, we aim to solve the following optimization problem:

$$\min_{\theta \in \mathbb{R}^d} \ell(y_1, \ldots, y_m, \theta^\top(\Omega^\top \mathbf{x}_1), \ldots, \theta^\top(\Omega^\top \mathbf{x}_m)) + \lambda\|\theta\|_1\,, \tag{3.16}$$

in which $\beta = \Omega\theta$ is the coefficient vector of the underlying linear model in terms of the original feature vectors. This problem can also be formulated as one in the regularization framework (3.5) with a penalty term that reads:

$$P^{\text{w-analysis}}(\beta) = \min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \quad \text{s.t. } \beta = \Omega\theta\,. \tag{3.17}$$

We term (3.16) as an analysis approach to wavelet smoothing or simply *wavelet-analysis* method, and hence we call (3.17) the wavelet-analysis penalty.

The synthesis approach (3.15) and the analysis approach (3.16) elaborated above are special cases of two popular alternatives when performing wavelet smoothing in the field of signal processing [Elad 2007]. If Ψ and Ω form a complete bi-orthogonal system of primal and dual wavelet basis of the graph, i.e., Ψ is invertible and $\Omega^\top = \Psi^{-1}$, it is easy to verify that both the wavelet-synthesis penalty (3.14) and the wavelet-analysis penalty (3.17) are special cases of the generalized lasso (3.13). In particular, the synthesis approach (3.15) and the analysis approach (3.16) are equivalent when $\Psi(=\Omega)$ form a complete orthogonal system of wavelet basis of the graph. However, the two approaches give very different results generally in practice, which we will empirically demonstrate in Section 3.3.

3.2.4 Implementation

Efficient algorithms for optimization problems of the form (3.5) depend on the particular choices of the loss function and the penalty term. In this study, we are interested in path algorithms that produce the entire solution path varying the regularization parameter λ, or path-wise algorithms that produce solutions over a grid of regularization parameters efficiently.[2] For example, linear regression (3.2) penalized with the elastic net (3.8), including the lasso (3.7), can be efficiently solved by the path algorithm such as Least Angle Regression (LARS) [Efron 2004]. Path-wise algorithms for a broad class of loss functions penalized by the elastic net have been extensively studied, among which many are implemented in the `R` CRAN package `glmnet` [Friedman 2010, Simon 2011]. Further, for generalized lasso penalties (3.13), including the graph-fused lasso (3.12), [Tibshirani 2011] proposed a path algorithm and the implementation is available via the `R` CRAN package `genlasso`. However, the implementation is subject to the squared loss function of linear regression (3.2) and relatively computationally intensive.

For network-based Laplacian regularization (3.9), we opted for a slightly different penalty by adding a small ridge term that reads

$$\widetilde{P}^{\text{lap}} = \beta^\top (L + \mu I)\beta = \|\theta\|_2^2 \quad \text{s.t. } \beta = X(\Lambda + \mu I)^{-\frac{1}{2}} X^\top \theta \,,$$

where $\mu = 10^{-3}$ is a small number added to the diagonal of the Laplacian matrix for numeric stability and better performance as suggested by [Zhang 2013], and the last equality is due to (3.10) and the fact that $(L + \mu I)$ is invertible. The optimization problem (3.5) with $\widetilde{P}^{\text{lap}}$ now becomes equivalent to

$$\min_{\theta \in \mathbb{R}^n} \ell(y_1, \ldots, y_m, (X(\Lambda + \mu I)^{-\frac{1}{2}} X^\top \theta)^\top \mathbf{x}_1, \ldots, (X(\Lambda + \mu I)^{-\frac{1}{2}} X^\top \theta)^\top \mathbf{x}_m) + \lambda \|\theta\|_2^2 \,,$$

where $\beta = X(\Lambda + \mu I)^{-\frac{1}{2}} X^\top \theta$ reconstructs the coefficient vector in the underlying linear model. Therefore, in case that an exact eigendecomposition of the Laplacian

[2] For penalty functions that involve an additional regularization parameter ν, ν is always determined by cross-validation on the training set and then used to generate the solution path varying only λ.

is affordable, an algorithm is straightforward where each feature vector $\mathbf{x}_i$ for $i = 1, \ldots, m$ is first transformed by left-multiplication of a "preconditioning" matrix $X(\Lambda + \mu I)^{-\frac{1}{2}} X^\top$ and then a linear model is fitted to the transformed data with the standard ridge. An implementation of such a two-step algorithm adapted for various loss functions is easy to build upon off-the-shelf `R` CRAN package `glmnet`. Further, a path-wise algorithm for the Laplacian lasso penalty (3.11) combined with linear regression and the Cox model is proposed and analyzed respectively by [Li 2008] and [Sun 2014] with implementation available from `R` CRAN packages `glmgraph` and `Coxnet`.

For the methods based on graph wavelet smoothing, after the graph wavelets Ψ and dual wavelets Ω on a given graph are obtained, both the synthesis approach (3.15) and analysis approach (3.16) to wavelet smoothing are essentially equivalent to a simple two-step procedure where each feature vector $\mathbf{x}_i$ for $i = 1, \ldots, m$ is first transformed by left-multiplication of a "preconditioning" matrix $\Psi^\top$ or $\Omega^\top$ respectively and then a linear model is fitted to the transformed data with the standard lasso. Therefore a path algorithm implementing both approaches is straightforward by modifying that for the standard lasso, and an implementation is easy to build upon `R` CRAN package `glmnet` for instance.

For the sake of self-containment of the chapter, we will briefly review two techniques developed for constructing wavelets on general graphs, namely the graph wavelet transform based on spectral graph theory by [Hammond 2011] or a lifting procedure by [Jansen 2009]. In fact, a wealth of studies in signal processing have been devoted to designing wavelets for data aligned on a uniform lattice such as time series data (1-dimensional line) or images (2-dimensional grid) [Mallat 1999]. However, it is a non-trivial task to construct graph wavelets that capture locally irregular structure on general graphs, a topic that has received much attention and been explored in many studies [Hammond 2011, Section 1.1. Related work]. As demonstrated partly in Section 3.3, different approaches of constructing graph wavelets usually result in distinct characteristics and behaviors in practice.

Spectral graph wavelets. Here we first define the dual wavelets and then obtain the corresponding (primal) wavelets. Spectral graph dual wavelets[3] were proposed by [Hammond 2011] based on defining translation in the graph vertex domain and scaling on the Fourier modes in the graph spectral domain. Intuitively, they are formed by applying a scaled spectral band-pass filter to indicator functions at every vertex of the graph such that: at small scales, the filter lets through high-frequency modes to good localization and the corresponding wavelets only reach to their close neighborhood on the graph; at large scales, the filter compresses around low-frequency modes and the corresponding wavelets encode coarser description of the local structure. Recall from Section 3.2.2 that the graph Laplacian is decomposed as $L = X \Lambda X^\top$ for an orthogonal matrix X with Fourier basis in

[3]The authors of [Hammond 2011] simply call them wavelets whereas we specifically call them dual wavelets following our definition above.

columns and a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ with respective frequencies $0 = \lambda_1 \leq \cdots \leq \lambda_n$. Given the wavelet generating function $g : \mathbb{R} \to \mathbb{R}$ and a scale $s > 0$, the stretched spectral band-pass filter at scale s has a matrix representation that is diagonal on the Fourier modes, i.e.,

$$\Lambda^{s,g} = \mathrm{diag}(g(s\lambda_1), \ldots, g(s\lambda_n)) .$$

Let us denote by $\Omega^{s,g} \in \mathbb{R}^{n \times n}$ the spectral graph dual wavelet basis at scale s defined by

$$\Omega^{s,g} = (\omega_1^{s,g} | \ldots | \omega_n^{s,g}) = X \Lambda^{s,g} X^\top ,$$

where the u-th column $\omega_u^{s,g}$ is the dual wavelet centered around vertex $u \in \mathcal{V}$ providing a local view of the graph structure, and when convolved with any graph vector in the wavelet transform, extracts its local behaviors. In order to ensure stability for reconstruction purpose, it is convenient to introduce a second class of waveforms that arise from a scaling function $h : \mathbb{R} \to \mathbb{R}$, analogous to low-pass residual scaling functions in classical wavelet analysis. In practice, it is advised to form an overcomplete system of dual wavelets by combining the dual wavelet basis corresponding to h at a single scale $s = 1$ and g at multiple scales. Suppose $S = \{s_1, \ldots, s_J\}$ are J scales that are adapted to the eigenspectrum of graph Laplacian L, an overcomplete system of *spectral graph dual wavelets* $\Omega^{\mathrm{spec}} \in \mathbb{R}^{n \times (J+1)n}$ are given by concatenating column-wisely all the underlying dual wavelets, i.e.,

$$\Omega^{\mathrm{spec}} = (\Omega^{1,h} | \Omega^{s_1,g} | \ldots | \Omega^{s_J,g}) .$$

An efficient algorithm of bypassing the eigendecomposition of the Laplacian and obtaining an approximation of the wavelets by using Chebychev polynomials to approximate the filters is proposed by [Hammond 2011]. Finally, the corresponding *spectral graph (primal) wavelets* $\Psi^{\mathrm{spec}} \in \mathbb{R}^{n \times (J+1)n}$ are defined by

$$(\Psi^{\mathrm{spec}})^\top = (\Omega^{\mathrm{spec}})^+ ,$$

and the wavelet transform as well as the inverse transform follows from the definition. Note that in all experiments in Section 3.3, we simply compute the spectral graph wavelets by performing an exact eigendecomposition of the graph Laplacian L and take the largest eigenvalue λ_n to determine the following parameters suggested by [Hammond 2011]: $S = \{s_1, \ldots, s_J\}$ where the maximum scale $s_1 = 200/\lambda_n$, the minimum scale $s_J = 1/\lambda_n$, the other scales in S are logarithmically equispaced between them for $J = 4$, and the wavelet generating function

$$g(x) = \begin{cases} x^2 & \text{if } x < 1, \\ -5 + 11x - 6x^2 + x^3 & \text{if } 1 \leq x \leq 2, \\ (2/x)^2 & \text{if } x > 2, \end{cases}$$

and the scaling function

$$h(x) = \left(1 + \frac{2\sqrt{3}}{9}\right) \exp\left(-\left(\frac{x}{0.006\lambda_n}\right)^4\right) .$$

For interested readers, we refer to [Hammond 2011] on details about how these parameters and wavelet generating function are constructed.

Lifting-based graph wavelets. A second approach to wavelet construction is based on the lifting scheme [Sweldens 1998], which allows to obtain a complete bi-orthogonal system of wavelets and dual wavelets. The distinguishing merit of lifting-based design of wavelets is that it provides a more intuitive interpretation of the wavelet transform as well as the inverse transform, and the implementation has linear complexity both in time and in storage. Intuitively, the lifting scheme factorizes the (discrete) wavelet transform of any graph vector into a sequence of so-called "lifting" steps: at each step, the current "scales" which are indexed by presently remaining vertices are divided into two sets, of which one is processed to give the "wavelet residuals" and thus lifted out and then the other is updated to give coarser "scales" for the next step. This way, the wavelet residuals found by the end of each step reflect details of locally irregular behavior of the underlying graph vector. An inverse transform is straightforward by essentially inverting all the lifting steps. When the lifting-based wavelet transform and inverse transform are applied to an indicator function at some vertex, we obtain the primal and dual wavelets centered around that vertex of the underlying graph. For a lifting-based design of wavelets on general graphs, we adopt the "lifting one at a time" method proposed by [Jansen 2009], which is summarized in Algorithm 3.1. Notably, the algorithm only involves arithmetic computations that consumes linear time and can be implemented fully in space. Following the algorithm, we can obtain the *lifting-based graph wavelets* $\Psi^{\text{lift}} \in \mathbb{R}^{n\times n}$ and correspondingly the *lifting-based graph dual wavelets* $\Omega^{\text{lift}} \in \mathbb{R}^{n\times n}$ for a given graph $\mathcal{G}$ that satisfy

$$(\Omega^{\text{lift}})^\top = (\Psi^{\text{lift}})^{-1}.$$

In other words, Ψ^{lift} and Ω^{lift} form a complete bi-orthogonal system of primal and dual wavelets of the graph. The wavelet transform and the inverse transform for any graph vector follow from the definition.

3.3 Results

3.3.1 Experiment Set-ups: Data, Network and Methods

We demonstrate the above-mentioned regularization methods by analyzing the gene expression data derived from breast tumors collected from participants of the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) trial [Curtis 2012]. The dataset contains expression data corresponding to the mRNA measurements of $24,771$ genes for $1,981$ breast cancer patients.

The biological network we consult in this study to guide the gene selection as well as subnetwork detection is a protein-protein interaction (PPI) network from HPRD. After keeping the maximally connected component of the network composed by the genes available from the METABRIC dataset, we obtained a network consisting of $9,117$ genes as vertices and $36,326$ pairwise interactions as undirected edges where the (unweighted) edge is assigned a weight 1 if there exists a known interaction between the connected genes and 0 otherwise. In the resulting network,

Algorithm 3.1 Lifting-based wavelets and dual wavelets on general graphs [Jansen 2009].

Input: An undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, A)$ where $\mathcal{V}$ is indexed by $\{1, \dots, n\}$ and A is the weighted adjacency matrix encoding the non-negative weights assigned to each pair of vertices. Assume that $\mathcal{G}$ has no self-loops or A has 0 on diagonal.

Initialize: Let $\mathcal{W}$ be the index set of wavelets already found, initialized to be the empty set, and let $\mathcal{S}$ be the index set of wavelets yet to be found, initialized to be $\mathcal{V}$. For $i = 1, \dots, n$, let $\iota_i \in \mathbb{R}$ be the "integral of scales" [Jansen 2009] associated to vertex i, initialized to be the vertex degree $\sum_{j=1}^{n} A_{ji}$, and let $\psi_i \in \mathbb{R}^n, \omega_i \in \mathbb{R}^n$ respectively be the graph wavelet and the corresponding dual wavelet centered around vertex i, both initialized to be indicator function at vertex i.

For $r = n, \dots, 1$, repeat:

1: Pick the next vertex to be lifted indexed by i_r such that it has the smallest current integral of scales in the remaining set, i.e., $i_r = \arg\min_{i \in \mathcal{S}} \iota_i$.

2: The "predict" and "update" equations for the primal and dual wavelets at step r are respectively

$$\begin{cases} \psi_j \leftarrow \psi_j + a_j^r \psi_{i_r} \quad \text{for } j \sim i_r\,, \\ \psi_{i_r} \leftarrow \psi_{i_r} - \sum_{j \sim i_r} b_j^r \psi_j\,, \end{cases} \quad \text{and} \quad \begin{cases} \omega_{i_r} \leftarrow \omega_{i_r} - \sum_{j \sim i_r} a_j^r \omega_j\,, \\ \omega_j \leftarrow \omega_j + b_j^r \omega_{i_r} \quad \text{for } j \sim i_r\,, \end{cases}$$

where $j \sim i_r$ denotes that j and i_r are currently direct neighbors connected by an edge, i.e., $A_{ji_r} > 0$, the "predict" weights a_j^r are user-defined such that the weighted average of direct neighbors detects locally irregular behavior for any graph vector, i.e.,

$$a_j^r = A_{ji_r} / \sum_{k \sim i_r} A_{ki_r} \quad \text{for } j \sim i_r\,,$$

and the "update" weights b_j^r must satisfy the requirement of vanishing moments of wavelet filters, i.e.,

$$b_j^r = \iota_{i_r} \iota_j / \sum_{k \sim i_r} \iota_k^2 \quad \text{for } j \sim i_r\,,$$

in which the integral of scales for the direct neighbors of i_r have been refined to be

$$\iota_j \leftarrow \iota_j + a_j^r \iota_{i_r} \quad \text{for } j \sim i_r\,.$$

and then

3: Lift the vertex indexed by i_r from the graph and reweight its direct neighborhood, i.e., modify the adjacency matrix A by assigning

$$\begin{aligned} A_{ji_r} = A_{i_r j} &\leftarrow 0 \quad \text{for } j \sim i_r\,, \\ A_{jk} = A_{kj} &\leftarrow \max\{A_{jk}, A_{ji_r} A_{ki_r}\} \quad \text{for } j \sim i_r \text{ and } k \sim i_r\,. \end{aligned}$$

4: Set $\mathcal{S} \leftarrow \mathcal{S} \setminus \{i_r\}$ and $\mathcal{D} \leftarrow \mathcal{D} \cup \{i_r\}$ meaning the wavelet indexed by i_r has been found.

Output: $\Psi^{\text{lift}} := (\psi_1 | \dots | \psi_n) \in \mathbb{R}^{n \times n}$ are the lifting-based graph wavelets and $\Omega^{\text{lift}} := (\omega_1 | \dots | \omega_n) \in \mathbb{R}^{n \times n}$ are the corresponding dual wavelets.

the distribution of the vertex degree ranges from 1 to 267 with the median at 4 and the two quantiles 25% and 75% at 2 and 8 respectively. Note that for network-based regularization methods, only the subset of genes found on the network can be used. Therefore for the sake of fair comparison of different methods, our numerical analysis is constrained to the genes underlying the given network.

Table 3.1: Summary of different regularization methods in our numerical experiments.

Label	Penalty function $J(\beta)$	Network-based	Feature selection
ridge	$\|\beta\|_2^2$		
lasso	$\|\beta\|_1$		✓
e-net	$\nu\|\beta\|_1 + (1-\nu)\|\beta\|_2^2$		✓
lap	$\sum_{i\sim j}(\beta_i - \beta_j)^2$	✓	
laplasso	$\nu\|\beta\|_1 + (1-\nu)\sum_{i\sim j}(\beta_i - \beta_j)^2$	✓	✓
gflasso	$\sum_{i\sim j}\|\beta_i - \beta_j\|$	✓	✓
w-synthesis	$\min_\theta \|\theta\|_1 \quad \text{s.t. } \beta = \Psi\theta$	✓	✓
w-analysis	$\min_\theta \|\theta\|_1 \quad \text{s.t. } \beta = \Omega\theta$	✓	✓

Our numerical experiments aim to provide a benchmark study that compares several above-mentioned regularization methods, with a particular focus on those based on wavelet smoothing. Specifically, we study the prediction performance and feature selection of different methods under the regularized predictive framework (3.5), where the loss function takes (3.2) for linear regression in simulation studies (Section 3.3.2) or (3.4) for breast cancer survival analysis (Section 3.3.3). Details on different regularization methods are found in Section 3.2 and summarized in Table 3.1. A few variants of the listed methods are also considered and will be denoted by a suffix appended to the label of the corresponding method. For all methods, genes underlying the given network are by default used, unless a suffix "org" is added to the label of network-free methods indicating that the entire set of genes in the METABRIC dataset are used instead. For methods involving graph Laplacian, the non-normalized graph Laplacian is used by default, unless a suffix "norm" is added to the label indicating that the normalized Laplacian is used instead. For wavelet-based methods, a suffix "spec" indicates spectral graph wavelets are used and a suffix "lift" indicates the lifting-based graph wavelets are used.

All numerical experiments in this chapter are performed in `R`.

3.3.2 Simulation Studies

Our simulation set-ups follow a simple linear regression framework where simulated responses are generated using the real biological network and real gene expression data. To start with, the network we used in the simulation studies is a subnetwork of the HPRD PPI network that has $n = 1,744$ genes and $15,911$ edges with a distribution of the vertex degrees ranging from 4 to 184 with the median at 12 and the two quantiles 25% and 75% at 9 and 20 respectively. The subnetwork was deduced from the complete HPRD PPI network by iteratively removing genes with the smallest vertex degree among those currently remaining on the network,

and the rationale for trimming the network in simulation studies is to drastically reduce the computation time by reducing the number of features for training some computationally intensive methods such as graph-fused lasso. Data generation then proceed in three steps as follows:

1. The coefficient vector $\beta \in \mathbb{R}^n$ is generated by $\beta = \Psi^{\text{spec}}\theta$, where the spectral graph wavelets Ψ^{spec} are obtained on the trimmed network while the wavelet representation of synthesized coefficients θ is designed to be sparse. Note that the number of non-zero coordinates in θ is a design parameter that takes 1, 10 or 100 intending for different levels of structured sparsity. The positions of these non-zero coordinates in θ are randomly sampled one-by-one following a categorical distribution where the probability of a coordinate being non-zero is proportional to the vertex degree of the corresponding gene on the network transformed by a logistic function, which aims to synthesize the fact that genes with more known interactions tend to have higher biological importance. The values of these non-zero coordinates in θ are designed to be either constant $+1$ or random $+1/-1$ following a Rademacher distribution, which takes on the assumption that the contribution from different gene modules might occur in the opposite direction. Finally β is reconstructed from the synthesized θ and normalized to have unit variance. Note that consequently the coefficients β are supposed to be globally smooth and localized on the network.

2. To resemble the "large n, small m" situation in most biomedical applications, we randomly sampled $m = 500$ patients from the METABRIC data, and the expression profiles of these samples are constrained on the $n = 1,774$ genes from the trimmed network. These feature data are further standardized to have zero mean and unit variance, denoted by $\{\mathbf{x}_i\}_{i=1}^m$ where $\mathbf{x}_i \in \mathbb{R}^n$.

3. For $i = 1, \ldots, m$, a responses $y_i \in \mathbb{R}$ is simulated by $y_i = \beta^\top \mathbf{x}_i + \varepsilon_i$, where ε_i is an additive i.i.d. random noise following a normal distribution of zero mean and variance σ^2 such that the signal-to-noise ratio for the cohort of samples is estimated to be 5.

For each combination of simulation set-up given a fixed number of non-zero coordinates in θ (1, 10 or 100) and a type of values in θ (constant $+1$ or Rademacher $+1/-1$), these data generation steps are repeated 20 times to address the randomness therein, and hence we obtained a total of 20 simulated datasets for each combination of simulation set-up.

On each simulated dataset, we performed 5-fold cross-validation and evaluated the prediction mean squared error (PMSE) on each test fold, while the regularization parameters (λ and ν) were determined by nested cross-validation on each training fold. Results are shown in Figure 3.1 where, under each combination of simulation set-up, boxplots present the PMSE over the $20 \times 5 = 100$ training and test splits of simulated data for a total of 10 regularization methods including related variants. For all simulation set-ups, the wavelet-synthesis method with spectral graph

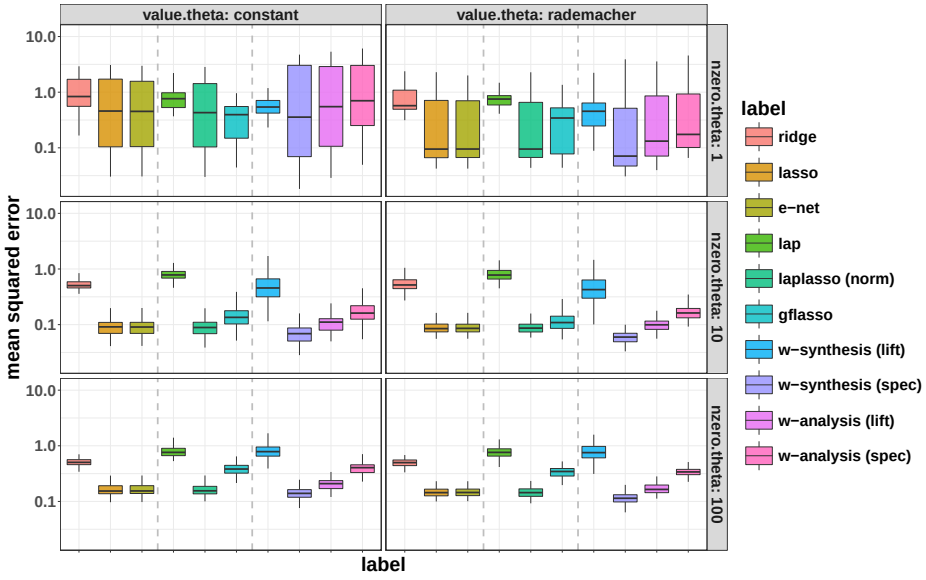
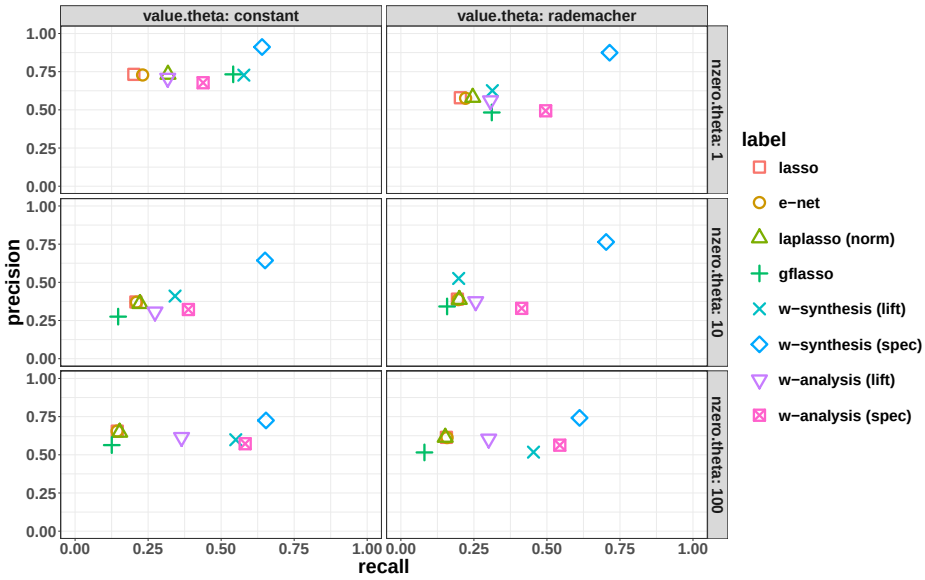


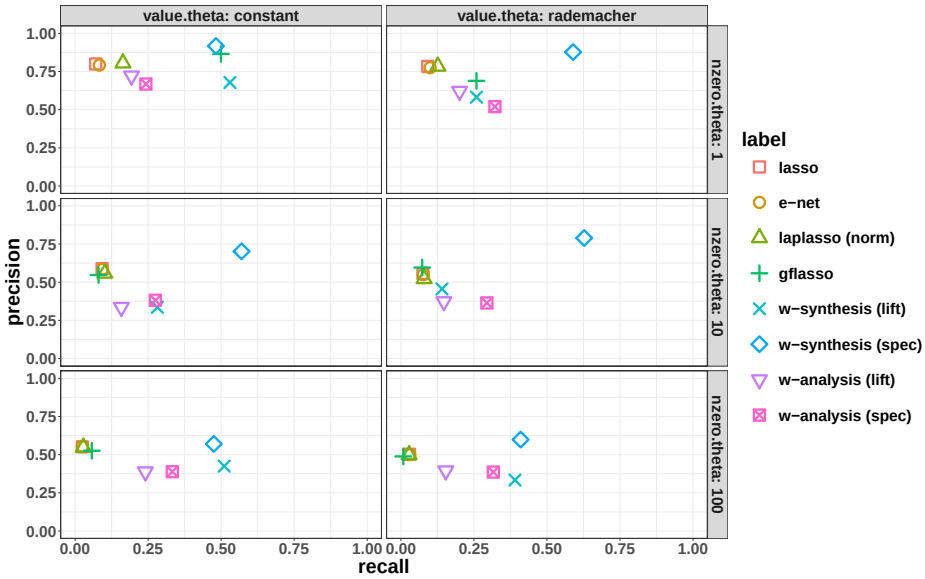
Figure 3.1: Boxplots on regression performance evaluated by prediction mean squared error over the 100 training and test splits of the simulated data.

wavelets denoted by "w-synthesis (spec)" is the consistent winner as expected, due to the fact that the simulated data are generated so that the informative features form locally connected subnetworks whose shapes and sizes are coherent with the spectral graph wavelets by design. The superiority is most striking when θ takes a reasonably intermediate number of non-zero coordinates, denoted by "nzero.theta: 10", and values that oscillate in sign, denoted by "value.theta: rademacher". Further, we observed that the same method with lifting-based graph wavelets instead of spectral graph wavelets, denoted by "w-synthesis (lift)", surprisingly exhibits the worst performance overall among all wavelet-based methods, which suggests that different approaches for constructing graph wavelets result in distinct characteristics and behaviors of underlying wavelets. However, the wavelet-analysis method is better suited with lifting-based graph wavelets than with spectral graph wavelets under our simulation set-ups, for which reasons remain unclear. Although the lasso and the elastic net are network-free methods, they still give competitive performance compared to the network-based Laplacian lasso or even outperform the network-based graph-fused lasso in general in terms of PMSE. The two methods that do not allow for feature selection, namely the ridge and the Laplacian regularization, yield the worst prediction performance, due to the fact that the simulated data are generated by a sparse model involving only a few informative features.

We further studied the sparsity-inducing methods which allow for feature selection regarding their ability to recover the informative features as well as the connecting edges over the network. Specifically, under a specific combination of simulation set-up, we applied each regularization method to each simulated dataset and obtained an estimate of β. We then compared the support as well as the con-



(a) Support recovery of the coefficient vector β.



(b) Support recovery of the connecting edges over the network.

Figure 3.2: Precision-recall plots on the recovery of simulated support of the coefficient vector β and the connecting edges over the network.

necting edges over the network underlying the estimated and true β in terms of precision (the fraction of selected features that are truly informative) and recall (the fraction of truly informative features that are selected). In fact, as most values in β are very small but not exactly zero in some cases, the "non-zero" support of β is defined as the coordinates whose values are greater than one-hundredth of the largest value in β. For ease of visualization, we fit an ellipse over all the precision-recall points over the total of 20 simulated datasets for each method, and the ellipse centers are shown in Figure 3.2 for a total of 8 methods including related variants. Again, the method denoted by "w-synthesis (spec)" is predominant as expected in terms of support recovery, most remarkably when θ takes a reasonably intermediate number of non-zero coordinates, denoted by "nzero.theta: 10", resulting in a modest number of connected subnetworks formed by the support of β. Compared to other methods, the wavelet-based methods generally tend to achieve competitive precision but much higher recall for the recovery of β and the connecting edges over the network. Surprisingly, when the number of informative features becomes large and hence more connected over the network, the network-based Laplacian lasso and graph-fused lasso does not distinguishingly outperform the network-free lasso and elastic net in terms of feature selection or even edge identification.

3.3.3 Breast Cancer Survival Analysis

For *in vivo* experiments, we performed survival analysis for breast cancer using the METABRIC data guided by the HPRD PPI network. In fact, for each breast tumor sample, the METABRIC dataset additionally provides clinical information on the overall survival time of the underlying patient, that is either the exact number of days dating from initial consultation until the patient passed away or the number of days until the patient is last seen alive. This clinical information was converted to right-censored survival data and we performed survival risk prediction for breast cancer.

The first objective of the *in vivo* investigation is to compare the survival risk prediction performance of different regularization methods. To this end, we performed 5-fold cross-validation repeated 10 times on the full METABRIC dataset and evaluated the concordance index (CI) scores on each test fold, while the regularization parameters (λ and ν) were determined by nested cross-validation on each training fold. Note that CI is a measure of rank-based consistency between the predicted survival risk and observed survival time for a cohort of patients, in the sense that it is an estimate of the probability that, given two randomly drawn patients, the patient who survives longer is predicted with a lower risk. Results are shown in Figure 3.3 with boxplots over the $10 \times 5 = 50$ splits of training and test data and in Table 3.2 with mean CI scores ($\pm$ standard deviation) for a total of 15 regularization methods including related variants. The ridge is the best-performing model in terms of mean CI scores but it does not allow for feature selection, and among methods that enable feature selection, the elastic net is the best-performing one. Notably, both methods do not make use of the prior knowledge encoded in the network. Among

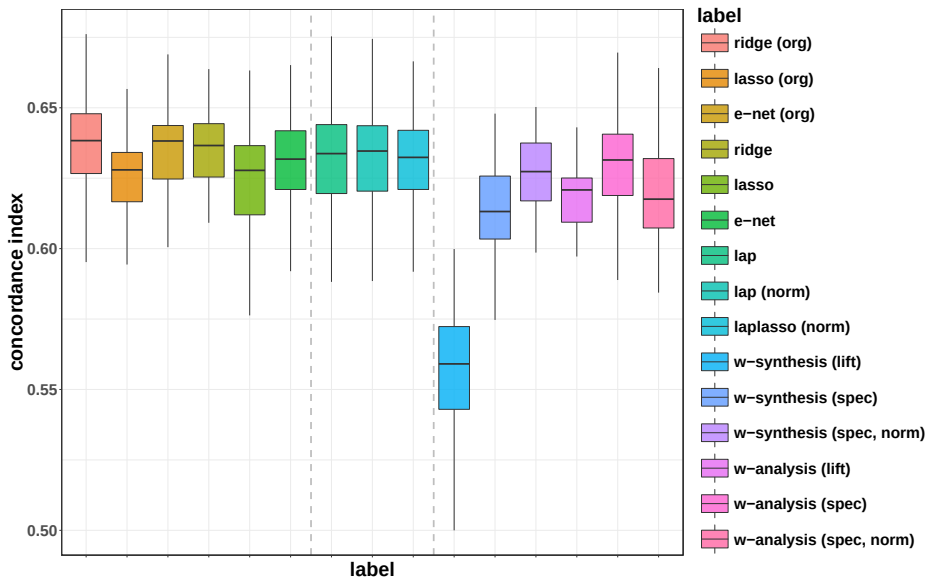


Figure 3.3: Boxplots on survival risk prediction performance evaluated by concordance index scores over 5-fold cross-validation repeated 10 times of the METABRIC data.

Table 3.2: Mean concordance index (CI) scores (± standard deviation) of survival risk prediction over 5-fold cross-validation repeated 10 times of the METABRIC data. Methods are ordered by decreasing mean CI scores.

Label	Mean CI scores (± SD)	Network-based	Feature selection
ridge (org)	0.6370 (±0.0178)		
ridge	0.6360 (±0.0180)		
e-net (org)	0.6345 (±0.0185)		✓
lap (norm)	0.6330 (±0.0196)	✓	
lap	0.6320 (±0.0193)	✓	
laplasso (norm)	0.6312 (±0.0185)	✓	✓
e-net	0.6304 (±0.0183)		✓
w-analysis (spec)	0.6295 (±0.0198)	✓	✓
w-synthesis (spec, norm)	0.6264 (±0.0180)	✓	✓
lasso	0.6260 (±0.0177)		✓
lasso (org)	0.6257 (±0.0172)		✓
w-analysis (spec, norm)	0.6216 (±0.0228)	✓	✓
w-analysis (lift)	0.6163 (±0.0182)	✓	✓
w-synthesis (spec)	0.6157 (±0.0184)	✓	✓
w-synthesis (lift)	0.5587 (±0.0232)	✓	✓

network-based methods that enable feature selection, the best-performing one is the Laplacian lasso, followed by the wavelet-analysis method with spectral graph dual wavelets. We observed that the wavelet-based methods provide relatively less accurate survival risk prediction in terms of CI scores compared to other methods. We performed two-sided t-tests to statistically quantify the difference of the cross-validation CI scores between each pair of methods. FDR-adjusted p-values suggest that, at the significance level 0.05, there is no significant decrease in the prediction performance for the best-performing wavelet-based method, namely the method denoted by "w-analysis (spec)", compared to any of the methods tested. Notably, network-free methods using the entire list of genes available from the METABRIC data does not significantly improve the survival risk prediction performance, compared to those using the subset of genes only found on the HPRD PPI network. This justifies that the loss is not significant when our analysis is restrained to the subset of genes found on the network. Another interesting observation regarding wavelet-based methods is that, as the spectral graph wavelets are constructed using graph Laplacian that appear in either normalized or non-normalized version, the wavelet-analysis method is better suited with the non-normalized version while the wavelet-synthesis method is better suited with the normalized version, for which reasons remain unclear.

The focus of *in vivo* experiments in this study is to select genes that are related to breast cancer survival, and in particular we favor methods that result in a list of selected genes that tend to form gene modules on the HPRD PPI network potentially by making use of their biological interaction known *a priori*. To this end, we then compared the goodness of gene selection of different methods from several aspects, namely stability, connectivity and interpretability. Note that only those 9 methods including related variants that enable feature selection are discussed for the remainder of this section.

Stability is an important concept that advocates reproducibility of selecting the significant features across independent studies. In order to compare stability of feature selection by different methods, we randomly split the full METABRIC dataset evenly into two halves and obtained a pair of models independently trained on the two disjoint subsets with the same method. Then constraining on each fixed number of genes selected, we computed the number of commonly selected genes between the two independent models as a score indicating stability. We repeatedly split the data 100 times to address the randomness in splitting the data. For ease of visualization, we applied locally weighted scatterplot smoothing (LOESS) [Cleveland 1979] to all the stability scores for each method and thus obtained a curve of stability scores along the solution path. Results are shown in Figure 3.4. We observed that the number of commonly selected genes between pairs of independent models tends to become larger as the number of selected genes increases. As we are most interested in selecting typically a few hundreds of genes that could interestingly form subnetworks of modest sizes, the wavelet-based method denoted by "w-analysis (spec, norm)" distinguishingly won in terms of stability at the scale of 10^2 genes selected, followed by two other wavelet-based methods denoted by

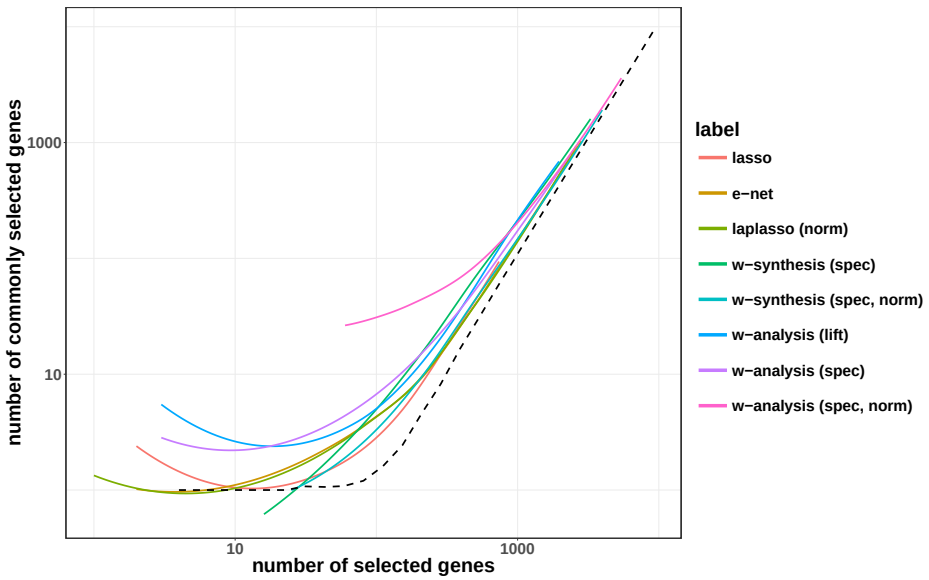


Figure 3.4: Stability performance of gene selection related to breast cancer survival, estimated over 100 random experiments. The black dotted curve denotes random selection.

"w-analysis (spec)" and "w-synthesis (spec)". Recall that the method denoted by "w-analysis (spec)" is the best-performing method in terms of CI scores for survival risk prediction. Last but not least, the network-free methods, namely the lasso and elastic net, and network-based Laplacian lasso provide feature selection procedures that are overall less stable. Further, as negative-control experiment, we randomly selected twice the same number of genes along the solution path and counted the number of commonly selected genes. At each fixed number of genes selected, the number of commonly selected genes is averaged over 100 repeats to address the randomness. A stability curve for random selection is shown by the black dotted curve in the figure. Note that all methods tested in this study outperform the random selection in terms of stability, especially strikingly at the scale of 10^2 genes selected.

Recall from the introduction that the motivation of this study is to encourage selected features to be connected given a network, and for that purpose, we would like to quantitatively compare feature selection in terms of connectivity over the given network. A model is trained by applying each method to the full METABRIC dataset and, constraining on each fixed number of genes selected, a connectivity score is defined as the number of connecting edges between the selected genes. Thus for each method we obtained a curve of connectivity scores along the solution path. Results are shown in Figure 3.5, where special marks indicate the number of genes and connecting edges that was determined by tuning the regularization parameter λ by cross-validation. We observed that the wavelet-based methods remarkably outperform other methods in terms of connectivity in general. In particular, the method denoted by "w-analysis (spec, norm)" that has stood out in terms of stability remains to be one of the top-performing methods in terms of connectivity,

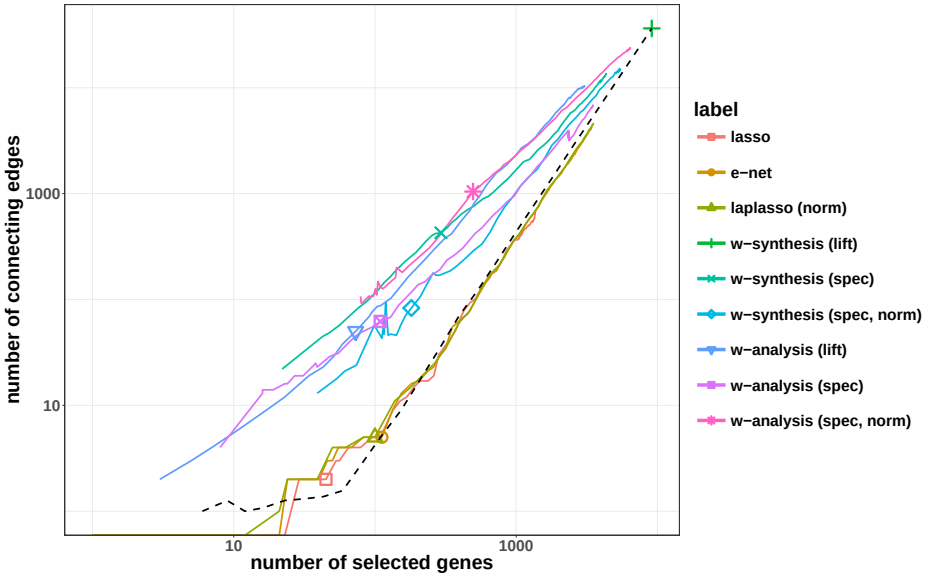
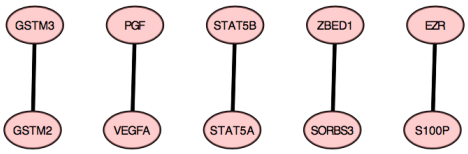


Figure 3.5: Connectivity performance of gene selection related to breast cancer survival, where special marks correspond to the number tuned by cross-validation. The black dotted curve denotes random selection.

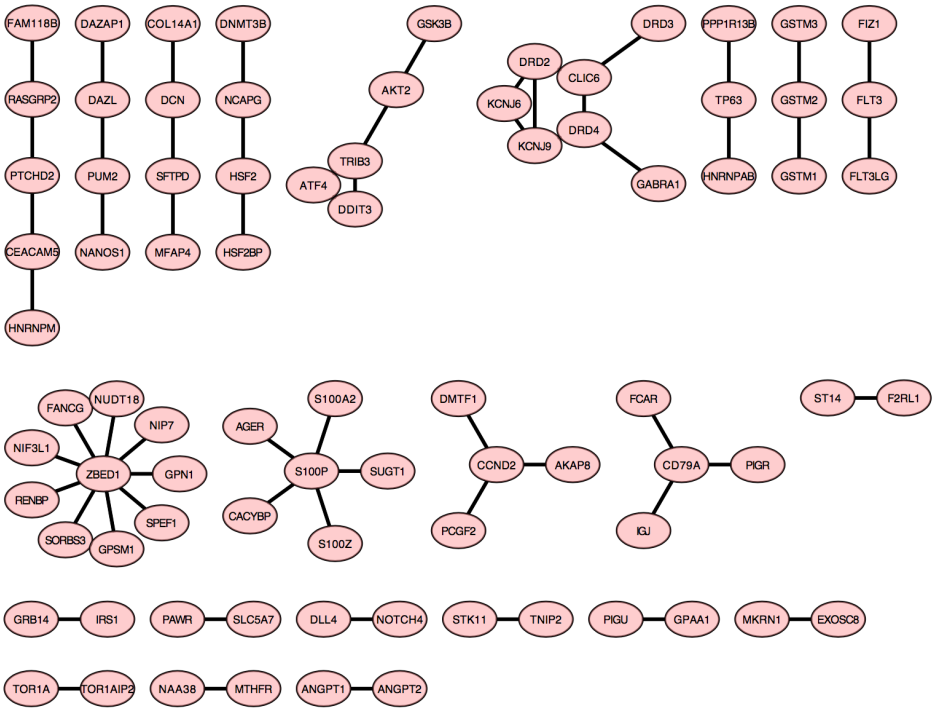
whereas it selected a relatively large number of genes by cross-validation. It is worth special mention that the method denoted by "w-analysis (spec)" selected around 10^2 genes by cross-validation that attain a reasonably good number of connecting edges, and it has been one of the top-performing methods in terms of stability in feature selection when this number of genes is selected, and it also remains competitive to all other methods in terms of CI scores for survival risk prediction. A side note on the strange performance of the method denoted by "w-synthesis (lift)" is that this method output an estimate of β that is constant[4] when fitted to the full METABRIC dataset particularly, and consequently the full list of genes were considered as selected. Further, as negative-control experiment, we randomly selected a certain number of genes regardless of a network, and counted the number of connecting edges over the given network. At each fixed number of genes selected, the number of connecting edges is averaged over 100 repeats to address the randomness. A connectivity curve for random selection is shown by the black dotted curve in the figure. Note that the network-free methods, namely the lasso and the elastic net, and network-based Laplacian lasso do not seem to significantly outperform the random selection in terms of connectivity.

It is interesting to investigate from a biological point of view the genes selected by our survival analysis for breast cancer using the METABRIC gene expression data and the HPRD PPI network, where the regularization parameter λ and thus the number of genes selected is determined by cross-validation. In particular, we

[4]Due to numerical instability in computation, there may exist infinitesimal variations in the values of estimated β output by the method.



(a) Gene subnetworks identified by the elastic net (10 genes connected out of 112 selected) or the Laplacian lasso (10 genes connected out of 100 selected).



(b) Gene subnetworks identified by the wavelet-analysis method with spectral graph dual wavelets (82 genes connected out of 109 selected).

Figure 3.6: Gene subnetworks related to breast cancer survival identified by regularization methods using the METABRIC data and HPRD PPI network.

focus on comparing the three methods denoted by "e-net", "laplasso (norm)" and "w-analysis (spec)" for the reminder of the section, each representing a class of methods. The network-free elastic net identified a total of 112 genes, among which only 10 genes are not isolated from one another on the HPRD PPI network that form 5 connected gene pairs (Figure 3.6a). The network-based Laplacian lasso identified a total of 100 genes, among which 10 genes are connected and they coincide exactly with the the connected genes identified by the elastic net (Figure 3.6a). Our network-based wavelet-analysis method with spectral graph dual wavelets identified a total of 109 genes, among which 82 genes are connected that form 23 gene modules of various sizes and shapes (Figure 3.6b). Here we focus on the genes selected by each method which are connected and thus form collaboratively functional gene modules.

The wavelet-based method is able to select more genes that form larger subnetworks than the other two methods. Specifically, there are two pairs of connected genes that are commonly selected by the three methods. First, while the other two methods detected the relation to cancer survival of GSTM2 and GSTM3 genes, the wavelet-based method was able to detect the involvement of three genes from the same gene family that are GSTM1, GSTM2 and GSTM3. In fact, glutathione metabolism is able to play both protective and pathogenic roles with respect to cancer [Balendiran 2004, Wu 2004], and the human GSTM gene family encodes the mu class of metabolic isoenzymes of glutathione S-transferase, consisting of five different but closely related isotypes GSTM1 to GSTM5. It has been reported that certain GSTM genes are correlated with the likelihood of breast cancer recurrence and functionally contribute prognostic information [Kiefer 2014]. In particular, GSTM1, which was detected only by the wavelet-based method, has been extensively studied in breast cancer risk especially due to its null genotype [Roodi 2004, Sull 2004], but the absence of a functional GSTM1 enzyme in a null variant can be meaningfully compensated for by GSTM2 [Bhattacharjee 2013]. Second, while the other two methods selected ZBED1 and SORBS3 genes simultaneously, the wavelet-based method detected through interactions documented in HPRD a star-shaped subnetwork in which the hub gene ZBED1 is centered around by 9 other genes SORBS3, FANCG, GPSM1, NUDT18, SPEF1, NIP7, GPN1, RENBP, NIF3L1. This is in fact the largest subnetwork identified by the wavelet-based method, in which some genes are of particular interest. In fact, the ZBED1 gene encodes a protein which binds to DNA elements found in the promoter regions of a number of genes related to cell proliferation [Matsukage 2008, Yamashita 2007]. The FANCG gene,, which was detected only by the wavelet-based method, provides instructions for making a protein complex involved in the Fanconi anemia (FA) pathway responsible for DNA repair and it has been reported to directly interact with BRCA2 gene that plays an important role in homologous recombination repair and survival risk in breast cancer [Hussain 2003, Wilson 2008].

Among the subnetworks identified exclusively by our wavelet-based method, some are of particular interest and the involvement of the connected genes in cancer biology was previously reported in literature. For example, an interest-

ing subnetwork is composed of 7 genes DRD2, DRD3, DRD4, CLIC6, KCNJ6, KCNJ9, GABRA1. In fact, the D2-like family of the dopamine receptors, encoded by genes DRD2, DRD3 and DRD4, are coupled to certain guanine nucleotide-binding proteins (G proteins) which directly inhibits adenylate cyclase activity and cyclic adenosine monophosphate (cAMP) formation [Neves 2002], and whose signaling has been linked to cancer progression and cancer risk [Murphy 2009, Mao 2015] leading to many preclinical studies on the antitumor effects sought by antagonizing DRD2 signaling [Pornour 2015, Hoeppner 2015]. Notably, all three genes DRD2, DRD3 and DRD4 are simultaneously selected due to their common interactor gene CLIC6. The second interesting subnetwork of interest is composed of 5 genes AKT2, ATF4, DDIT3, GSK3B, TRIB3. In this subnetwork, three genes AKT2, ATF4 and DDIT3 are present in the mitogen-activated protein kinase (MAPK) signaling pathway whose relevance to cancer has been profoundly studied and we refer to [Dhillon 2007] for an overview; three genes AKT2, ATF4 and GSK3B are found in the PI3K/Akt signaling pathway whose role in breast cancer has been reported in [Fresno Vara 2004, Paplomata 2014] for instance, and two genes AKT2 and GSK3B are included in the KEGG pathways in cancer. The third interesting subnetwork is a star-shaped gene module composed of 4 genes CCND2, DMTF1, AKAP8, PCGF2. In fact, the hub gene CCND2 encodes cyclin D2, a protein belonging to a highly conserved family of cyclin proteins that control cell progression through regulating cell cycle. There exists an extensive body of literature on the role of D-type cyclins as a biomarker in cancer phenotype and progression, see [Musgrove 2011] for a recent review. Connected to gene CCND2 is gene DMTF1 which encodes a cyclin D-binding myb-like transcription factor. DMTF1 was known for its tumor suppressive role linked to the regulation of many signaling pathways involving the tumor protein 53 (TP53) as well as CCND1, see [Tian 2017] for a recent review. The last subnetwork worth special mention is the gene triplet of FLT3, FLT3LG and FIZ1, among which two genes FLT3LG and FLT3 are included in the KEGG pathways in cancer. The gene FLT3 encodes a class III receptor tyrosine kinase that regulates hematopoiesis whose role in the pathogenesis of acute myeloid leukemia (AML) in particular has been long recognized [Levis 2003]. Besides, FLT3LG, namely the FLT3 ligand, also plays a role in the immune response, and hence it was investigated in the pursuit of promising immuno-therapy against cancer by means of vaccine adjuvant [Lynch 1997, Kreiter 2011]. Notably, gene FLT3 was selected by all three methods under consideration, but its connected gene FLT3LG was selected exclusively by the wavelet-based method.

3.4 Discussion

In the present chapter, we have studied network-based regularization methods under a predictive framework with linear models in order to incorporate relationships between features presumably encoded by a known network, and proposed to use network-based wavelet smoothing in order for subnetwork detection by structured

feature selection. Notably, the proposed methods are essentially a class of penalty terms that are readily combined with any loss function appropriately chosen depending on the application in fitting linear models, and path-wise algorithms for solving the underlying optimization problems are straightforwardly available by modifying those solving the standard lasso. Finally, we demonstrated the proposed methods by performing survival analysis for breast cancer using METABRIC gene expression data guided by a PPI network obtained from HPRD. Results show that, compared to several state-of-the-art methods, the wavelet-analysis method with spectral graph dual wavelets, as a representative of wavelet-based methods, was able to improve gene selection in terms of stability, connectivity and interpretability, while achieving competitive performance of survival risk prediction. In particular, the wavelet-based method identified larger subnetworks involving more connected genes, and the relevance of many genes to breast cancer survival have been previously reported by independent studies.

Key insights into the superiority of network-based wavelet smoothing to other network-based regularization methods regarding feature selection lie in the properties of graph wavelets. Recall that graph wavelets are graph vectors that are localized on the graph and fully determined by the local structure of the graph. In particular, the construction of graph wavelets conceals an automated procedure of designating subnetworks whose location, size and shape are inherently specified by the underlying graph wavelets. When any graph vector is decomposed into a linear combination of graph wavelets, we obtain a new representation of the graph vector that are decorrelated and modularized with respect to graph wavelets. The idea of wavelet smoothing originates from seeking for a coefficient vector of the linear model that is sparse in its wavelet representation, and the modularized sparsity of the coefficient vector consequently enables direct identification of subnetworks adapted to optimizing the prediction. Contrariwise, the sparsity-inducing term for feature selection in the Laplacian lasso is a standard lasso term that regards features rather individually, despite an additional Laplacian term that controls the global smoothness of the coefficients and thus encourages features closer on the network to be selected simultaneously. Likewise, the spirit of feature selection for the graph-fused lasso is indeed direct identification of subnetworks. This is achieved, however, through an estimate of the coefficient vector that is forced piece-wise constant over the network. Therefore, the resulting subnetwork is only data-adaptive but not adapted to the locally irregular structure of the network. Besides, when we expect that the relationships between features conform to the network structure, the compulsory constrain seems too strong an assumption that all features in each subnetwork should share exactly the same coefficient.

When performing network-based analysis of gene expression data, we highlighted the benefits of using a PPI network to guide gene selection related to breast cancer survival. An important issue is that our knowledge of protein-protein interaction is undoubtedly incomplete and the edges of the known biological network can possibly be subject to errors or misspecifications, especially when it comes to the biology of cancer mechanism. Future research of pressing need would be to investigate how

sensitive the results are to perturbation of the network structure. A potentially helpful trick to improve our trust in the external knowledge provided by the network could be to adapt the given network to data by modifying edges from the network, for instance removing certain edges if the correlation of the expression levels between the two connected genes is very small. Another issue prior to employing network-based analysis is to decide which biological network to use, which in principle depends on domain expertise. In fact, the methods considered in this study, including the wavelet-based methods, can be applied with various biological networks such as coexpression networks. However, an open question is to compare the list of selected genes and detected subnetworks when guided by different biological networks or from different databases. Finally, we would like to point out that, despite the findings of this study and many others that have demonstrated improved gene selection in breast cancer outcome prediction (see, e.g., [Allahyar 2015]), the rationale for network-guided genomic data analysis for improving the prediction performance remains a controversial topic [Staiger 2013]. Comprehensive studies benchmarking the breast cancer survival analysis with more datasets, networks, regularization methods and prediction tasks, such as those that follow the evaluation pipeline of [Staiger 2013], is called for in future research.

For the methods of network-based wavelet smoothing, many variants exist and have been empirically tested in the numerical experiments of this study, raising a few points worth discussion and further investigation. First, we observed distinct performance with respect to which version of graph Laplacian (normalized or non-normalized) is used for the construction of spectral graph wavelets. In fact, there are many theoretical guarantees that favor the normalized Laplacian [Chung 1997] but a debate is ongoing over which version should be used in practice. Although we opted for the non-normalized Laplacian by default, we do not conclude definitively on the choice of Laplacian that is better suited for the construction of spectral graph wavelets integrated in network-based wavelet smoothing. Second, this study provides an empirical benchmark comparing the performance of two particular types of graph wavelets, that are spectral graph wavelets and lifting-based graph wavelets. Evidences from all experiments strongly advocate the use of spectral graph wavelets over lifting-based graph wavelets, and the substantially deficient performance of latter is somewhat surprising. In fact, if data are time series that reside on 1-dimensional chain graph, it has been proven that any discrete wavelet transform with all classical wavelet filter banks can be factored into a sequence of lifting steps [Daubechies 1998]. Therefore, it calls for theoretical studies that provide a unifying overview of different techniques of performing wavelet transform on general graphs. Third, we proposed to use two approaches to wavelet smoothing, namely the synthesis approach (3.15) and the analysis approach (3.16). Despite the similarity in their mathematical formulation, the motivation underlying both approaches differ fundamentally. The synthesis approach seeks a reconstruction of the coefficient vector as a sparse combination of graph wavelets, while the analysis approach aspires to build sparse predictive models on the wavelet-transformed feature data. Our real-data experiments on survival analysis for breast cancer particularly suggest that the

analysis approach usually outperforms the synthesis approach, as observed also by [Elad 2007] concerning applications in signal processing.

There are many interesting extensions of the current work. One direction would be to perform wavelet smoothing on directed graphs or graphs with some edge attributes. This is particularly relevant when we would like to explore relationships between features associated with irreversible direction and meaningful attributes. For example, in a signaling pathway network, each edge is associated with a direction (indicating cell signaling is transduced from one gene to another but cannot be reversed) as well as an annotated type of interaction (activation or inhibition). Another direction would be to explore the possibility of adopting, besides graph wavelet transform, other types of localized or multiscale transforms specifically designed to analyze data on graphs, such as windowed graph Fourier transform [Shuman 2012] or multiscale graph pyramid transform [Shuman 2016]. In particular when the two directions engage in one application, deeper understanding of the network utilized can be enlightening for the applicability of the methods and transforms employed.

Chapter 4

Signaling Pathway Activities Improve Prognosis for Breast Cancer

---

Publication and Dissemination: *The work in this chapter has been submitted as joint work with Marta R. Hidalgo, Cankut Çubuk, Alicia Amadoz, José Carbonell-Caballero, Jean-Philippe Vert and Joaquín Dopazo in [Jiao 2017a].*

Abstract: *With the advent of high-throughput technologies for genome-wide expression profiling, a large number of methods have been proposed to discover gene-based signatures as biomarkers to guide cancer prognosis. However, it is often difficult to interpret the list of genes in a prognostic signature regarding the underlying biological processes responsible for disease progression or therapeutic response. A particularly interesting alternative to gene-based biomarkers is mechanistic biomarkers, derived from signaling pathway activities, which are known to play a key role in cancer progression and thus provide more informative insights into cellular functions involved in cancer mechanism. In this chapter, we demonstrate that a pathway-level feature, such as the activity of signaling circuits, outperform conventional gene-level features in prediction performance in breast cancer prognosis. We also show that the proposed classification scheme can even suggest, in addition to relevant signaling circuits related to disease outcome, a list of genes that do not code for signaling proteins whose contribution to cancer prognosis potentially supplements the mechanisms detected by pathway analysis.*

Résumé : *Avec l'avènement des technologies à haut débit pour le profilage d'expression génomique, un grand nombre de méthodes ont été proposées pour découvrir des signatures basées sur les gènes en tant que biomarqueurs pour aider le pronostic du cancer. Cependant, il est souvent difficile d'interpréter la liste des gènes dans une signature pronostique et ce à cause des processus biologiques sous-jacents responsables de la progression de la maladie*

ou de la réponse thérapeutique. Une alternative particulièrement intéressante aux biomarqueurs génétiques est le biomarqueur mécanique, dérivé des activités des voies de signalisation, qui est connu pour jouer un rôle clé dans la progression du cancer et ainsi fournir des informations plus pertinentes sur les fonctions cellulaires impliquées dans le mécanisme du cancer. Dans ce chapitre, nous démontrons que les variables issues du réseau, comme l'activité des circuits de signalisation, surpasse les variables classiques au niveau du gène en termes de prédiction du pronostic du cancer du sein. Nous montrons également que notre méthode de classification permet de proposer, en plus de la pertinence des variables issues du réseau liées au résultat de la maladie, une liste des gènes qui ne codent pas de protéines de signalisation dont la contribution au pronostic du cancer peut compléter les mécanismes détectés par l'analyse du réseau.

4.1 Introduction

Over the past decades, many efforts have been addressed to the identification of gene-based signatures to predict patient prognosis using gene expression data [van 't Veer 2002, Paik 2004, Wang 2005a, Sotiriou 2009, Reis-Filho 2011]. Despite the success of its use, gene expression signatures have not been exempt of problems [Ein-Dor 2006, Iwamoto 2010]. Specifically, one major drawback of multi-gene biomarkers is that they often lack proper interpretation in terms of mechanistic link to the fundamental cell processes responsible for disease progression or therapeutic response [van 't Veer 2008, Dopazo 2010]. Actually, it is increasingly recognized that complex traits, such as disease or drug response, are better understood as alterations in the operation of functional modules caused by different combinations of gene perturbations [Barabási 2004, Oti 2007, Barabási 2011]. To address this inherent complexity different methodologies have tried to exploit several functional module conceptual representations, such as protein interaction networks or pathways, to interpret gene expression data within a systems biology context [Barabási 2011, Vidal 2011, Hood 2013, Fryburg 2014].

Here we focus on consulting prior knowledge of signaling pathways to guide cancer prognosis. It is well understood that cell signaling is a system of within-cell communication and signal transduction process between gene products, mostly proteins, that coordinates cell activities to perceive and correctly respond to microenvironment, resulting in signaling pathways that form a particular type of functional gene modules and play a key role in disease progression (Figure 4.1). Consequently as a tempting solution to the limitation of conventional analysis at the level of individual genes, analysis at the level of pathways renders great interest in providing informative insights into cellular functions that facilitates understanding of the disease mechanism. Actually, it has recently been shown that the pathway-

level representation generates clinically relevant stratifications and outcome predictors for glioblastoma and colorectal cancer [Drier 2013] and also breast cancer [Livshits 2015]. Moreover, mathematical models of the activity of a pathway have demonstrated a significantly better association to poor prognosis in neuroblastoma patients than the activity of their constituent genes, including MICN, a conventional biomarker [Fey 2015]. This observation has recently been extended to other cancers [Hidalgo 2017] and to the prediction of drug effects [Amadoz 2015].

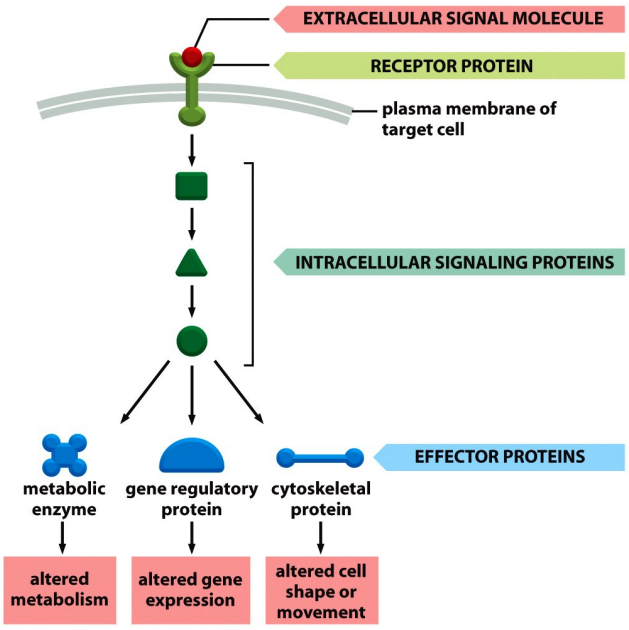


Figure 4.1: An illustration of cell signaling process. Typically the signal transduction begins at receptor proteins that receive molecular stimuli from cell microenvironment and ends at effector proteins that execute specific actions in response to the stimulation.

Given that the inferred activity of the pathway should be closely related to its cellular mechanism for disease progression, its use to guide cancer prognosis seems promising. Recently, a number of pathway activity inference methods have been proposed [Hidalgo 2017, Jacob 2012, Li 2015, Martini 2013]. Here, we use the *hiPathia* method proposed in [Hidalgo 2017], as it has been demonstrated to have a superior performance finding significant associations of specific circuit[1] activities, directly responsible for triggering the prominent cancer hallmarks [Hanahan 2011], to patient survival. This method recodes gene expression values into measurements of signaling circuit activities that ultimately account for cell responses to specific stimuli. Such activity values can be considered multigenic mechanistic biomarkers that can be used as features for cancer prognosis.

In this chapter, we demonstrate that the activity of signaling circuits yields comparable or even better prediction in breast cancer prognosis than the expression

[1]Circuits can be understood as sub-pathways with specific structure in stimulus-response signaling pathways, while definitions are postponed to Section 4.2.2.

of individual genes, while detected mechanistic biomarkers enjoy the compelling advantage of readily available interpretation in terms of the corresponding cellular functions they trigger. Moreover, we show that the proposed prediction scheme can even suggest, in addition to interesting signaling circuits related to disease outcome, a list of genes that do not code for signaling proteins whose contribution to cancer prognosis potentially supplements the mechanism included in the pathways modeled. All numerical results are produced with `R` and code for reproducing the experiments is available in the online supplementaries at https://github.com/YunlongJiao/hipathiaCancerPrognosis.

4.2 Methods

4.2.1 Data Source and Processing

Our interest in this study lies in predicting the overall survival outcome of breast cancer patients making use of gene expression data. The breast cancer gene expression and survival data here were downloaded from The Cancer Genome Atlas (TCGA), release No. 20 of the International Cancer Genome Consortium (ICGC) data portal under project name BRCA-US[2]. This dataset provides the RNA-seq counts of 18,708 genes for 879 tumor samples in which we also have records of the vital status of corresponding donors, namely the overall survival outcome of the cancer patients being alive or deceased at the end of clinical treatment (Table 4.1). This way we deal with a binary classification problem distinguishing good vs poor prognosis based on gene expression measurements of breast tumor samples. Since TCGA cancer data are collected from different origins and underwent different management processes, non-biological experimental variations, commonly known as batch effect, associated to Genome Characterization Center (GCC) and plate ID must be removed from the RNA-seq data. The COMBAT method [Johnson 2007] was used for this purpose. We then applied the trimmed mean of M-values normalization method (TMM) method [Robinson 2010] for data normalization which is essential in applying the *hiPathia* method. The resulting normalized values were finally entered to the pathway analysis method.

Table 4.1: Summary of survival outcome of the breast cancer patients in the TCGA dataset.

Donor vital status	Pseudo label	No. of samples	Percentage
Deceased (poor prognosis)	Positive	124	14.1%
Alive (good prognosis)	Negative	755	85.9%
	Total	879	100.0%

In order to explore the potential of utilizing external knowledge of cell signaling to enhance prognosis, we consulted Kyoto Encyclopedia of Genes and Genomes

[2]More information can be found at https://dcc.icgc.org/releases/release_20/Projects/BRCA-US.

(KEGG) repository [Kanehisa 2012] to retrieve relationships between proteins within signaling pathways. A total of 60 KEGG pathways were used (Table 4.2), comprehending 2,212 gene products that participate in 3,379 nodes. Note that most gene products are proteins, and two types of nodes are defined in KEGG: plain nodes which may contain one or more proteins and complex nodes. These pathways each compose into a directed network where nodes are connected with edges labeled by either activation or inhibition depending on the action in transmitting signals along the path. In particular, input nodes that have no incoming edges represent receptor proteins which receive molecular stimuli from cell microenvironment, and output nodes that have no outgoing edges represent effector proteins which carry out specific cellular functions. We will elaborate in the following subsection on how to decompose the complex structure of KEGG pathways in order to effectively apply the *hiPathia* method.

Table 4.2: The 60 KEGG pathways for which signaling activity is modeled.

KEGG identifier	**Pathway name**
hsa04014	Ras signaling pathway
hsa04015	Rap1 signaling pathway
hsa04010	MAPK signaling pathway
hsa04012	ErbB signaling pathway
hsa04310	Wnt signaling pathway
hsa04330	Notch signaling pathway
hsa04340	Hedgehog signaling pathway
hsa04350	TGF-beta signaling pathway
hsa04390	Hippo signaling pathway
hsa04370	VEGF signaling pathway
hsa04630	Jak-STAT signaling pathway
hsa04064	NF-kappa B signaling pathway
hsa04668	TNF signaling pathway
hsa04066	HIF-1 signaling pathway
hsa04068	FoxO signaling pathway
hsa04020	Calcium signaling pathway
hsa04071	Sphingolipid signaling pathway
hsa04024	cAMP signaling pathway
hsa04022	cGMP-PKG signaling pathway
hsa04151	PI3K-Akt signaling pathway
hsa04152	AMPK signaling pathway
hsa04150	mTOR signaling pathway
hsa04110	Cell cycle
hsa04114	Oocyte meiosis
hsa04210	Apoptosis

Continued on next page

Table 4.2 – continued from previous page

KEGG identifier	Pathway name
hsa04115	p53 signaling pathway
hsa04510	Focal adhesion
hsa04520	Adherens junction
hsa04530	Tight junction
hsa04540	Gap junction
hsa04611	Platelet activation
hsa04620	Toll-like receptor signaling pathway
hsa04621	NOD-like receptor signaling pathway
hsa04622	RIG-I-like receptor signaling pathway
hsa04650	Natural killer cell mediated cytotoxicity
hsa04660	T cell receptor signaling pathway
hsa04662	B cell receptor signaling pathway
hsa04664	Fc epsilon RI signaling pathway
hsa04666	Fc gamma R-mediated phagocytosis
hsa04670	Leukocyte transendothelial migration
hsa04062	Chemokine signaling pathway
hsa04910	Insulin signaling pathway
hsa04922	Glucagon signaling pathway
hsa04920	Adipocytokine signaling pathway
hsa03320	PPAR signaling pathway
hsa04912	GnRH signaling pathway
hsa04915	Estrogen signaling pathway
hsa04914	Progesterone-mediated oocyte maturation
hsa04921	Oxytocin signaling pathway
hsa04919	Thyroid hormone signaling pathway
hsa04916	Melanogenesis
hsa04261	Adrenergic signaling in cardiomyocytes
hsa04270	Vascular smooth muscle contraction
hsa04722	Neurotrophin signaling pathway
hsa05200	Pathways in cancer
hsa05231	Choline metabolism in cancer
hsa05202	Transcriptional misregulation in cancer
hsa05205	Proteoglycans in cancer
hsa04971	Gastric acid secretion
hsa05160	Hepatitis C

4.2.2 Modeling Framework for Signaling Pathways

We applied the *hiPathia* method[3] proposed by [Hidalgo 2017] in pursuit of modeling signaling activity. Overall, *hiPathia* is a method that estimates the level of activity within a signaling circuit by modeling cell signaling process in order to recode gene expression values into measurements that ultimately account for cell responses caused by pathway activities. Essentially the *hiPathia* method computes an activity value for each stimulus-response sub-pathway within signaling circuits. This way, the sub-pathways which associate naturally with cell functionalities can be considered as mechanistic features that are modularized from multigenic signatures, and their activity values connected to the activation or deactivation of specific cellular functions thus provide quantitative clues to understand disease mechanisms when further related to phenotypes of interest such as cancer survival.

Recall that in cell signaling processes represented in KEGG pathways, cell signal arrives to an initial input node and starts to transmit along any path following the direction of the edges until it reaches an output node that finally triggers a cellular action. In particular, from different input nodes the signal may follow different routes to reach different output nodes. Within the modeling context, a *circuit* is naturally defined as all possible routes the signal can traverse to be transmitted from a particular input node to a particular output node (Figure 4.2, A). A total of 6,101 circuits are identified and modeled in this study. Now we take efforts to describe first how *hiPathia* estimates the signaling activity of a circuit.

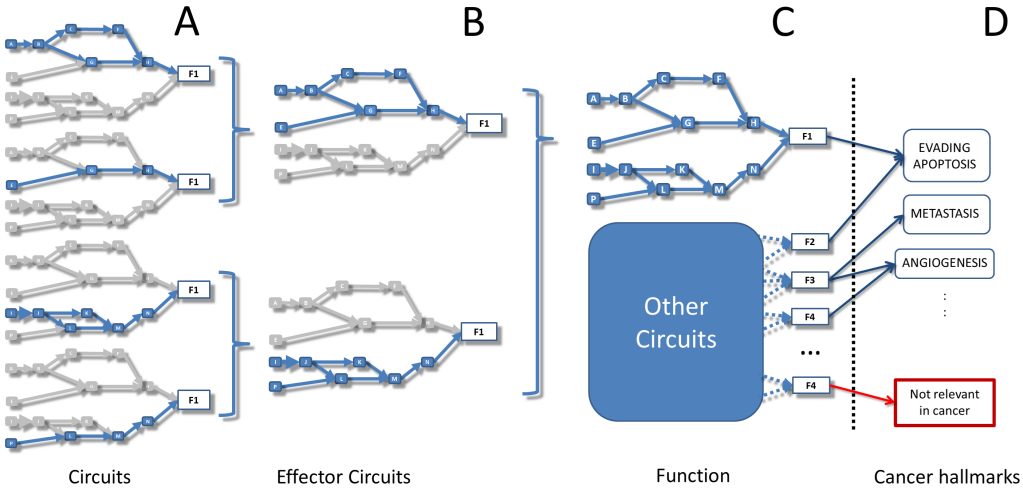


Figure 4.2: The different levels of abstraction within pathways: A) Circuits that communicate one receptor to one effector; B) Effector circuits that communicate all the receptors that signal a specific effector; C) Function circuits that collect the signal from all the effectors that trigger a specific function (according to UniProt or GO keywords); D) Cancer hallmarks, a sub-selection of only those functions related to cancer hallmarks.

[3] Available as an `R` package at `https://github.com/babelomics/hipathia` and via a web interface at `http://hipathia.babelomics.org/`.

In a signaling circuit, the transmission of the signal depends on the integrity of the chain of nodes that connect the receptor to the effector and the capability of transmitting signals of each node involved intuitively depends on two folds: the abundance of the proteins corresponding to that node and its activity status due to the interaction with its parent nodes. First, we need to estimate a value for each node in the pathways in regard to the presence of proteins involved in the node. Following the convention of [Bhardwaj 2005, Efroni 2007, Montaner 2009, Sebastian-Leon 2014], the presence of the mRNA (the normalized RNA-seq counts rescaled between 0 and 1) is taken as a proxy for the presence of the proteins involved in each node. Notably, for different types of nodes defined in KEGG, the value of a plain node in the pathways is defined as the ninetieth percentile of the values of the proteins contained, and the value of a complex node is taken as the minimum value of the proteins contained (the limiting component of the complex). Then, the degree of integrity of the circuit is estimated by modeling the signal flow across it, transmitting node-by-node following the path while its intensity value gets propagated along the way taking into account the current node value and the intensity of the signals arriving to it. Specifically, we initialize an incoming signal of intensity value of 1 received by the input (receptor) node of the circuit, and then for each node n of the circuit, the signal value s_n is updated by the following rule:

$$s_n = v_n \cdot \left(1 - \prod_{a \in A_n} (1 - s_a)\right) \cdot \prod_{i \in I_n} (1 - s_i)\,,$$

where A_n denotes the set of signals arriving to the current node n from activation edges, I_n denotes the set of signals arriving to the current node n from inhibition edges, and v_n is the (normalized) value of the current node n. In case of loops present in the circuit, a node may be visited multiple times, until the difference in the updates of the signal value at that node is below certain threshold, before the signal exits the loop and continues to propagate down the cascade. Finally, the activity value for the circuit is defined by the signal intensity transmitted through the last (effector) protein of the circuit which quantifies the cell function ultimately activated by the circuit. See Figure 4.3 for an example of deducing the activity value of an artificial circuit by the *hiPathia* method.

Besides, the *hiPathia* method straightforwardly allows to explore pathway-level analysis at different levels of abstraction by applying to different notions of signaling circuits. As the output nodes at the end of circuits are the ultimate responsible to carry out specific cellular actions, an *effector circuit* is defined from a functional viewpoint as a higher-level signaling entity that compose all circuits ending at the same output node (Figure 4.2, B). When applied to an effector circuit, the *hiPathia* method returns the joint intensity of the signal arriving to the corresponding effector node. Furthermore, the known functions triggered in cell by each effector protein can be derived from their functional annotations. Here we use UniProt [Consortium 2015b] and Gene Ontology (GO) [Consortium 2015a] annotations. Finally, inferred signaling activity values of those effector circuits ending at proteins

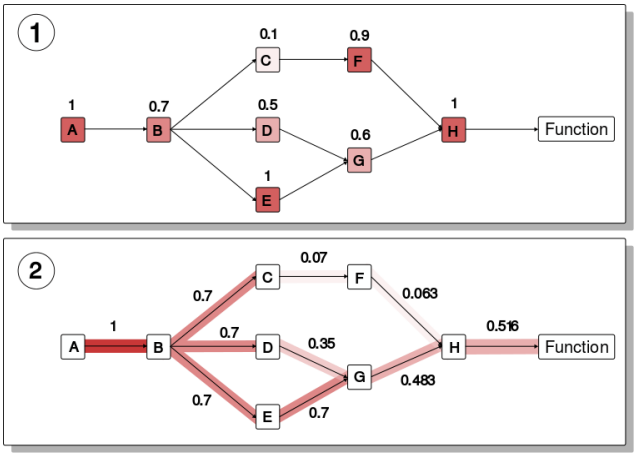


Figure 4.3: An example of computing the activity value of an artificial circuit by the *hiPathia* method. In Step 1, node values are derived from the normalized mRNA measurements. In Step 2, signal is propagated along the path while its intensity value gets updated according to the rule of the *hiPathia* method. Finally, The signal value attained after the last protein is visited accounts for the signaling activity of the circuit.

with the same annotated functions are averaged to quantify the activity of the function realized in cell. This way we obtain estimated activity values directly connected to a list of cellular functions (Figure 4.2, C). Figure 4.2 depicts the different levels of abstraction from circuits, to effector circuits and finally functions. Eventually for the sake of interpretation, a subset of curated functions can be used for a specific scenario in which the relevant functions are known to interpret the cancer biology, for which we use cancer hallmarks [Hanahan 2011] (Figure 4.2, D).

4.2.3 Cancer Prognosis with Inferred Signaling Pathway Activity

In this study, we are interested in comparing the prognostic power of pathway-level mechanistic features and gene-based features, both separately and in combination, in order to distinguish good vs poor prognosis. Using the *hiPathia* method, we recoded the list of gene expression values of each tumor sample into the corresponding lists of signaling activity values for the three levels of abstraction: circuits, effector circuits and functions, as described in UniProt and GO annotations. Therefore for each tumor sample, we end up with a profile of gene expression, a profile of circuit signaling activity, a profile of effector circuit signaling activity, a profile of UniProt-based cellular function activity and a profile of GO-based cellular function activity. These profiles are sample-specific, or so-called *personalized*, profiles that can be straightforwardly used as prognostic features for cancer prognosis following any off-the-shelf classification algorithm. Note that pathway-level profiles are derived with no regard to any information provided by the genes whose products do not participate in cell signaling, and the prognostic power of pathway-level profiles may thus be limited by the coverage of genes in known biological pathways. In order to understand the relative contribution to the pathway-level profiles and

gene-level profiles to the accurate separation between good vs poor prognosis, we devised 4 artificial profiles: path-gene expression profile containing only genes that are involved in the KEGG signaling pathways, other-gene expression profile containing only genes that are absent from the KEGG pathways, a combined profile consisting of signaling activity of effector circuits and expression of other-genes, and a combined profile consisting of signaling activity of circuits and expression of other-genes. Thus we obtained a total of 9 types of profiles (detailed in Table 4.3).

Table 4.3: Summary of 9 different types of profiles used as predictive features for breast cancer prognosis.

Alias	**Profile type**	**No. of features**	**Analysis level**
fun.vals	UniProt-based functions	81	Pathway-level cellular function values
go.vals	GO-based functions	370	
eff.vals	Effector circuits	1,038	Pathway-level signaling activity values
path.vals	Circuits	6,101	
path.genes.vals	Path-genes	2,212	Gene-level expression values
other.genes.vals	Other-genes	16,496	
genes.vals	All genes	18,708	
eff.and.other.genes.vals	Effector circuits and other-genes	17,534	Combination of pathway-level signaling activity values and gene-level expression values
path.and.other.genes.vals	Circuits and other-genes	22,597	

From the viewpoint of machine learning, this study is formulated as a typical binary classification problem where we determine a positive or negative pseudo label for each sample. Based on the data available in this study (Table 4.1), we perform a 5-fold cross-validation repeated 10 times on the dataset and report the mean performance over the $5 \times 10 = 50$ splits to assess the prognostic power for each type of profile. The performance is evaluated by the Area Under the ROC Curve (AUROC) criteria [Sing 2005]. Note that usually a classifier returns a continuous prediction between 0 and 1 for each sample denoting the probability of that sample being in the positive class rather than in the negative class, and then assigns either label to the sample according to some cutoff value thresholding the prediction. AUROC is a cutoff-free score that measures the probability that the classifier will score a randomly drawn positive sample higher than a randomly drawn negative sample.

In this study, we considered a total of 12 classification algorithms as candidate classifiers, most of which are state-of-the-art (Table 4.4). When we assess the

prognosis performance for a specific type of profile on a specific (external) cross-validation split of the data, we perform an internal 5-fold cross-validation on the training set to determine which classifier returns the highest cross-validated performance and the best classifier is then used on the test set to obtain the performance score. The rationale behind the nested cross-validation is that, although any classification algorithm from the machine learning literature can be used to discriminate good vs poor prognosis with any profile type considered as predictive features, in practice, however, we do not have a definitive concept of which classifier suits the best universally for all types of profiles. In other words, it will be a confusing factor if we predetermine just one classifier throughout the study. In fact, the underlying hypotheses of different classifiers vary, for instance linear or non-linear relationships can be assumed between features and labels, and some classifiers can be particularly sensitive to the presence of a large number of noisy features. As a consequence, the procedure of choosing the best suited algorithm for different types of profiles by a nested cross-validation guarantees that the prediction performance is evaluated in an impartial fashion.

4.3 Results

4.3.1 Signaling Pathway Activities Lead to Improved Prognosis for Breast Tumor Samples

The performance of using different types of profiles (Table 4.3) as predictive features to classify survival outcome for breast cancer patients is shown in Figure 4.4 while mean scores with standard deviation are reported in Table 4.5. Under the criterion of AUROC to evaluate the classification performance, we observe that the activity values of signaling circuits, denoted by *path.vals*, yield the best performance overall. In particular, they outperform the profiles based solely on gene expression values, denoted by *path.genes.vals*, *other.genes.vals* and *genes.vals*. In other words, we are able to integrate the expression values of path-genes with the prior knowledge of cell signaling to obtain pathway-level features that achieve improved prognosis. Interestingly, these pathway-level features relate to biological processes and cellular functions *per se*. Although the pathway-level features are derived from the expression of path-genes and thus agnostic to the expression of other-genes, the inclusion of other-genes to the signaling circuits, inducing the profiles denoted by *eff.and.other.genes.vals* and *path.and.other.genes.vals*, does not significantly improve the performance by performing a two-sided t-test comparing the difference between the cross-validation AUROC scores obtained by each pair of profiles further FDR-adjusted for multiple testing [Benjamini 1995] (Table 4.6).

When comparing the prognostic power between pathway-level and gene-level profiles, we have also derived cellular function activity profiles, denoted by *fun.vals* and *go.vals* (Table 4.3), and observed that the performance of these profiles are slightly worse than other pathway-level profiles (Figure 4.4). This is probably due to the excessively simplistic procedure that basically averages the signaling activity

Table 4.4: The 12 candidate classifiers used to discriminate prognosis classes for breast tumor samples.

Alias	Classifier	Reference
LDA	Linear discriminant analysis	[Venables 2002, Ripley 2007]
LogitLasso	L1-regularized logistic regression	[Friedman 2010]
LinearSVM	Support Vector Machines with linear kernel	[Chang 2011]
RadialSVM	Support Vector Machines with Gaussian RBF kernel	[Chang 2011]
KendallSVM	Support Vector Machines with Kendall kernel	[Zeileis 2004, Jiao 2015]
KNN	k-nearest neighbor classifier	[Venables 2002, Ripley 2007]
NB	Naive Bayes classifier	[Ripley 2007]
GBM	Gradient Boosting Machines	[Friedman 2001]
RF	Random Forests	[Liaw 2002, Breiman 2001]
SparseSVM	L1-regularized L2-loss Support Vector Machines	[Fan 2008]
PAM	Nearest shrunken centroid classifier	[Tibshirani 2002]
Constant	Majority voting classifier	Outputs constant label of the dominant class (negative-control)

Table 4.5: Mean AUROC scores with standard deviation (SD) and the top 2 most frequently selected classifiers by internal cross-validation for each type of prognostic profile in classifying breast cancer prognosis.

Profile alias	Mean	SD	Classifier 1	Classifier 2
fun.vals	0.6962	0.05438	*RadialSVM*	*GBM*
go.vals	0.6807	0.06095	*RadialSVM*	*LinearSVM*
eff.vals	0.7087	0.05099	*RadialSVM*	*LinearSVM*
path.vals	0.7211	0.06316	*RadialSVM*	*LinearSVM*
path.genes.vals	0.6938	0.05636	*RadialSVM*	*LinearSVM*
other.genes.vals	0.7075	0.05254	*LinearSVM*	*RadialSVM*
genes.vals	0.7075	0.05272	*LinearSVM*	*RadialSVM*
eff.and.other.genes.vals	0.7127	0.05838	*LinearSVM*	*RadialSVM*
path.and.other.genes.vals	0.7246	0.05359	*LinearSVM*	*RadialSVM*

Table 4.6: FDR-adjusted p-values comparing the difference between the corresponding AUROC scores of profiles in columns and profiles in rows over 50 cross-validation splits. See Table 4.5 for the mean scores of each profile individually. Significant p-values are boldfaced and marked with asterisks.

	go.vals	*eff.vals*	*path.vals*	*path.genes.-vals*	*other.genes.-vals*	*genes.vals*	*eff.and.-other.-genes.vals*	*path.and.-other.-genes.vals*
fun.vals	0.1467	0.1306	**0.0119***	0.8267	0.1908	0.1908	0.0815	**0.0020****
go.vals		**0.0024****	**0.0004*****	0.1557	**0.0120***	**0.0119***	**0.0046****	**<0.0001*****
eff.vals			0.0815	**0.0255***	0.8743	0.8743	0.6422	**0.0235***
path.vals				**0.0046****	0.1408	0.1394	0.3702	0.6422
path.genes.-vals					**0.0255***	**0.0255***	**0.0046****	**0.0002*****
other.genes.-vals						0.9483	0.2167	**0.0039****
genes.vals							0.2000	**0.0032****
eff.and.-other.-genes.vals								**0.0473***

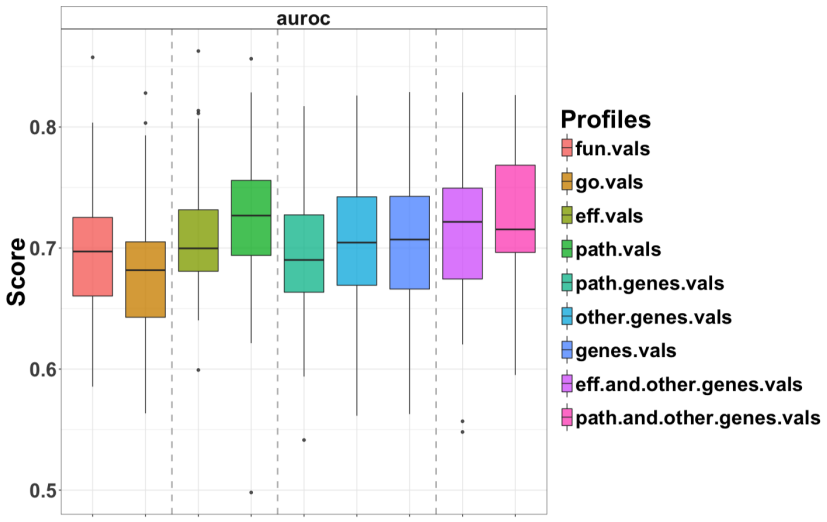


Figure 4.4: The AUROC performance of using different types of profiles as predictive features to classify survival outcome for breast cancer patients. Boxplot represents the variance of the performance on 50 cross-validation splits. Dotted vertical lines separate profiles by the underlying analysis levels.

values of effector circuits ending at proteins with the same annotated keywords according to UniProt or GO [Hidalgo 2017], annotations that can be incomplete and ambiguous to some extent.

Table 4.5 summarizes the best-performing classifiers for each type of prognostic profile in the sense that they are most frequently selected by internal cross-validation. Notably, it evidences that Support Vector Machines with various kernels are recurrently selected as the competent classifier in breast cancer prognosis that suits well for both gene-level and pathway-level features.

4.3.2 Signaling Circuits Selected as Features Relevant for Cancer Prognosis Account for Cancer Hallmarks

From the clinical standpoint of cancer prognosis, we are interested in identifying a small set of biomarkers that can guide decision making in cancer prognosis. As our analysis is made at the level of pathways, we would like to detect a few signaling circuits whose activity, and thus the underlying cell functionality, has a significant impact on discriminating the prognosis classes of cancer patients. We opted for the Random Forests classifier to perform this analysis, since it simultaneously predicts the survival outcome of tumor samples and scores the importance of each feature that is ultimately used in the prediction. We focus on the feature importance measure returned by fitting a Random Forest which accounts for the mean decrease in classification performance if we randomly permute the data of the corresponding feature.

Table 4.7 lists the 5 top-scored circuits by fitting Random Forests with the profiles of circuit activities, denoted by *path.vals*. The role played by each signaling

circuit in cancer progression can be inferred from the underlying cellular functions (taken from GO annotations) triggered by the last (effector) protein on the circuit. Thus, the first circuit, belonging to the HIF-1 signaling pathway, starts with the TLR4 receptor, which is known to be related to progression of several cancers (breast, ovarian, prostate and head and neck) via lipopolysaccharide stimulation [Yang 2014] and ends in the EDN1 effector, an hypoxia-inducible factor that mediates cancer progression [Semenza 2012]. Another relevant circuit belongs to the NF-kappa B signaling pathway and has the IL1B protein as receptor and the CXCL2 as effector. Polymorphisms in the receptor have been linked to several cancers in different populations [El-Omar 2000, Lu 2005] and it has been demonstrated the role of CXCL2 in tumor growth and angiogenesis [Keane 2004]. Similarly, polymorphisms in the LEP protein, the receptor of another circuit in the Adipocytokine signaling pathway, have been linked to cancer [Cleveland 2010], and its effector, the tyrosine phosphatase Shp2 (PTPN11), contributes to the pathogenesis of many cancers and other human diseases [Chan 2008]. The Cell cycle signaling pathway contains another relevant circuit whose receptor TTK transmits the signal until the cohesin complex. This four proteins complex is essential for chromosome segregation and DNA repair and mutations in its component genes have recently been identified in several types of tumors [Losada 2014]. Finally, the last relevant circuit, belonging to the Tight junction pathway, contains the AKT3 serine/threonine kinase with a known role in tumorigenesis [Testa 2001], is signaled by the receptor ACTN4, a protein which has been related to cell invasion and metastasis [Honda 2015]. An expanded list of top-scored 50 circuits can be found in online supplementaries.

Table 4.8 lists the 5 top-scored effector circuits by fitting Random Forests with the profiles of effector circuit activities, denoted by *eff.vals*. Although the cohesion complex effector is again selected, the effector circuit level analysis provided a slightly different perspective of relevant aspects of signaling in cancer patient survival. Thus, two effector circuits with effector proteins LEPR and PPARα, from the AMPK and the Adipocytokine signaling pathways, respectively, are activators of the fatty acid metabolism. Two more effector pathways ending in the Interleukin 6 (IL6), related to inflammatory processes and immune response in the Toll-like receptor pathway, seem more likely to be involved in blocking the cell differentiation through the Pathways in cancer (KEGG ID hsa05200). Actually, it has been described that IL6 blocks apoptosis in cells during the inflammatory process, keeping them alive in toxic environments, but the same process protects cells from apoptosis and chemotherapeutic drugs during neoplastic growth [Hodge 2005]. An expanded list of top-scored 50 effector circuits can be found in the online supplementaries.

4.3.3 The Classification Algorithm Suggests Additional Prognostic Genes That Do Not Code for Signaling Proteins

In order to find genes that could be relevant for patient survival that are not in the signal pathways, we have constructed a profile by combining signaling circuit activity profiles and gene expression profiles corresponding to other-genes absent

Table 4.7: Top 5 circuits with the highest feature importance measure by fitting Random Forests with *path.vals* in classifying breast cancer prognosis, along their functions as annotated in Gene Ontology (GO).

Rank	Pathway name	Receptor genes	Effector genes	Effector protein GO function
1	HIF-1 signaling pathway	TLR4	EDN1	Growth/survival factor in cancer
2	NF-kappa B signaling pathway	IL1B	CXCL2	Inflammatory response and angiogenesis
3	Adipocytokine signaling pathway	LEP	PTPN11	Protein phosphatase
4	Cell cycle	TTK	Cohesin complex (SMC1B, SMC3, STAG1, RAD21)	Chromosome segregation and DNA repair
5	Tight junction	ACTN4, MAGI3	AKT3	Cell invasion and metastasis

Table 4.8: Top 5 effector circuits with the highest feature importance measure by fitting Random Forests with *eff.vals* in classifying breast cancer prognosis, along their functions as annotated in Gene Ontology (GO).

Rank	Pathway name	Effector genes	Effector protein GO function
1	AMPK signaling pathway	LEPR	Regulation of fatty acid metabolism
2	Adipocytokine signaling pathway	PPARα	Peroxisome proliferation and fatty acid metabolism
3	Pathways in cancer	IL6	Blockage of differentiation, Anti-apoptosis
4	Cell cycle	Cohesin complex (SMC1B, SMC3, STAG1, RAD21)	Chromosome segregation and DNA repair
5	Toll-like receptor signaling pathway	IL6	Inflamation, Immune response, Anti-apoptosis

from signaling pathways, denoted by *path.and.other.genes.vals*. A feature selection procedure in breast cancer prognosis based on such a profile can select signaling circuits along with genes whose activity is unrelated to cell signaling but nonetheless related to patient survival. To this end, Random Forests was again used to score feature importance when fitted with the *path.and.other.genes.vals* profile in the classification of breast cancer survival outcome.

Table 4.9 lists the 5 top-scored other-genes that are part of the *path.and.other.genes.vals* composed profile. These genes are of particular interest given that they might represent relevant cancer processes not included in cell signaling. Notably, the gene ABCB5 belongs to the ATP-binding cassette subfamily B which is well known to be involved in multiple drug resistance in cancer therapy [Dean 2001], probably due to its functionality of efflux transmembrane transporter. It has also been reported that ABCB5 could mediate cell-to-cell fusion and contribute to breast cancer chemoresistance in expressing breast tumors [Frank 2003, Frank 2005]. In addition, ABCB5, as a "pro-survival" gene, has been suggested to be a potential target against drug resistant breast cancer cells [Yang 2010]. Besides, ABCB5 has been linked to melanoma [Wilson 2014]. LMO4 encodes a LIM-domain protein that has been reported as an essential mediator of cell cycle progression in ErbB2/HER2/Neu-induced breast cancer which is characterized by poor survival due to high proliferation and metastasis rates [Montañez-Wiscovich 2009, Matthews 2013]. It has been reported that LMO4 interacts with the renowned tumor suppressor BRCA1 and inhibits BRCA1 activity [Sum 2002, Sutherland 2003]. OPA1 encodes a mitochondrial fusion protein which might be a target for mitochondrial apoptotic effectors [Olichon 2003], such as sorafenib [Zhao 2013]. The role in cancer survival played by two most important genes according to the results, VPS72 and CHADL, is not as clear from the literature. It is worth mentioning that a mutation in VPS72 in cervix cancer with a high FATHMM pathogenicity score [Shihab 2015] is described in the COSMIC database (entry COSM458603). Regarding CHADL, it has been related to chondrocyte differentiation [Tillgren 2015] and extracellular matrix remodeling [Barallobre-Barreiro 2012]. Therefore, both genes are potentially involved in cancer processes, which suggest that further investigation of the complete list of top-ranked other-genes could render new cancer drivers and potential therapeutic targets. An expanded list containing the top 50 most important features among the other-genes can be found in online supplementaries, in which many genes with cancer-related functions[4] can be seen.

4.4 Discussion

In this study we have proposed a novel scheme to classify survival outcome for breast cancer patients based on mechanistic features consisting of signaling pathway activity profiles. We applied a pathway activity analysis method *hiPathia*

[4]Functions for those genes were taken from their UniProt annotations and, when absent, from GeneCards annotations [Stelzer 2016].

Table 4.9: Top 5 other-genes (genes unrelated to cell signaling) with the highest feature importance measure by fitting Random Forests with *path.and.other.genes.vals* in classifying breast cancer prognosis, along their functions as annotated in Gene Ontology (GO).

Rank	Gene ID	Gene symbol	Full name	GO function
1	6944	VPS72	Vacuolar protein sorting 72 homolog	DNA binding
2	150356	CHADL	Chondroadherin like	Collagen binding
3	340273	ABCB5	ATP binding cassette subfamily B member 5	ATP binding, Efflux transmembrane transporter activity
4	8543	LMO4	LIM domain only 4	Transcription factor activity, Sequence-specific DNA binding
5	4976	OPA1	OPA1, mitochondrial dynamin like GTPase	GTPase activity

[Hidalgo 2017] to recode gene expression profiles into activity values of signaling circuits, and demonstrated that, making use of the state-of-the-art computational tools, signaling circuit activity yields better prediction in breast cancer prognosis than gene expression. An additional advantage is that the identified pathway-level biomarkers are mechanistic signatures whose contribution to cancer progression can be readily interpreted in terms of the underlying cellular functions and biological processes.

The three feature sets *path.genes.vals*, *eff.vals* and *path.vals* are composed by the same set of genes, path-genes that are present in the pathways. However, pathway activity values recoded from these genes with the *hiPathia* method, *eff.vals* and *path.vals*, clearly outperforms (see Table 4.6) the original path-genes, *path.genes.vals*, in terms of prediction performance. Moreover, compared to the prediction performance with features based on all the genes, *genes.vals*, that indeed carry more information than the subset of path-genes, features based on path-genes, *path.genes.vals*, are significantly worse while features based on circuits of path-genes, *eff.vals* and *path.vals*, are competitive (see Table 4.6). It is worth noting that genes in the circuits assume only 12% of the total number of genes. Therefore, it suggests that combining the genes into circuits provides a real added value for prediction purposes.

Although a significant improvement of the performance was not observed when the expression values of other-genes were concatenated to the activity values of signaling circuits, the analysis based on the combination of the two provides an

interesting perspective regarding the interpretation of the biomarkers detected. In fact, the selected genes from the category of other-genes represent other aspects of the mechanism of the disease not explained by cell signaling. This approach allows expanding the scope of the analysis beyond the processes included in the pathways modeled.

Central to this study is the idea of promoting gene-level analysis to pathway-level analysis by obtaining personalized profiles of signaling circuit activity by applying the *hiPathia* method. We deem that reliable models of pathway activity have the potential be used to derive robust multigenic biomarkers, in the spirit of renowned MammaPrint [van 't Veer 2008], which in addition account properly for the underlying disease mechanisms or mechanisms of drug action.

Chapter 5

Conclusion and Perspectives

To summarize, the work presented in this thesis has been driven mainly by the development and investigation of machine learning methods to address the computational challenges confronted in genomic data analysis for breast cancer prognosis: ranked-based approaches for improved molecular prognosis and network-guided approaches for enhanced biomarker discovery. In fact, it is noteworthy that the theoretical and methodological contribution is significant and lies fundamentally in several branches of machine learning concerning applications across but not limited to cancer biology and social choice theory, specifically:

Learning with Rank Data. We have proposed two computationally attractive positive definite kernels between permutations, namely the Kendall and Mallows kernels, and further extended these kernels to rank data of complex structure that prevail in real-world applications including partial rankings, multivariate rankings and uncertain rankings (Chapter 2). The significance of this work is of at least two folds:

1. Thanks to these kernels, many kernel machines serve as off-the-shelf alternatives available to solve various problems pertaining to learning from rankings, or learning to rank, and can yield state-of-the-art performance that were demonstrated with an unsupervised cluster analysis of heterogeneous voting data and supervised biomedical classification tasks (Chapter 2).

2. The Kendall embedding of the symmetric group brings forth novel incentives of learning on the symmetric group from unprecedented aspects. For instance, the Kendall embedding has motivated a geometric interpretation of the combinatorial problem of Kemeny aggregation based on which a tractable approximation bound was derived (Appendix A) and can offer the opportunity of studying permutation problems such as seriation with yet another embedding following [Fogel 2013, Lim 2014].

Learning on Graphs. Given a network, we focused on network-guided feature selection coherent with the presumed network structure in two cases:

1. In case that the network is represented by an undirected graph and it encodes codependent relationships between features, assuming that neighboring features are encouraged to be selected simultaneously, we formalized the use

of network-based wavelet smoothing as a regularization method for inducing structured sparsity with network-adaptive modularity in linear predictive models (Chapter 3).

2. In case that the network is represented by a directed graph and each circuit in it encodes the transduction of signal between features, assuming that circuit-level groups of features, if selected, are always selected simultaneously, we investigated the idea of first transforming the original representation of feature data into a circuit-level representation based on mathematical modeling of the network structure and then applying any standard feature selection algorithm to select relevant circuits which now becomes straightforwardly viable (Chapter 4).

Proof-of-methodology survival analysis of breast cancer was performed guided by a protein-protein interaction network (undirected-graph case, Chapter 3) or a signaling pathway network (directed-graph case, Chapter 4), and in both cases empirical superiority was demonstrated where feature selection for molecular prognosis is enhanced using a biological network as prior knowledge.

While the investigation of these machine learning topics covered in the present thesis and their applications in cancer prognosis is certainly unfinished as remarked in the discussion sections in each corresponding chapter, many interesting perspectives that were not covered in the present thesis remain to be explored. For example, while the thesis work concerns general prognosis for all breast cancer patients, there exist molecular subtypes of breast cancer based on specific genomic defects for which distinct prognostic tests or treatment strategies apply, and computational approaches such as unsupervised clustering or factor analysis [Hastie 2009] can be used to stratify patients based on their genomic data beforehand, which would bring us one step further towards less costly and more effective personalized prognosis, personalized medicine and personalized treatment. Notably, all prognostic signatures elaborated in Section 1.2 only apply to patients under specific clinico-pathological conditions. As another example, while the thesis work deals primarily with gene expression data, many other types of genomic data are available for analysis, among which DNA copy number variation (CNV) in array comparative genomic hybridization (aCGH) analysis and single-nucleotide polymorphism (SNP) in genome-wide association study (GWAS) are particularly widespread in active research in cancer biology, along with many other types of "omics" data including, to name just a few, epigenomics, proteomics, transcriptomics, metabolomics and microbiomics, and standard clinico-pathological parameters which incontrovertibly still dominate clinical practice of breast cancer prognosis until today. To make use of multi-omics data in an integrative analysis, multi-view learning [Sun 2013] is such a branch in machine learning that studies how to combine different and heterogeneous views of a sample. In particular, within the paradigm of kernel learning, if a kernel is defined for each view of the data, multiple kernel learning [Gönen 2011] has been shown relevant for genomic data fusion [Lanckriet 2004b].

One thing that needs to be explicitly pointed out is that the contribution of the thesis work to genomic data analysis for breast cancer prognosis has been mainly theoretical and methodological with efforts to propose new ideas of improving molecular prognosis and designing new forms of biomarkers, but the numerical results are far from reaching clinical significance. In particular, we do not claim to have identified any multigene signature for breast cancer prognosis, and we treat the lists of prognostic genes identified from our studies with certain extent of skepticism. A somehow discouraging fact in the field of computational cancer research is that a voluminous literature of more than 150,000 papers documenting thousands of claimed biomarkers has been produced in medicine of which fewer than 100 have been validated for routine clinical practice [Poste 2011], and even fewer than 20 are recognized with variable levels of evidence in the 2014 European Society of Medical Oncology (ESMO) clinical practice guidelines for lung, breast, colon and prostate cancer [Schneider 2015]. Compared to the number of research findings in this area, the very few number of gene expression-based breast cancer prognostic signatures successfully implemented in clinical routines (Section 1.2) has inevitably raised controversies on the practical validity of molecular signatures. This is mainly because the vast majority of those findings lack a proper validation procedure, not to mention validation oriented for clinical implication, resulting in an exaggeration of trivial findings and their clinical utility so that a large number of claimed signatures could very likely fail to add significantly incremental values assisting prognosis assessment and therapeutic decision making in addition to the use of standard clinico-pathological parameters. Notwithstanding, the research community generally holds an optimistic prospective towards the future, as long as proper validation pipelines are taken systematically in all forthcoming research [Michiels 2016]. Since objectives oriented towards clinical utility were not at all accounted for in the first place and neither meta-analysis-based validation with multiple datasets nor cross-study validation was performed during the course of my doctoral studies, we cannot conclude with any convincing prognostic signatures. However, we will try our best to venture some caveats suggesting pitfalls in analyzing genomic data specifically for biomarker discovery in cancer prognosis, in line with many previous attempts [Ambroise 2002, Simon 2003, Issaq 2011, Weigelt 2012]:

Should We Engage in Biomarker Discovery to Improve Prognosis? Regarding prognosis improvement, a currently held belief rationalizing biomarker discovery by the research community is that a prognostic model based on only a few molecular features should better capture and explain the biological complexity related to cancer survival. However, an anti-intuitive yet intriguing phenomenon already observed by previous studies [Haury 2011] arises that inference based on a few selected ones compared to the use of all molecular features available underlying a much larger spectrum of genome often does *not* lead to drastic improvement and sometimes even lead to slight deterioration, as reported across all three studies in this thesis: in Section 2.6 it did not seem to be beneficial to perform feature se-

lection in the biomedical applications thereof when Support Vector Machines with the Kendall kernel is used to classify genomic profiles, in Section 3.3.3 the best-performing model with the highest accuracy of survival risk prediction was the simplest ridge regression in which no feature selection was carried out, in Section 4.3.1 the most frequently selected or best suited classifiers for all predictive profiles in the breast cancer prognosis classification task are Support Vector Machines with various kernels none of which involve feature selection. Fortunately, model performance did not degrade significantly when the initiative of performing feature selection is supplemented in the predictive models at least in the latter two cases. In particular, another disagreeably striking observation is that in the benchmark study in Section 3.3.3, although they indeed show superiority with respect to several evaluation criteria of feature selection efficacy such as stability, connectivity and interpretability, all tested network-guided feature selection methods performed worse in terms of survival risk prediction, albeit insignificantly, than the simple and network-free elastic net. In a nutshell, it is not supported by existing evidence that feature selection in genomic data analysis could guarantee to make prognosis more accurate, and this convention is merely an assumption still awaiting to be confirmed, which therefore should be taken cautiously.

Should We Trust the Biomarkers Discovered? The trust we should invest in the biological values of the biomarkers discovered from cancer research can be limited by many factors and thus their merits in the development of clinical prognostic assays (or to further suggest therapeutic targets) should be taken with extreme caution. One issue regarding insignificant improvement of prognosis accuracy due to feature selection has been discussed in the last paragraph. Other factors that can influence the validity of feature selection and should also be considered as indispensable requirements for the candidate molecular features to be considered biomarkers of interest include cross-study reproducibility and functional interpretability of the identified biomarkers. Unfortunately, these issues are indeed demanding challenges. It is rarely the case that two prognostic signatures identified with different analytical methods and/or based on different datasets have a significant overlap, for instance only three genes are in common in the now famous 70-gene signature of [van 't Veer 2002] versus the 76-gene signature of [Wang 2005a]. Even surprisingly, [Venet 2011] reported that most random gene expression signatures are significantly associated with breast cancer outcome, criticizing on a hypothesis that the performance of prognostic models using deliberately selected features can be within the range of likely values based on random selection of features. Several studies analyzed the difficulty in selecting robust signatures, and overall concluded that the lack of robustness is mainly due to the fact that many different sets of genes with little overlap can nonetheless collectively have similar predictive power and the situation should be expected to be ameliorated when in the future we can gather a much larger number of samples to draw conclusion on [Michiels 2005, Ein-Dor 2006, Haury 2011]. A major drawback is that nowadays numerous discoveries that are based on small

and unrepresentative datasets hardly sustain independent validation so that their clinical utility remains out of reach. In particular, the numerical results presented in this thesis requires cross-study validation too as already mentioned above, subject to many impending issues arising in cross-study validation such as test set bias that could affect reproducibility and needs meticulous attention as well [Patil 2015].

Some Last Words...

Our knowledge and understanding of cancer biology is still far incomplete but we are given the extraordinary opportunity in the era of "omics" revolution and data science to study cancer with machine learning. Just bear in mind that opportunities come with caveats that it necessarily calls for comprehensive study and proper validation as well as concerns such as clinical utility and cost-effectiveness of the computational findings on the road to breakthrough discoveries and success in the fight against cancer.

Appendix A

A Tractable Bound on Approximating Kemeny Aggregation

Publication and Dissemination: *The work in this chapter has been published as joint work with Anna Korba and Eric Sibony in [Jiao 2016b] and orally presented at ICML 2016 by Anna Korba.*

Abstract: *Due to its numerous applications, rank aggregation has become a problem of major interest across many fields of the computer science literature. In the vast majority of situations, Kemeny consensus is considered as the "golden" solution. It is however well known that its computation is NP-hard. Much contribution have thus been devoted to establishing various results to apprehend this complexity. In this chapter, we introduce a practical method to predict, given a dataset and a ranking typically output by some approximate procedure, how close this ranking is to the Kemeny consensus of the dataset. A major strength of the proposed method is its generality: it requires little assumption on the dataset nor the ranking. Furthermore, it relies on a geometric interpretation of Kemeny aggregation that we believe could paves way to other results beyond those presented in this chapter.*

Résumé : *En raison de ses nombreuses applications, l'agrégation de classements est devenue un problème d'intérêt majeur dans de nombreux domaines de la littérature en science informatique. Dans la grande majorité des situations, le consensus de Kemeny est considéré comme la solution «dorée». Il est cependant bien connu que son calcul est NP-difficile. Par conséquent, de nombreuses contributions ont été consacrées à l'établissement de divers résultats pour appréhender cette complexité. Dans ce chapitre, nous introduisons une méthode pratique pour prédire, compte tenu d'un ensemble de données et d'un classement généralement produit par une procédure approximative, quelle est la proximité de ce classement au consensus de Kemeny sur l'ensemble de données. Une force majeur de la méthode proposée est sa généralité : elle*

nécessite peu d'hypothèses sur l'ensemble de données, ni sur le classement. En outre, il repose sur une interprétation géométrique de l'agrégation de Kemeny que nous croyons pouvoir ouvrir la voie à d'autres résultats au-delà de ceux présentés dans ce chapitre.

A.1 Introduction

Given a collection of rankings on a set of alternatives, how to aggregate them into one ranking? This rank aggregation problem has gained a major interest across many research fields. Starting from elections in social choice theory [de Borda 1781, Condorcet 1785, Arrow 1950, Xia 2015], it has been applied to meta-search engines [Dwork 2001, Renda 2003, Desarkar 2016], competitions ranking [Davenport 2005, Deng 2014], analysis of biological data [Kolde 2012, Patel 2013] or natural language processing [Li 2014, Zamani 2014] among others.

Among many ways of formulating the problem of rank aggregation stands out the Kemeny aggregation [Kemeny 1959]. Defined as the problem of minimizing a cost function over the symmetric group (see Section A.2 for definition), its solutions, called Kemeny consensuses, have been shown to satisfy desirable properties from many points of view [Young 1978].

Computing a Kemeny consensus is however NP-hard, even for only four rankings [Bartholdi 1989, Cohen 1999, Dwork 2001]. This fact has motivated the research community to introduce many approximate procedures and to evaluate them on datasets (see, for instance, [Schalekamp 2009, Ali 2012]). It has also triggered a tremendous amount of work of obtaining theoretical guarantees on these procedures and more generally in understanding the complexity of Kemeny aggregation from various perspectives. Some have proven bounds on the approximation cost of the approximate procedures [Diaconis 1977, Coppersmith 2006, Van Zuylen 2007, Ailon 2008, Freund 2015], while some have established recovery properties [Saari 2000, Procaccia 2012]. Notably it has been shown that exact Kemeny aggregation is tractable if some quantity is known on the dataset [Betzler 2008, Betzler 2009, Cornaz 2013] or if the dataset satisfies some conditions [Brandt 2015]. Besides, some contributions have established approximation bounds that can be computed on the dataset [Davenport 2004, Conitzer 2006, Sibony 2014].

In this chapter we introduce a novel approach to apprehend the complexity of Kemeny aggregation. Consider the following question: *Given a dataset and a ranking typically output by some approximate procedure, can we predict how close the ranking is to any Kemeny consensus without computing the latter?* We exhibit a tractable quantity that allows to give a positive answer to this question. The main contribution of our results is a simple and practical method to obtain such a guarantee for the outcome of an aggregation procedure on any dataset. A major strength of our approach is its generality: it applies to all aggregation procedures and to any dataset. Further, our results are based on a geometric analysis of Kemeny aggregation (see Section A.3) that has been unprecedentedly exploited in

the literature but constitutes a powerful tool. We thus take efforts to explain it in details. We believe that it could pave way to many other results on Kemeny aggregation beyond those presented here.

The chapter is structured as follows. Section A.2 introduces the general notations and states the question of interest. The geometric analysis is detailed in Section A.3 and further studied in Section A.5 while our main result is presented in Section A.4. Finally, numerical experiments are reported in Section A.6 to address the efficacy and usefulness of our method on datasets from real-world applications.

A.2 Kemeny Aggregation Problem

Let $[\![n]\!] = \{1, \dots, n\}$ be a set of alternatives to be ranked. In this study, we focus only on total rankings. A total ranking $a_1 \succ \cdots \succ a_n$ on $[\![n]\!]$ is interchangeably seen as the permutation σ of $[\![n]\!]$ that maps an item to its rank: $\sigma(a_i) = i$ for $i \in [\![n]\!]$. The set of all permutations of $[\![n]\!]$ is called the symmetric group and denoted by $\mathbb{S}_n$. Given a collection of N permutations $\mathcal{D}_N = (\sigma_1, \dots, \sigma_N) \in \mathbb{S}_n^N$, Kemeny aggregation aims at solving

$$\min_{\sigma \in \mathbb{S}_n} \sum_{t=1}^{N} d(\sigma, \sigma_t)\,, \tag{A.1}$$

where d is the Kendall's tau distance defined for $\sigma, \sigma' \in \mathbb{S}_n$ as the number of their pairwise disagreements: $d(\sigma, \sigma') = \sum_{1 \le i < j \le n} \mathbb{1}\{(\sigma(j) - \sigma(i))(\sigma'(j) - \sigma'(i)) < 0\}$. The objective function in (A.1) denotes the cost, and a permutation σ^* solving (A.1) is called a Kemeny consensus. We denote by $\mathcal{K}_N$ the set of Kemeny consensuses on the dataset $\mathcal{D}_N$.

Exact Kemeny aggregation is NP-hard: it cannot be solved efficiently with a general procedure. This does not mean however that nothing can be done. For example, it is clear that on a dataset where all permutations are equal to a $\sigma_0 \in \mathbb{S}_n$, the Kemeny consensus is trivially given by σ_0. Many contributions from the literature have thus focused on particular approaches to apprehending some part of the complexity of Kemeny aggregation. The examples given in the introduction generally fall in three categories:

- **General guarantees for approximate procedures.** These results provide a bound on the cost of a specific voting rule, valid for any dataset [Diaconis 1977, Coppersmith 2006, Van Zuylen 2007, Ailon 2008, Freund 2015].
- **Bounds on the approximation cost computed from the dataset.** These results provide a bound, either on the cost of a consensus or on the cost of the outcome of a specific voting rule, that depends on a quantity computed from the dataset [Davenport 2004, Conitzer 2006, Sibony 2014].
- **Conditions for the exact Kemeny aggregation to become tractable.** These results ensure the tractability of exact computation of Kemeny aggre-

gation if the dataset satisfies some condition or if some quantity is known from the dataset [Betzler 2008, Betzler 2009, Cornaz 2013, Brandt 2015].

In this chapter, we introduce a novel approach, which falls into the second category above, to apprehend the complexity of Kemeny aggregation by considering the following question (henceforth referred to as The Question):

The Question. Let $\sigma \in \mathbb{S}_n$ be a permutation, typically output by a computationally efficient aggregation procedure on $\mathcal{D}_N$. Can we use tractable quantities to give an upper bound on the distance $d(\sigma, \sigma^*)$ between σ and any Kemeny consensus σ^* on $\mathcal{D}_N$?

The answer to The Question is positive as we will elaborate in this study. We propose an upper bound, guaranteed by Theorem A.1, that reads: for any σ and $\mathcal{D}_N$, we have $d(\sigma, \sigma^*) \leq k_{\min}$ for all consensuses $\sigma^* \in \mathcal{K}_N$, where $k_{\min}$ is defined in (A.5). We would like to stress a major strength of our method compared to those previously studied in literature that is the generality: it can be applied to any dataset $\mathcal{D}_N$ and any permutation σ with little assumption on neither. This is because it exploits a powerful geometric framework for the analysis of Kemeny aggregation.

A.3 Geometric Analysis of Kemeny Aggregation

Because of its rich mathematical structure, Kemeny aggregation can be analyzed from many different point of views. For instance, some contributions deal directly with the combinatorics of the symmetric group [Diaconis 1977, Blin 2011], some work on the pairwise comparison graph [Coppersmith 2006, Conitzer 2006, Jiang 2011], and some exploit the geometry of the Permutahedron [Saari 2000]. In this chapter, we analyze it via the Kendall embedding [Jiao 2015, Theorem 1]. For the self-containment of this chapter, recall from the proof of Theorem 2.1 that we have used the following definition.[1]

Definition A.1 (Kendall embedding)**.** The Kendall embedding is the mapping $\phi : \mathbb{S}_n \to \mathbb{R}^{\binom{n}{2}}$ defined by

$$\phi : \sigma \mapsto \begin{pmatrix} \vdots \\ \operatorname{sgn}(\sigma(i) - \sigma(j)) \\ \vdots \end{pmatrix}_{1 \leq i < j \leq n},$$

where the sign function $\operatorname{sgn}(x) = 1$ if $x > 0$ and -1 if $x < 0$ and 0 otherwise.

The Kendall embedding ϕ maps a permutation to a vector in $\mathbb{R}^{\binom{n}{2}}$ where each coordinate is indexed by an (unordered) pair $\{i, j\} \subset [\![n]\!]$ (we choose $i < j$ by

[1] We will omit the scaling constant $\frac{1}{\binom{n}{2}}$ due to its irrelevance in this study and use ϕ instead of Φ to distinguish this trivial difference.

convention). Though this vector representation is equivalent to representing a permutation as a flow on the complete graph on $[\![n]\!]$, it allows us to perform a geometric analysis of Kemeny aggregation in the Euclidean space $\mathbb{R}^{\binom{n}{2}}$. Denoting by $\langle \cdot, \cdot \rangle$ the canonical inner product and $\|\cdot\|$ the Euclidean norm, the starting point of our analysis is the following result, already brought up by [Barthelemy 1981] and rephrased in the proof of Theorem 2.1.

Proposition A.1 (Background results)**.** *For all $\sigma, \sigma' \in \mathbb{S}_n$,*

$$\|\phi(\sigma)\| = \sqrt{\frac{n(n-1)}{2}} := R \text{ and } \|\phi(\sigma) - \phi(\sigma')\|^2 = 4d(\sigma, \sigma'),$$

and for any dataset $\mathcal{D}_N = (\sigma_1, \dots \sigma_N) \in \mathbb{S}_n^N$, Kemeny aggregation (A.1) is equivalent to the minimization problem

$$\min_{\sigma \in \mathbb{S}_n} C_N(\sigma) := \|\phi(\sigma) - \phi(\mathcal{D}_N)\|^2, \tag{A.2}$$

where

$$\phi(\mathcal{D}_N) := \frac{1}{N} \sum_{t=1}^{N} \phi(\sigma_t). \tag{A.3}$$

Proposition A.1 leads to the following geometric point of view of Kemeny aggregation, illustrated by Figure A.1. First, as $\|\phi(\sigma)\| = \sqrt{n(n-1)/2}$ for all $\sigma \in \mathbb{S}_n$, the embeddings of all the permutations in $\mathbb{S}_n$ lie on a sphere centered at 0 with radius $R = \sqrt{n(n-1)/2}$. Notice that $\|\phi(\sigma) - \phi(\sigma')\|^2 = 4d(\sigma, \sigma')$ for all $\sigma, \sigma' \in \mathbb{S}_n$ implies that ϕ is injective, in other words that it maps two different permutations to two different points on the sphere. A dataset $\mathcal{D}_N = (\sigma_1, \dots, \sigma_N) \in \mathbb{S}_n^N$ is thus mapped to a weighted point cloud on this sphere, where for any $\sigma \in \mathbb{S}_n$, the weight of $\phi(\sigma)$ is the number of times σ appears in $\mathcal{D}_N$. The vector $\phi(\mathcal{D}_N)$, defined by (A.3), is then equal to the barycenter of this weighted point cloud. We call it the *mean embedding* of $\mathcal{D}_N$. Now, the reformulation of Kemeny aggregation given by (A.2) means that a Kemeny consensus is a permutation σ^* whose embedding $\phi(\sigma^*)$ is closest to $\phi(\mathcal{D}_N)$, *with respect* to the Euclidean norm in $\mathbb{R}^{\binom{n}{2}}$.

From an algorithmic point of view, Proposition A.1 naturally decomposes problem (A.1) of Kemeny aggregation in two steps: first compute the mean embedding $\phi(\mathcal{D}_N)$ in the space $\mathbb{R}^{\binom{n}{2}}$, and then find a consensus σ^* as a solution of problem (A.2). The first step is naturally performed in $O(Nn^2)$ operations. The NP-hardness of Kemeny aggregation thus stems from the second step. In this regard, one may argue that having $\phi(\mathcal{D}_N)$ does not greatly alleviate the complexity in identifying an exact Kemeny consensus. However, a closer look at the problem leads us to asserting that $\phi(\mathcal{D}_N)$ contains rich information about the location of the Kemeny consensuses. More specifically, we show in Theorem A.1 that the knowledge of $\phi(\mathcal{D}_N)$ helps to provide an upper bound for the distance between a given permutation $\sigma \in \mathbb{S}_n$ and any Kemeny consensus σ^*.

In fact, Proposition A.1 implies that Kemeny's rule is a "Mean Proximity Rule", a family of voting rules introduced in [Zwicker 2008] and further studied

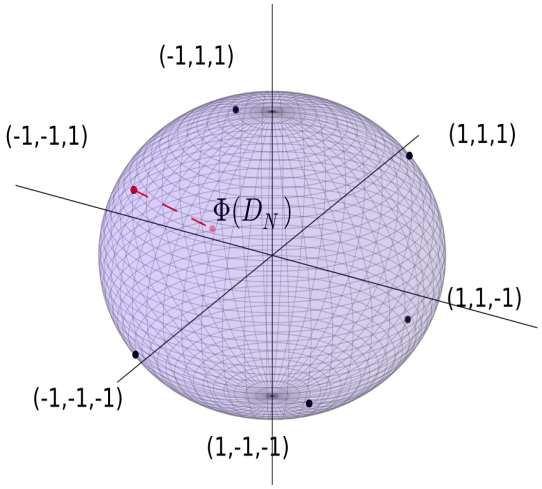


Figure A.1: Kemeny aggregation for $n = 3$.

in [Lahaie 2014]. Our approach actually applies more generally to other voting rules from this class but we focus our discussion on Kemeny's rule in this study for the sake of clarity.

A.4 Controlling the Distance to Kemeny Consensus

In this section, we now state our main results and demonstrate with an illustrative example how our proposed method addresses The Question. For a permutation $\sigma \in \mathbb{S}_n$, let us define the angle $\theta_N(\sigma)$ between $\phi(\sigma)$ and $\phi(\mathcal{D}_N)$ by

$$\cos(\theta_N(\sigma)) = \frac{\langle \phi(\sigma), \phi(\mathcal{D}_N) \rangle}{\|\phi(\sigma)\| \, \|\phi(\mathcal{D}_N)\|}, \tag{A.4}$$

with $0 \leq \theta_N(\sigma) \leq \pi$ by convention.

Theorem A.1. *Let $\mathcal{D}_N \in \mathbb{S}_n^N$ be a dataset and $\sigma \in \mathbb{S}_n$ a permutation. For any integer $0 \leq k \leq \binom{n}{2} - 1$, one has the following implication:*

$$\cos(\theta_N(\sigma)) > \sqrt{1 - \frac{k+1}{\binom{n}{2}}} \quad \Longrightarrow \quad \max_{\sigma^* \in \mathcal{K}_N} d(\sigma, \sigma^*) \leq k.$$

The proof of Theorem A.1 along with its geometric interpretation are postponed to Section A.5. Broadly speaking, Theorem A.1 ensures that if the angle $\theta_N(\sigma)$ between the embedding $\phi(\sigma)$ of a permutation $\sigma \in \mathbb{S}_n$ and the mean embedding $\phi(\mathcal{D}_N)$ is small, then the Kemeny consensuses cannot be too far from σ. Its application in practice is straightforward. Assume that one applies an aggregation procedure on $\mathcal{D}_N$, say the Borda Count for instance, that outputs σ. A natural question is then: how far is it from the Kemeny consensuses in terms of

Kendall's tau distance? It is well known that the Kendall's tau distance takes values in $\{0, \dots, \binom{n}{2}\}$ [Stanley 1986]. Consequently, the distance is at most equal to $\max_{\sigma', \sigma'' \in \mathbb{S}_n} d(\sigma', \sigma'') = \binom{n}{2}$. But if one computes the quantity $\cos(\theta_N(\sigma))$, a better bound can be allowed due to Theorem A.1. More specifically, the best bound is given by the minimal $k \in \{0, \dots, \binom{n}{2} - 1\}$ such that $\cos(\theta_N(\sigma)) > \sqrt{1 - (k+1)/\binom{n}{2}}$. Denoting by $k_{\min}(\sigma; \mathcal{D}_N)$ this integer, it is easy to see that

$$k_{\min}(\sigma; \mathcal{D}_N) = \begin{cases} \left\lfloor \binom{n}{2} \sin^2(\theta_N(\sigma)) \right\rfloor & \text{if } 0 \leq \theta_N(\sigma) \leq \frac{\pi}{2} \\ \binom{n}{2} & \text{if } \frac{\pi}{2} \leq \theta_N(\sigma) \leq \pi. \end{cases} \tag{A.5}$$

where $\lfloor x \rfloor$ denotes the integer part of the real x. We formalize this method (henceforth referred to as The Method) in the following description.

The Method. Let $\mathcal{D}_N \in \mathbb{S}_n^N$ be a dataset and let $\sigma \in \mathbb{S}_n$ be a permutation considered as an approximation of Kemeny's rule. In practice σ is the consensus returned by a tractable voting rule. Then by Theorem A.1, we have $d(\sigma, \sigma^*) \leq k_{\min}(\sigma; \mathcal{D}_N)$ for any Kemeny consensus $\sigma^* \in \mathcal{K}_N$, where $k_{\min}(\sigma; \mathcal{D}_N)$ is obtained by (A.5).

The following proposition ensures that The Method has tractable complexity.

Proposition A.2 (Complexity of The Method)**.** *The application of The Method has complexity in time $O(Nn^2)$.*

With a concrete example, we demonstrate the applicability and the generality of The Method.

Example A.1 (Application of The Method to the sushi dataset)*.* We report here the results of a case-study on the sushi dataset provided by [Kamishima 2003] to illustrate our method. The dataset consists of $N = 5000$ total rankings given by different individuals of the preference order on $n = 10$ sushi dishes such that a brute-force search for the Kemeny consensus is already very computationally intensive. To apply our method, we selected seven tractable voting rules, denoted by σ, as approximate candidates to Kemeny's rule to provide an initial guess (details of voting rules can be found in Section A.6). Table A.1 summarizes the values of $\cos(\theta_N(\sigma))$ and $k_{\min}(\sigma)$, respectively given by (A.4) and (A.5). Results show that on this particular dataset, if we use for instance Borda Count to approximate a Kemeny consensus, we are confident that any exact consensus has a distance of at most 14 to the approximate ranking. We detail empirical analysis of the results further in Section A.6.

A.5 Geometric Interpretation Revisit and Proof of Theorem A.1

This section details the proof of Theorem A.1 based the geometric interpretation introduced in Section A.3. We deem that our proof has indeed a standalone interest, and that it could pave way to other profound results on Kemeny aggregation.

Table A.1: Summary of a case-study on the applicability of The Method with the sushi dataset ($N = 5000, n = 10$). Rows are ordered by increasing $k_{\min}$ (or decreasing cosine) value.

Voting rule	$\cos(\theta_N(\sigma))$	$k_{\min}(\sigma)$
Borda	0.820	14
Copeland	0.822	14
QuickSort	0.822	14
Plackett-Luce	0.80	15
2-approval	0.745	20
1-approval	0.710	22
Pick-a-Perm	$0.383^{\dagger}$	$34.85^{\dagger}$
Pick-a-Random	$0.377^{\dagger}$	$35.09^{\dagger}$

†For randomized methods such as Pick-a-Perm and Pick-a-Random, results are averaged over 10,000 computations.

Recall that the Kemeny consensuses of a dataset $\mathcal{D}_N$ are the solutions of the problem (A.2):

$$\min_{\sigma \in \mathbb{S}_n} C_N(\sigma) = \|\phi(\sigma) - \phi(\mathcal{D}_N)\|^2 .$$

This is an optimization problem on the discrete set $\mathbb{S}_n$, naturally hard to analyze. In particular the shape of the cost function C_N is not easy to understand. However, since all the vectors $\phi(\sigma)$ for $\sigma \in \mathbb{S}_n$ lie on the sphere

$$\mathbb{S} := \left\{ x \in \mathbb{R}^{\binom{n}{2}} | \, \|x\| = R \right\} ,$$

where radius R is the equal norm of the embedded point of any permutation and by Proposition A.1,

$$R = \sqrt{\frac{n(n-1)}{2}} .$$

It is natural to consider the relaxed problem on $\mathbb{S}$ that reads

$$\min_{x \in \mathbb{S}} \mathcal{C}_N(x) := \|x - \phi(\mathcal{D}_N)\|^2 .$$

We call $\mathcal{C}_N$ the extended cost function with domain $\mathbb{S}$. The advantage of $\mathcal{C}_N$ is that it has a very simple shape. We denote by $\theta_N(x)$ the angle between a vector $x \in \mathbb{S}$ and $\phi(\mathcal{D}_N)$ (with the slight abuse of notations that $\theta_N(\phi(\sigma)) \equiv \theta_N(\sigma)$). For any $x \in \mathbb{S}$, one has

$$\mathcal{C}_N(x) = R^2 + \|\phi(\mathcal{D}_N)\|^2 - 2R \, \|\phi(\mathcal{D}_N)\| \cos(\theta_N(x)) .$$

This means that the extended cost $\mathcal{C}_N(x)$ of a vector $x \in \mathbb{S}$ only depends on the angle $\theta_N(x)$. The level sets of $\mathcal{C}_N$ are thus of the form $\{x \in \mathbb{S} \mid \theta_N(x) = \alpha\}$, for $0 \leq \alpha \leq \pi$. If $n = 3$, these level sets are circles in planes orthogonal to $\phi(\mathcal{D}_N)$, each centered around the projection of the latter on the plane (Figure A.2). This property implies the following result.

Lemma A.1. *A Kemeny consensus of a dataset $\mathcal{D}_N$ is a permutation σ^* such that:*

$$\theta_N(\sigma^*) \leq \theta_N(\sigma) \qquad \text{for all } \sigma \in \mathbb{S}_n\,.$$

Lemma A.1 means that the problem of Kemeny aggregation translates into finding permutations σ^* that have minimal angle $\theta_N(\sigma^*)$. This reformulation is crucial to our approach.

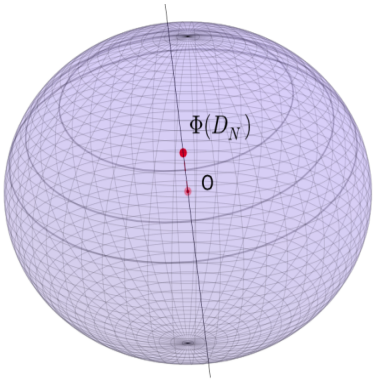


Figure A.2: Level sets of the extended cost function $\mathcal{C}_N$ over $\mathbb{S}$ for $n = 3$.

A.5.1 Interpretation of the Condition in Theorem A.1

The second element of our approach is motivated by the following observation. Let $x \in \mathbb{S}$ be a point on the sphere and let $r \geq 0$. If r is large enough, then all the points $x' \in \mathbb{S}$ on the sphere that have distance $\|x' - x\|$ greater than r will have a greater angle $\theta_N(x')$. Formally, we denote by $\mathcal{B}(x, r) = \{x' \in \mathbb{R}^{\binom{n}{2}} \mid \|x' - x\| < r\}$ the (open) ball of center x and radius r. Then one has the following result.

Lemma A.2. *For $x \in \mathbb{S}$ and $r \geq 0$, one has the following implication:*

$$\cos(\theta_N(x)) > \sqrt{1 - \frac{r^2}{4R^2}} \implies \min_{x' \in \mathbb{S} \setminus \mathcal{B}(x,r)} \theta_N(x') > \theta_N(x)\,.$$

Proof. Let $\bar{\phi}(\mathcal{D}_N) = \frac{\phi(\mathcal{D}_N)}{\|\phi(\mathcal{D}_N)\|}$. We discuss over two cases.
Case I: $\left\|\bar{\phi}(\mathcal{D}_N) - x\right\| \geq r$. By laws of cosines, this case is equivalent to:

$$2R^2(1 - \cos(\theta_N(x))) = \left\|\bar{\phi}(\mathcal{D}_N) - x\right\|^2 \geq r^2$$
$$\iff \cos(\theta_N(x)) \leq 1 - \frac{r^2}{2R^2} \leq 1 - \frac{r^2}{4R^2}\,.$$

Note also that in this case, we have $\bar{\phi}(\mathcal{D}_N) \in \mathbb{S} \setminus \mathcal{B}(x, r)$ and hence $\min_{x' \in \mathbb{S} \setminus \mathcal{B}(x,r)} \theta_N(x') = \min_{x' \in \mathbb{S}} \theta_N(x') = 0 \leq \theta_N(x)$ always holds, where the minimum is attained at $x' = \bar{\phi}(\mathcal{D}_N)$.

Case II: $\|\bar{\phi}(\mathcal{D}_N) - x\| < r$, that is $\bar{\phi}(\mathcal{D}_N) \in \mathcal{B}(x, r)$. As the function $x' \mapsto \theta_N(x')$ is convex with global minimum in $\mathcal{B}(x, r)$, its minimum over $\mathbb{S} \setminus \mathcal{B}(x, r)$ is attained at the boundary $\mathbb{S} \cap \partial\mathcal{B}(x, r) = \{x' \in \mathbb{R}^{\binom{n}{2}} \mid \|x'\| = R \text{ and } \|x' - x\| = r\}$, which is formed by cutting $\mathbb{S}$ with the $(\binom{n}{2} - 1)$-dimensional hyperplane written as

$$\mathbb{L} := \left\{ x' \in \mathbb{R}^{\binom{n}{2}} \;\middle|\; \langle x', x \rangle = \frac{2R^2 - r^2}{2} \right\}.$$

Straightforwardly one can verify that $\mathbb{S} \cap \partial\mathcal{B}(x, r)$ is in fact a $(\binom{n}{2} - 1)$-dimensional sphere lying in $\mathbb{L}$, centered at $c = \frac{2R^2 - r^2}{2R^2} x$ with radius $\gamma = r\sqrt{1 - \frac{r^2}{4R^2}}$. Now we take effort to identify:

$$x^* = \operatorname*{arg\,min}_{x' \in \mathbb{S} \cap \partial\mathcal{B}(x,r)} \theta_N(x') = \operatorname*{arg\,min}_{x' \in \mathbb{S} \cap \partial\mathcal{B}(x,r)} \mathcal{C}_N(x').$$

Note that $\phi(\mathcal{D}_N)$ projected onto $\mathbb{L}$ is the vector $(\phi(\mathcal{D}_N))_{\mathbb{L}} := \phi(\mathcal{D}_N) - \frac{\langle \phi(\mathcal{D}_N), x \rangle}{R^2} x$. One can easily verify by Pythagoras rule that, for any set $\mathbb{K} \subseteq \mathbb{L}$,

$$\operatorname*{arg\,min}_{x' \in \mathbb{K}} \|x' - \phi(\mathcal{D}_N)\| = \operatorname*{arg\,min}_{x' \in \mathbb{K}} \|x' - (\phi(\mathcal{D}_N))_{\mathbb{L}}\| .$$

Therefore we have:

$$\begin{aligned} x^* = \operatorname*{arg\,min}_{x' \in \mathbb{S} \cap \partial\mathcal{B}(x,r)} \|x' - (\phi(\mathcal{D}_N))_{\mathbb{L}}\| &= c + \gamma \frac{(\phi(\mathcal{D}_N))_{\mathbb{L}}}{\|(\phi(\mathcal{D}_N))_{\mathbb{L}}\|} \\ &= \frac{2R^2 - r^2}{2R^2} x + r\sqrt{1 - \frac{r^2}{4R^2}} \frac{\phi(\mathcal{D}_N) - \frac{\langle \phi(\mathcal{D}_N), x \rangle}{R^2} x}{\sqrt{\|\phi(\mathcal{D}_N)\|^2 - \frac{\langle \phi(\mathcal{D}_N), x \rangle^2}{R^2}}} . \end{aligned}$$

Tedious but undemanding calculation leads to

$$\theta_N(x^*) > \theta_N(x) \iff \langle x^*, \phi(\mathcal{D}_N) \rangle > \langle x, \phi(\mathcal{D}_N) \rangle \iff \cos(\theta_N(x)) > \sqrt{1 - \frac{r^2}{4R^2}} .$$

□

It is interesting to look at the geometric interpretation of Lemma A.2. In fact, it is clear from the proof that x^* should lie in the 2-dimensional subspace spanned by $\phi(\mathcal{D}_N)$ and x. We are thus able to properly define multiples of an angle by summation of angles on such linear space $2\theta_N(x) := \theta_N(x) + \theta_N(x)$. Figure A.3 provides an illustration of Lemma A.2 in this 2-dimensional subspace from the geometric point of view. In words, provided that $\theta_N(x) \leq \pi/2$, x^* has a smaller angle than x is equivalently written using laws of cosines as

$$\begin{aligned} r^2 = \|x - x^*\|^2 > 2R^2(1 - \cos(2\theta_N(x))) \\ \iff \cos(2\theta_N(x)) > 1 - \frac{r^2}{2R^2} \iff \cos(\theta_N(x)) > \sqrt{1 - \frac{r^2}{4R^2}} . \end{aligned}$$

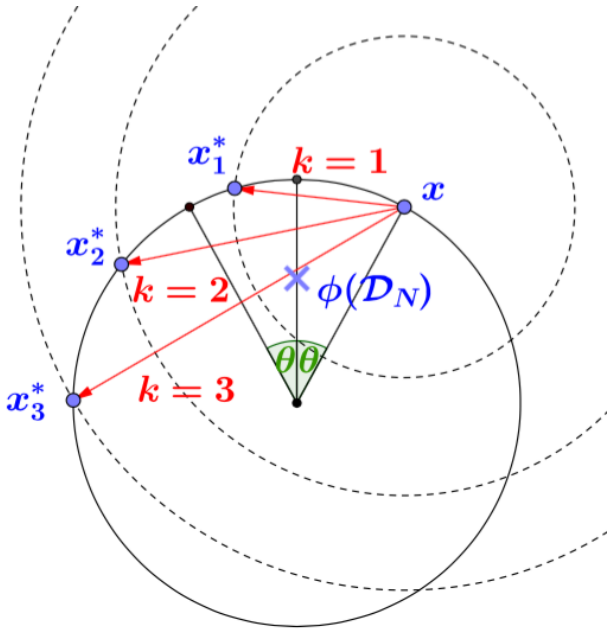


Figure A.3: Geometric illustration of the bound in Lemma A.2 with $x = \phi(\sigma)$ and $k = \frac{r^2}{4}$ taking integer values (representing possible Kendall's tau distance). The smallest integer value for k such that these inequalities hold is $k = 2$.

This recovers exactly the condition stated in Lemma A.2.

A final lemma necessary for the proof of Theorem A.1 is on the embedding of a ball in the Euclidean space. For $\sigma \in \mathbb{S}_n$ and $k \in \{0, \dots, \binom{n}{2}\}$, we denote by $B(\sigma, k)$ the (closed) ball for the Kendall's tau distance of center σ and radius k, i.e. $B(\sigma, k) = \{\sigma' \in \mathbb{S}_n \mid d(\sigma, \sigma') \leq k\}$. The following is a direct consequence of Proposition A.1.

Lemma A.3. *For $\sigma \in \mathbb{S}_n$ and $k \in \{0, \dots, \binom{n}{2}\}$,*

$$\phi\left(\mathbb{S}_n \setminus B(\sigma, k)\right) \subset \mathbb{S} \setminus \mathcal{B}(\phi(\sigma), 2\sqrt{k+1})\,.$$

A.5.2 Proof of Theorem A.1

We can now prove Theorem A.1 by combining the previous results and observations.

Proof of Theorem A.1. Let $\mathcal{D}_N \in \mathbb{S}_n^N$ be a dataset and $\sigma \in \mathbb{S}_n$ a permutation. By Lemma A.2, one has for any $r > 0$,

$$\cos(\theta_N(\sigma)) > \sqrt{1 - \frac{r^2}{4R^2}} \implies \min_{x \in \mathbb{S} \setminus B(\phi(\sigma), r)} \theta_N(x) > \theta_N(\sigma)\,.$$

We take $r = 2\sqrt{k+1}$. The left-hand term becomes $\cos(\theta_N(\sigma)) > \sqrt{1 - \frac{k+1}{R^2}}$, which is the condition in Theorem A.1. The right-hand term becomes:

$$\min_{x \in \mathbb{S} \setminus B(\phi(\sigma), 2\sqrt{k+1})} \theta_N(x) > \theta_N(\sigma)\,,$$

which implies by Lemma A.3 that

$$\min_{\sigma' \in \mathbb{S}_n \setminus B(\sigma,k)} \theta_N(\sigma') > \theta_N(\sigma)\,.$$

This means that for all $\sigma' \in \mathbb{S}_n$ with $d(\sigma, \sigma') > k$, $\theta_N(\sigma') > \theta_N(\sigma)$. Now, by Lemma A.1, any Kemeny consensus σ^* necessarily satisfies $\theta_N(\sigma^*) \leq \theta_N(\sigma)$. One therefore has $d(\sigma, \sigma^*) \leq k$, and the proof is concluded. □

A.6 Numerical Experiments

In this section we study the tightness of the bound in Theorem A.1 and the applicability of The Method through numerical experiments. We first elaborate in detail the voting rules used in the chapter to approximate Kemeny's rule. Note that if multiple consensuses are returned from a rule on a given dataset, we always randomly pick one from these consensuses.

- **Positional scoring rules**. Given a scoring vector $w = (w_1, ..., w_n) \in \mathbb{R}^n$ of weights respectively for each alternative in $[\![n]\!]$, the ith alternative in a vote scores w_i. A total ranking is given by sorting the averaged scores over all votes, for example, the winner is the alternative with highest total score over all the votes. The **plurality** rule has the weight vector $(1, 0, ..., 0)$, the k**-approval** rule has $(1, ..., 1, 0..., 0)$ containing 1s in the first k positions, and the **Borda** rule [de Borda 1781] has $(n, n-1, ..., 1)$.

- **Copeland** [Copeland 1951]. A total ranking is given by sorting the Copeland scores averaged over all votes, for which the score of alternative i is the number of pairwise beats, or $\#\{j \neq i : i \text{ beats } j\}$. For example, the Copeland winner is the alternative that wins the most pairwise elections.

- **QuickSort** [Ali 2012]. QuickSort recursively divides an unsorted list into two lists – one list comprising alternatives that occur before a chosen index (called the *pivot*), and another comprising alternatives that occur after, and then sorts each of the two lists. The pivot is always chosen as the first alternative.

- **Pick-a-Perm** [Ali 2012]. A total ranking is picked randomly from $\mathbb{S}_n$ according to the empirical distribution of the dataset $\mathcal{D}_N$.

- **Plackett-Luce**. A Plackett-Luce ranking model defined for any $\sigma \in \mathbb{S}_n$ by $p_w(\sigma) = \prod_{i=1}^n w_{\sigma(i)} / \left(\sum_{j=i}^n w_{\sigma(j)}\right)$ parameterized by $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$, fitted to $\mathcal{D}_N$ by means of the MM algorithm [Hunter 2004]. A total ranking is then given by sorting w.

- **Pick-a-Random**. A total ranking is picked randomly from $\mathbb{S}_n$ according to uniform law (independent from $\mathcal{D}_N$).

Notably, Pick-a-Random is expected as a negative control experiment. To intuitively understand the rationale behind Pick-a-Random, let us consider the case conditioned on that the output of a voting rule has (at least) certain Kendall's tau distance to the Kemeny consensus. Compared to what Pick-a-Random would blindly pick any permutation without accessing to the dataset $\mathcal{D}_N$ at all, a sensible voting rule should have a better chance to output one permutation with a smaller angle θ with $\phi(\mathcal{D}_N)$ among all the permutations that share the same distance to Kemeny consensus. As we have reasoned in the geometric proof of The Method that the smaller the angle θ is, the more applicable our method will be, Pick-a-Random is expected to perform worse than other voting rules in terms of applicability of our method.

A.6.1 Tightness of the Bound

Recall that we denote by n the number of alternatives, by $\mathcal{D}_N \in \mathbb{S}_n^N$ any dataset, by r any voting rule, and by $r(\mathcal{D}_N)$ a consensus of $\mathcal{D}_N$ given by r. For ease of notation convenience, we assume that $\mathcal{K}_N$ contains a single consensus (otherwise we pick one randomly as we do in all experiments). The approximation efficiency of r to Kemeny's rule is exactly measured by $d(r(\mathcal{D}_N), \mathcal{K}_N)$. Applying our method with $r(\mathcal{D}_N)$ would return an upper bound for $d(r(\mathcal{D}_N), \mathcal{K}_N)$, that is:

$$d(r(\mathcal{D}_N), \mathcal{K}_N) \leq k_{\min} .$$

Notably here we are not interested in studying the approximation efficiency of a particular voting rule, but we are rather interested in studying the approximation efficiency specific to our method indicated by the tightness of the bound, i.e.,

$$s\left(r, \mathcal{D}_N, n\right) := k_{\min} - d(r(\mathcal{D}_N), \mathcal{K}_N) .$$

In other words, $s\left(r, \mathcal{D}_N, n\right)$ quantifies how confident we are when we use $k_{\min}$ to "approximate" the approximation efficiency $d(r(\mathcal{D}_N), \mathcal{K}_N)$ of r to Kemeny's rule on a given dataset $\mathcal{D}_N$. The smaller $s\left(r, \mathcal{D}_N, n\right)$ is, the better our method works when it is combined with the voting rule r to pinpoint the Kemeny consensus on a given dataset $\mathcal{D}_N$. Note that our notation stresses on the fact that s depends typically on $(r, \mathcal{D}_N, n)$.

We empirically investigate the efficiency of our proposed method by experimenting $s\left(r, \mathcal{D}_N, n\right)$ with various voting rules r, on different datasets $\mathcal{D}_N$, implicitly involving n as well. For that purpose, in each experiment we test six prevalent voting rules plus one negative-control method as approximate candidates to Kemeny's rule: three scoring rules that are Borda Count, k-approval, Copeland; two algorithmic approaches that are QuickSort and Pick-a-Perm; one statistical approach based on Plackett-Luce ranking model; one baseline method serving a negative control that is Pick-a-Random.

We first look at the the effect of different voting rules r on $s\left(r; \mathcal{D}_N, n\right)$ with the APA dataset. In the 1980 American Psychological Association (APA) presidential

election, voters were asked to rank $n = 5$ candidates in order of preference and a total of $N = 5738$ complete ballots were reported. With the original collection of ballots introduced by [Diaconis 1989], We created 500 bootstrapped pseudo-samples following [Popova 2012]. As shown in Figure A.4, $s\,(r; \mathcal{D}_N, n)$ varies across different voting rules and our method works typically well combined with Borda Count or Plackett-Luce, a phenomenon that constantly occurs in many experiments. For example for Borda Count the median tightness being 3 means that our method provides a bound that tolerates an approximation within a Kendall's tau distance up to 3. We also observe that on the contrary, the boxplot of Pick-a-Random always shows a wider range and larger median as expected.

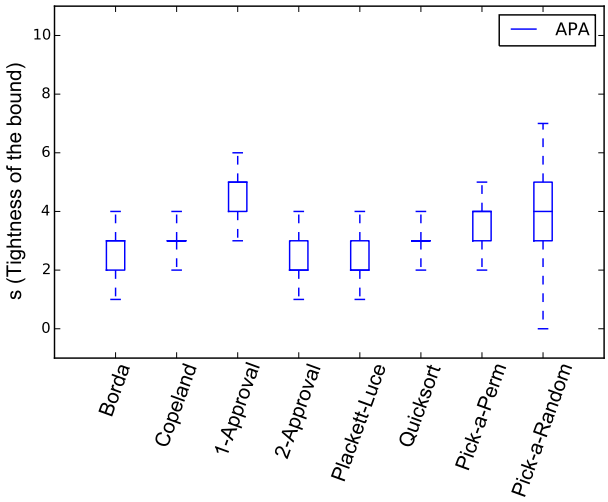


Figure A.4: Boxplots of $s\,(r, \mathcal{D}_N, n)$ over sampling collections of datasets shows the effect from different voting rules r with 500 bootstrapped pseudo-samples of the APA dataset $(n = 5, N = 5738)$.

The effect of datasets $\mathcal{D}_N$ on the measure $s\,(\mathcal{D}_N; r, n)$ is tested with the Netflix data provided by [Mattei 2012]. We set $n = 3$ the number of ranked alternatives and take two types of data with distinct characteristics to contrast their impact: we took the 100 datasets with a Condorcet winner and randomly selected 100 datasets from those with no Condorcet winner. The rationale for this experiment is that Kemeny's rule is a Condorcet method, i.e., Kemeny's rule always yields a Condorcet winner if it exists. Therefore we suppose that the efficiency of our method should also depend on this particular social characteristic present in data. As expected, it is interesting to note the clear difference shown by the two types of data shown by Figure A.5. In words, our method is more efficient in case that a Condorcet winner is present in the dataset than the other case that a Condorcet winner is absent in the sense that s is generally smaller in the former case.

We finally study how the $s\,(n; r, \mathcal{D}_N)$ grows with the size of the alternative set n using the sushi dataset found in [Kamishima 2003], originally provided as a

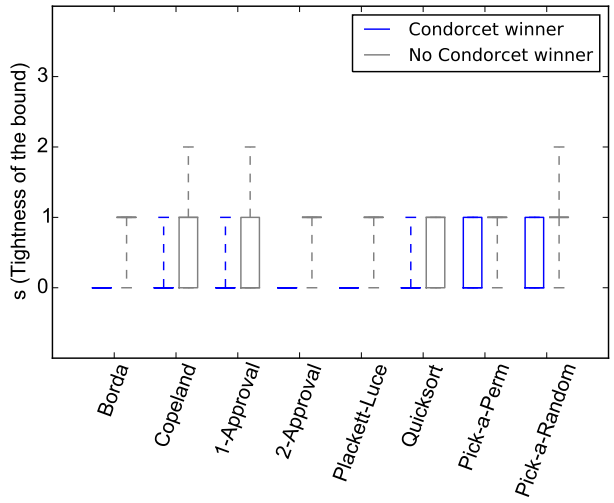


Figure A.5: Boxplots of $s\left(r, \mathcal{D}_N, n\right)$ over sampling collections of datasets shows the effect from datasets $\mathcal{D}_N$. 100 Netflix datasets with the presence of Condorcet winner and 100 datasets with no Condorcet winner ($n = 4$ and N varies for each sample).

dataset of $N = 5000$ total rankings of 10 sushi dishes. As evaluating s requires exact Kemeny consensus which can quickly become intractable when n is large, we strict in this study the number of sushi dishes n to be relatively small, and generate collections of datasets, indexed by combinations of n sushi dishes out of $\{1, \ldots, 10\}$, by counting the total occurrences of such order present in the original dataset. For example, when $n = 3$ we have a total of $\binom{10}{3} = 120$ different combinations of alternatives (hence 120 collections of datasets) each generated by counting the total occurrences of preference orders of individuals restricted to these 3 alternatives. Therefore we have a total of $120; 210; 252$ datasets respectively for $n = 3; 4; 5$. Figure A.6 shows that $s\left(r, \mathcal{D}_N, n\right)$ increases as n grows, a trend that is dominant and consistent across all voting rules. Since the maximal distance $\binom{n}{2}$ in $\mathbb{S}_n$ grows quadratically with respect to n, an interesting question would remain to specify explicitly the dependency of $k_{\min}$ on n, or the dependency of $s\left(r, \mathcal{D}_N, n\right)$ on n, for a given voting rule.

A.6.2 Applicability of The Method

We have so far focused on small n ($n \leq 5$) case, and verified that our method is efficient in using $k_{\min}$ to approximate $d(r(\mathcal{D}_N), \mathcal{K}_N)$. We are now mostly interested in the usefulness of our method when $k_{\min}$ is directly combined with voting rules in pinpointing Kemeny consensus $\mathcal{K}_N$ particularly when n is large. Now we employ our method by using $k_{\min}$ for each dataset to upper bound the approximation performance of $r(\mathcal{D}_N)$ to Kemeny's rule. Moreover, suppose that we are still interested in finding the exact Kemeny consensus despite a good approximation $r(\mathcal{D}_N)$. Once we have computed an approximated ranking $r(\mathcal{D}_N)$ and $k_{\min}$ is iden-

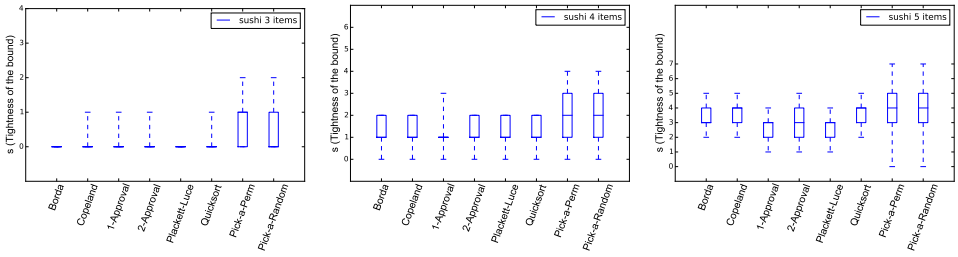


Figure A.6: Boxplots of $s\,(r, \mathcal{D}_N, n)$ over sampling collections of datasets shows the effect from different size of alternative set n with restricted sushi datasets ($n = 3; 4; 5, N = 5000$).

tified via our method, the search scope for the exact Kemeny consensuses can be narrowed down to those permutations within a distance of $k_{\min}$ to $r(\mathcal{D}_N)$. Notably [Wang 2013, Lemma 1] proved that the total number of such permutations in $\mathbb{S}_n$ is upper bounded by $\binom{n+k_{\min}-1}{k_{\min}}$ which can be smaller than $|\mathbb{S}_n| = n!$ by orders.

We took the original sushi dataset consisting of $N = 5000$ individual votes on $n = 10$ sushi dishes and created 500 bootstrapped pseudo-samples following the same empirical distribution. Note that $k_{\min}$ should also depend on $(r, \mathcal{D}_N, n)$. Since our bound is established in general with any $\sigma \in \mathbb{S}_n$ and does take into consideration the approximation efficiency of specific voting rules to Kemeny's rule, the predicted $k_{\min}$ should significantly rely on the approximate voting rules utilized and should be biased more in favor to voting rules with good approximation to Kemeny's rule since $k_{\min}$ can never be inferior to $d(r(\mathcal{D}_N), \mathcal{K}_N)$. As shown in Figure A.7, Pick-a-Random and Pick-a-Perm typically performs poorly, but this is largely due to the fact that the two voting rules are too naive to well approximate Kemeny's rule *per se*. On the contrary, we observe that Borda, Copeland and QuickSort combined with our method best pinpoint Kemeny consensuses with $k_{\min}$ of a median distance 14. This further means that in order to obtain all the exact Kemeny consensuses now, on average we need to search through at most $\binom{10+14-1}{14} = 817,190$ permutations instead of $10! = 3,628,800$ permutations, where 77% of permutations in $\mathbb{S}_{10}$ are removed from consideration.

A.7 Discussion

In this chapter, we have established a theoretical result that allows to control the Kendall's tau distance between a permutation and the Kemeny consensuses of any dataset. In practice, this provides a simple and general method to predict, for any ranking aggregation procedure, how close its output on a dataset is from the Kemeny consensuses. From a broader perspective, it constitutes a novel approach to apprehend the complexity of Kemeny aggregation.

Our results rely on some geometric properties of the Kendall embedding. Although they have rarely been used in the literature, the geometric properties have proved to provide a powerful framework to analyze Kemeny aggregation. We there-

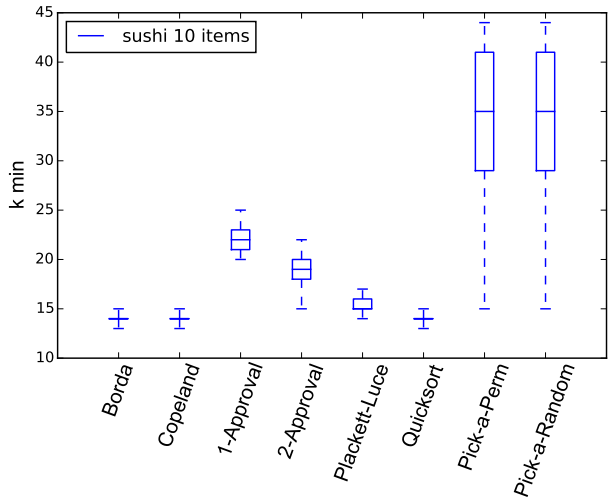


Figure A.7: Boxplots of $k_{\min}$ over 500 bootstrapped pseudo-samples of the sushi dataset ($n = 10, N = 5000$).

fore believe that it could pave way to other profound results. In particular, we deem that an analysis of how the embeddings of the permutation spread on the sphere could lead to a finer condition in Theorem A.1, which is left as future work.

Another interesting direction would certainly be to extend our method to rank aggregation from partial orders, such as pairwise comparisons or top-k rankings. Two main approaches can be followed. In the first one, a partial order would be identified with the set $S \subset \mathbb{S}_n$ of its linear extensions and its distance to a permutation $\sigma \in \mathbb{S}_n$ defined by the average $(1/|S|)\sum_{\sigma' \in S} d(\sigma, \sigma')$. The Kendall embedding would then naturally be extended to S as $(1/|S|)\sum_{\sigma' \in S} \phi(\sigma')$, the barycenter of embeddings of its linear extensions. In the second approach, one would see a partial order as a collection of pairwise comparisons $\{i_1 \succ j_1, \dots, i_m \succ j_m\}$ and define its distance to a permutation $\sigma \in \mathbb{S}_n$ by the average number of pairwise disagreements $(1/m)\sum_{r=1}^{m} \mathbb{1}\{\sigma(i_r) > \sigma(j_r)\}$. The Kendall embedding would then naturally be extended to $\{i_1 \succ j_1, \dots, i_m \succ j_m\}$ as the embedding of any linear extension σ where the coordinate on $\{i, j\}$ is put equal to 0 if $\{i, j\}$ does not appear in the collection. In both cases, our approach would apply with slight changes to exploit related geometric properties.

Bibliography

[Ailon 2008] N. Ailon, M. Charikar and A. Newman. *Aggregating inconsistent information: Ranking and clustering.* Journal of the ACM (JACM), vol. 55, no. 5, pages 23:1–23:27, 2008. (Cited on pages 18, 110 and 111.)

[Ali 2012] A. Ali and M. Meilă. *Experiments with Kemeny ranking: What works when?* Mathematical Social Sciences, vol. 64, no. 1, pages 28–40, 2012. (Cited on pages 14, 110 and 120.)

[Allahyar 2015] A. Allahyar and J. de Ridder. *FERAL: network-based classifier with application to breast cancer outcome prediction.* Bioinformatics, vol. 31, no. 12, pages i311–i319, 2015. (Cited on page 81.)

[Alon 1999] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack and A. J. Levine. *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.* Proceedings of the National Academy of Sciences of the United States of America (PNAS), vol. 96, no. 12, pages 6745–6750, 1999. (Cited on page 46.)

[Amadoz 2015] A. Amadoz, P. Sebastian-Leon, E. Vidal, F. Salavert and J. Dopazo. *Using activation status of signaling pathways as mechanism-based biomarkers to predict drug sensitivity.* Scientific Reports, vol. 5, 2015. (Cited on page 85.)

[Ambroise 2002] C. Ambroise and G. J. McLachlan. *Selection bias in gene extraction on the basis of microarray gene-expression data.* Proceedings of the National Academy of Sciences of the United States of America (PNAS), vol. 99, no. 10, pages 6562–6566, 2002. (Cited on page 105.)

[Arrow 1950] K. J. Arrow. *A difficulty in the concept of social welfare.* The Journal of Political Economy, pages 328–346, 1950. (Cited on page 110.)

[Arrow 2012] K. J. Arrow. Social choice and individual values, volume 12. Yale Univ Press, 2012. (Cited on page 18.)

[Arteaga 2012] C. L. Arteaga, M. X. Sliwkowski, C. K. Osborne, E. A. Perez, F. Puglisi and L. Gianni. *Treatment of HER2-positive breast cancer: current status and future perspectives.* Nature Reviews Clinical Oncology, vol. 9, no. 1, pages 16–32, 2012. (Cited on page 4.)

[Ashburner 2000] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler *et al. Gene Ontology: tool for the unification of biology.* Nature Genetics, vol. 25, no. 1, pages 25–29, 2000. (Cited on page 54.)

[Azencott 2016] C.-A. Azencott. *Network-Guided Biomarker Discovery*. In Machine Learning for Health Informatics, pages 319–336. Springer, 2016. (Cited on pages 12 and 55.)

[Bach 2004] F. R. Bach, G. R. G. Lanckriet and M. I. Jordan. *Multiple Kernel Learning, Conic Duality, and the SMO Algorithm*. In Proceedings of the Twenty-first International Conference on Machine Learning (ICML-04), page 6, New York, NY, USA, 2004. ACM. (Cited on page 29.)

[Bakir 2007] G. H. Bakir, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar and S. V. N. Vishwanathan. Predicting structured data. MIT Press, 2007. (Cited on page 52.)

[Balcan 2008] M.-F. Balcan, N. Bansal, A. Beygelzimer, D. Coppersmith, J. Langford and G. B. Sorkin. *Robust reductions from ranking to classification*. Machine Learning, vol. 72, no. 1–2, pages 139–153, 2008. (Cited on page 18.)

[Balendiran 2004] G. K. Balendiran, R. Dabur and D. Fraser. *The role of glutathione in cancer*. Cell Biochemistry and Function, vol. 22, no. 6, pages 343–352, 2004. (Cited on page 78.)

[Barabási 2004] A.-L. Barabási and Z. N. Oltvai. *Network biology: understanding the cell's functional organization*. Nature Reviews Genetics, vol. 5, no. 2, pages 101–113, 2004. (Cited on page 84.)

[Barabási 2011] A.-L. Barabási, N. Gulbahce and J. Loscalzo. *Network medicine: a network-based approach to human disease*. Nature Reviews Genetics, vol. 12, no. 1, pages 56–68, 2011. (Cited on page 84.)

[Barallobre-Barreiro 2012] J. Barallobre-Barreiro, A. Didangelos, F. A. Schoendube, I. Drozdov, X. Yin *et al.* *Proteomics analysis of cardiac extracellular matrix remodeling in a porcine model of ischemia-reperfusion injury*. Circulation, 2012. (Cited on page 99.)

[Barillot 2012] E. Barillot, L. Calzone, P. Hupé, J.-P. Vert and A. Zinovyev. Computational systems biology of cancer. CRC Press, 2012. (Cited on page 7.)

[Barthelemy 1981] J. P. Barthelemy and B. Monjardet. *The median procedure in cluster analysis and social choice theory*. Mathematical Social Sciences, vol. 1, pages 235–267, 1981. (Cited on page 113.)

[Bartholdi III 1989] J. Bartholdi III, C. A. Tovey and M. A. Trick. *Voting schemes for which it can be difficult to tell who won the election*. Social Choice and Welfare, vol. 6, no. 2, pages 157–165, 1989. (Cited on pages 14 and 37.)

[Bartholdi 1989] J. J. Bartholdi, C. A. Tovey and M. A. Trick. *The computational difficulty of manipulating an election*. Social Choice and Welfare, vol. 6, pages 227–241, 1989. (Cited on page 110.)

[Beer 2002] D. G. Beer, S. L. R. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin *et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma.* Nature Medicine, vol. 8, no. 8, pages 816–824, Aug 2002. (Cited on page 46.)

[Belkin 2004] M. Belkin, I. Matveeva and P. Niyogi. *Regularization and semi-supervised learning on large graphs.* In International Conference on Computational Learning Theory, pages 624–638. Springer, 2004. (Cited on pages 55 and 56.)

[Benjamini 1995] Y. Benjamini and Y. Hochberg. *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society: Series B (Methodological), pages 289–300, 1995. (Cited on page 93.)

[Bernard 2017] E. Bernard, Y. Jiao, E. Scornet, V. Stoven, T. Walter and J.-P. Vert. *Kernel Multitask Regression for Toxicogenetics.* Molecular Informatics, 2017. In press. bioRxiv preprint bioRxiv-171298. (Cited on page 15.)

[Betzler 2008] N. Betzler, M. R. Fellows, J. Guo, R. Niedermeier and F. A. Rosamond. *Fixed-parameter algorithms for Kemeny scores.* In Algorithmic Aspects in Information and Management, pages 60–71. Springer, 2008. (Cited on pages 110 and 112.)

[Betzler 2009] N. Betzler, M. R. Fellows, J. Guo, R. Niedermeier and F. A. Rosamond. *How similarity helps to efficiently compute Kemeny rankings.* In Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems–Volume 1, pages 657–664. International Foundation for Autonomous Agents and Multiagent Systems, 2009. (Cited on pages 110 and 112.)

[Bhardwaj 2005] N. Bhardwaj and H. Lu. *Correlation between gene expression profiles and protein-protein interactions within and across genomes.* Bioinformatics, vol. 21, no. 11, pages 2730–2738, 2005. (Cited on page 90.)

[Bhattacharjee 2013] P. Bhattacharjee, S. Paul, M. Banerjee, D. Patra, P. Banerjee *et al. Functional compensation of glutathione S-transferase M1 (GSTM1) null by another GST superfamily member, GSTM2.* Scientific Reports, vol. 3, page 2704, 2013. (Cited on page 78.)

[Bilal 2013] E. Bilal, J. Dutkowski, J. Guinney, I. S. Jang, B. A. Logsdon *et al. Improving breast cancer survival analysis through competition-based multidimensional modeling.* PLoS Computational Biology, vol. 9, no. 5, page e1003047, 2013. (Cited on pages xi, 10 and 11.)

[Blin 2011] G. Blin, M. Crochemore, S. Hamel and S. Vialette. *Median of an odd number of permutations.* Pure Mathematics and Applications, vol. 21, no. 2, pages 161–175, 2011. (Cited on page 112.)

[Bottou 2008] L. Bottou and O. Bousquet. *The Tradeoffs of Large Scale Learning.* In J.C. Platt, D. Koller, Y. Singer and S. Roweis, editors, Advances in Neural Information Processing Systems (NIPS-08), volume 20, pages 161–168. Curran Associates, Inc., 2008. (Cited on page 34.)

[Boucheron 2013] S. Boucheron, G. Lugosi and P. Massart. Concentration inequalities. Oxford Univ Press, 2013. (Cited on page 32.)

[Brandt 2015] F. Brandt, M. Brill, E. Hemaspaandra and L. A. Hemaspaandra. *Bypassing combinatorial protections: Polynomial-time algorithms for single-peaked electorates.* Journal of Artificial Intelligence Research, pages 439–496, 2015. (Cited on pages 110 and 112.)

[Breiman 1984] L. Breiman, J. Friedman, C. J. Stone and R. A. Olshen. Classification and regression trees. CRC Press, 1984. (Cited on page 8.)

[Breiman 2001] L. Breiman. *Random forests.* Machine Learning, vol. 45, no. 1, pages 5–32, 2001. (Cited on pages 8 and 94.)

[Chan 2008] G. Chan, D. Kalaitzidis and B. G. Neel. *The tyrosine phosphatase Shp2 (PTPN11) in cancer.* Cancer and Metastasis Reviews, vol. 27, no. 2, pages 179–192, 2008. (Cited on page 97.)

[Chang 2011] C.-C. Chang and C.-J. Lin. *LIBSVM: A library for support vector machines.* ACM Transactions on Intelligent Systems and Technology, vol. 2, pages 27:1–27:27, 2011. (Cited on page 94.)

[Chen 2001] S. S. Chen, D. L. Donoho and M. A. Saunders. *Atomic decomposition by basis pursuit.* SIAM review, vol. 43, no. 1, pages 129–159, 2001. (Cited on page 62.)

[Cheng 2013a] W.-Y. Cheng, T.-H. Ou Yang and D. Anastassiou. *Biomolecular events in cancer revealed by attractor metagenes.* PLoS Computational Biology, vol. 9, no. 2, page e1002920, 2013. (Cited on page 10.)

[Cheng 2013b] W.-Y. Cheng, T.-H. Ou Yang and D. Anastassiou. *Development of a prognostic model for breast cancer survival in an open challenge environment.* Science Translational Medicine, vol. 5, no. 181, pages 181ra50–181ra50, 2013. (Cited on page 10.)

[Chia 2008] S. Chia, B. Norris, C. Speers, M. Cheang, B. Gilks *et al. Human epidermal growth factor receptor 2 overexpression as a prognostic factor in a large tissue microarray series of node-negative breast cancers.* Journal of Clinical Oncology, vol. 26, no. 35, pages 5697–5704, 2008. (Cited on page 4.)

[Chung 1997] F. Chung. Spectral graph theory, volume 92. American Mathematical Society, 1997. (Cited on pages 59 and 81.)

[Cleveland 1979] W. S. Cleveland. *Robust locally weighted regression and smoothing scatterplots.* Journal of the American Statistical Association, vol. 74, no. 368, pages 829–836, 1979. (Cited on page 74.)

[Cleveland 2010] R. J. Cleveland, M. D. Gammon, C.-M. Long, M. M. Gaudet, S. M. Eng *et al. Common genetic variations in the LEP and LEPR genes, obesity and breast cancer incidence and survival.* Breast Cancer Research and Treatment, vol. 120, no. 3, pages 745–752, 2010. (Cited on page 97.)

[Cohen 1999] W. W. Cohen, R. E. Schapire and Y. Singer. *Learning to Order Things.* Journal of Artificial Intelligence Research, vol. 10, no. 1, pages 243–270, may 1999. (Cited on page 110.)

[Commons 2017] Wikimedia Commons. *File:Heatmap.png — Wikimedia Commons, the free media repository.* https://commons.wikimedia.org/w/index.php?title=File:Heatmap.png&oldid=233617968, 2017. Online; accessed June 30, 2017. (Cited on pages xi and 5.)

[Condorcet 1785] N. Condorcet. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. L'imprimerie royale, 1785. (Cited on page 110.)

[Conitzer 2006] V. Conitzer, A. Davenport and J. Kalagnanam. *Improved bounds for computing Kemeny rankings.* In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06)–Volume 1, volume 6, pages 620–626, 2006. (Cited on pages 110, 111 and 112.)

[Consortium 2015a] The Gene Ontology Consortium. *Gene ontology consortium: going forward.* Nucleic Acids Research, vol. 43, no. Database Issue, pages D1049–D1056, 2015. (Cited on page 90.)

[Consortium 2015b] The UniProt Consortium. *UniProt: a hub for protein information.* Nucleic Acids Research, vol. 43, no. Database Issue, pages D204–D212, 2015. (Cited on page 90.)

[Copeland 1951] A. H. Copeland. *A reasonable social welfare function.* In Mimeographed notes from a Seminar on Applications of Mathematics to the Social Sciences. University of Michigan, 1951. (Cited on pages 37 and 120.)

[Coppersmith 2006] D. Coppersmith, L. Fleischer and A. Rudra. *Ordering by Weighted Number of Wins Gives a Good Ranking for Weighted Tournaments.* In Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithm, SODA '06, pages 776–782, 2006. (Cited on pages 110, 111 and 112.)

[Cormen 2009] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein. Introduction to algorithms. MIT Press, 3rd edition, 2009. (Cited on page 52.)

[Cornaz 2013] D. Cornaz, L. Galand and O. Spanjaard. *Kemeny Elections with Bounded Single-Peaked or Single-Crossing Width.* In The 23rd International Joint Conference on Artificial Intelligence (IJCAI-13), volume 13, pages 76–82. Citeseer, 2013. (Cited on pages 110 and 112.)

[Cortes 1995] C. Cortes and V. Vapnik. *Support-Vector Networks.* Machine Learning, vol. 20, no. 3, pages 273–297, 1995. (Cited on pages 8 and 18.)

[Cox 1958] D. R. Cox. *The Regression Analysis of Binary Sequences.* Journal of the Royal Statistical Society: Series B (Methodological), vol. 20, no. 2, pages 215–242, 1958. (Cited on page 8.)

[Cox 1972] D. R. Cox. *Regression Models and Life-Tables.* Journal of the Royal Statistical Society: Series B (Methodological), vol. 34, no. 2, pages 187–220, 1972. (Cited on pages 8, 14 and 57.)

[Crammer 2002] K. Crammer and Y. Singer. *On the algorithmic implementation of multiclass kernel-based vector machines.* Journal of Machine Learning Research, vol. 2, pages 265–292, 2002. (Cited on page 52.)

[Creixell 2015] P. Creixell, J. Reimand, S. Haider, G. Wu, T. Shibata *et al. Pathway and network analysis of cancer genomes.* Nature Methods, vol. 12, no. 7, pages 615–621, 2015. (Cited on page 12.)

[Critchlow 1985] D. E. Critchlow. Metric methods for analyzing partially ranked data. Springer, 1985. (Cited on page 20.)

[Curtis 2012] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda *et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.* Nature, vol. 486, no. 7403, pages 346–352, 2012. (Cited on page 66.)

[Czajkowski 2011] M. Czajkowski and M. Kretowski. *Top scoring pair decision tree for gene expression data analysis.* In Software Tools and Algorithms for Biological Systems, pages 27–35. Springer, 2011. (Cited on page 11.)

[Daubechies 1998] I. Daubechies and W. Sweldens. *Factoring wavelet transforms into lifting steps.* Journal of Fourier Analysis and Applications, vol. 4, no. 3, pages 247–269, 1998. (Cited on page 81.)

[Davenport 2004] A. Davenport and J. Kalagnanam. *A computational study of the Kemeny rule for preference aggregation.* In Proceedings of the 19th National Conference on Artifical Intelligence (AAAI-04), volume 4, pages 697–702, 2004. (Cited on pages 110 and 111.)

[Davenport 2005] A. Davenport and D. Lovell. *Ranking Pilots in Aerobatic Flight Competitions.* Technical report, IBM Research Report RC23631 (W0506-079), TJ Watson Research Center, NY, 2005. (Cited on page 110.)

[de Borda 1781] J. C. de Borda. *Mémoire sur les élections au scrutin.* Histoire de l'Academie Royale des Sciences, 1781. (Cited on pages 37, 110 and 120.)

[Dean 2001] M. Dean, Y. Hamon and G. Chimini. *The human ATP-binding cassette (ABC) transporter superfamily.* Journal of Lipid Research, vol. 42, no. 7, pages 1007–1017, 2001. (Cited on page 99.)

[Deng 2014] K. Deng, S. Han, K. J. Li and J. S. Liu. *Bayesian aggregation of order-based rank data.* Journal of the American Statistical Association, vol. 109, no. 507, pages 1023–1039, 2014. (Cited on page 110.)

[Desarkar 2016] M. S. Desarkar, S. Sarkar and P. Mitra. *Preference relations based unsupervised rank aggregation for metasearch.* Expert Systems with Applications, vol. 49, pages 86–98, 2016. (Cited on page 110.)

[Desmedt 2007] C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand *et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series.* Clinical Cancer Research, vol. 13, no. 11, pages 3207–3214, 2007. (Cited on pages 7 and 46.)

[Dhillon 2007] A. S. Dhillon, S. Hagan, O. Rath and W. Kolch. *MAP kinase signalling pathways in cancer.* Oncogene, vol. 26, no. 22, pages 3279–3290, 2007. (Cited on page 79.)

[Diaconis 1977] P. Diaconis and R. L. Graham. *Spearman's footrule as a measure of disarray.* Journal of the Royal Statistical Society: Series B (Methodological), pages 262–268, 1977. (Cited on pages 110, 111 and 112.)

[Diaconis 1988] P. Diaconis. Group representations in probability and statistics, volume 11 of *Lecture Notes–Monograph Series.* Institut of Mathematical Statistics, Hayward, CA, 1988. (Cited on pages 18 and 40.)

[Diaconis 1989] P. Diaconis. *A generalization of spectral analysis with application to ranked data.* The Annals of Statistics, pages 949–979, 1989. (Cited on page 122.)

[Dopazo 2010] J. Dopazo. *Functional profiling methods in Cancer.* Cancer Gene Profiling: Methods and Protocols, pages 363–374, 2010. (Cited on page 84.)

[Drier 2013] Y. Drier, M. Sheffer and E. Domany. *Pathway-based personalized analysis of cancer.* Proceedings of the National Academy of Sciences of the United States of America (PNAS), vol. 110, no. 16, pages 6388–6393, 2013. (Cited on pages 12 and 85.)

[Drutu 2017] C. Drutu and M. Kapovich. *Geometric Group Theory.* Technical report, UC Davis, 2017. Preprint version: January 16, 2017, a book to be published in the AMS series "Colloquium Publications" in 2017. (Cited on page 35.)

[Dwork 2001] C. Dwork, R. Kumar, M. Naor and D. Sivakumar. *Rank aggregation methods for the web.* In Proceedings of the Tenth International Conference on World Wide Web (WWW-01), pages 613–622. ACM, 2001. (Cited on pages 18 and 110.)

[Eduati 2015] F. Eduati, L. Mangravite, T. Wang, H. Tang, J. Bare *et al. Prediction of human population responses to toxic compounds by a collaborative competition.* Nature Biotechnology, vol. 33, no. 9, pages 933–940, 2015. (Cited on page 15.)

[Efron 2004] B. Efron, T. Hastie, I. Johnstone and R. Tibshirani. *Least angle regression.* The Annals of Statistics, vol. 32, no. 2, pages 407–499, 2004. (Cited on page 63.)

[Efroni 2007] S. Efroni, C. F. Schaefer and K. H. Buetow. *Identification of key processes underlying cancer phenotypes using biologic pathway analysis.* PloS One, vol. 2, no. 5, page e425, 2007. (Cited on page 90.)

[Ein-Dor 2006] L. Ein-Dor, O. Zuk and E. Domany. *Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer.* Proceedings of the National Academy of Sciences of the United States of America (PNAS), vol. 103, no. 15, pages 5923–5928, 2006. (Cited on pages 84 and 106.)

[El-Omar 2000] E. M. El-Omar, M. Carrington, W.-H. Chow, K. E. L. McColl, J. H. Bream *et al. Interleukin-1 polymorphisms associated with increased risk of gastric cancer.* Nature, vol. 404, no. 6776, pages 398–402, 2000. (Cited on page 97.)

[Elad 2007] M. Elad, P. Milanfar and R. Rubinstein. *Analysis versus synthesis in signal priors.* Inverse Problems, vol. 23, no. 3, page 947, 2007. (Cited on pages 63 and 82.)

[Fan 2008] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang and C.-J. Lin. *LIBLINEAR: A library for large linear classification.* Journal of Machine Learning Research, vol. 9, no. Aug, pages 1871–1874, 2008. (Cited on page 94.)

[Ferlay 2013] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser *et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11.* `http://globocan.iarc.fr`, 2013. Online; accessed June 30, 2017. Lyon, France: International Agency for Research on Cancer. (Cited on page 2.)

[Fey 2015] D. Fey, M. Halasz, D. Dreidax, S. P. Kennedy, N. Rauch *et al. Signaling pathway models as biomarkers: Patient-specific simulations of JNK activity predict the survival of neuroblastoma patients.* Science Signaling, vol. 8, no. 408, pages ra130–ra130, 2015. (Cited on page 85.)

[Filipits 2011] M. Filipits, M. Rudas, R. Jakesz, P. Dubsky, F. Fitzal *et al.* *A new molecular predictor of distant recurrence in ER-positive, HER2-negative breast cancer adds independent information to conventional clinical risk factors.* Clinical Cancer Research, vol. 17, no. 18, pages 6012–6020, 2011. (Cited on page 6.)

[Filippone 2008] M. Filippone, F. Camastra, F. Masulli and S. Rovetta. *A survey of kernel and spectral methods for clustering.* Pattern Recognition, vol. 41, no. 1, pages 176–190, 2008. (Cited on page 51.)

[Fisher 1936] R. A. Fisher. *The use of multiple measurements in taxonomic problems.* Annals of Human Genetics, vol. 7, no. 2, pages 179–188, 1936. (Cited on page 8.)

[Fligner 1986] M. A. Fligner and J. S. Verducci. *Distance based ranking models.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 48, no. 3, pages 359–369, 1986. (Cited on page 20.)

[Foekens 2006] J. A. Foekens, D. Atkins, Y. Zhang, F. C.G.J. Sweep, N. Harbeck *et al.* *Multicenter Validation of a Gene Expression-Based Prognostic Signature in Lymph Node-Negative Primary Breast Cancer.* Journal of Clinical Oncology, vol. 24, no. 11, pages 1665–1671, 2006. (Cited on page 7.)

[Fogel 2013] F. Fogel, R. Jenatton, F. Bach and A. D'Aspremont. *Convex Relaxations for Permutation Problems.* In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems (NIPS-13), pages 1016–1024. Curran Associates, Inc., 2013. (Cited on page 103.)

[Frank 2003] N. Y. Frank, S. S. Pendse, P. H. Lapchak, A. Margaryan, D. Shlain *et al.* *Regulation of progenitor cell fusion by ABCB5 P-glycoprotein, a novel human ATP-binding cassette transporter.* Journal of Biological Chemistry, vol. 278, no. 47, pages 47156–47165, 2003. (Cited on page 99.)

[Frank 2005] N. Y. Frank, A. Margaryan, Y. Huang, T. Schatton, A. M. Waaga-Gasser *et al.* *ABCB5-mediated doxorubicin transport and chemoresistance in human malignant melanoma.* Cancer Research, vol. 65, no. 10, pages 4320–4333, 2005. (Cited on page 99.)

[Fresno Vara 2004] J. Á. Fresno Vara, E. Casado, J. de Castro, P. Cejas, C. Belda-Iniesta and M. González-Barón. *PI3K/Akt signalling pathway and cancer.* Cancer Treatment Reviews, vol. 30, no. 2, pages 193–204, 2004. (Cited on page 79.)

[Freund 2015] D. Freund and D. P. Williamson. *Rank Aggregation: New Bounds for MCx.* Technical report, Cornell University, 2015. arXiv preprint arXiv:1510.00738. (Cited on pages 110 and 111.)

[Friedman 2001] J. H. Friedman. *Greedy function approximation: a gradient boosting machine.* Annals of Statistics, pages 1189–1232, 2001. (Cited on pages 8 and 94.)

[Friedman 2010] J. Friedman, T. Hastie and R. Tibshirani. *Regularization paths for generalized linear models via coordinate descent.* Journal of Statistical Software, vol. 33, no. 1, page 1, 2010. (Cited on pages 63 and 94.)

[Fryburg 2014] D. A. Fryburg, D. H. Song, D. Laifenfeld and D. de Graaf. *Systems diagnostics: anticipating the next generation of diagnostic tests based on mechanistic insight into disease.* Drug Discovery Today, vol. 19, no. 2, pages 108–112, 2014. (Cited on page 84.)

[Fukumizu 2008] K. Fukumizu, A. Gretton, B. Schölkopf and B. K. Sriperumbudur. *Characteristic Kernels on Groups and Semigroups.* In D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, editors, Advances in Neural Information Processing Systems (NIPS-08), volume 21, pages 473–480. Curran Associates, Inc., 2008. (Cited on page 19.)

[Gärtner 2004] T. Gärtner, J.W. Lloyd and P.A. Flach. *Kernels and Distances for Structured Data.* Machine Learning, vol. 57, no. 3, pages 205–232, 2004. (Cited on pages 14 and 18.)

[Geman 2004] D. Geman, C. d'Avignon, D. Q. Naiman and R. L. Winslow. *Classifying gene expression profiles from pairwise mRNA comparisons.* Statistical Applications in Genetics and Molecular Biology, vol. 3, no. 1, page Article19, 2004. (Cited on pages 11, 18, 29, 46 and 47.)

[Girolami 2002] M. Girolami. *Mercer kernel-based clustering in feature space.* IEEE Transactions on Neural Networks, vol. 13, no. 3, pages 780–784, 2002. (Cited on page 37.)

[Gönen 2011] M. Gönen and E. Alpaydın. *Multiple kernel learning algorithms.* Journal of Machine Learning Research, vol. 12, pages 2211–2268, 2011. (Cited on pages 29 and 104.)

[Gordon 2002] G. J. Gordon, R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumenstock *et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma.* Cancer Research, vol. 62, no. 17, pages 4963–4967, 2002. (Cited on page 46.)

[Gormley 2006] I. C. Gormley and T. B. Murphy. *Analysis of Irish third-level college applications data.* Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 169, no. 2, pages 361–379, 2006. (Cited on page 37.)

[Gormley 2008] I. C. Gormley and T. B. Murphy. *Exploring voting blocs within the Irish electorate: A mixture modeling approach.* Journal of the American

Statistical Association, vol. 103, no. 483, pages 1014–1027, 2008. (Cited on page 37.)

[Guyon 2002] I. Guyon, J. Weston, S. Barnhill and V. Vapnik. *Gene selection for cancer classification using support vector machines.* Machine Learning, vol. 46, no. 1, pages 389–422, 2002. (Cited on page 10.)

[Guyon 2003] I. Guyon and A. Elisseeff. *An introduction to variable and feature selection.* Journal of Machine Learning Research, vol. 3, no. Mar, pages 1157–1182, 2003. (Cited on page 10.)

[Győrffy 2015] B. Győrffy, C. Hatzis, T. Sanft, E. Hofstatter, B. Aktas and L. Pusztai. *Multigene prognostic tests in breast cancer: past, present, future.* Breast Cancer Research, vol. 17, no. 1, page 11, 2015. (Cited on page 7.)

[Hammond 2011] D. K. Hammond, P. Vandergheynst and R. Gribonval. *Wavelets on graphs via spectral graph theory.* Applied and Computational Harmonic Analysis, vol. 30, no. 2, pages 129–150, 2011. (Cited on pages 61, 64 and 65.)

[Hanahan 2000] D. Hanahan and R. A. Weinberg. *The hallmarks of cancer.* Cell, vol. 100, no. 1, pages 57–70, 2000. (Cited on page 4.)

[Hanahan 2011] D. Hanahan and R. A. Weinberg. *Hallmarks of cancer: the next generation.* Cell, vol. 144, no. 5, pages 646–674, 2011. (Cited on pages 4, 12, 85 and 91.)

[Hastie 2009] T. Hastie, R. Tibshirani and J. Friedman. The elements of statistical learning: Data mining, inference, and prediction. Springer-Verlag, New York, 2nd edition, 2009. (Cited on pages 8 and 104.)

[Hastie 2015] T. Hastie, R. Tibshirani and M. Wainwright. Statistical learning with sparsity: The lasso and generalizations. CRC Press, 2015. (Cited on page 9.)

[Haury 2011] A.-C. Haury, P. Gestraud and J.-P. Vert. *The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures.* PloS One, vol. 6, no. 12, page e28210, 2011. (Cited on pages 105 and 106.)

[Haussler 1999] D. Haussler. *Convolution kernels on discrete structures.* Technical report UCSC-CRL-99-10, UC Santa Cruz, 1999. (Cited on pages 18 and 22.)

[Helmbold 2009] D. P. Helmbold and M. K. Warmuth. *Learning Permutations with Exponential Weights.* Journal of Machine Learning Research, vol. 10, pages 1705–1736, 2009. (Cited on page 18.)

[Hidalgo 2017] M. R. Hidalgo, C. Çubuk, A. Amadoz, F. Salavert, J. Carbonell-Caballero and J. Dopazo. *High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes.*

Oncotarget, vol. 8, no. 3, pages 5160–5178, 2017. (Cited on pages 12, 15, 85, 89, 96 and 100.)

[Hira 2015] Z. M. Hira and D. F. Gillies. *A review of feature selection and feature extraction methods applied on microarray data.* Advances in Bioinformatics, vol. 2015, 2015. (Cited on page 10.)

[Hocking 1976] R. R. Hocking. *The analysis and selection of variables in linear regression.* Biometrics, vol. 32, no. 1, pages 1–49, 1976. (Cited on page 9.)

[Hodge 2005] D. R. Hodge, E. M. Hurt and W. L. Farrar. *The role of IL-6 and STAT3 in inflammation and cancer.* European Journal of Cancer, vol. 41, no. 16, pages 2502–2512, 2005. (Cited on page 97.)

[Hoeppner 2015] L. H. Hoeppner, Y. Wang, A. Sharma, N. Javeed, V. P. Van Keulen *et al. Dopamine D2 receptor agonists inhibit lung cancer progression by reducing angiogenesis and tumor infiltrating myeloid derived suppressor cells.* Molecular Oncology, vol. 9, no. 1, pages 270–281, 2015. (Cited on page 79.)

[Hoerl 1970] A. E. Hoerl and R. W. Kennard. *Ridge regression: Biased estimation for nonorthogonal problems.* Technometrics, vol. 12, no. 1, pages 55–67, 1970. (Cited on page 57.)

[Honda 2015] K. Honda. *The biological role of actinin-4 (ACTN4) in malignant phenotypes of cancer.* Cell & Bioscience, vol. 5, no. 1, page 41, 2015. (Cited on page 97.)

[Hood 2013] L. Hood. *Systems Biology and P4 Medicine: Past, Present, and Future.* Rambam Maimonides Medical Journal, vol. 4, no. 2, page e0012, 2013. (Cited on page 84.)

[Hosmer 1999] D. W. Hosmer and S. Lemeshow. Applied survival analysis: Regression modeling of time to event data. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1999. (Cited on page 8.)

[Huang 2009] J. Huang, C. Guestrin and L. Guibas. *Fourier Theoretic Probabilistic Inference over Permutations.* Journal of Machine Learning Research, vol. 10, pages 997–1070, 2009. (Cited on page 18.)

[Hubert 1985] L. Hubert and P. Arabie. *Comparing Partitions.* Journal of Classification, vol. 2, no. 1, pages 193–218, 1985. (Cited on page 42.)

[Hunter 2004] D. R. Hunter. *MM algorithms for generalized Bradley-Terry models.* Annals of Statistics, pages 384–406, 2004. (Cited on page 120.)

[Hussain 2003] S. Hussain, E. Witt, P. A. Huber, A. L. Medhurst, A. Ashworth and C. G. Mathew. *Direct interaction of the Fanconi anaemia protein FANCG*

with BRCA2/FANCD1. Human Molecular Genetics, vol. 12, no. 19, pages 2503–2510, 2003. (Cited on page 78.)

[Issaq 2011] H. J. Issaq, T. J. Waybright and T. D. Veenstra. *Cancer biomarker discovery: opportunities and pitfalls in analytical methods.* Electrophoresis, vol. 32, no. 9, pages 967–975, 2011. (Cited on page 105.)

[Iwamoto 2010] T. Iwamoto and L. Pusztai. *Predicting prognosis of breast cancer with gene signatures: are we lost in a sea of data?* Genome Medicine, vol. 2, no. 11, page 81, 2010. (Cited on page 84.)

[Jacob 2009] L. Jacob, G. Obozinski and J.-P. Vert. *Group lasso with overlap and graph lasso.* In Proceedings of the 26th Annual International Conference on Machine Learning, pages 433–440. ACM, 2009. (Cited on pages 12, 55 and 58.)

[Jacob 2012] L. Jacob, P. Neuvial and S. Dudoit. *More power via graph-structured tests for differential expression of gene networks.* The Annals of Applied Statistics, pages 561–600, 2012. (Cited on page 85.)

[Jacques 2014] J. Jacques and C. Biernacki. *Model-based clustering for multivariate partial ranking data.* Journal of Statistical Planning and Inference, vol. 149, pages 201–217, 2014. (Cited on page 44.)

[Jansen 2009] M. Jansen, G. P. Nason and B. W. Silverman. *Multiscale methods for data on graphs and irregular multidimensional situations.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 71, no. 1, pages 97–125, 2009. (Cited on pages 64, 66 and 67.)

[Jiang 2011] X. Jiang, L.-H. Lim, Y. Yao and Y. Ye. *Statistical ranking and combinatorial Hodge theory.* Mathematical Programming, vol. 127, no. 1, pages 203–244, 2011. (Cited on page 112.)

[Jiao 2015] Y. Jiao and J.-P. Vert. *The Kendall and Mallows Kernels for Permutations.* In D. Blei and F. Bach, editors, Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pages 1935–1944, 2015. (Cited on pages 17, 94 and 112.)

[Jiao 2016a] Y. Jiao. ***kernrank***, *version 1.0.2.* `https://github.com/YunlongJiao/kernrank`, 2016. Online; accessed April 2016. Open-source `R` package publicly available on GitHub. (Cited on page 19.)

[Jiao 2016b] Y. Jiao, A. Korba and E. Sibony. *Controlling the Distance to a Kemeny Consensus without Computing It.* In Proceedings of the 33rd International Conference on Machine Learning (ICML-16), pages 2971–2980, 2016. (Cited on page 109.)

[Jiao 2016c] Y. Jiao, J.-P. Vert, F. Heinemann, S. Dahlmanns and S. Kobel. *Failure State Prediction for Automated Analyzers for Analyzing a Biological Sample*, 2016. Pending European patent filed by Roche Diagnostics GmbH, F. Hoffmann–La Roche AG, December 2016. (Cited on page 16.)

[Jiao 2017a] Y. Jiao, M. R. Hidalgo, C. Çubuk, A. Amadoz, J. Carbonell-Caballero, J.-P. Vert and J. Dopazo. *Signaling Pathway Activities Improve Prognosis for Breast Cancer*. 2017. Submitted. bioRxiv preprint bioRxiv-132357. (Cited on page 83.)

[Jiao 2017b] Y. Jiao and J.-P. Vert. *The Kendall and Mallows Kernels for Permutations*. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2017. In press. HAL preprint HAL-01279273. (Cited on page 17.)

[Jiao 2017c] Y. Jiao and J.-P. Vert. *Network-based Wavelet Smoothing for Analysis of Genomic Data*. Technical report, École nationale supérieure des mines de Paris, 2017. (Cited on page 53.)

[Johnson 2007] W. E. Johnson, C. Li and A. Rabinovic. *Adjusting batch effects in microarray expression data using empirical Bayes methods*. Biostatistics, vol. 8, no. 1, pages 118–127, 2007. (Cited on page 86.)

[Kamishima 2003] T. Kamishima. *Nantonac collaborative filtering: Recommendation based on order responses*. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-03), pages 583–588. ACM, 2003. (Cited on pages 29, 115 and 122.)

[Kanehisa 2000] M. Kanehisa and S. Goto. *KEGG: Kyoto encyclopedia of genes and genomes*. Nucleic Acids Research, vol. 28, no. 1, pages 27–30, 2000. (Cited on page 54.)

[Kanehisa 2012] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi and M. Tanabe. *KEGG for integration and interpretation of large-scale molecular data sets*. Nucleic Acids Research, vol. 40, no. Database Issue, page D109, 2012. (Cited on page 87.)

[Kaplan 1958] E. L. Kaplan and P. Meier. *Nonparametric estimation from incomplete observations*. Journal of the American Statistical Association, vol. 53, no. 282, pages 457–481, 1958. (Cited on page 8.)

[Kashima 2003] H. Kashima, K. Tsuda and A. Inokuchi. *Marginalized Kernels between Labeled Graphs*. In T. Faucett and N. Mishra, editors, Proceedings of the Twentieth International Conference on Machine Learning (ICML-03), pages 321–328, New York, NY, USA, 2003. AAAI Press. (Cited on page 18.)

[Keane 2004] M. P. Keane, J. A. Belperio, Y. Y. Xue, M. D. Burdick and R. M. Strieter. *Depletion of CXCR2 inhibits tumor growth and angiogenesis in a*

murine model of lung cancer. The Journal of Immunology, vol. 172, no. 5, pages 2853–2860, 2004. (Cited on page 97.)

[Kemeny 1959] J. G. Kemeny. *Mathematics without numbers*. Daedalus, vol. 88, pages 571–591, 1959. (Cited on pages 14 and 110.)

[Kemeny 1962] J. G. Kemeny and J. L. Snell. Mathematical models in the social sciences, volume 9. Ginn New York, 1962. (Cited on page 37.)

[Kendall 1938] M. G. Kendall. *A new measure of rank correlation*. Biometrika, vol. 30, no. 1/2, pages 81–93, 1938. (Cited on page 20.)

[Kendall 1948] M. G. Kendall. Rank correlation methods. Griffin, 1948. (Cited on page 20.)

[Keshava Prasad 2009] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar *et al. Human Protein Reference Database–2009 update*. Nucleic Acids Research, vol. 37, no. suppl.1, page D767, 2009. (Cited on page 54.)

[Kiefer 2014] M. Kiefer, K. Hoyt, J. Hackett, M. Walker and J. Baker. *Multiple GSTM gene family members are recurrence risk markers in breast cancer*. Cancer Research, vol. 66, no. 8 Supplement, pages 846–846, 2014. (Cited on page 78.)

[Kim 2014] W. H. Kim, V. Singh, M. K. Chung, C. Hinrichs, D. Pachauri *et al. Multi-resolutional shape features via non-Euclidean wavelets: Applications to statistical analysis of cortical thickness*. NeuroImage, vol. 93, pages 107–123, 2014. (Cited on page 62.)

[Kirkpatrick 1983] S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi. *Optimization by Simulated Annealing*. Science, vol. 220, no. 4598, pages 671–680, 1983. (Cited on page 9.)

[Knight 1966] W. R. Knight. *A computer method for calculating Kendall's tau with ungrouped data*. Journal of the American Statistical Association, vol. 61, no. 314, pages 436–439, 1966. (Cited on pages 21 and 30.)

[Kolde 2012] R. Kolde, S. Laur, P. Adler and J. Vilo. *Robust rank aggregation for gene list integration and meta-analysis*. Bioinformatics, vol. 28, no. 4, pages 573–580, 2012. (Cited on page 110.)

[Kondor 2002] I. R. Kondor and J. Lafferty. *Diffusion kernels on graphs and other discrete input spaces*. In Proceedings of the Nineteenth International Conference on Machine Learning (ICML-02), volume 2, pages 315–322, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. (Cited on page 35.)

[Kondor 2008] I. R. Kondor. *Group Theoretical Methods in Machine Learning*. PhD thesis, Columbia University, 2008. (Cited on page 19.)

[Kondor 2010] R. I. Kondor and M. S. Barbosa. *Ranking with Kernels in Fourier space*. In A. T. Kalai and M. Mohri, editors, The 23rd Conference on Learning Theory (COLT-10), pages 451–463. Omnipress, June 2010. (Cited on pages 19, 35, 36 and 51.)

[Kourou 2015] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis and D. I. Fotiadis. *Machine learning applications in cancer prognosis and prediction*. Computational and Structural Biotechnology Journal, vol. 13, pages 8–17, 2015. (Cited on page 10.)

[Kreiter 2011] S. Kreiter, M. Diken, A. Selmi, J. Diekmann, S. Attig *et al. FLT3 ligand enhances the cancer therapeutic potency of naked RNA vaccines*. Cancer Research, vol. 71, no. 19, pages 6132–6142, 2011. (Cited on page 79.)

[Lahaie 2014] S. Lahaie and N. Shah. *Neutrality and geometry of mean voting*. In Proceedings of the Fifteenth ACM Conference on Economics and Computation, pages 333–350. ACM, 2014. (Cited on page 114.)

[Lai 2000] P. L. Lai and C. Fyfe. *Kernel and nonlinear canonical correlation analysis*. International Journal of Neural Systems, vol. 10, no. 05, pages 365–377, 2000. (Cited on page 51.)

[Lanckriet 2004a] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui and M. I. Jordan. *Learning the kernel matrix with semidefinite programming*. Journal of Machine Learning Research, vol. 5, pages 27–72, 2004. (Cited on page 29.)

[Lanckriet 2004b] G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan and W. S. Noble. *A statistical framework for genomic data fusion*. Bioinformatics, vol. 20, no. 16, pages 2626–2635, Nov 2004. (Cited on pages 29 and 104.)

[Lebanon 2008] G. Lebanon and Y. Mao. *Non-Parametric Modeling of Partially Ranked Data*. Journal of Machine Learning Research, vol. 9, pages 2401–2429, 2008. (Cited on pages 18 and 20.)

[Levis 2003] M. Levis and D. Small. *FLT3: ITDoes matter in leukemia*. Leukemia, vol. 17, no. 9, pages 1738–1752, 2003. (Cited on page 79.)

[Li 2003] J. Li, H. Liu and L. Wong. *Mean-entropy discretized features are effective for classifying high-dimensional biomedical data*. In M. J. Zaki, J. T.-L. Wang and H. Toivonen, editors, Proceedings of the 3nd ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD-03), pages 17–24, August 2003. (Cited on page 46.)

[Li 2008] C. Li and H. Li. *Network-constrained regularization and variable selection for analysis of genomic data.* Bioinformatics, vol. 24, no. 9, pages 1175–1182, 2008. (Cited on pages 55, 60 and 64.)

[Li 2010] C. Li and H. Li. *Variable selection and regression analysis for graph-structured covariates with an application to genomics.* The Annals of Applied Statistics, vol. 4, no. 3, page 1498, 2010. (Cited on pages 12, 55 and 60.)

[Li 2014] H. Li. *Learning to rank for information retrieval and natural language processing.* Synthesis Lectures on Human Language Technologies, vol. 7, no. 3, pages 1–121, 2014. (Cited on page 110.)

[Li 2015] X. Li, L. Shen, X. Shang and W. Liu. *Subpathway analysis based on signaling-pathway impact analysis of signaling pathway.* PloS One, vol. 10, no. 7, page e0132813, 2015. (Cited on page 85.)

[Li 2016] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang and H. Liu. *Feature selection: A data perspective.* Technical report, Arizona State University, 2016. arXiv preprint arXiv:1601.07996. (Cited on page 10.)

[Liaw 2002] A. Liaw and M. Wiener. *Classification and regression by randomForest.* R News, vol. 2, no. 3, pages 18–22, 2002. (Cited on page 94.)

[Lim 2014] C. H. Lim and S. Wright. *Beyond the Birkhoff Polytope: Convex Relaxations for Vector Permutation Problems.* In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems (NIPS-14), pages 2168–2176. Curran Associates, Inc., 2014. (Cited on page 103.)

[Lin 2009] X. Lin, B. Afsari, L. Marchionni, L. Cope, G. Parmigiani, D. Naiman and D. Geman. *The ordering of expression among a few genes can provide simple cancer biomarkers and signal BRCA1 mutations.* BMC Bioinformatics, vol. 10, page 256, 2009. (Cited on pages 11, 18, 29, 46 and 47.)

[Livshits 2015] A. Livshits, A. Git, G. Fuks, C. Caldas and E. Domany. *Pathway-based personalized analysis of breast cancer expression data.* Molecular Oncology, vol. 9, no. 7, pages 1471–1483, 2015. (Cited on page 85.)

[Lockhart 1996] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo *et al. Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nature Biotechnology, vol. 14, no. 13, pages 1675–1680, 1996. (Cited on page 5.)

[Losada 2014] A. Losada. *Cohesin in cancer: chromosome segregation and beyond.* Nature Reviews Cancer, vol. 14, no. 6, pages 389–393, 2014. (Cited on page 97.)

[Lu 2005] W. Lu, K. Pan, L. Zhang, D. Lin, X. Miao and W. You. *Genetic polymorphisms of interleukin (IL)-1B, IL-1RN, IL-8, IL-10 and tumor necrosis factor α and risk of gastric cancer in a Chinese population.* Carcinogenesis, vol. 26, no. 3, pages 631–636, 2005. (Cited on page 97.)

[Lynch 1997] D. H. Lynch, A. Andreasen, E. Maraskovsky, J. Whitmore, R. E. Miller and J. C. L. Schuh. *FLT3 ligand induces tumor regression and antitumor immune responses in vivo.* Nature Medicine, vol. 3, no. 6, pages 625–631, 1997. (Cited on page 79.)

[Ma 2008] X.-J. Ma, R. Salunga, S. Dahiya, W. Wang, E. Carney *et al. A five-gene molecular grade index and HOXB13:IL17BR are complementary prognostic factors in early stage breast cancer.* Clinical Cancer Research, vol. 14, no. 9, pages 2601–2608, 2008. (Cited on page 6.)

[Mallat 1999] S. Mallat. A wavelet tour of signal processing. Academic Press, 1999. (Cited on pages 56 and 64.)

[Mallows 1957] C. L. Mallows. *Non-null ranking models. I.* Biometrika, vol. 44, no. 1/2, pages 114–130, 1957. (Cited on pages 20 and 38.)

[Mania 2016] H. Mania, A. Ramdas, M. J. Wainwright, M. I. Jordan and B. Recht. *Universality of Mallows' and degeneracy of Kendall's kernels for rankings.* Technical report, UC Berkeley, 2016. arXiv preprint arXiv:1603.08035. (Cited on page 52.)

[Mao 2015] M. Mao, T. Yu, J. Hu and L. Hu. *Dopamine D2 receptor blocker thioridazine induces cell death in human uterine cervical carcinoma cell line SiHa.* Journal of Obstetrics and Gynaecology Research, vol. 41, no. 8, pages 1240–1245, 2015. (Cited on page 79.)

[Marden 1996] J. I. Marden. Analyzing and modeling rank data. CRC Press, 1996. (Cited on pages 18 and 37.)

[Margolin 2013] A. A. Margolin, E. Bilal, E. Huang, T. C. Norman, L. Ottestad *et al. Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer.* Science Translational Medicine, vol. 5, no. 181, pages 181re1–181re1, 2013. (Cited on page 10.)

[Martini 2013] P. Martini, G. Sales, M. S. Massa, M. Chiogna and C. Romualdi. *Along signal paths: an empirical gene set approach exploiting pathway topology.* Nucleic Acids Research, vol. 41, no. 1, pages e19–e19, 2013. (Cited on page 85.)

[Matsukage 2008] A. Matsukage, F. Hirose, M.-A. Yoo and M. Yamaguchi. *The DRE/DREF transcriptional regulatory system: a master key for cell proliferation.* Biochimica et Biophysica Acta (BBA)–Gene Regulatory Mechanisms, vol. 1779, no. 2, pages 81–89, 2008. (Cited on page 78.)

[Mattei 2012] N. Mattei, J. Forshee and J. Goldsmith. *An empirical study of voting rules and manipulation with large datasets.* In Proceedings of the Sixth International Workshop on Computational Social Choice (COMSOC-12). Citeseer, 2012. (Cited on page 122.)

[Matthews 2013] J. M. Matthews, K. Lester, S. Joseph and D. J. Curtis. *LIM-domain-only proteins in cancer.* Nature Reviews Cancer, vol. 13, no. 2, pages 111–122, 2013. (Cited on page 99.)

[McCullagh 1989] P. McCullagh and J. A. Nelder. Generalized linear models, volume 37. CRC Press, 1989. (Cited on page 14.)

[Meilă 2007] M. Meilă, K. Phadnis, A. Patterson and J. Bilmes. *Consensus ranking under the exponential model.* In Proceedings of the Twenty-third Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-07), pages 285–294, Corvallis, Oregon, 2007. AUAI Press. (Cited on page 20.)

[Michiels 2005] S. Michiels, S. Koscielny and C. Hill. *Prediction of cancer outcome with microarrays: a multiple random validation strategy.* The Lancet, vol. 365, no. 9458, pages 488–492, 2005. (Cited on page 106.)

[Michiels 2016] S. Michiels, N. Ternès and F. Rotolo. *Statistical controversies in clinical research: prognostic gene signatures are not (yet) useful in clinical practice.* Annals of Oncology, vol. 27, no. 12, pages 2160–2167, 2016. (Cited on page 105.)

[Mika 1999] S. Mika, G. Rätsch, J. Weston, B. Schölkopf and K.R. Müller. *Fisher discriminant analysis with kernels.* In Y.-H. Hu, J. Larsen, E. Wilson and S. Douglas, editors, Neural Networks for Signal Processing IX, pages 41–48. IEEE, 1999. (Cited on page 46.)

[Moffat 2014] F. L. Moffat. *Clinical and Pathologic Prognostic and Predictive Factors.* In J. R. Harris, M. E. Lippman, M. Morrow and C. K. Osborne, editors, Diseases of the Breast, chapter 28. Lippincott Williams & Wilkins, 5th edition, 2014. (Cited on page 3.)

[Montaner 2009] D. Montaner, P. Minguez, F. Al-Shahrour and J. Dopazo. *Gene set internal coherence in the context of functional profiling.* BMC Genomics, vol. 10, no. 1, page 197, 2009. (Cited on page 90.)

[Montañez-Wiscovich 2009] M. E. Montañez-Wiscovich, D. D. Seachrist, M. D. Landis, J. Visvader, B. Andersen and R. A. Keri. *LMO4 is an essential mediator of ErbB2/HER2/Neu-induced breast cancer cell cycle progression.* Oncogene, vol. 28, no. 41, pages 3608–3618, 2009. (Cited on page 99.)

[Muandet 2012] K. Muandet, K. Fukumizu, F. Dinuzzo and B. Schölkopf. *Learning from distributions via support measure machines.* In F. Pereira, C. J. C.

Burges, L. Bottou and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems (NIPS-12), volume 25, pages 10–18. Curran Associates, Inc., 2012. (Cited on pages 30 and 34.)

[Murphy 2003] T. B. Murphy and D. Martin. *Mixtures of distance-based models for ranking data.* Computational Statistics & Data Analysis, vol. 41, no. 3, pages 645–655, 2003. (Cited on pages 39 and 42.)

[Murphy 2009] G. Murphy, A. J. Cross, L. S. Sansbury, A. Bergen, A. O. Laiyemo *et al. Dopamine D2 receptor polymorphisms and adenoma recurrence in the Polyp Prevention Trial.* International Journal of Cancer, vol. 124, no. 9, pages 2148–2151, 2009. (Cited on page 79.)

[Musgrove 2011] E. A. Musgrove, C. E. Caldon, J. Barraclough, A. Stone and R. L. Sutherland. *Cyclin D as a therapeutic target in cancer.* Nature Reviews Cancer, vol. 11, no. 8, pages 558–572, 2011. (Cited on page 79.)

[Neves 2002] S. R. Neves, P. T. Ram and R. Iyengar. *G protein pathways.* Science, vol. 296, no. 5573, pages 1636–1639, 2002. (Cited on page 79.)

[Olichon 2003] A. Olichon, L. Baricault, N. Gas, E. Guillou, A. Valette, P. Belenguer and G. Lenaers. *Loss of OPA1 perturbates the mitochondrial inner membrane structure and integrity, leading to cytochrome c release and apoptosis.* Journal of Biological Chemistry, vol. 278, no. 10, pages 7743–7746, 2003. (Cited on page 99.)

[on Cancer 2010] American Joint Committee on Cancer. AJCC Cancer Staging Manual. Springer, New York, NY, 7th edition, 2010. (Cited on page 3.)

[Oti 2007] M. Oti and H. G. Brunner. *The modular nature of genetic diseases.* Clinical Genetics, vol. 71, no. 1, pages 1–11, 2007. (Cited on page 84.)

[Paik 2004] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker *et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer.* New England Journal of Medicine, vol. 351, no. 27, pages 2817–2826, 2004. (Cited on pages 6 and 84.)

[Pan 2010] W. Pan, B. Xie and X. Shen. *Incorporating predictor network in penalized regression with application to microarray data.* Biometrics, vol. 66, no. 2, pages 474–484, 2010. (Cited on page 55.)

[Paplomata 2014] E. Paplomata and R. O'Regan. *The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers.* Therapeutic Advances in Medical Oncology, vol. 6, no. 4, pages 154–166, 2014. (Cited on page 79.)

[Parker 2009] J. S. Parker, M. Mullins, M. C. U. Cheang, S. Leung, D. Voduc *et al. Supervised risk predictor of breast cancer based on intrinsic subtypes.* Journal of Clinical Oncology, vol. 27, no. 8, pages 1160–1167, 2009. (Cited on page 6.)

[Patel 2013] T. Patel, D. Telesca, R. Rallo, S. George, T. Xia and A. E. Nel. *Hierarchical Rank Aggregation with Applications to Nanotoxicology.* Journal of Agricultural, Biological, and Environmental Statistics, vol. 18, no. 2, pages 159–177, 2013. (Cited on page 110.)

[Patil 2015] P. Patil, P.-O. Bachant-Winner, B. Haibe-Kains and J. T. Leek. *Test set bias affects reproducibility of gene signatures.* Bioinformatics, vol. 31, no. 14, pages 2318–2323, 2015. (Cited on page 107.)

[Petricoin 2002] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro *et al. Use of proteomic patterns in serum to identify ovarian cancer.* The Lancet, vol. 359, no. 9306, pages 572–577, 2002. (Cited on page 46.)

[Pharoah 1999] P. D. P. Pharoah, N. E. Day and C. Caldas. *Somatic mutations in the p53 gene and prognosis in breast cancer: a meta-analysis.* British Journal of Cancer, vol. 80, no. 12, page 1968, 1999. (Cited on page 4.)

[Pomeroy 2002] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo *et al. Prediction of central nervous system embryonal tumour outcome based on gene expression.* Nature, vol. 415, no. 6870, pages 436–442, 2002. (Cited on page 46.)

[Popova 2012] A. Popova. The robust beauty of APA presidential elections: an empty-handed hunt for the social choice conundrum. Master's thesis, University of Illinois at Urbana-Champaign, 2012. (Cited on page 122.)

[Pornour 2015] M. Pornour, G. Ahangari, S. H. Hejazi and A. Deezagi. *New perspective therapy of breast cancer based on selective dopamine receptor D2 agonist and antagonist effects on MCF-7 cell line.* Recent Patents on Anti-cancer Drug Discovery, vol. 10, no. 2, pages 214–223, 2015. (Cited on page 79.)

[Poste 2011] G. Poste. *Bring on the biomarkers.* Nature, vol. 469, no. 7329, pages 156–157, 2011. (Cited on page 105.)

[Procaccia 2012] A. D. Procaccia, S. J. Reddi and N. Shah. *A Maximum Likelihood Approach For Selecting Sets of Alternatives.* In Proceedings of the Conference on Uncertainly in Artificial Intelligence (UAI-12), volume 695, page 704. Citeseer, 2012. (Cited on page 110.)

[Rapaport 2007] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot and J.-P. Vert. *Classification of microarray data using gene networks.* BMC Bioinformatics, vol. 8, no. 1, page 35, 2007. (Cited on pages xi, 12, 13 and 60.)

[Reis-Filho 2011] J. S. Reis-Filho and L. Pusztai. *Gene expression profiling in breast cancer: classification, prognostication, and prediction.* The Lancet, vol. 378, no. 9805, pages 1812–1823, 2011. (Cited on page 84.)

[Renda 2003] M. E. Renda and U. Straccia. *Web metasearch: rank vs. score based rank aggregation methods.* In Proceedings of the 2003 ACM Symposium on Applied Computing, pages 841–846. ACM, 2003. (Cited on page 110.)

[Ripley 2007] B. D. Ripley. Pattern recognition and neural networks. Cambridge Univ Press, 2007. (Cited on page 94.)

[Robinson 2010] M. D. Robinson and A. Oshlack. *A scaling normalization method for differential expression analysis of RNA-seq data.* Genome Biology, vol. 11, no. 3, page 1, 2010. (Cited on page 86.)

[Roodi 2004] N. Roodi, W. D. Dupont, J. H. Moore and F. F. Parl. *Association of homozygous wild-type glutathione S-transferase M1 genotype with increased breast cancer risk.* Cancer Research, vol. 64, no. 4, pages 1233–1236, 2004. (Cited on page 78.)

[Saari 2000] D. G. Saari and V. R. Merlin. *A geometric examination of Kemeny's rule.* Social Choice and Welfare, vol. 17, no. 3, pages 403–438, 2000. (Cited on pages 110 and 112.)

[Saeys 2007] Y. Saeys, I. Inza and P. Larrañaga. *A review of feature selection techniques in bioinformatics.* Bioinformatics, vol. 23, no. 19, pages 2507–2517, 2007. (Cited on page 10.)

[Schalekamp 2009] F. Schalekamp and A. van Zuylen. *Rank aggregation: Together we're strong.* In Proceedings of the Meeting on Algorithm Engineering & Expermiments, pages 38–51. Society for Industrial and Applied Mathematics, 2009. (Cited on page 110.)

[Schneider 2015] D. Schneider, G. Bianchini, D. Horgan, S. Michiels, W. Witjes *et al. Establishing the evidence bar for molecular diagnostics in personalised cancer care.* Public Health Genomics, vol. 18, no. 6, pages 349–358, 2015. (Cited on page 105.)

[Schnitt 2010] S. J. Schnitt. *Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy.* Modern Pathology, vol. 23, pages S60–S64, 2010. (Cited on page 4.)

[Schoenberg 1938] I. J. Schoenberg. *Metric spaces and positive definite functions.* Transactions of the American Mathematical Society, vol. 44, no. 3, pages 522–536, 1938. (Cited on page 21.)

[Schölkopf 1999a] B. Schölkopf, A. J. Smola and K. R. Müller. *Kernel principal component analysis.* In Advances in Kernel Methods, pages 327–352. MIT Press, 1999. (Cited on page 52.)

[Schölkopf 1999b] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor and J. C. Platt. *Support Vector Method for Novelty Detection.* In Advances

in Neural Information Processing Systems (NIPS-12), volume 12, pages 582–588. MIT Press, 1999. (Cited on page 52.)

[Schölkopf 2002] B. Schölkopf and A. J. Smola. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, 2002. (Cited on pages 18 and 52.)

[Schölkopf 2004] B. Schölkopf, K. Tsuda and J.-P. Vert. Kernel Methods in Computational Biology. MIT Press, The MIT Press, Cambridge, Massachussetts, 2004. (Cited on pages 14, 18 and 29.)

[Schroeder 2011] M. Schroeder, B. Haibe-Kains, A. Culhane, C. Sotiriou, G. Bontempi and J. Quackenbush. *breastCancerTRANSBIG: Gene expression dataset published by Desmedt et al. [2007] (TRANSBIG)*, 2011. R package version 1.2.0. (Cited on page 46.)

[Sebastian-Leon 2014] P. Sebastian-Leon, E. Vidal, P. Minguez, A. Conesa, S. Tarazona *et al. Understanding disease mechanisms with models of signaling pathway activities.* BMC Systems Biology, vol. 8, no. 1, page 1, 2014. (Cited on page 90.)

[Semenza 2012] G. L. Semenza. *Hypoxia-inducible factors: mediators of cancer progression and targets for cancer therapy.* Trends in Pharmacological Sciences, vol. 33, no. 4, pages 207–214, 2012. (Cited on page 97.)

[Shawe-Taylor 2004] J. Shawe-Taylor and N. Cristianini. Kernel Methods for Pattern Analysis. Cambridge Univ Press, New York, NY, USA, 2004. (Cited on pages 14 and 18.)

[Shi 2011] P. Shi, S. Ray, Q. Zhu and M. A. Kon. *Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction.* BMC Bioinformatics, vol. 12, page 375, 2011. (Cited on pages 11, 46 and 47.)

[Shihab 2015] H. A. Shihab, M. F. Rogers, J. Gough, M. Mort, D. N. Cooper *et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation.* Bioinformatics, vol. 31, no. 10, pages 1536–1543, 2015. (Cited on page 99.)

[Shuman 2012] D. I. Shuman, B. Ricaud and P. Vandergheynst. *A windowed graph Fourier transform.* In Statistical Signal Processing Workshop (SSP), pages 133–136. IEEE, 2012. (Cited on page 82.)

[Shuman 2013] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega and P. Vandergheynst. *The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains.* IEEE Signal Processing Magazine, vol. 30, no. 3, pages 83–98, 2013. (Cited on pages 15 and 56.)

[Shuman 2016] D. I. Shuman, M. J. Faraji and P. Vandergheynst. *A multiscale pyramid transform for graph signals*. IEEE Transactions on Signal Processing, vol. 64, no. 8, pages 2119–2134, 2016. (Cited on page 82.)

[Sibony 2014] E. Sibony. *Borda count approximation of Kemeny's rule and pairwise voting inconsistencies*. In NIPS-2014 Workshop on Analysis of Rank Data: Confluence of Social Choice, Operations Research, and Machine Learning. Curran Associates, Inc., 2014. (Cited on pages 110 and 111.)

[Simon 2003] R. Simon, M. D. Radmacher, K. Dobbin and L. M. McShane. *Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification*. Journal of the National Cancer Institute, vol. 95, no. 1, pages 14–18, 2003. (Cited on page 105.)

[Simon 2011] N. Simon, J. Friedman, T. Hastie and R. Tibshirani. *Regularization paths for Cox's proportional hazards model via coordinate descent*. Journal of Statistical Software, vol. 39, no. 5, page 1, 2011. (Cited on page 63.)

[Sing 2005] T. Sing, O. Sander, N. Beerenwinkel and T. Lengauer. *ROCR: visualizing classifier performance in R*. Bioinformatics, vol. 21, no. 20, pages 3940–3941, 2005. (Cited on page 92.)

[Singh 2002] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola *et al. Gene expression correlates of clinical prostate cancer behavior*. Cancer Cell, vol. 1, no. 2, pages 203–209, 2002. (Cited on page 46.)

[Smola 2007] A. Smola, A. Gretton, L. Song and B. Schölkopf. *A Hilbert Space Embedding for Distributions*. In M. Hutter, R. A. Servedio and E. Takimoto, editors, Proceedings of the 18th International Conference on Algorithmic Learning Theory (ALT-07), pages 13–31, Berlin, Heidelberg, October 2007. Springer Berlin Heidelberg. (Cited on page 34.)

[Sonnenburg 2006] S. Sonnenburg, G. Rätsch, C. Schäfer and B. Schölkopf. *Large scale multiple kernel learning*. Journal of Machine Learning Research, vol. 7, pages 1531–1565, 2006. (Cited on page 29.)

[Sotiriou 2006] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox *et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis*. Journal of the National Cancer Institute, vol. 98, no. 4, pages 262–272, 2006. (Cited on page 6.)

[Sotiriou 2009] C. Sotiriou and L. Pusztai. *Gene-expression signatures in breast cancer*. New England Journal of Medicine, vol. 360, no. 8, pages 790–800, 2009. (Cited on pages 6 and 84.)

[Staiger 2013] C. Staiger, S. Cadot, B. Györffy, L. Wessels and G. Klau. *Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis*. Frontiers in Genetics, vol. 4, page 289, 2013. (Cited on page 81.)

[Stanley 1986] R. P. Stanley. Enumerative combinatorics. Wadsworth Publishing Company, Belmont, CA, USA, 1986. (Cited on page 115.)

[Steinwart 2005] I. Steinwart. *Consistency of support vector machines and other regularized kernel classifiers.* IEEE Transactions on Information Theory, vol. 51, no. 1, pages 128–142, 2005. (Cited on page 34.)

[Stelzer 2016] G. Stelzer, N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik *et al. The genecards suite: from gene data mining to disease genome sequence analyses.* Current Protocols in Bioinformatics, pages 1–30, 2016. (Cited on page 99.)

[Stuart 2003] J. M. Stuart, E. Segal, D. Koller and S. K. Kim. *A gene-coexpression network for global discovery of conserved genetic modules.* Science, vol. 302, no. 5643, pages 249–255, 2003. (Cited on page 55.)

[Subramanian 2005] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert *et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences of the United States of America (PNAS), vol. 102, no. 43, pages 15545–15550, 2005. (Cited on page 12.)

[Sull 2004] J. W. Sull, H. Ohrr, D. R. Kang and C. M. Nam. *Glutathione S-transferase M1 status and breast cancer risk: a meta-analysis.* Yonsei Medical Journal, vol. 45, pages 683–689, 2004. (Cited on page 78.)

[Sum 2002] E. Y. M. Sum, B. Peng, X. Yu, J. Chen, J. Byrne, G. J. Lindeman and J. E. Visvader. *The LIM domain protein LMO4 interacts with the cofactor CtIP and the tumor suppressor BRCA1 and inhibits BRCA1 activity.* Journal of Biological Chemistry, vol. 277, no. 10, pages 7849–7856, 2002. (Cited on page 99.)

[Sun 2013] S. Sun. *A survey of multi-view machine learning.* Neural Computing and Applications, vol. 23, no. 7-8, pages 2031–2038, 2013. (Cited on page 104.)

[Sun 2014] H. Sun, W. Lin, R. Feng and H. Li. *Network-regularized high-dimensional Cox regression for analysis of genomic data.* Statistica Sinica, vol. 24, no. 3, page 1433, 2014. (Cited on page 64.)

[Sutherland 2003] K. D. Sutherland, J. E. Visvader, D. Y. H. Choong, E. Y. M. Sum, G. J. Lindeman and I. G. Campbell. *Mutational analysis of the LMO4 gene, encoding a BRCA1-interacting protein, in breast carcinomas.* International Journal of Cancer, vol. 107, no. 1, pages 155–158, 2003. (Cited on page 99.)

[Sweldens 1998] W. Sweldens. *The lifting scheme: A construction of second generation wavelets.* SIAM Journal on Mathematical Analysis, vol. 29, no. 2, pages 511–546, 1998. (Cited on page 66.)

[Tan 2005] A. C. Tan, D. Q. Naiman, L. Xu, R. L. Winslow and D. Geman. *Simple decision rules for classifying human cancers from gene expression profiles.* Bioinformatics, vol. 21, no. 20, pages 3896–3904, 2005. (Cited on pages 11, 18, 29, 46 and 47.)

[Tarca 2008] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal *et al. A novel signaling pathway impact analysis.* Bioinformatics, vol. 25, no. 1, pages 75–82, 2008. (Cited on page 12.)

[Tax 2004] D. M. Tax and R. P. Duin. *Support vector data description.* Machine Learning, vol. 54, no. 1, pages 45–66, 2004. (Cited on page 52.)

[Testa 2001] J. R. Testa and A. Bellacosa. *AKT plays a central role in tumorigenesis.* Proceedings of the National Academy of Sciences of the United States of America (PNAS), vol. 98, no. 20, pages 10983–10985, 2001. (Cited on page 97.)

[Tian 2013] X. Tian, X. Wang and J. Chen. *Network-constrained group lasso for high-dimensional multinomial classification with application to cancer subtype prediction.* Cancer Informatics, vol. 13, no. Suppl 6, pages 25–33, 2013. (Cited on pages 55 and 60.)

[Tian 2017] N. Tian, J. Li, J. Shi and G. Sui. *From General Aberrant Alternative Splicing in Cancers and Its Therapeutic Application to the Discovery of an Oncogenic DMTF1 Isoform.* International Journal of Molecular Sciences, vol. 18, no. 3, page 191, 2017. (Cited on page 79.)

[Tibshirani 1996] R. Tibshirani. *Regression shrinkage and selection via the lasso.* Journal of the Royal Statistical Society: Series B (Methodological), pages 267–288, 1996. (Cited on pages 54 and 58.)

[Tibshirani 2002] R. Tibshirani, T. Hastie, B. Narasimhan and G. Chu. *Diagnosis of multiple cancer types by shrunken centroids of gene expression.* Proceedings of the National Academy of Sciences of the United States of America (PNAS), vol. 99, no. 10, pages 6567–6572, 2002. (Cited on pages 10 and 94.)

[Tibshirani 2005] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu and K. Knight. *Sparsity and smoothness via the fused lasso.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 67, no. 1, pages 91–108, 2005. (Cited on pages 55 and 60.)

[Tibshirani 2011] R. J. Tibshirani and J. Taylor. *The solution path of the generalized lasso.* The Annals of Statistics, vol. 39, no. 3, pages 1335–1371, 2011. (Cited on pages 60 and 63.)

[Tillgren 2015] V. Tillgren, J. C. S. Ho, P. Önnerfjord and S. Kalamajski. *The novel small leucine-rich protein chondroadherin-like (CHADL) is expressed*

in cartilage and modulates chondrocyte differentiation. Journal of Biological Chemistry, vol. 290, no. 2, pages 918–925, 2015. (Cited on page 99.)

[Torre 2015] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent and A. Jemal. *Global Cancer Statistics, 2012.* CA: A Cancer Journal for Clinicians, vol. 65, no. 2, pages 87–108, 2015. (Cited on page 2.)

[Tremblay 2014] N. Tremblay and P. Borgnat. *Graph wavelets for multiscale community mining.* IEEE Transactions on Signal Processing, vol. 62, no. 20, pages 5227–5239, 2014. (Cited on page 62.)

[Tsochantaridis 2005] I. Tsochantaridis, T. Joachims, T. Hofmann and Y. Altun. *Large margin methods for structured and interdependent output variables.* Journal of Machine Learning Research, vol. 6, pages 1453–1484, 2005. (Cited on page 52.)

[van 't Veer 2002] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart *et al. Gene expression profiling predicts clinical outcome of breast cancer.* Nature, vol. 415, no. 6871, pages 530–536, 2002. (Cited on pages 6, 46, 84 and 106.)

[van 't Veer 2008] L. J. van 't Veer and R. Bernards. *Enabling personalized cancer medicine through analysis of gene-expression patterns.* Nature, vol. 452, no. 7187, pages 564–570, 2008. (Cited on pages 3, 84 and 101.)

[Van Zuylen 2007] A. Van Zuylen and D. P. Williamson. *Deterministic algorithms for rank aggregation and other ranking and clustering problems.* In Approximation and Online Algorithms, pages 260–273. Springer, 2007. (Cited on pages 110 and 111.)

[Vandin 2011] F. Vandin, E. Upfal and B. J. Raphael. *Algorithms for detecting significantly mutated pathways in cancer.* Journal of Computational Biology, vol. 18, no. 3, pages 507–522, 2011. (Cited on page 12.)

[Vapnik 1998] V. N. Vapnik. Statistical Learning Theory. Wiley, New York, 1998. (Cited on page 18.)

[Venables 2002] W. N. Venables and B. D. Ripley. Modern applied statistics with S. Springer, New York, fourth edition, 2002. (Cited on page 94.)

[Venet 2011] D. Venet, J. E. Dumont and V. Detours. *Most random gene expression signatures are significantly associated with breast cancer outcome.* PLoS Computational Biology, vol. 7, no. 10, page e1002240, 2011. (Cited on page 106.)

[Vidal 2011] M. Vidal, M. E. Cusick and A.-L. Barabási. *Interactome networks and human disease.* Cell, vol. 144, no. 6, pages 986–998, 2011. (Cited on page 84.)

[Vishwanathan 2009] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor and K. M. Borgwardt. *Graph Kernels*. Journal of Machine Learning Research, vol. 10, pages 1–41, 2009. (Cited on page 18.)

[Vogelstein 2004] B. Vogelstein and K. W. Kinzler. *Cancer genes and the pathways they control*. Nature Medicine, vol. 10, no. 8, pages 789–799, 2004. (Cited on pages 4 and 12.)

[Wang 2003] J. Wang, J. Lee and C. Zhang. *Kernel trick embedded Gaussian mixture model*. In Proceedings of the 14th International Conference on Algorithmic Learning Theory (ALT-03), pages 159–174. Springer, 2003. (Cited on page 40.)

[Wang 2005a] Y. Wang, J. G. M. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look *et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer*. The Lancet, vol. 365, no. 9460, pages 671–679, 2005. (Cited on pages 7, 84 and 106.)

[Wang 2005b] Y. Wang, F. S. Makedon, J. C. Ford and J. Pearlman. *HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data*. Bioinformatics, vol. 21, no. 8, pages 1530–1537, 2005. (Cited on page 46.)

[Wang 2009] Z. Wang, M. Gerstein and M. Snyder. *RNA-Seq: a revolutionary tool for transcriptomics*. Nature Reviews Genetics, vol. 10, no. 1, pages 57–63, 2009. (Cited on page 5.)

[Wang 2013] D. Wang, A. Mazumdar and G. W. Wornell. *A rate-distortion theory for permutation spaces*. In Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on, pages 2562–2566. IEEE, 2013. (Cited on page 124.)

[Weigel 2010] M. T. Weigel and M. Dowsett. *Current and emerging biomarkers in breast cancer: prognosis and prediction*. Endocrine-Related Cancer, vol. 17, no. 4, pages R245–R262, 2010. (Cited on page 4.)

[Weigelt 2012] B. Weigelt, L. Pusztai, A. Ashworth and J. S. Reis-Filho. *Challenges translating breast cancer gene signatures into the clinic*. Nature Reviews Clinical Oncology, vol. 9, no. 1, pages 58–64, 2012. (Cited on page 105.)

[Wilson 2008] J. B. Wilson, K. Yamamoto, A. S. Marriott, S. Hussain, P. Sung *et al. FANCG promotes formation of a newly identified protein complex containing BRCA2, FANCD2 and XRCC3*. Oncogene, vol. 27, no. 26, pages 3641–3652, 2008. (Cited on page 78.)

[Wilson 2014] B. J. Wilson, K. R. Saab, J. Ma, T. Schatton, P. Pütz *et al. ABCB5 maintains melanoma-initiating cells through a proinflammatory cytokine signaling circuit*. Cancer Research, vol. 74, no. 15, pages 4196–4207, 2014. (Cited on page 99.)

[Wu 2004] G. Wu, Y. Z. Fang, S. Yang, J. R. Lupton and N. D. Turner. *Glutathione metabolism and its implications for health.* The Journal of Nutrition, vol. 134, no. 3, pages 489–492, 2004. (Cited on page 78.)

[Xia 2015] L. Xia. *Generalized Decision Scoring Rules: Statistical, Computational, and Axiomatic Properties.* In Proceedings of the Sixteenth ACM Conference on Economics and Computation, EC '15, pages 661–678, New York, NY, USA, 2015. ACM. (Cited on page 110.)

[Xing 2001] E. P. Xing, M. I. Jordan and R. M. Karp. *Feature Selection for High-dimensional Genomic Microarray Data.* In Proceedings of the 18th International Conference on Machine Learning (ICML-01), pages 601–608, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. (Cited on page 9.)

[Xu 2005] L. Xu, A. C. Tan, D. Q. Naiman, D. Geman and R. L. Winslow. *Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data.* Bioinformatics, vol. 21, no. 20, pages 3905–3911, 2005. (Cited on pages 11, 18, 29, 46 and 47.)

[Yamashita 2007] D. Yamashita, Y. Sano, Y. Adachi, Y. Okamoto, H. Osada *et al. hDREF regulates cell proliferation and expression of ribosomal protein genes.* Molecular and Cellular Biology, vol. 27, no. 6, pages 2003–2013, 2007. (Cited on page 78.)

[Yang 2010] J. Y. Yang, S.-A. Ha, Y.-S. Yang and J. W. Kim. *p-Glycoprotein ABCB5 and YB-1 expression plays a role in increased heterogeneity of breast cancer cells: correlations with cell fusion and doxorubicin resistance.* BMC Cancer, vol. 10, no. 1, page 388, 2010. (Cited on page 99.)

[Yang 2014] H. Yang, B. Wang, T. Wang, L. Xu, C. He *et al. Toll-like receptor 4 prompts human breast cancer cells invasiveness via lipopolysaccharide stimulation and is overexpressed in patients with lymph node metastasis.* PloS One, vol. 9, no. 10, page e109980, 2014. (Cited on page 97.)

[Young 1978] H. P. Young and A. Levenglick. *A consistent extension of Condorcet's election principle.* SIAM Journal on Applied Mathematics, vol. 35, no. 2, pages 285–300, 1978. (Cited on page 110.)

[Yuan 2006] M. Yuan and Y. Lin. *Model selection and estimation in regression with grouped variables.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 68, no. 1, pages 49–67, 2006. (Cited on pages 55 and 58.)

[Zamani 2014] H. Zamani, A. Shakery and P. Moradi. *Regression and learning to rank aggregation for user engagement evaluation.* In Proceedings of the 2014 Recommender Systems Challenge, page 29. ACM, 2014. (Cited on page 110.)

[Zeileis 2004] A. Zeileis, K. Hornik, A. Smola and A. Karatzoglou. *kernlab – an S4 package for kernel methods in R.* Journal of Statistical Software, vol. 11, no. 9, pages 1–20, 2004. (Cited on page 94.)

[Zhang 2002] R. Zhang and A. Rudnicky. *A large scale clustering scheme for kernel k-means.* In Proceedings of the Sixteenth International Conference on Pattern Recognition (ICPR-02), volume 4, pages 289–292. IEEE, 2002. (Cited on page 37.)

[Zhang 2009] Y. Zhang, A. M. Sieuwerts, M. McGreevy, G. Casey, T. Cufer *et al. The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy.* Breast Cancer Research and Treatment, vol. 116, no. 2, pages 303–309, Jul 2009. (Cited on page 7.)

[Zhang 2013] W. Zhang, T. Ota, V. Shridhar, J. Chien, B. Wu and R. Kuang. *Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment.* PLoS Computational Biology, vol. 9, no. 3, page e1002975, 2013. (Cited on page 63.)

[Zhao 2013] X. Zhao, C. Tian, W. M. Puszyk, O. O. Ogunwobi, M. Cao *et al. OPA1 downregulation is involved in sorafenib-induced apoptosis in hepatocellular carcinoma.* Laboratory Investigation, vol. 93, no. 1, pages 8–19, 2013. (Cited on page 99.)

[Zhong 2015] Q. Zhong, H.-L. Peng, X. Zhao, L. Zhang and W.-T. Hwang. *Effects of BRCA1- and BRCA2-Related Mutations on Ovarian and Breast Cancer Survival: A Meta-analysis.* Clinical Cancer Research, vol. 21, no. 1, pages 211–220, 2015. (Cited on page 4.)

[Zhu 2016] Y. Zhu, J. Wu, C. Zhang, S. Sun, J. Zhang *et al. BRCA mutations and survival in breast cancer: an updated systematic review and meta-analysis.* Oncotarget, vol. 7, no. 43, page 70113, 2016. (Cited on page 4.)

[Zou 2005] H. Zou and T. Hastie. *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 67, no. 2, pages 301–320, 2005. (Cited on pages 54 and 58.)

[Zwicker 2008] W. S. Zwicker. *Consistency without neutrality in voting rules: When is a vote an average?* Mathematical and Computer Modelling, vol. 48, no. 9, pages 1357–1373, 2008. (Cited on page 113.)

Résumé

Le cancer du sein est le deuxième cancer le plus répandu dans le monde et la principale cause de décès due à un cancer chez les femmes. L'amélioration du pronostic du cancer a été l'une des principales préoccupations afin de permettre une meilleure gestion et un meilleur traitement clinique des patients. Avec l'avancement rapide des technologies de profilage génomique durant ces dernières décennies, la disponibilité aisée d'une grande quantité de données génomiques pour la recherche médicale a motivé la tendance actuelle qui consiste à utiliser des outils informatiques tels que l'apprentissage statistique dans le domaine de la science des données afin de découvrir les biomarqueurs moléculaires en lien avec l'amélioration du pronostic. Cette thèse est conçue suivant deux directions d'approches destinées à répondre à deux défis majeurs dans l'analyse de données génomiques pour le pronostic du cancer du sein d'un point de vue méthodologique de l'apprentissage statistique : les approches basées sur le classement pour améliorer le pronostic moléculaire et les approches guidées par un réseau donné pour améliorer la découverte de biomarqueurs. D'autre part, les méthodologies développées et étudiées dans cette thèse, qui concernent respectivement l'apprentissage à partir de données de classements et l'apprentissage sur un graphe, apportent une contribution significative à plusieurs branches de l'apprentissage statistique, concernant au moins les applications à la biologie du cancer et la théorie du choix social.

Mots Clés

Cancer du sein, pronostic moléculaire, découverte de biomarqueurs, réseau biologique, apprentissage statistique, analyse de données génomiques

Abstract

Breast cancer is the second most common cancer worldwide and the leading cause of women's death from cancer. Improving cancer prognosis has been one of the problems of primary interest towards better clinical management and treatment decision making for cancer patients. With the rapid advancement of genomic profiling technologies in the past decades, easy availability of a substantial amount of genomic data for medical research has been motivating the currently popular trend of using computational tools, especially machine learning in the era of data science, to discover molecular biomarkers regarding prognosis improvement. This thesis is conceived following two lines of approaches intended to address two major challenges arising in genomic data analysis for breast cancer prognosis from a methodological standpoint of machine learning: rank-based approaches for improved molecular prognosis and network-guided approaches for enhanced biomarker discovery. Furthermore, the methodologies developed and investigated in this thesis, pertaining respectively to learning with rank data and learning on graphs, have a significant contribution to several branches of machine learning, concerning applications across but not limited to cancer biology and social choice theory.

Keywords

Breast Cancer, Molecular Prognosis, Biomarker Discovery, Biological Network, Machine Learning, Genomic Data Analysis